LIKELIHOOD TRAINING OF SCHRÖDINGER BRIDGE USING FORWARD-BACKWARD SDEs THEORY

Roland Andrews, Dorian Gailhard, Vincent Garot, Julien Nguyen Van

June 27, 2023

Abstract

Your abstract.

1 Introduction

The Schrödinger Bridge (SB) problem, an entropy-regularized optimal transport formulation, has gained attention in deep generative modeling due to its mathematical flexibility compared to Score-Based Generative Models (SGM). However, the relationship between the SB optimization principle and modern deep generative model training based on log-likelihood objectives remains unclear. This raises questions about the suitability of SB models as principled alternatives for generative applications. In this report, we aim to explore the novel computational framework for likelihood training of SB models using Forward-Backward Stochastic Differential Equations Theory, introduced by [TC22].

2 Theory

2.1 Score-Based Generative Modeling

Score-Based Generative Modeling is a framework that addresses the task of generating realistic samples from a given distribution by matching the gradients of a model's log-likelihood with those of the target distribution. Unlike traditional generative models that aim to directly estimate the underlying density function, Score-Based Generative Models focus on capturing the score function, which provides the gradient information of the target distribution. By estimating the score function, these models can generate samples that capture the essential characteristics of the original data distribution.

The key advantage of Score-Based Generative Models lies in their flexibility and applicability to a wide range of data types. They do not require explicit density estimation, making them particularly suitable for complex and high-dimensional datasets. Score-Based Generative Models forward pass is based on the Euler-Maruyama discretization of the Ornstein-Ulhenbeck process :

$$\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2\gamma}\mathrm{d}\mathbf{B}_t$$

where $\mathbf{B}_t$ is the standard Wiener process.

2.2 The Schrodinger Bridge problem

The Schrödinger bridge problem, formulated by Erwin Schrödinger in 1931 [Sch32], addresses the task of transforming one probability distribution into another while minimizing a specific cost or distance metric. It combines concepts from optimal transport theory and quantum mechanics, offering a unique perspective on the efficient transformation of probability distributions. The problem involves finding the most probable path for transitioning from an initial probability density function (PDF) to a target PDF within a given time frame. By leveraging mathematical techniques such as functional optimization, partial differential equations, and stochastic processes, the Schrödinger bridge problem provides a principled approach to solving optimal transport problems in a quantum mechanical framework.

The Schrödinger bridge problem is rooted in the principles of quantum mechanics, incorporating both classical and quantum dynamics into the framework of optimal transport. It involves solving the Schrödinger equation, a partial differential equation that describes the evolution of quantum systems, to determine the optimal transition probability function. This function characterizes the probability of transitioning from the initial PDF to the target PDF over time. By finding the optimal trajectory that minimizes the overall cost, the Schrödinger bridge problem reveals the most efficient path for transforming distributions, offering valuable insights for tasks such as image processing, signal analysis, and machine learning.

With more details, we are looking to solve the following minimization problem :

$$\min_{\mathbb{Q} \in \mathcal{P}(p_{data}, p_{prior})} D_{KL}(\mathbb{Q}||\mathbb{P}) \tag{1}$$

where D_{KL} corresponds to the Kulback-Leibler divergence between two distribution paths, and $\mathcal{P}(p_{data}, p_{prior})$ corresponds to the set of distribution paths such that marginals at $t = 0$ and $t = T$ are equal to respectively p_{data} and p_{prior}.

This framework is contrasting with SGM in term of time horizon, which is finite in the first case and theoretically infinite in the latter.

Moreover, the solution of 1 has been shown [CH19] to be expressed by the path mesure of the following forward 2, or equivalently backward 3, SDE :

$$\mathrm{d}\mathbf{X}_t = \left[f + g^2 \nabla_x \log \Psi(t, \mathbf{X}_t)\right] \mathrm{d}t + g\,\mathrm{d}\mathbf{W}_t,\ \mathbf{X}_0 \sim p_{data} \tag{2}$$

$$\mathrm{d}\mathbf{X}_t = \left[f - g^2 \nabla_x \log \widehat{\Psi}(t, \mathbf{X}_t)\right] \mathrm{d}t + g\,\mathrm{d}\mathbf{W}_t,\ \mathbf{X}_0 \sim p_{data} \tag{3}$$

where Ψ and $\widehat{\Psi}$ are given by the following PDEs :

$$\begin{cases} \frac{\partial \Psi}{\partial t} & = -\nabla_x \Psi^\intercal f - \frac{1}{2}\mathrm{Tr}(g^2 \nabla_x^2 \Psi) \\ \frac{\partial \widehat{\Psi}}{\partial t} & = -\nabla_x \cdot (\widehat{\Psi} f) + \frac{1}{2}\mathrm{Tr}(g^2 \nabla_x^2 \widehat{\Psi}) \end{cases} \text{s.t. } \Psi(0, \cdot)\widehat{\Psi}(0, \cdot) = p_{data}, \Psi(T, \cdot)\widehat{\Psi}(T, \cdot) = p_{prior} \tag{4}$$

Even though theorically we obtained the exact solution of our problem 1, solving these PDEs 4 are hard and we can not obtain numerically the path measure $\mathbb{Q}$ we are looking for. However, it has been shown in [TC22] that, thanks to a substitution trick and considering the Ito's process associated to $\log \Psi(t, \mathbf{X}_t)$, we have the following SDEs:

$$\begin{cases} \mathrm{d}\mathbf{X}_t & = (f + g\mathbf{Z}_t)\mathrm{d}t + g\mathrm{d}\mathbf{W}_t \\ \mathrm{d}\mathbf{Y}_t & = \frac{1}{2}\mathbf{Z}_t^T \mathbf{Z}_t \mathrm{d}t + \mathbf{Z}_t^T \mathrm{d}\mathbf{W}_t \\ \mathrm{d}\hat{\mathbf{Y}}_t & = \left(\frac{1}{2}\hat{\mathbf{Z}}_t^T \hat{\mathbf{Z}}_t + \nabla_x \cdot (g\hat{\mathbf{Z}}_t - f) + \hat{\mathbf{Z}}_t^T \mathbf{Z}_t\right) \mathrm{d}t + \hat{\mathbf{Z}}_t^T \mathrm{d}\mathbf{W}_t \end{cases} \tag{5}$$

where

$$\mathbf{Y}_t = \log \Psi(t, \mathbf{X}_t) \qquad \mathbf{Z}_t = g\nabla_x \log \Psi(t, \mathbf{X}_t)$$
$$\hat{\mathbf{Y}}_t = \log \widehat{\Psi}(t, \mathbf{X}_t) \qquad \widehat{\mathbf{Z}}_t = g\nabla_x \log \widehat{\Psi}(t, \mathbf{X}_t)$$

Furthermore, we have the relation $\mathbf{Y}_t + \widehat{\mathbf{Y}}_t = \log p_t^{\mathrm{SB}}(\mathbf{X}_t)$ where $p_t^{\mathrm{SB}}(\mathbf{X}_t)$ is the density of the Schrodinger bridge.

Thanks to these equation we are now looking for $\mathbf{Z}(t, \mathbf{X}_t)$ and $\widehat{\mathbf{Z}}(t, \mathbf{X}_t)$. Using the equations 5 we obtain that the log-likelihood of the SB model given by $(\mathbf{Z}_t, \widehat{\mathbf{Z}}_t)$ can be written as $\mathbb{E}\left[\mathbf{Y}_0 + \widehat{\mathbf{Y}}_0 | \mathbf{X}_0 = x_0\right]$. We obtain the following formula

$$\log p_0^{\mathrm{SB}}(x_0) = \mathbb{E}[\log p_T(\mathbf{X}_T)] - \int_0^T \mathbb{E}\left[\frac{1}{2}\|\mathbf{Z}_t\|^2 + \frac{1}{2}\|\widehat{\mathbf{Z}}_t\|^2 + \nabla_x \cdot (g\widehat{\mathbf{Z}} - f) + \widehat{\mathbf{Z}}_t^\intercal \mathbf{Z}_t \middle| \mathbf{X}_0 = x_0\right] \mathrm{d}t \tag{6}$$

that can be used to derive a loss to train two neural network $s(t, x, \theta)$ and $\widehat{s}(t, x, \phi)$ such that $s(t, \mathbf{X}_t, \theta)$ approximates $\mathbf{Z}_t$ and similarly for letters with hats.

These two neural networks s_t and $\hat{s}_t$ can be trained with the following losses, derived from the log-likelihood of samples above :

$$\tilde{\mathcal{L}}_{SB}(x_0;\phi) = -\int_0^T \mathbb{E}_{\mathbf{X}_t\sim(7a)}\left[\frac{1}{2}\|\hat{\mathbf{Z}}(t,\mathbf{X}_t;\phi)\|^2 + g\nabla_x\cdot\hat{\mathbf{Z}}(t,\mathbf{X}_t;\phi) + \mathbf{Z}_t^T\hat{\mathbf{Z}}(t,\mathbf{X}_t;\phi)\right]\mathrm{d}t \tag{7}$$

$$\tilde{\mathcal{L}}_{SB}(x_T;\phi) = -\int_0^T \mathbb{E}_{\mathbf{X}_t\sim(7b)}\left[\frac{1}{2}\|\mathbf{Z}(t,\mathbf{X}_t;\phi)\|^2 + g\nabla_x\cdot\mathbf{Z}(t,\mathbf{X}_t;\phi) + \hat{\mathbf{Z}}_t^T\mathbf{Z}(t,\mathbf{X}_t;\phi)\right]\mathrm{d}t \tag{8}$$

By taking $(\mathbf{Z}_t, \widehat{\mathbf{Z}}_t) = (0, g\cdot s_t)$, where s_t is the neural network trained with SGM - we recover the SGM training objective from the formula above. This shows again that the SB problem is a generalization of SGM problem.

2.3 Benefits of the formulation over classical diffusion models

In the SB formulation, the forward drift is learned which allows to use a broader class of SDE - including non-linear ones - which greatly speeds up convergence. As said before, this framework is contrasting with SGM in term of time horizon, which is finite in the first case and theoretically infinite in the latter.

With the forward convergence much quicker than in SGM, the denoising also needs a much lower number of passes, resulting in a sampling time decreased by 80% in the implementation of the authors.

It can also be noted that the sampling density is not necessarily gaussian in SB.

2.4 Stochastic control

The goal here is to resolve the following problem :

$$\min_{\mathbb{Q}\in\mathcal{P}(p_{data},p_{prior})} D_{KL}(\mathbb{Q}||\mathbb{P})$$

$$\text{s.t.}\quad \begin{cases} \mathrm{d}\mathbf{X}_t = \left[f + g^2\nabla_x\log\Psi(t,\mathbf{X}_t)\right]\mathrm{d}t + g\,\mathrm{d}\mathbf{W}_t \\ \mathrm{d}\mathbf{X}_t = \left[f - g^2\nabla_x\log\widehat{\Psi}(t,\mathbf{X}_t)\right]\mathrm{d}t + g\,\mathrm{d}\mathbf{W}_t \\ \mathbf{X}_0\sim p_{data} \end{cases}$$

This can be seen through a stochastic control perspective : we are searching for a time-varying control policy - $\nabla_x\log\hat{\Psi}(t,\mathbf{X}_t)$ - that minimizes an objective - $D_{KL}(\mathbb{Q}||\mathbb{P})$ - while being subjected to control-affine SDEs.

Stochastic control theory then gives the optimality conditions of such a problem, and applying the Hopf-Cole transformation to the set of ODE obtained allows to reformulate the problem in terms of SB theory :

$$\begin{cases} \dfrac{\partial\psi}{\partial t} = -\nabla_x\psi^T f - \dfrac{1}{2}Tr(g^2\nabla_x^2\psi) \\ \dfrac{\partial\hat{\psi}}{\partial t} = -\nabla_x\cdot(\hat{\psi}f) + \dfrac{1}{2}Tr(g^2\nabla_x^2\hat{\psi}) \end{cases} \quad \text{s.t. } \psi(0,\cdot)\hat{\psi}(0,\cdot) = p_{data},\ \psi(T,\cdot)\hat{\psi}(T,\cdot) = p_{prior}$$

Then Forward-Backward SDEs theory can be applied on the above set of equations to work on SDEs instead of PDEs, which is the main contribution of this paper.

2.5 Quasi parabolic PDEs

3 Theoretical improvement ideas

We recall that the SB problem has two important benefits compared to the standard diffusion problem. For the latter, as we are using affine drifts, we obtain the convergence toward Gaussian distribution only when time tends to infinity. However, using non linear drift permits to obtain this Gaussian distribution in finite time. That brings us to the second point: with SB problem we can theoretically diffuse towards any distribution, which would allow to have more diversity in the way of sampling, even though considering the usefulness of the Gaussian distribution, we are more likely going to sample under this distribution.

As a counter part of these advantages, when we take a look at the log likelihood and the lower bounds we can obtain, we are not able to compute it analytically as we were able to with SGM likelihood. Indeed, the term $\nabla_x \log p_t$ was a Gaussian distribution (conditioned on the initial input) so we were able to compute analytically the term $\mathbb{E}\left[\lambda(t)\|s(t,x,\theta) - \nabla_x \log p(t,x)\|^2\right]$. As with SB we do not know the laws of the term in the expectation.

We wondered if it was possible to restraint the functions Z_t such that the log likelihood could be analytically computed, even if we are expecting these laws to be more general than Gaussian distributions. It is clearly linked to the general question of improving the speed of convergence in the standard diffusion model, and the idea would be to still learn this drift.

Looking at the current state of the art, this question have been demonstrated to be really hard. We have been interested in the following SDE:

$$\mathrm{d}\mathbf{X}_t = P(\mathbf{X}_t)\mathrm{d}t + g(t)\dot{d}\mathbf{W}_t \tag{9}$$

However, it has led to nowhere as it is still a research field. We decided to investigate other aspect but we keep this idea in mind and will still investigate this and we are looking forward to use this theory to speed up the convergence towards Gaussian distribution as with SB problem, but while keeping the analytical aspect that speed up computations.

4 Experiments

4.1 Methodology

4.2 Results

5 Conclusion

$d\mathbf{X}_t = f(t,\mathbf{X}_t)dt + g(t)d\mathbf{W}_t,\ X_0 \sim p_{data}$

$d\mathbf{X}_t = \left[f - g^2\nabla_x logp_t^{(1)}(\mathbf{X}_t)\right]dt + gd\mathbf{W}_t,\ \mathbf{X}_T \sim p_T^{(1)}$

$logp_0^{SGM}(x_0) \geq \mathcal{L}_{SGM}(x_0;\theta) = \mathbb{E}\left[logp_T(\mathbf{X}_T)\right] - \int_0^T \mathbb{E}\left[\frac{1}{2}g^2\|s_t\|^2 + \nabla_x \cdot (g^2 s_t - f)\right]dt = \mathbb{E}\left[logp_T(\mathbf{X}_T)\right] - \int_0^T \mathbb{E}\left[\frac{1}{2}g^2\|s_t - \nabla_x logp_{t|x_0}\|^2 - \frac{1}{2}\|g\nabla_x logp_{t|x_0}\|^2 - \nabla_x \cdot f\right]dt$

$d\mathbf{X}_t = \left[f - g^2 s(t,\mathbf{X}_t;\theta)\right]dt + gd\mathbf{W}_t,\ \mathbf{X}_t \sim p_{prior}$

$$\begin{cases} \dfrac{\partial\psi}{\partial t} = -\nabla_x\psi^T f - \dfrac{1}{2}Tr(g^2\nabla_x^2\psi) \\ \dfrac{\partial\hat{\psi}}{\partial t} = -\nabla_x \cdot (\hat{\psi}f) + \dfrac{1}{2}Tr(g^2\nabla_x^2\hat{\psi}) \end{cases} \text{s.t. } \psi(0,\cdot)\hat{\psi}(0,\cdot) = p_{data},\ \psi(T,\cdot)\hat{\psi}(T,\cdot) = p_{prior}$$

$d\mathbf{X}_t = \left[f + g^2\nabla_x log\psi(t,\mathbf{X}_t)\right]dt + g\mathbf{W}_t,\ X_0 \sim p_{data}$

$d\mathbf{X}_t = \left[f - g^2\nabla_x log\hat{\psi}(t,\mathbf{X}_t)\right]dt + g\mathbf{W}_t,\ X_T \sim p_{prior}$

$\frac{\partial v}{\partial t} + \frac{1}{2}Tr(\nabla_x^2 vGG^T) + \nabla_x v^T f + h(t,x,v,G^T\nabla_x v) = 0,\ v(T,x) = \phi(x)$

$v(t,\mathbf{X}_t) = \mathbf{Y}_t$ and $G(t,\mathbf{X}_t)^T\nabla_x v(t,\mathbf{X}_t) = \mathbf{Z}_t$

$$\begin{cases} d\mathbf{X}_t = (f + g\mathbf{Z}_t)dt + gd\mathbf{W}_t \\ d\mathbf{Y}_t = \dfrac{1}{2}\mathbf{Z}_t^T\mathbf{Z}_t dt + \mathbf{Z}_t^T d\mathbf{W}_t \\ d\hat{\mathbf{Y}}_t = \left(\dfrac{1}{2}\hat{\mathbf{Z}}_t^T\hat{\mathbf{Z}}_t + \nabla_x \cdot (g\hat{\mathbf{Z}}_t - f) + \hat{\mathbf{Z}}_t^T\mathbf{Z}_t\right)dt + \hat{\mathbf{Z}}_t^T d\mathbf{W}_t \end{cases}$$

$\mathbf{Y}_t = log\psi(t,\mathbf{X}_t),\ \mathbf{Z}_t = g\nabla_x log(\psi(t,\mathbf{X}_t)$

$\hat{\mathbf{Y}}_t = log\hat{\psi}(t,\mathbf{X}_t),\ \hat{\mathbf{Z}}_t = g\nabla_x log(\hat{\psi}(t,\mathbf{X}_t)$

$\mathbf{Y}_t + \hat{\mathbf{Y}}_t = logp_t^{SB}(\mathbf{X}_t)$

$logp_0^{SB}(x_0) = \mathbb{E}\left[logp_T(\mathbf{X}_T)\right] - \int_0^T \mathbb{E}\left[\frac{1}{2}\|\mathbf{Z}_t\|^2 + \frac{1}{2}\|\hat{\mathbf{Z}}_T - g\nabla_x logp_t^{SB} + \mathbf{Z}_t\|^2 - \frac{1}{2}\|g\nabla_x logp_t^{SB} - \mathbf{Z}_t\|^2 - \nabla_x \cdot f\right]dt$

$= \mathbb{E}\left[logp_T(\mathbf{X}_T)\right] - \int_0^T \mathbb{E}\left[\frac{1}{2}\|\mathbf{Z}_t\|^2 + \frac{1}{2}\|\hat{\mathbf{X}}_t\|^2 + \nabla_x \cdot (g\hat{\mathbf{Z}}_t - f) + \hat{\mathbf{Z}}_t^T\mathbf{Z}_t\right]dt$

$d\mathbf{X}_t = \left[f + g\mathbf{Z}(t,\mathbf{X}_t) - \frac{1}{2}g(\mathbf{Z}(t,\mathbf{X}_t) + \hat{\mathbf{Z}}(t,\mathbf{X}_t))\right]dt$

$$\tilde{\mathcal{L}}_{SB}(x_0;\phi) = -\int_0^T \mathbb{E}_{\mathbf{X}_t \sim (7a)} \left[\frac{1}{2} \|\hat{\mathbf{Z}}(t,\mathbf{X}_t;\phi)\|^2 + g\nabla_x \cdot \hat{\mathbf{Z}}(t,\mathbf{X}_t;\phi) + \mathbf{Z}_t^T \hat{\mathbf{Z}}(t,\mathbf{X}_t;\phi) \right] dt$$

$$\tilde{\mathcal{L}}_{SB}(x_T;\phi) = -\int_0^T \mathbb{E}_{\mathbf{X}_t \sim (7b)} \left[\frac{1}{2} \|\mathbf{Z}(t,\mathbf{X}_t;\phi)\|^2 + g\nabla_x \cdot \mathbf{Z}(t,\mathbf{X}_t;\phi) + \hat{\mathbf{Z}}_t^T \mathbf{Z}(t,\mathbf{X}_t;\phi) \right] dt$$

References

[CH19] Kenneth F. Caluya and Abhishek Halder. Wasserstein proximal algorithms for the schrödinger bridge problem: Density control with nonlinear drift. *IEEE Transactions on Automatic Control*, 67:1163–1178, 2019.

[Sch32] Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Annales de l'institut Henri Poincaré, Volume 2*, 1932.

[TC22] Evangelos A. Theodorou Tianrong Chen, Guan-Horng Liu. Likelihood training of schrödinger bridge using forward-backward sdes theory. *ICLR*, 2022.