The data-driven Schrödinger bridge

Michele Pavon Esteban G. Tabak Giulio Trigila

October 22, 2019

Abstract

Erwin Schrödinger posed, and to a large extent solved in 1931/32 the problem of finding the most likely random evolution between two continuous probability distributions. This article considers this problem in the case when only samples of the two distributions are available. A novel iterative procedure is proposed, inspired by Fortet-Sinkhorn type algorithms. Since only samples of the marginals are available, the new approach features constrained maximum likelihood estimation in place of the nonlinear boundary couplings, and importance sampling to propagate the functions φ and $\hat{\varphi}$ solving the Schrödinger system. This method mitigates the curse of dimensionality, compared to the introduction of grids which in high dimensions lead to numerically unfeasible methods. The methodology is illustrated in two applications: entropic interpolation of two-dimensional Gaussian mixtures, and the estimation of integrals through a variation of importance sampling.

1 Introduction

This article proposes a methodology for solving the following problem: given m and n independent samples $\{x_i\}$ and $\{y_j\}$ from two distributions with probability densities $\rho_0(x)$ and $\rho_1(y)$ respectively, and a prior probability $p(t_1, x, t_2, y)$ that a "particle" at position x at time t_1 will end up at position y at time t_2, find the most likely intermediate evolution $\rho(z, t)$, $t \in [0, 1]$ satisfying $\rho(x, 0) = \rho_0(x)$ and $\rho(y, 1) = \rho_1(y)$. This is a data-driven version of the Schrödinger Bridge Problem, which we describe below. In addition to the evolving density $\rho(z, t)$, the solution provides the posterior transition density $p^*(t_1, x, t_2, y)$ most consistent with the observed initial and final distributions, useful for model improvement.

1.1 Motivation, examples and extensions

Many problems of practical and theoretical interest can be directly formulated as data-driven Schrödinger bridges. Consider the following two examples, arising in climate studies and evolutionary biology:

1. With the current knowledge of oceanic or atmospheric flows described in terms of a velocity field $v(x,t)$ and a diffusion operator D, the corresponding Fokker-Planck evolution equation yields the prior $p(t_1,x,t_2,y)$ for the trajectories of tracers. If at any point in time a cloud of particles is released into the fluid such as a volcanic eruption, or its current concentration ρ_0 is sampled, and at some other time its distribution ρ_1 is sampled again, the data-driven bridge problem provides an estimate for the most likely intermediate evolution $\rho(z,t)$ of the tracer cloud and to an improved model for the currents v.

2. Given the distribution of traits (genomic or phenomic) for a species at two points in time, and a stochastic model for their evolution, the problem asks for the most likely intermediate evolutionary stages, and provides as additional output an improved stochastic evolutionary model.

In other problems, it is not an intermediate evolution that one is after, but the probabilistic matching $\pi(x,y)$ between two distributions $\rho_0(x)$ and $\rho_1(y)$ under a prior matching model $p(y|x)$. In this case, both the problem and the methodology proposed for solving it extend without changes to situations where the variables x and y do not have the same dimensions, arising frequently in practice. For instance, in applications to the employment market, there is no reason for the number of variables characterizing employers and employees to be the same.

In a third type of scenarios, there is only one data-given distribution $\rho_1(y)$; the other distribution $\rho_0(x)$ and the prior $p(t_1,x,t_2,y)$ are introduced for convenience by the modeler, so as to perform $\rho_1(y)$-related tasks. As an example, in Section 4 we apply the Schrödinger bridge to develop a variation of importance sampling where the distribution over which expected values of a function are sought is known only through samples.

In other applications, one has only $\rho_1(y)$ and the prior $p(t_1,x,t_2,y)$, and would like to determine $\rho_t(z)$ for $t<1$. Two prototypal examples are inverse problems, such as describing the most likely previous temperature distribution of a system given its current one, and large deviation problems: if the stochastic process described by p has a statistically steady state ρ_{eq}, what are the most likely paths that will lead to a ρ_1 different from ρ_{eq}, such as the one corresponding to a strong storm or a drought in applications to weather and climate.

Most of these examples involve prior transition probabilities suited to the particular application being considered. This article focuses on a methodology for the case where the prior is the Wiener process (the one that Schrödinger considered originally.) The extension to more general priors is work in progress. Nevertheless, we foresee that the extension to the case when the prior measure corresponds to a *reversible* Markov diffusion presents no difficulty. In that case, by Kolmogorov's characterization, the (forward) Ito differential is of the form

$$dX_t = -\nabla H(x_t)dt + dW_t$$

where $H:\mathbb{R}^n \to \mathbb{R}$ is sufficiently smooth and such that $\exp[-2H(x)]$ is integrable. In that case, X is reversible with stationary probability density

$\rho(x) = c\exp[-2H(x)]$, where c is a normalizing constant. Moreover, the backward (reverse-time) drift in the sense of Nelson [47] is $\bar{b}(x) = \nabla H(x)$ and the reverse time transition density $\bar{p}$ satisfies

$$\bar{p}(x, t, y, s) = p(x, s, y, t), \quad \forall s < t, \quad \forall x, y \in \mathbb{R}^n.$$

This can be used in simulations. Another family of non translation invariant prior measures to which our methodology can be extended is the case when the prior is a general Gauss-Markov process as considered in [15, 16]. Both extensions will be considered in future work.

1.2 The methodology

The solution to the Schrödinger Bridge Problem can be factorized in the form (see (14) in Section 2 below)

$$\rho_t(x) = \varphi_t(x)\hat{\varphi}_t(x),$$

where $\rho_t(x)$ represents the distribution at time t, and φ and $\hat{\varphi}$ evolve from $t = 1$ and $t = 0$ respectively, following the prior:

$$\hat{\varphi}_t(y) = \int p(0, x, t, y)\hat{\varphi}_0(x)dx,$$
$$\varphi_t(x) = \int p(t, x, 1, y)\varphi_1(y)dy.$$

One can therefore, starting from an arbitrary $\hat{\varphi}_0(x)$, propagate it into the corresponding $\hat{\varphi}_1(y)$, and write

$$\varphi_1(y) = \frac{\rho_1(y)}{\hat{\varphi}_1(y)}.$$

Then, evolving $\varphi_1(y)$ back into the corresponding $\varphi_0(x)$, we write

$$\hat{\varphi}_0(x) = \frac{\rho_0(x)}{\varphi_0(x)},$$

and repeat. This idea underlies iteration schemes that, under suitable assumptions, converge to the solution of the Schrödinger Bridge Problem [32, 12].

Yet this procedure assumes that the initial and final distributions ρ_0 and ρ_1, as well as the transition probability p, are known explicitly, and that the integrals propagating φ and $\hat{\varphi}$ between $t = 0$ and $t = 1$ can be evaluated in closed form. By contrast, in applications ρ_0 and ρ_1 are typically only known through samples. In addition, it is often the case that the transition probability p can be sampled through the integration of a stochastic differential equation, but not evaluated, which would require solving the corresponding Fokker-Plank equation. Moreover, even if ρ_0, ρ_1 and p are known, one still needs to estimate the integrals propagating φ and $\hat{\varphi}$ numerically.

The methodology developed in this article mimics the iterative procedure above, but replacing each step by a sample-based equivalent. Thus the statements that

$$\varphi_0(x)\hat{\varphi}_0(x) = \rho_0(x) \quad \text{and} \quad \varphi_1(y)\hat{\varphi}_1(y) = \rho_1(y)$$

are interpreted as density estimations and implemented via maximum likelihood, and the propagators for φ and $\hat{\varphi}$ are estimated via importance sampling. Both tasks involve elements unique to the Schrödinger Bridge Problem, described in Section 3.

1.3 Prior work

Schrödinger's statistical mechanical thought experiment (large deviations problem) was motivated by analogies with quantum mechanics. On the other hand, since Boltzmann's fundamental work [7], and then through Sanov's theorem [52], we know that finding the most likely *Zustandverteilung* (macrostate) is equivalent to solving a maximum entropy problem. This connection provides a second important motivation for Schrödinger bridges, as an inference methodology that prescribes a posterior distribution making the fewest number of assumptions beyond the available information. This approach has been developed over the years, thanks in particular to the work of Jaynes, Burg, Dempster and Csiszár [39, 40, 8, 9, 26, 19, 20, 21]. A more recent third motivation for studying Schrödinger bridges is that they can be viewed as regularizations of the Optimal Mass Transport (OMT) problem [44, 45, 46, 41, 42, 10] which mitigates its computational challenges [3, 4, 50]. A large number of papers have since appeared on computational regularized OMT (Sinkhorn-type algorithms), see e.g. [22, 5, 14, 12, 17, 43, 2, 18]. While most of the classical work concentrates on the continuous problem, see e.g. the bibliography in [42] and Section 2 below, these papers concern the discrete Schrödinger Bridge Problem [48, 34]. Hardly any attention, however, has been given to the case when only samples of continuous marginals are available. One exception is [27] which deals with using regularized optimal transport for hard and soft clustering. One might think that the latter case may be readily treated by discretizing the spatial variables through grids. As we argue in the beginning of Section 3, such an approach is often numerically unfeasible and/or not reliable. Thus, in this paper we provide what appears to be the first numerically viable approach to the data-driven continuous Schrödinger Bridge Problem.

As discussed at the end of Subsection 2.4, this approach permits finding a map from ρ_0 to ρ_1, relating this work to [59] and [57] developed in the context of optimal transport.

1.4 Organization of the article

The paper is organized as follows. In Section 2, we provide an introduction to Schrödinger bridges. We include a concise description of Schrödinger's original motivation, and elements of the connection between the large deviation problem and a path space maximum entropy problem, and with Optimal Transport. We also sketch derivations of the Schrödinger system and of the stochastic control and fluid dynamic formulations, focusing on the case when the prior transition density is the heat kernel.

In Subsection 2.5, we outline Fortet's iterative algorithm, dating back to 1940, which represents a sort of guideline for the numerical methods we develop in the rest of the paper. Section 3 features the novel methodology to attack the data-driven bridge problem, motivated by numerical, statistical and optimization considerations. First, the so-called half-bridge problem is treated, and then the full bridge, leading to the algorithm of Subsection 3.3. In Section 4, we illustrate the methodology in two relevant applications: the entropic interpolation between two Gaussian mixtures on $\mathbb{R}^2$ and a new application of Schrödinger bridges to a variation of Importance Sampling. Finally, in Section 5 we summarize the results and propose future avenues of research.

2 Background on Schrödinger bridges

2.1 Schrödinger's hot gas experiment and maximum entropy formulation

In the early 1930s, Erwin Schrödinger proposed the following *Gedankenexperiment* [53, 54]. Consider the evolution of a cloud of N independent Brownian particles in $\mathbb{R}^n$. This cloud of particles has been observed to have at the initial time $t = 0$ an empirical distribution approximately equal to $\rho_0(x)dx$. At time $t = 1$, an empirical distribution is observed approximately equal to $\rho_1(y)dy$ which considerably differs from what it should be according to the law of large numbers (N is large, typically of the order of Avogadro's number), namely

$$\rho_1(y) \neq \int_{\mathbb{R}^3} p(0, x, 1, y)\rho_0(x)dx,$$

where

$$p(s, y, t, x) = [2\pi(t-s)]^{-\frac{n}{2}} \exp\left[-\frac{|x-y|^2}{2(t-s)}\right], \quad s < t \tag{1}$$

is the transition density of the Wiener process. It is apparent that the particles have been transported in an unlikely way. But of the many unlikely ways in which this could have happened, which one is the most likely? In modern probabilistic terms, this is a problem of *large deviations of the empirical distribution* as observed by Föllmer [31]. Thanks to Sanov's theorem [52], Schrödinger's problem can be turned into a maximum entropy problem for distributions on trajectories. Let $C([0, 1]; \mathbb{R}^n)$ be the space of $\mathbb{R}^n$ valued continuous functions and let W be Wiener measure on $C([0, 1]; \mathbb{R}^n))$. Then Sanov's theorem roughly asserts that the most likely random evolution between two given marginals is the solution of the Schrödinger Bridge Problem:

Problem 1.

$$\text{Minimize} \quad \mathbb{D}(P\|W) \quad \text{over} \quad P \in \mathcal{D}(\rho_0, \rho_1). \tag{2}$$

where $\mathcal{D}(\rho_0,\rho_1)$ are distributions on $C([0,1];\mathbb{R}^n)$ having marginal densities ρ_0 and ρ_1 at times $t=0$ and $t=1$, respectively, and

$$\mathbb{D}(P\|W)=\begin{cases}\mathbb{E}_P\left(\log\frac{\mathrm{d}P}{\mathrm{d}W}\right), & \text{if } P\ll W\\ +\infty & \text{otherwise}\end{cases}$$

is the relative entropy functional or Kullback-Leibler divergence between P and W. The optimal solution is called the *Schrödinger Bridge* between ρ_0 and ρ_1 over W, and its marginal flow (ρ_t) is the *entropic interpolation.* Two good surveys on Schrödinger bridges are [61, 42]. Let

$$W_x^y=W\left[\,\cdot\mid X_0=x,X_1=y\right]$$

be the disintegration W with respect to the initial and final positions. Then the solution of 2 can be shown [31] to have the form

$$P^*(\cdot)=\int_{\mathbb{R}^n\times\mathbb{R}^n}W_x^y(\cdot)\pi^*(x,y)dxdy,$$

where $\pi^*(x,y)$ is the joint initial-final time density under P^* solving the static problem:

Problem 2. *Given the joint initial-final time density π^W under W, minimize over densities π on $\mathbb{R}^n\times\mathbb{R}^n$ the index*

$$\mathbb{D}(\pi\|\pi^W)=\int\int\left[\log\frac{\pi(x,y)}{\pi^W(x,y)}\right]\pi(x,y)dxdy \tag{3}$$

subject to the (linear) constraints

$$\int\pi(x,y)dy=\rho_0(x),\qquad\int\pi(x,y)dx=\rho_1(y). \tag{4}$$

Consider now the case when the prior is W_γ, namely Wiener measure with variance γ, so that

$$p(0,x,1,y)=[2\pi\gamma]^{-\frac{n}{2}}\exp\left[-\frac{|x-y|^2}{2\gamma}\right].$$

It can be shown [42, 29] that the initial marginal density of the prior can WLOG always be taken equal to ρ_0 and that Problem 2 of minimizing $\mathbb{D}(\pi\|\pi^{\mathbf{W}_\gamma})$ over $\Pi(\rho_0,\rho_1)$, namely the "couplings" of ρ_0 and $\rho_1$is equivalent to

$$\inf_{\pi\in\Pi(\rho_0,\rho_1)}\int\frac{|x-y|^2}{2}\pi(x,y)\mathrm{d}x\mathrm{d}y+\gamma\int\pi(x,y)\log\pi(x,y)\mathrm{d}x\mathrm{d}y. \tag{5}$$

This is just a regularization of Optimal Mass Transport (OMT) [60] with quadratic cost.

2.2 The Schrödinger system

Using Lagrange multipliers for the linear constraints (4), Schrödinger showed that the optimal $\pi^*(\cdot,\cdot)$ in the form

$$\pi^*(x,y) = \hat{\varphi}(x)p(0,x,1,y)\varphi(y), \tag{6}$$

where φ and $\hat{\varphi}$ must satisfy

$$\begin{aligned} \hat{\varphi}(x)\int p(0,x,1,y)\varphi(y)dy &= \rho_0(x), &(7)\\ \varphi(y)\int p(0,x,1,y)\hat{\varphi}(x)dx &= \rho_1(y). &(8)\end{aligned}$$

Define $\hat{\varphi}_0(x) = \hat{\varphi}(x), \quad \varphi_1(y) = \varphi(y)$ and

$$\hat{\varphi}_1(y) := \int p(0,x,1,y)\hat{\varphi}_0(x)dx, \quad \varphi_0(x) := \int p(0,x,1,y)\varphi_1(y)dy.$$

Then, (7)-(8) can be replaced by the system

$$\begin{aligned} \hat{\varphi}_1(y) &= \int p(0,x,1,y)\hat{\varphi}_0(x)dx, &(9)\\ \varphi_0(x) &= \int p(0,x,1,y)\varphi_1(y)dy, &(10)\end{aligned}$$

coupled with the boundary conditions

$$\varphi_0(x)\cdot\hat{\varphi}_0(x) = \rho_0(x), \quad \varphi_1(y)\cdot\hat{\varphi}(1,y) = \rho_1(y). \tag{11}$$

Notice that dividing both sides of (6) by $\rho_0(x)$, we get

$$p^*(0,x,1,y) = \frac{1}{\varphi_0(x)}p(0,x,1,y)\varphi_1(y), \tag{12}$$

where φ, in Doob's language, is the *space time harmonic* satisfying (10) or, equivalently,

$$\frac{\partial\varphi}{\partial t} + \frac{1}{2}\Delta\varphi = 0. \tag{13}$$

The solution is namely obtained from the prior distribution via a *multiplicative functional transformation* of the prior Markov processes [37]. The question of existence and uniqueness of positive functions $\hat{\varphi}$, φ satisfying (9, 10, 11), left open by Schrödinger, is a highly nontrivial one and has been settled in various degrees of generality by Fortet, Beurlin, Jamison and Föllmer [32, 6, 38, 31], see also [42, Proposition 2.5]. The pair $(\varphi, \hat{\varphi})$ is unique up to multiplication of φ by a positive constant c and division of $\hat{\varphi}$ by the same constant. At each time t, the marginal ρ_t factorizes as

$$\rho_t(x) = \varphi_t(x)\cdot\hat{\varphi}_t(x). \tag{14}$$

Schrödinger observes: “Merkwürdige Analogien zur Quantenmechanik, die mir sehr des Hindenkens wert erscheinen.”[1] Indeed (14) resembles Born’s relation $\rho_t(x)=\psi_t(x)\cdot\bar{\psi}_t(x)$ with ψ and $\bar{\psi}$ satisfying two adjoint equations like φ and $\hat{\varphi}$. Moreover, the solution of Problem 2 enjoes the following remarkable *reversibility property*: Exchanging the two marginal densities ρ_0 and ρ_1, the new solution is the time reversal of the previous one. This explains the title “On the reversal of natural laws” of [53].

2.3 “Half bridges”

Consider the following variant of Problem 2 with prior distribution W_γ:

Problem 3.

$$\text{Minimize} \quad \mathbb{D}(P\|W_\gamma) \quad \text{over} \quad P\in\mathcal{D}(\rho_1), \tag{15}$$

namely, we only impose the final marginal. The same argument as before shows that Problem 15 reduces to the following variant of Problem 2:

Problem 4. *Minimize over densities π on $\mathbb{R}^n\times\mathbb{R}^n$ the index*

$$\mathbb{D}(\pi\|\pi^{W_\gamma})=\int\int\left[\log\frac{\pi(x,y)}{\pi^{W_\gamma}(x,y)}\right]\pi(x,y)dxdy \tag{16}$$

subject to the (linear) constraint

$$\int\pi(x,y)dx=\rho_1(y). \tag{17}$$

A simplified variational analysis, with μ as Lagrange multiplier for the constraint (17), gives the optimality condition

$$1+\log\pi^*(x,y)-\log p(0,x,1,y)-\log\rho_0^W(x)+\mu(y)=0,$$

where $\rho_0^W(x)$ is the initial marginal for the reference measure. We then get

$$\frac{\pi^*(x,y)}{p(0,x,1,y)}=\exp\left[\log\rho_0^W(x)-1-\mu(y)\right]=\rho_0^W(x)\exp\left[-1-\mu(y)\right]. \tag{18}$$

Thus, in the previous notation, we can set $\hat{\varphi}(x)=\rho_0^W(x)$ and $\varphi(y)=\exp\left[-1-\mu(y)\right]$. Let

$$\rho_1^{W_\gamma}(y)=\int[2\pi\gamma]^{-\frac{n}{2}}\exp\left[-\frac{|x-y|^2}{2\gamma}\right]\rho_0^W(x)dx$$

which replaces (9) with $\hat{\varphi}_0(x)=\rho_0^W(x)$ and $\hat{\varphi}_1(y)=\rho_1^{W_\gamma}(y)$.Then (8) gives immediately

$$\varphi(y)=\frac{\rho_1(y)}{\rho_1^{W_\gamma}(y)}. \tag{19}$$

[1] Remarkable analogies to quantum mechanics which appear to me very worth of reflection.

We now get the form of the optimal initial-final joint distribution of the half-bridge:

$$\begin{aligned}\pi^*(x,y) &= \rho_0^W(x)p(0,x,1,y)\frac{\rho_1(y)}{\rho_1^{W_\gamma}(y)}\\ &= \rho_0^W(x)\,[2\pi\gamma]^{-\frac{n}{2}}\exp\left[-\frac{|x-y|^2}{2\gamma}\right]\frac{\rho_1(y)}{\rho_1^{W_\gamma}(y)} = \pi^{W_\gamma}(x,y)\frac{\rho_1(y)}{\rho_1^{W_\gamma}(y)}.\end{aligned}$$

Finally, let

$$\varphi_0(y) := \int (2\pi)^{-\frac{n}{2}}\,[2\pi\gamma]^{-\frac{n}{2}}\exp\left[-\frac{|x-y|^2}{2\gamma}\right]\varphi(y)dy. \tag{20}$$

Then, the initial marginal of the solution is given by

$$\rho_0(x) = \varphi_0(y)\rho_0^W(x).$$

Notice that here there is no delicate question about existence and uniqueness for the Schrödinger system as $\hat{\varphi}$ coincides at all times with the prior one-time marginal. This, in turn, provides the terminal condition for the φ function at time $t = 1$ which then only needs to be propagated backward through (20) to provide the full solution. In the special case when $\rho_0^W(x) = \delta(x)$, we have $\rho_t^{W_\gamma}(x) = (2\pi\gamma t)^{-\frac{n}{2\gamma}}\exp\left[-\frac{|x|^2}{2\gamma t}\right]$ and, in particular, $\rho_1^{W_\gamma}(y) = (2\pi\gamma)^{-\frac{n}{2}}\exp\left[-\frac{|y|^2}{2\gamma}\right]$.

An immediate application of the half-bridge problem is the reconstruction of the past of a system given its current state and a prior model for its evolution. The availability of a prior here is crucial, as without a prior or other regularization such inverse problems are typically ill-posed. Another application concerns deviations from equilibrium. Consider a stochastic system whose dynamics $p(t_1,x_1,t_2,x_2)$ has a statistically steady state $\rho_{eq}(\mathrm{x})$, possibly modulated in time. What is the most likely path that would take us at time t to a state $\rho_1(x)$ away from equilibrium? For example, one may want to anticipate the likely path of strong storms or large waves, so as to be able to forecast them.

2.4 Stochastic control and fluid-dynamic formulations

In addition to the formulations above, there exist also dynamic versions of the problem such as the following stochastic control formulation originating with [23, 24, 49]: Problem 2 (when the prior has variance γ) is equivalent to

Problem 5.

$$\begin{gathered}\text{Minimize}_{u\in\mathcal{U}}\ J(u) = \mathbb{E}\left[\int_0^1 \frac{1}{2\gamma}\|u_t\|^2 dt\right],\\ \text{subject to } dX_t = u_t dt + \sqrt{\gamma}dW_t,\quad X_0 \sim \rho_0(x)dx,\quad X_1 \sim \rho_1(y)dy,\end{gathered} \tag{21}$$

where the family $\mathcal{U}$ consists of adapted, finite-energy control functions.

The optimal control is of the feedback type

$$u_t = \gamma \nabla \log \varphi_t(X_t), \tag{22}$$

where $(\varphi, \hat{\varphi})$ solve the Schrödinger system (9, 10, 11). These formulations are particularly relevant in applications where the prior distribution on paths is not simply the Wiener measure, but is associated to the uncontrolled ("free") evolution of a dynamical system, see e.g [15, 16, 13] and in image morphing/interpolation [12, Subsection 5.3]. In the case of the half bridge, (22) still holds with φ satisfying

$$\frac{\partial \varphi}{\partial t} + \frac{\gamma}{2}\Delta\varphi = 0, \quad \varphi(1,\cdot) = \frac{\rho_1(\cdot)}{\rho_1^{\mathbf{W}_\gamma}(\cdot)}.$$

Problem 21 leads immediately to the following fluid dynamic problem:

Problem 6.

$$\inf_{(\rho,b)} \int_{\mathbb{R}^n} \int_0^1 \frac{1}{2}\|b(x,t)\|^2 \rho(t,x) dt dx, \tag{23a}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (b\rho) - \frac{\gamma}{2}\Delta\rho = 0, \tag{23b}$$

$$\rho(0,x) = \rho_0(x), \quad \rho(1,y) = \rho_1(y). \tag{23c}$$

where $b(\cdot,\cdot)$ varies over continuous functions on $\mathbb{R}^n \times [0,1]$. This problem is not equivalent to Problems 2, 2 and 21 in that it only reproduces the optimal *entropic interpolating flow* $\{\rho_t; 0 \leq t \leq 1\}$. Information about correlations at different times and smoothness of the trajectories is here lost. As $\gamma \searrow 0$, the solution to this problem converges to the solution of the Benamou-Brenier Optimal Mass Transport problem [4, 44, 45, 46, 42, 41]:

$$\inf_{(\rho,v)} \int_{\mathbb{R}^n} \int_0^1 \frac{1}{2}\|v(x,t)\|^2 \rho(t,x) dt dx, \tag{24a}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (v\rho) = 0, \tag{24b}$$

$$\rho(0,x) = \rho_0(x), \quad \rho(1,y) = \rho_1(y). \tag{24c}$$

Let (ρ, b) be optimal for Problem 6 and define the *current velocity field* [47]

$$\begin{aligned} v(x,t) &= b(x,t) - \frac{\gamma}{2}\nabla \log \rho_t(x) \\ &= \gamma \nabla \log \varphi_t(x) - \frac{\gamma}{2}\nabla \log \rho_t(x) = \frac{\gamma}{2}\nabla \log \frac{\varphi_t(x)}{\hat{\varphi}_t(x)}, \end{aligned} \tag{25}$$

where we have used (22) and (14). Assume that v guarantees existence and uniqueness of the initial value problem on $[0,1]$ for any deterministic initial condition and consider

$$\dot{X}(t) = v(X(t),t), \quad X(0) \sim \rho_0 dx. \tag{26}$$

Then the probability density $\rho_t(x)$ of $X(t)$ satisfies (weakly) the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (v\rho) = 0$$

as well as (23b) with the same initial condition and therefore coincides with $\rho(x,t)$. This suggests that an alternative fluid-dynamic problem characterizing the entropic interpolation flow $\{\rho_t; 0 \leq t \leq 1\}$ may be possible. Indeed, such time-symmetric problem was derived in [14]:

Problem 7.

$$\inf_{(\rho,v)} \int_{\mathbb{R}^n} \int_0^1 \left[\frac{1}{2}\|v(x,t)\|^2 + \frac{\gamma}{8}\|\nabla \log \rho\|^2\right] \rho(t,x) dt dx, \tag{27a}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (v\rho) = 0, \tag{27b}$$

$$\rho(0,x) = \rho_0(x), \quad \rho(1,y) = \rho_1(y). \tag{27c}$$

The two criteria differ by $(\gamma/8)\mathcal{I}(\rho)$ where the Fisher information functional $\mathcal{I}$ is given by

$$\mathcal{I}(\rho) = \int \|\nabla \log \rho_t\|^2 \rho_t(x) dx$$

while the Fokker-Planck equation (23b) has been replaced by the continuity equation (27b). Both Problems 6 and 7 can be thought of as regularizations of the Benamou-Brenier problem (24) and as dynamic counterparts of (5). Also notice that, precisely as in Problem (24), the optimal current velocity (25) in Problem 7 is of the gradient type.

Finally, consider the family of diffeomorphisms $\{T_t; 0 \leq t \leq 1\}$ satisfying

$$\frac{dT_t}{dt}(x) = v(T_t(x), t), \quad T_0 = I, \tag{28}$$

where v is defined by (25). Then, in analogy to the *displacement interpolation* of Optimal Mass Transport, we have the following relation for the entropic interpolation flow

$$\rho_t(x)dx = T_t \# \rho_0(x) dx, \tag{29}$$

namely $\rho_t(x)dx$ is the *push-forward* of the measure $\rho_0(x)dx$ under the map T_t. In particular, the map $T_\gamma = T_1$ pushes $\rho_0(x)dx$ onto $\rho_1(x)dx$ and represents therefore the entropic counterpart of the map solving the original Monge problem. It may be called the *Monge-Schrödinger map.*

2.5 Fortet's iterative algorithm

The oldest proof of existence and uniqueness for the Schrödinger system (9, 10, 11), due to Fortet [32], is *algorithmic* in nature, establishing convergence of successive approximations. More explicitly, let $g(x,y)$ be a nonnegative, continuous function bounded from above. Suppose $g(x,y) > 0$ except possibly for a zero measure set for each fixed value of x or of y. Suppose that $\rho_0(x)$ and $\rho_1(y)$ are continuous, nonnegative functions such that

$$\int \rho_0(x)dx = \int \rho_1(y)dy.$$

Suppose, moreover, that the integral

$$\int \frac{\rho_1(y)}{\int g(z,y)\rho_0(z)dz}dy$$

is finite. Then, [32, Theorem 1], the system

$$\begin{aligned} \phi(x)\int g(x,y)\psi(y)dy &= \rho_0(x), \qquad (30)\\ \psi(y)\int g(x,y)\phi(x)dx &= \rho_1(y) \qquad (31) \end{aligned}$$

admits a solution $(\phi(x), \psi(y))$ with $\phi \geq 0$ continuous and $\psi \geq 0$ measurable. Moreover, $\phi(x) = 0$ only where $\rho_0(x) = 0$ and $\psi(y) = 0$ only where $\rho_1(y) = 0$.

The result is proven by setting up a complex approximation scheme to show that equation

$$h(x) = \Omega(h) = \int g(x,y)\frac{\rho_1(y)dy}{\int g(z,y)\frac{\rho_0(z)}{h(z)}dz}. \qquad (32)$$

has a positive solution. Notice that

$$g(x,y) = p(0,x,1,y) = [2\pi\gamma]^{-\frac{n}{2}}\exp\left[-\frac{|x-y|^2}{2\gamma}\right].$$

satisfies all assumptions of Fortet's theorem. Uniqueness, in the sense described in Subsection 2.2, namely uniqueness of rays, is much easier to establish. In the recent paper [29], the bulk of Fortet's paper has been rewritten filling in all the missing steps and providing explanations for the rationale behind the various articulations of his approach.

Independently, at about the same time and in the discrete setting, an *iterative proportional fitting* (IPF) procedure, was proposed in the statistical literature on contingency tables [25]. Convergence for the IPF algorithm was first established (in a special case) by Richard Sinkhorn in 1964 [55]. The iterates were shortly afterwards shown to converge to a "minimum discrimination information" [36, 30, 51], namely to a minimum entropy distance. This line of research, usually called *Sinkhorn algorithms*, continues to this date, see e.g. [22, 2, 58]. It is apparent that an iterative scheme can be designed based on (32) which, in the previous notation, reads

$$\Omega(\varphi_0(x)) = \int p(0,x,1,y)\frac{\rho_1(y)dy}{\int p(0,z,1,y)\frac{\rho_0(z)}{\varphi_0(z)}dz}. \qquad (33)$$

This was accomplished in [12], showing convergence of the iterates in a suitable projective metric, but only for the case when both marginals have compact support.

Setting up an iterative scheme based on (33) when only samples of the two marginals are available is obviously much more challenging: This is the main topic of this paper which we shall pursue starting from the next section. This will also provide an approach to data-driven Optimal Mass Transport alternative to [59] since, as observed at the end of Subsection 2.1, the Schrödinger Bridge Problem may be viewed as a regularization of OMT.

3 Numerical methodology

This section develops a sample-based numerical methodology for the solution of the Schrödinger Bridge Problem. This is the case, ubiquitous in applications, where the distributions ρ_0 and ρ_1 are only known through the finite sample sets $\{x_i\}$ and $\{y_j\}$ of cardinality m and n respectively.

One could propose a scheme whereby one first estimates ρ_0 and ρ_1 from the samples provided, and then solves the regular Schrödinger Bridge Problem between these two estimates. Yet there are a number of reasons why a procedure based directly on the sample sets is preferable. In particular,

1. Density estimation adds an extra computational layer to the algorithm, and hence a source of additional potential approximation errors. Notice that the formula (6) for the posterior joint probability $\pi^*(x,y)$ does not involve the two marginal densities explicitly, only the potentials φ and $\hat{\varphi}$.

2. Even with estimations for ρ_0 and ρ_1 known in closed form, the solution to the Schrödinger Bridge Problem requires the calculation of integrals that in most cases cannot be performed in closed form.

For conciseness, we shall denote $p(y|x)$ the prior transition density $p(0,x,1,y)$, and write $\hat{\varphi}_0(x)$ and $\varphi_1(y)$ instead of $\hat{\varphi}(x)$ and $\varphi(y)$, respectively. Then (6) reads

$$\pi^*(x,y) = \hat{\varphi}_0(x)\ p(y|x)\ \varphi_1(y).$$

The entropic interpolation between ρ_0 and ρ_1 is given by $\rho_t(z) = \varphi_t(z)\hat{\varphi}_t(z)$, where

$$\varphi_t(z) = \int p(t,z,1,y)\varphi_1(y)dy \qquad \hat{\varphi}_t(z) = \int p(0,x,t,z)\hat{\varphi}_0(x)dx. \tag{34}$$

In particular, one needs to solve the system

$$\rho_0(x) = \varphi_0(x)\hat{\varphi}_0(x), \quad \rho_1(y) = \varphi_1(y)\hat{\varphi}_1(y),$$

with

$$\varphi_0(x) = \int p(y|x)\varphi_1(y)dy \qquad \hat{\varphi}_1(y) = \int p(y|x)\hat{\varphi}_0(x)dx.$$

To begin, we need to reformulate the problem so that it involves the distributions only through their available samples.

3.1 The half-bridge problem through maximal likelihood

We develop first an algorithm for the half-bridge problem. Even though this is much simpler than the full bridge, it includes some of its main ingredients. Recalling that $\hat{\varphi}_1$ is known (see Subsection 2.3), the equality $\hat{\varphi}_1(y)\varphi_1(y) = \rho_1(y)$ can be reformulated in a data-friendly way by minimizing the Kullback-Leibler divergence between $\rho_1(y)$ and $\hat{\varphi}_1(y)\varphi_1(y)$, leading to the following constrained optimization problem:

$$\varphi_1 = \arg\max_{\varphi_1(y)\geq 0} \int \log\left(\varphi_1(y)\right)\rho_1(y)\ dy, \qquad \int \left(\hat{\varphi}_1(y)\varphi_1(y)\right) dy = 1,$$

where two functions φ_1 are considered equal if they differ on a set of measure zero.

We can satisfy the positivity constraint automatically by proposing an exponential form for φ_1:

$$\varphi_1(y) = e^{g(y)},$$

which yields

$$\max_g \int g(y)\rho_1(y)dy \quad \text{s.t.} \quad \int \hat{\varphi}_1(y)e^{g(y)}dy = 1,$$

or, introducing a Lagrange multiplier λ for the constraint,

$$\max_g \min_\lambda L(g,\lambda) = \int g(y)\rho_1(y)\ dy - \lambda\left(\int \hat{\varphi}_1(y)e^{g(y)}dy - 1\right).$$

Maximizing over g first yields

$$\frac{\delta L}{\delta g} = \rho_1(y) - \lambda\hat{\varphi}_1(y)e^{g(y)} = 0 \quad \text{resulting in} \quad g(y) = \log\left(\frac{\rho_1(y)}{\lambda\hat{\varphi}_1(y)}\right).$$

Then the minimization over λ becomes

$$\min_\lambda\left[-\log(\lambda) + \lambda\right] \Rightarrow \lambda = 1.$$

Hence the value of the optimal λ is known explicitly, and the estimation problem becomes:

$$\max_g L(g) = \int g(y)\rho_1(y)\ dy - \int \hat{\varphi}_1(y)e^{g(y)}dy + 1. \tag{35}$$

Notice that the solution to (35) is

$$g(y) = \log\left(\frac{\rho_1(y)}{\hat{\varphi}_1(y)}\right) \Rightarrow \varphi_1(y) = \frac{\rho_1(y)}{\hat{\varphi}_1(y)},$$

the exact answer to the problem. Yet in the true problem $\rho_1(y)$ is only known through samples $\{y_j\}$, so the first integral in (35) must be replaced by its empirical counterpart:

$$\int g(y)\rho_1(y)\ dy \to \frac{1}{n}\sum_j g(y_j).$$

Then, introducing a rough estimate $\tilde{\rho}_1$ of ρ_1 that one can sample, such as a Gaussian, and drawing $\tilde{n}$ samples $\tilde{y}_k$ from it, we can replace the second integral above by its Monte Carlo simulation:

$$\int \hat{\varphi}_1(y)e^{g(y)}dy = \int \frac{\hat{\varphi}_1(y)e^{g(y)}}{\tilde{\rho}_1(y)}\tilde{\rho}_1(y)dy \to \frac{1}{\tilde{n}}\sum_k \frac{\hat{\varphi}_1(\tilde{y}_k)e^{g(\tilde{y}_k)}}{\tilde{\rho}_1(\tilde{y}_k)}.$$

(Notice that for $\tilde{\rho}_1 = \rho_1$ and g the true maximizer, this is an estimation with zero variance.) Regarding the use of Monte Carlo in high dimensions, one must be aware of the results derived in [11] and [1], showing that the sample size needed to obtain accurate importance sampling estimates can grow exponentially with the dimension.

Finally, proposing a parameterization of the unknown $g(y)$, such as

$$g(y,\beta) = \sum_l \beta_l F_l(y),$$

where the F_l are functions externally provided, we end up with the following algorithm for estimating $\varphi_1(y)$. Let $\varphi_1(y,\beta)$ be the parametrization of $\varphi_1(y)$ given by

$$\varphi_1(y,\beta) = e^{\sum_l \beta_l F_l(y)},$$

where β solves

$$\beta = \arg\max L = \sum_l \left(\frac{1}{n}\sum_j F_l(y_j)\right)\beta_l - \frac{1}{\tilde{n}}\sum_k \frac{\hat{\varphi}_1(\tilde{y}_k)e^{\sum_l \beta_l F_l(\tilde{y}_k)}}{\tilde{\rho}_1(\tilde{y}_k)}.$$

Notice that L is concave, since the

$$\frac{\partial^2 L}{\partial\beta_i\beta_j} = -\frac{1}{\tilde{n}}\sum_k \frac{\hat{\varphi}_1(\tilde{y}_k)e^{\sum_l \beta_l F_l(\tilde{y}_k)}}{\tilde{\rho}_1(\tilde{y}_k)}F_i(\tilde{y}_k)F_j(\tilde{y}_k)$$

form a negative definite matrix.

More generally, we could have adopted a parametrization $\varphi_1(y,\beta)$ for $\varphi_1(y)$ different from the exponential, while still guaranteeing positivity, such as

$$\varphi_1(y,\beta) = g(y,\beta)^2,$$

where $g(y,\beta)$ is any family of real functions with parameters β. Then the problem above would have become

$$\beta = \arg\max L = \frac{1}{n}\sum_j \log(\varphi(y_j,\beta)) - \frac{1}{\tilde{n}}\sum_k \frac{\hat{\varphi}_1(\tilde{y}_k)\varphi\left(\tilde{y}_k,\beta\right)}{\tilde{\rho}_1(\tilde{y}_k)}.$$

3.2 The full bridge problem

Since the solution of the Schrödinger problem is given in (6) by

$$\pi^*(x,y) = \hat{\varphi}_0(x)p(y|x)\varphi_1(y),$$

it is natural to parameterize in closed form only the functions $\hat{\varphi}_0(x)$ and $\varphi_1(x)$. As in the half-bridge problem, we guarantee the positivity of these two functions directly through their parameterization $\hat{\varphi}_0(x,\hat{\beta}), \varphi_1(y,\beta)$, for instance writing them as the exponential or square of some other real functions.

If $\hat{\varphi}_1$ were given, we would find the coefficients β defining φ_1 by solving an optimization problem entirely analogous to the half-bridge problem before:

$$\beta = \arg\max L_1 = \frac{1}{n}\sum_j \log\left(\varphi_1(y_j,\beta)\right) - \int \hat{\varphi}_1(y)\varphi_1(y,\beta)dy.$$

However, at every step in the algorithm, only $\hat{\varphi}_0(x)$ is available in closed form; in order to find $\hat{\varphi}_1(y)$ we need to propagate the former through

$$\hat{\varphi}_1(y) = \int p(y|x)\hat{\varphi}_0(x,\hat{\beta})dx.$$

Then

$$\begin{aligned}\int \hat{\varphi}_1(y)\varphi_1(y,\beta)dy &= \int \left[\int p(y|x)\hat{\varphi}_0(x,\hat{\beta})dx\right]\varphi_1(y,\beta)dy \\ &= \int \left[\int p(y|x)\varphi_1(y,\beta)dy\right]\hat{\varphi}_0(x,\hat{\beta})dx.\end{aligned}$$

Since the inner integral equals $\varphi_0(x)$, and $\varphi_0(x)\hat{\varphi}_0(x) = \rho_0(x)$, we can multiply and divide by a sampleable estimator $\tilde{\rho}_0$ of ρ_0 with $\tilde{m}$ samples $\{\tilde{x}_i\}$, and write

$$\int \hat{\varphi}_1(y)\varphi_1(y,\beta)dy \approx \frac{1}{\tilde{m}}\sum_i \left[\int p(y|\tilde{x}_i)\varphi_1(y,\beta)dy\right]\frac{\hat{\varphi}_0(\tilde{x}_i,\hat{\beta})}{\tilde{\rho}_0(\tilde{x}_i)},$$

an estimation with zero variance at the exact solution if $\tilde{\rho}_0 = \rho_0$. Since the $\tilde{x}_i$ are fixed throughout the algorithm, we can at little expense extract, for each i, $\hat{n}$ samples $\hat{y}_i^j$ from the prior $p(y|\tilde{x}_i)$, and write the final estimator

$$\int \hat{\varphi}_1(y)\varphi_1(y,\beta)dy \approx \frac{1}{\tilde{m}\hat{n}}\sum_{i,j}\varphi_1(\hat{y}_i^j,\beta)\frac{\hat{\varphi}_0(\tilde{x}_i,\hat{\beta})}{\tilde{\rho}_0(\tilde{x}_i)},$$

so the problem for β becomes

$$\beta = \arg\max \frac{1}{n}\sum_j \log\left(\varphi_1(y_j,\beta)\right) - \frac{1}{\tilde{m}\hat{n}}\sum_{i,j}\varphi_1(\hat{y}_i^j,\beta)\frac{\hat{\varphi}_0(\tilde{x}_i,\hat{\beta})}{\tilde{\rho}_0(\tilde{x}_i)}. \qquad (36)$$

For the parameters $\hat{\beta}$, we have

$$\hat{\beta} = \arg\max L_0 = \frac{1}{m}\sum_i \log\left(\hat{\varphi}_0(x_i,\hat{\beta})\right) - \int \hat{\varphi}_0(x,\hat{\beta})\varphi_0(x)dx,$$

where

$$\varphi_0(x) = \int p(y|x)\varphi_1(y,\beta)dy.$$

Then

$$\begin{aligned}\int \hat{\varphi}_0(x,\hat{\beta})\varphi_0(x)dx &= \int \left[\int p(y|x)\varphi_1(y,\beta)dy\right]\hat{\varphi}_0(x,\hat{\beta})dx \\ &\approx \frac{1}{\tilde{m}\hat{n}}\sum_{i,j}\varphi_1(\hat{y}_i^j,\beta)\frac{\hat{\varphi}_0(\tilde{x}_i,\hat{\beta})}{\tilde{\rho}_0(\tilde{x}_i)}.\end{aligned}$$

(The fact that this is exactly the same estimation as for the integral $\int \hat{\varphi}_1(y)\varphi_1(y,\beta)dy$ should not be entirely surprising, as both equal one and involve the same parameters.) Finally,

$$\hat{\beta} = \arg\max \frac{1}{m}\sum_i \log\left(\varphi_0(x_i,\hat{\beta})\right) - \frac{1}{\tilde{m}\hat{n}}\sum_{i,j}\varphi_1(\hat{y}_i^j,\beta)\frac{\hat{\varphi}_0(\tilde{x}_i,\hat{\beta})}{\tilde{\rho}_0(\tilde{x}_i)}. \tag{37}$$

3.3 The algorithm

Summarizing the results above, we have developed the following algorithm:

1. **Data:** We are provided with m samples $\{x_i\}$ of $\rho_0(x)$, n samples $\{y_j\}$ of $\rho_1(y)$, and a prior conditional probability density $p(y|x)$. The latter needs not be known in closed form, but one should be able to sample it for any value of x (if the opposite is true, i.e. we know $p(y|x)$ in closed form but cannot sample it, an alternative algorithm presented below should be applied.)

2. **Goal:** To find the most likely joint distribution $\pi(x,y)$ under the prior $p(y|x)$ consistent with the two marginals, and the corresponding posterior $p^*(y|x)$. When $p(y|x)$ is the end result of the prior $p(t_1,x,t_2,y)$ for a time dependent process, we also seek the more detailed posterior $p^*(t_1,x,t_2,y)$ for this process, as well as the intermediate distributions $\rho_t(z)$ for $t \in [0,1]$.

3. **Preliminary work:** Based on the samples $\{x_i\}$, we need to produce a first estimate $\tilde{\rho}_0$ of $\rho_0(x)$ and $\tilde{m}$ independent samples $\{\tilde{x}_i\}$ drawn from it. More specifically, we will need these $\tilde{m}$ samples and the values $\tilde{\rho}_0(\tilde{x}_i)$ of $\tilde{\rho}_0$ on them. For instance, one can use the Gaussian kernel density estimator

$$\tilde{\rho}_0(x) = \frac{1}{m}\sum_i G(x - x_i),$$

where G is an isotropic Gaussian with suitable bandwidth. In building this estimate, we can use, in addition to the samples $\{x_i\}$, any additional prior information that we may have on $\rho_0(x)$. For instance, its support may be known to be contained within some set Ω, typically not to include unrealistic negative values of some components of x. One simple way to address this particular case is to multiply the unconstrained estimator $\tilde{\rho}_0$

by the characteristic function of Ω, reject any sample outside of Ω, and normalize the resulting distribution through division by the factor

$$\frac{\tilde{m}+m_r}{\tilde{m}},$$

where m_r is the total number of rejections that occurred.

For each sample $\tilde{x}_i$, we need to produce $\hat{n}$ samples $\hat{y}_i^j$ drawn independently from $p(y|\tilde{x}_i)$. For instance, if p is the result of a diffusive process between $t=0$ and $t=1$, with drift $u(x,t)$ and diffusivity $\nu(x,t)$, we would simulate the stochastic process

$$dx = u(x,t)dt + \nu(x,t)dW, \quad x(0) = x_i, \quad y_i^j = x(1).$$

If $p(y|x)$ is known in closed form but is not easily sampled, one can propose another conditional probability $q(y|x)$ not very far from p but sampleable, and produce weighted samples y_i^j from $q(y|\tilde{x}_i)$, with weights

$$w_i^j = \frac{p(y_i^j|\tilde{x}_i)}{q(y_i^j|\tilde{x}_i)},$$

to be included as extra factors under the second sum in problems (36) and (37).

4. **Model selection and initialization:** We need to propose a parametric family of non-negative real functions $\varphi(z,\beta)$. Examples are

$$\varphi(z,\beta) = e^{\sum_k \beta_k F_k(z)} \quad \text{and} \quad \varphi(z,\beta) = \left(\sum_k \beta_k F_k(z)\right)^2, \qquad (38)$$

where the F_k are a given set of functions (monomials, Legendre functions, sines and cosines, splines, etc.) In high dimensions, we may want to use instead a low-rank tensor factorization as in [35, 56]. The final estimated joint density will adopt the form

$$\pi(x,y) = \varphi(x,\hat{\beta})\ p(y|x)\ \varphi(y,\beta),$$

and the estimated posterior conditional probability will be

$$P^*(y|x) = \frac{p(y|x)\ \varphi(y,\beta)}{\int p(z|x)\ \varphi(z,\beta)dz},$$

where the integral in the denominator can be estimated for each desired value of x by simulating $p(z|x)$. In the notation above,

$$\hat{\varphi}_0(x) = \varphi(x,\hat{\beta}) \quad \text{and} \quad \varphi_1(y) = \varphi(y,\beta).$$

We initialize the algorithm with an initial guess for β, such as the β that yields the default $\varphi_1(y) = 1$ (i.e. $\beta = 0$ when using the first of the parametrizations in (38). This is typically easier than starting with a guess for $\hat{\beta}$ approximating the corresponding default $\hat{\varphi}_0(x) = \rho_0(x)$). When using the quadratic parametrization in (38), we start with a choice of $\hat{\beta}$ that, depending on the chosen basis functions F_l, yields to the biggest effective support of $\phi(x)$.

5. **Main loop:** We alternate between the updates (37) for $\hat{\beta}$ and (36) for β iteratively until a convergence criterion is met. Some choices for the family $\varphi(z, \beta)$, such as

$$\varphi(z, \beta) = e^{\sum_k \beta_k F_k(z)}$$

yield automatically convex optimization problems for $\hat{\beta}$ and β.

4 Numerical examples

This section illustrates the proposed methodology on two examples relevant in applications: the interpolation of probability distributions, and a variation on importance sampling in the context of Monte Carlo estimates of integrals.

4.1 Interpolation between two Gaussian mixtures

Figure 1 displays the two marginal distributions of a two dimensional numerical example, where ρ_0 and ρ_1 are Gaussian mixtures given by

$$\rho_0 = \frac{1}{3} \sum \left[\mathcal{N}(\mu_1, \Sigma_1) + \mathcal{N}(\mu_2, \Sigma_2) + \mathcal{N}(\mu_3, \Sigma_3)\right]$$

$$\rho_1 = \frac{1}{3} \left[\mathcal{N}(\mu_4, \Sigma_4) + \mathcal{N}(\mu_5, \Sigma_5) + \mathcal{N}(\mu_6, \Sigma_6)\right]$$

with parameters

$$\mu_1 = \begin{bmatrix} -2 \\ 1.5 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0.2 \\ 1.2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.6 & -0.4 \\ -0.4 & 0.6 \end{bmatrix},$$
$$\mu_3 = \begin{bmatrix} 0.5 \\ -1 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 0.5 & 0.4 \\ 0.4 & 0.7 \end{bmatrix}, \mu_4 = \begin{bmatrix} -1.8 \\ 1.1 \end{bmatrix}, \Sigma_4 = \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix},$$
$$\mu_5 = \begin{bmatrix} -0.2 \\ 1.2 \end{bmatrix}, \Sigma_5 = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.8 \end{bmatrix} \mu_6 = \begin{bmatrix} -0.5 \\ 0.9 \end{bmatrix}, \Sigma_6 = \begin{bmatrix} 0.6 & 0.2 \\ 0.2 & 0.6 \end{bmatrix}. \quad (39)$$

Figure 2 displays the interpolation between ρ_0 and ρ_1 obtained by computing $\rho_t(z) = \varphi_t(z)\hat{\varphi}_t(z)$ for each time $t \in [0, 1]$ at the data points $z(t)$ obtained by integrating the equation (25, 26) with $\gamma = 2$. In this example, both φ and $\hat{\varphi}$ were represented as the square of linear combinations of the first 10 Hermite functions.

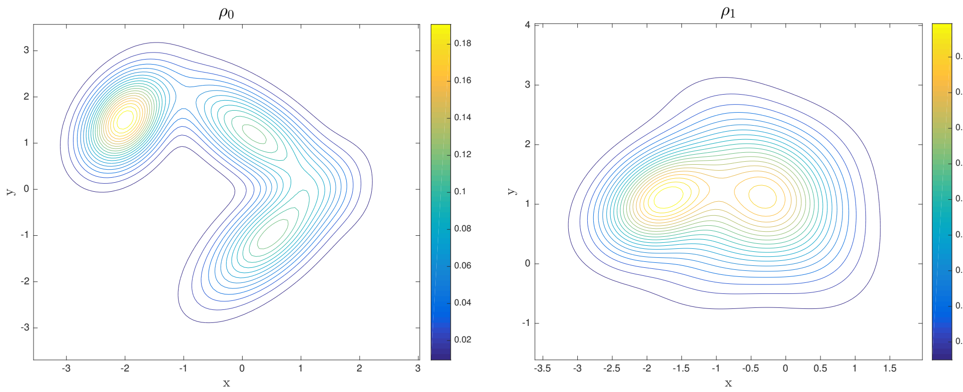


Figure 1: Initial and final probability density distribution from which points x_i and y_j where sampled respectively. This is the only input used by the algorithm.

4.2 A variation on importance sampling

The methodology of this article turns out to be particularly well suited to improve Monte Carlo estimates of the quantity

$$I = \int f(y)\rho_1(y)dy, \tag{40}$$

when $\rho_1(y)$ is only known through n sample points drawn from it. It is known that ordinary Monte Carlo estimates suffer of a slow convergence rate as a function of n. Moreover, when the support of f is localized in regions where the value of ρ_1 is small, we may have very few points where f is substantially different from zero. If ρ_1 where known in closed form, we could remedy these problems through importance sampling, whereby we would rewrite (40) in the form

$$I = \int f(y)\rho_1(y)dy = \int \frac{f(y)\rho_1(y)}{\mu(y)}\mu(y)dy,$$

where $\mu(y)$ is a distribution easy to sample and such that $f\rho_1/\mu$ has small variance (This variance can be made arbitrarily small when f has a definite sign. If f is bounded, one can always achieve this by adding a constant to f.) One then estimates I via Monte Carlo:

$$I \approx \frac{1}{n}\sum_{i=1}^{n} \frac{f(z_i)\rho_1(z_i)}{\mu(z_i)},$$

where the z_i are samples drawn independently from μ. Yet this procedure requires the capacity to evaluate ρ_1 at the given points. We are considering instead the frequently occurring situation where ρ_1 is only known through a fixed set of n samples $\{y_j\}$.

In this case, we propose to use the sample points $\{y_j\}$ to solve the Schrödinger Bridge Problem between $\rho_1(y)$ and a distribution $\rho_0(x)$ of our choice. This allows us to map arbitrary points in y-space to x-space. In particular, we can chose points $\tilde{y}_j$ that resolve f well, and use them to estimate the integral I through the following steps:

1. Sample points $\tilde{y}_j$ from a distribution $\rho_f(y)$ spanning the support of f. A simple choice is to draw them uniformly from the support of f if this is finite; another is to draw them form a sampleable estimate ρ_f for $|f|/(\int |f| dy)$. The distribution ρ_f must be chosen so that it can be both evaluated and sampled from. In the example below, we have adopted points from a uniform grid spanning the effective support of f, i.e. an interval outside of which $|f|$ was comparable to machine error. This corresponds to adopting a uniform ρ_f on the support of f.

2. Compute φ_1 and $\hat{\varphi}_0$ solving the Schrödinger bridge between $\rho_1(y)$ and any chosen distribution $\rho_0(x)$ through the procedure described in section 3.3. The distribution ρ_0 can be selected arbitrarily; in the examples below we have used a standard Gaussian.

3. For each point $\tilde{y}_j$ obtained in the first step, one would like to sample $P(x|\tilde{y}_j)$. Since
$$P(x|y) = P_\gamma(x|y)\frac{\hat{\varphi}_0(x)}{\hat{\varphi}_1(y)}$$
and
$$\hat{\varphi}_1(y) = \int P_\gamma(y|x)\hat{\varphi}_0(x)dx = \int P_\gamma(x|y)\hat{\varphi}_0(x)dx,$$
one draws Q samples $P_\gamma(x|\tilde{y}_j)$ instead and assign to each such sample x_j^l a weight $q_j^l = Q\frac{\hat{\varphi}_0(x_j^l)}{\sum_l \hat{\varphi}_0(x_j^l)}$.

4. Perform a Gaussian mixture density estimation $\nu(x)$ of the distribution underlying the points x_j^l with weights $q_j^l * \rho_0(x_j^l) * |f(\tilde{y}_j)|/\rho_f(\tilde{y}_j)$. This can be achieved through a modified EM algorithm that takes the weights into account. By construction, once transferred back via the bridge to y-space, ν will approximate $|f|\rho_1$, as can be seen in expression (41) below. Then sample N new points $\tilde{x}_i$ from ν.

5. Now for each $\tilde{x}_i$ we would like to sample $P(y|\tilde{x}_i)$. Again, since
$$P(y|x) = P_\gamma(x|y)\frac{\varphi_1(y)}{\varphi_0(x)},$$
and
$$\varphi_0(x) = \int P_\gamma(y|x)\varphi_1(y)dy,$$
produce instead M samples y_h^i from $P_\gamma(y|\tilde{x}_i)$ and assign to each such sample a weight $w_h^i = M\varphi_1(y_h^i)/\sum_h \varphi_1(y_h^i)$.

$$I_{\mathrm{R}} = 0.09894$$
$$I_{\mathrm{MC}} = 0.09941 \pm 0.0099117$$
$$I_{\mathrm{S}} = 0.09888 \pm 0.0014477$$

Table 1: I_{R} indicates the reference value for I, I_{MC} is the Monte Carlo estimates of I and I_{S} is the estimate of I obtained with the procedure described above. I_{MC} and I_S are computed by averaging 100 runs each one obtained using a $N = 1000$ sample points from $\rho_1(y)$. The error is estimated by estimating the standard deviation over 100 runs.

The integral in (40) is then estimated through

$$\int f(y)\rho_1(y)dy = \int f(y)\left[\int \rho_0(x)P(y|x)dx\right] dy = \int\int f(y)\frac{\rho_0(x)P(y|x)}{\nu(x)}\nu(x)\ dy\ dx \approx \approx \frac{1}{NM}\sum_{i,h} w_h^i f(y_h^i)\frac{\rho_0(\tilde{x}_i)}{\nu(\tilde{x}_i)}. \quad (41)$$

Hence we have used the Schrödinger bridge to transfer importance sampling from y to the auxiliary x-space. Notice that, by construction, ν is roughly proportional to $|f| * \rho_0$ and w is close to 1, so this estimate has small variance when f has a definite sign.

In the numerical experiment in Figure 3, we chose ρ_1 to be the equal weight mixture of the three Gaussians: $\mathcal{N}(-1.4, 0.8^2), \mathcal{N}(2.2, 0.4^2), \mathcal{N}(0.2, 0.1^2)$, and $f(y)$ a mixture of the two Gaussians $\mathcal{N}(-0.8, 0.02^2)$, $\mathcal{N}(1, 0.03^2)$, again with equal weights. We compute the reference value I_{R} for the integral $I = \int f(y)\rho_1(y)dy$ using a uniform grid of step size $h = 10^{-4}$ and compare, over 100 independent evaluations of I, this value with plain MC estimates of I obtained with 1000 points sampled from ρ_1 and with our procedure. As it can be seen from Table 1, the procedure described above gives a better estimates in terms of both the error with respect the reference value and the uncertainty associated with the estimate. Since in practice one has access only to one sample set of $\rho_1(y)$, not the 100 we have displayed here, the relevant numbers to use to compare straightforward MC and our bridge-based procedure is the mean square error

$$e_{\mathrm{MC}} = \frac{1}{100}\sqrt{\sum_i \left(I^i_{\mathrm{MC}} - I_{\mathrm{R}}\right)^2} = 0.0099228,$$

$$e_{\mathrm{S}} = \frac{1}{100}\sqrt{\sum_i \left(I^i_{\mathrm{S}} - I_{\mathrm{R}}\right)^2} = 0.0014489.$$

The fact that e_{S} is more than 6 times smaller than e_{MC} shows that the procedure does indeed improve the estimation significantly, much as conventional importance sampling does for distributions known in close form.

Other procedures for mapping a known distribution into another known only from samples have been proposed in the literature, see for instance [59], where a data-driven dual formulation of optimal transport is used, extended in [33] to the entropic regularized case. Yet to the best of our knowledge they have not been combined with a procedure for importance sampling. In principle, the procedure that we propose here for importance sampling can be modified so that it is applicable to those scenarios as well.

5 Conclusions

In this article, we have posed the sample-based Schrödinger Bridge Problem and developed a methodology for its numerical solution. Characterizing the initial and final distributions of the bridge in terms of samples is well-suited for applications and also natural from a theoretical perspective, since what is a large-deviation problem for a large but finite set of particles becomes a true impossibility as the number of particles grows unboundedly. One must distinguish though between the sample-based formulation, where $\{x_i\}$ and $\{y_j\}$ are regarded as samples of underlying distributions ρ_0 and ρ_1, from the discrete Schrödinger problem, where the latter are replaced by the empirical distributions $\frac{1}{m}\sum_{i=1}^{m}\delta\left(x-x_i\right)$ and $\frac{1}{n}\sum_{j=1}^{n}\delta\left(y-y_j\right)$. This article studies the former, finding the joint distribution $\pi^*(x,y)$ for all values of (x,y), not just the sample points, and characterizing the intermediate distributions $\rho_t(z)$ also for all z.

The methodology of this article mimics the iterative scheme developed for the classical bridge problem, but replacing some of its key ingredients by data analogues. Thus the boundary conditions at $t=0$ and $t=1$ are re-interpreted in a maximum likelihood sense, thus giving rise to optimization problems, and the integrals defining the propagation of the two factors of ρ_t are estimated via importance sampling.

The data-based Schrödinger problem has a broad scope of applicability. Potential applications include the estimation of atmospheric winds and oceanic currents from tracers, the solution of inverse diffusive problems, the reconstruction of the intermediate evolution of species between well-documented stages, and many more. However, these applications require further development of the procedure. In particular, they require the ability to sample from the transition probability associated to the backward prior process. Even though this is straightforward when the prior is the Wiener process, many applications require more general priors, which require an extension of the presented methodology, currently under development. Since this article is concerned with the development of a general methodology, we have not dwelled into any application in particular, but just illustrated the procedure with two relatively simple examples.

Acknowledgments

M.P. would like to thank the Courant Institute of Mathematical Sciences of the New York University for the hospitality during the time this research was carried out. Tabak's work was partially supported by NSF grant DMS-1715753 and ONR grant N00014-15-1-2355.

References

[1] Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., Stuart, A., et al. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017.

[2] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, 31, 1961–1971, 2017.

[3] Sigurd Angenent, Steven Haker, and Allen Tannenbaum. Minimizing flows for the Monge-Kantorovich problem. *SIAM J. Math. Anal.*, 35(1):61–97, 2003.

[4] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[5] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[6] Arne Beurling. An automorphism of product measures. *Annals of Mathematics*, 72:189–200, 1960.

[7] L Boltzmann. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung resp. den Sätzen über das Wärmegleichgewicht. Wiener Berichte 76, 373-435, 1877. *Reprinted in F. Hasenoehrl (ed.): Wissenschaftliche Abhandlungen.* Leipzig: J. A. Barth 1909, Vol. 2, 164-223.

[8] John Parker Burg. Maximum entropy spectral analysis. In *37th Annual International Meeting Soc. of Explor. Geophys.*, Oklahoma City, Okla., Oct. 31, 1967, 1967. Reprinted in *Modern SpectrumAnalysis*, D. G. Childers, Ed. New York: IEEE Press, 34–41, 1978.

[9] John Parker Burg, David G Luenberger, and Daniel L Wenger. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9):963–974, 1982.

[10] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.*, 49(2):1385–1418, 2017.

[11] Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2): 1099-1135, 2018.

[12] Yongxin Chen, Tryphon Georgiou, and Michele Pavon. Entropic and displacement interpolation: a computational approach using the Hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396, 2016.

[13] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Fast cooling for a system of stochastic oscillators. *Journal of Mathematical Physics*, 56(11):113302, 2015.

[14] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2):671–691, 2016.

[15] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal steering of a linear stochastic system to a final probability distribution, Part I. *IEEE Transactions on Automatic Control*, 61(5):1158–1169, 2016.

[16] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal steering of a linear stochastic system to a final probability distribution, Part II. *IEEE Transactions on Automatic Control*, 61(5):1170–1180, 2016.

[17] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal transport over a linear dynamical system. *IEEE Transactions on Automatic Control*, 62(5):2137–2152, 2017.

[18] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *Mathematics of computation*, 87 (314):2563–2609, 2018.

[19] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[20] Imre Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, 12(3): 768–793, 1984.

[21] Imre Csiszar. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

[22] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems* 26, 2292–2300, 2013.

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

[24] Paolo Dai Pra and Michele Pavon. On the Markov processes of Schrödinger, the Feynman-Kac formula and stochastic control. In *Realization and Modelling in System Theory - Proc. 1989 MTNS Conf.*, M.A.Kaashoek, J.H. van Schuppen, A.C.M. Ran Eds., Birkäuser, 497–504, 1990.

[25] W Edwards Deming and Frederick F Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[26] Arthur P Dempster. Covariance selection. *Biometrics*, 28(1): 157–175, 1972.

[27] Arnaud Dessein, Nicolas Papadakis, and Charles-Alban Deledalle. Parameter estimation in finite mixture models by regularized optimal transport: A unified framework for hard and soft clustering. *arXiv preprint arXiv:1711.04366*, 2017.

[28] Richard S Ellis. *Entropy, large deviations, and statistical mechanics.* Springer, 2007.

[29] Montacer Essid and Michele Pavon. Traversing the Schrödinger bridge strait: Robert Fortet's marvelous proof redux. *J. Optimiz. Theory Appl.*, 181(1): 23–60, 2019.

[30] Stephen E Fienberg. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3): 907–917, 1970.

[31] Hans Föllmer. Random fields and diffusion processes. In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, pages 101–203. Springer, 1988.

[32] R Fortet. Résolution d'un système d'equations de M. Schrödinger. *J. Math. Pure Appl.*, IX: 83–105, 1940.

[33] Aude Genevay, Marco Cuturi, Gabriel Peyré and Francis Bach, Stochastic Optimization for Large-scale Optimal Transport. *Advances in neural information processing systems* 29, 3440–3448, 2016.

[34] Tryphon T Georgiou and Michele Pavon. Positive contraction mappings for classical and quantum Schrödinger systems. *Journal of Mathematical Physics*, 56(3):033301, 2015.

[35] Thomas Gerstner and Michael Griebel. Numerical integration using sparse grids. *Numerical algorithms*, 18(3-4):209, 1998.

[36] C Terrance Ireland and Solomon Kullback. Contingency tables with given marginals. *Biometrika*, 55(1):179–188, 1968.

[37] Kiyosi Ito and Shinzo Watanabe. Transformation of Markov processes by multiplicative functionals. *Ann. Inst. Fourier*, 15(1):13–30, 1965.

[38] Benton Jamison. The Markov processes of Schrödinger. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32(4):323–331, 1975.

[39] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[40] Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

[41] Christian Léonard. From the Schrödinger problem to the Monge-Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[42] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst. A*, 2014, 34(4): 1533–1574, 2014. .

[43] Wuchen Li, Penghang Yin, and Stanley Osher. Computations of optimal transport distance with Fisher information regularization. *Journal of Scientific Computing*, 75(3):1581–1595, 2018.

[44] Toshio Mikami. Monges problem with a quadratic cost by the zero-noise limit of h-path processes. *Probability theory and related fields*, 129(2):245–260, 2004.

[45] Toshio Mikami and Michèle Thieullen. Duality theorem for the stochastic optimal control problem. *Stochastic processes and their applications*, 116(12):1815–1835, 2006.

[46] Toshio Mikami and Michèle Thieullen. Optimal transportation problem by stochastic optimal control. *SIAM Journal on Control and Optimization*, 47(3):1127–1139, 2008.

[47] Edward Nelson. *Dynamical theories of Brownian motion.* Princeton university press, 1967.

[48] Michele Pavon and Francesco Ticozzi. Discrete-time classical and quantum markovian evolutions: Maximum entropy problems on path space. *Journal of Mathematical Physics*, 51(4):042104, 2010.

[49] Michele Pavon and Anton Wakolbinger. On free energy, stochastic control, and Schrödinger processes. In *Modeling, Estimation and Control of Systems with Uncertainty*, pages 334–348. Birkhäuser, 1991.

[50] Gabriel Peyre and Marco Cuturi. Computational optimal transport. *arXiv preprint arXiv:1803.00567*, 2018.

[51] L. Ruschendorf, Convergence of the Iterative Proportional Fitting Procedure, em Ann. Statist., **23** (4), 1160–1174, 1995.

[52] Ivan N Sanov. On the probability of large deviations of random magnitudes (in Russian) *Mat. Sb. N. S.*, **42** (84) (1957), 11-44, Select. Transl. Math. Statist. Probab., 1, 213–244, 1961.

[53] Erwin Schrödinger. Über die Umkehrung der Naturgesetze. *Sitzungsberichte der Preuss Akad. Wissen. Berlin, Phys. Math. Klasse*, 144-153, 1931.

[54] Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Ann. Inst. H. Poincaré*, 2(4):269–310, 1932.

[55] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964.

[56] Esteban G Tabak and Giulio Trigila. Conditional expectation estimation through attributable components. *Information and Inference: A Journal of the IMA*, 7(4): 727-754, 2018.

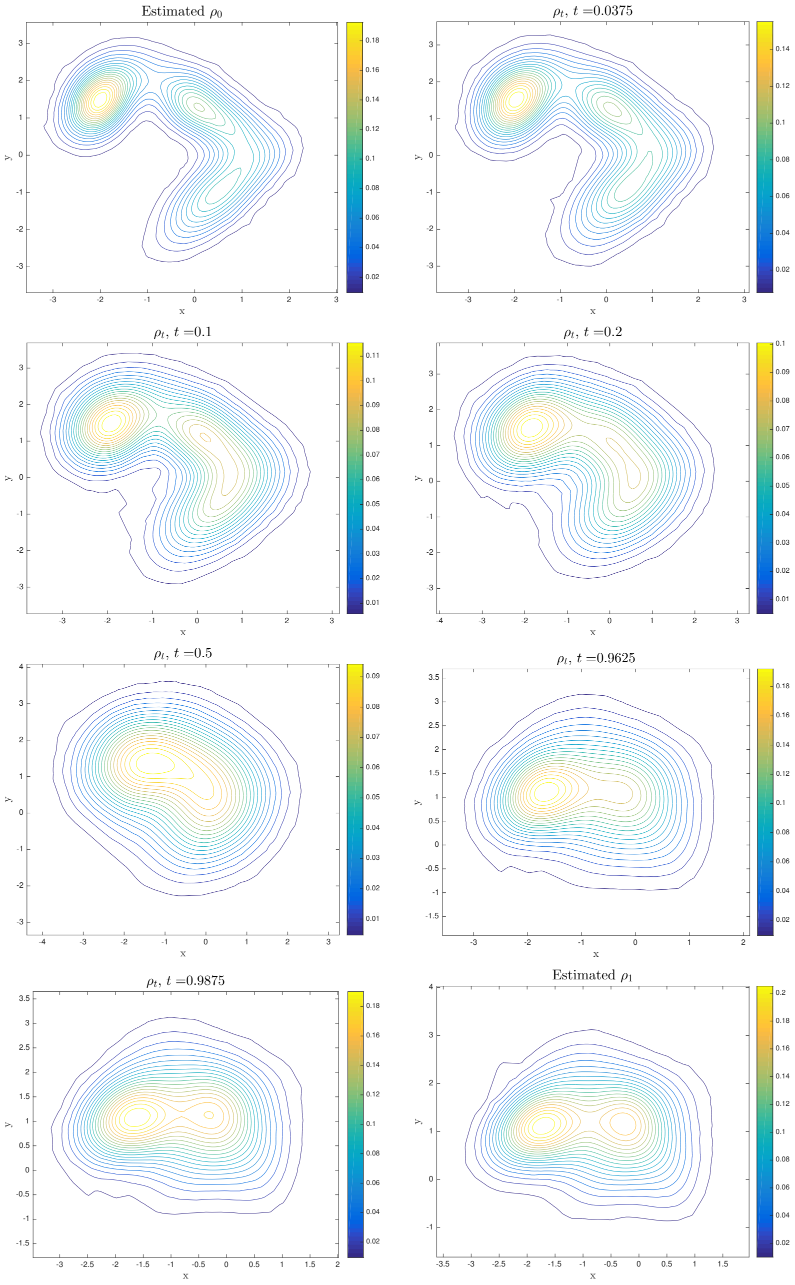
[57] Esteban G Tabak and Giulio Trigila. Explanation of variability and removal of confounding factors from data through optimal transport. *Communications on Pure and Applied Mathematics*, 71(1):163–199, 2018.

[58] Alexis Thibault, Lénaïc Chizat, Charles Dossal, and Nicolas Papadakis. Overrelaxed Sinkhorn-Knopp algorithm for regularized optimal transport. *arXiv preprint arXiv:1711.01851*, 2017.

[59] Giulio Trigila and Esteban G Tabak. Data-driven optimal transport. *Communications on Pure and Applied Mathematics*, 69(4):613–648, 2016.

[60] Cédric Villani. *Topics in optimal transportation.* Vol. 58. American Mathematical Soc., 2003.

[61] A Wakolbinger. Schrödinger bridges from 1931 to 1991. In *Proc. of the 4th Latin American Congress in Probability and Mathematical Statistics, Mexico City 1990, Contribuciones en probabilidad y estadistica matematica*, E. Cabaña et al. (eds), 3 (1992), pages 61–79.



Estimated ρ_0

ρ_t, t =0.0375

ρ_t, t =0.1

ρ_t, t =0.2

ρ_t, t =0.5

ρ_t, t =0.9625

ρ_t, t =0.9875

Estimated ρ_1

Figure 2: Interpolation between ρ_0 and ρ_1. Each image is obtained by interpolating $\rho_t(z)$ on the points $z(t)$ representing the solution of (25, 26). Both φ and $\hat{\varphi}$ were represented as the square of linear combinations of the first 10 Hermite functions.

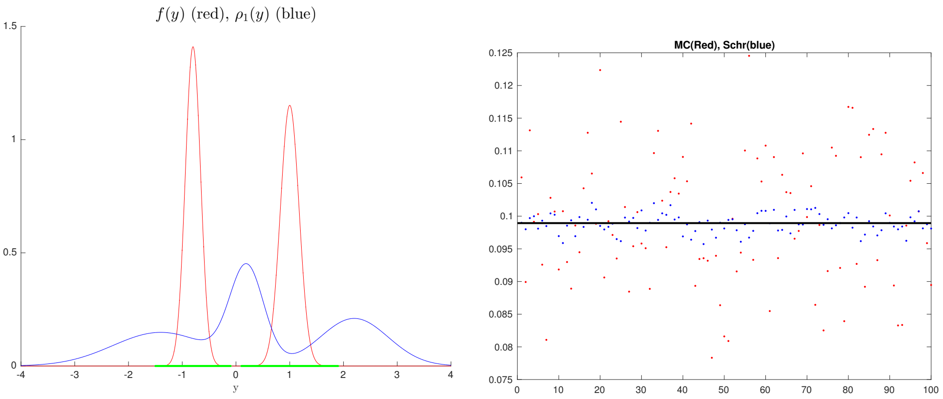


Figure 3: Left panel: The density $\rho_1(y)$ is plotted in blue while the function $f(y)$ is plotted in red. Notice that the support of $f(y)$ is substantially different from zero where the two local minima of ρ_1 are placed. The green points y_j, appearing one the x axis are points on a regular grid that were selected based on the value of f being bigger than a certain threshold. Right panel: estimates of I for 100 different sample sets from ρ_1 each one containing 1000 points. Results from the bridge-based procedure are in blue and from the plain MC simulation in red. The solid black line represents the true value of I.