Unit 13: Schrödinger Bridge and Iterative Proportional Fitting

13.1 Introduction

Recall that in denoising diffusion models, we construct a diffusion process $(X_t)_{0\le t\le T}$ such that $X_T \sim \pi$. We achieved this by utilizing the reverse-time SDE theory. That is, we first construct a diffusion process $(Y_t)_{0\le t\le T}$ with $Y_0 \sim \pi$ and then reverse it in time. Note that we cannot pick an arbitrary SDE for constructing Y_t, because we need to evaluate the distribution of Y_T (which is also the distribution of X_0) so that we can simulate the process X_t. In Theorem 12.1 (a result used by some popular denoising diffusion models), we require $X_0 \sim \pi * \phi_\sigma$, where ϕ_σ denotes $N(0, \sigma^2 I)$, and assuming σ is large enough, we can simply draw X_0 from ϕ_σ. Of course this approximation may not always be satisfactory.

In this unit, we introduce a general approach for constructing a diffusion process moving between two arbitrary distributions. Suppose that we are given a diffusion process $(X_t)_{0\le t\le T}$ with an arbitrary initial distribution $X_0 \sim q_0$. We can apply an algorithm, known as iterative proportional fitting (IPF), to modify the dynamics of X_t so that $X_T \sim \pi$. To explain how it works, we begin by first considering the bivariate case, where we only need to simulate (X_0, X_T). This leads to a classical problem that has been well studied in statistics. See Section 13.2. Next, we generalize the result to the discrete-time process $(X_0, X_{t_1}, X_{t_2}, \dots, X_T)$; see Section 13.3. Finally, we present the continuous-time theory; see Section 13.4. In Sections 13.2 and 13.3, we will use slightly different notation for generality. The methodology in this unit heavily relies on the concept of Kullback–Leibler divergence.

Definition 13.1. Given two probability measures μ, ν absolutely continuous with respect to each other, the Kullback–Leibler (KL) divergence is defined by

$$\mathrm{KL}(\mu, \nu) = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu.$$

We will also use $\mathrm{KL}(p, q)$ to denote the KL divergence between μ and ν, where p, q are the density functions of μ, ν, respectively.

Exercise 13.1. Prove that KL divergence is always non-negative.

Exercise 13.2. Prove that the KL divergence between the standard normal distribution and standard Cauchy distribution is infinite.

Exercise 13.3. Let p, q be density functions. Prove that $\mathrm{KL}(p, q)$ equals the supremum of

$$\int f(x)p(x)\mathrm{d}x - \log \int q(x)e^{f(x)}\mathrm{d}x,$$

taken over any function f such that $\int q(x)e^{f(x)}\mathrm{d}x < \infty$.

13.2 Iterative Proportional Fitting

Consider two random vectors $X, Y \in \mathbb{R}^d$ with joint Lebesgue density function $q(x, y)$. Denote the conditional density functions by $q_{X|Y}(x \,|\, y)$ and $q_{Y|X}(y \,|\, x)$. Let $\pi_X(x), \pi_Y(y)$ be two Lebesgue density functions on $\mathbb{R}^d$, and define

$$\mathcal{M}(\pi_X, \pi_Y) = \left\{ p(x, y) \geq 0 \colon \int p(x, y) \mathrm{d}y = \pi_X(x), \int p(x, y) \mathrm{d}x = \pi_Y(y) \right\}$$

to be the collection of all joint density functions with marginals π_X and π_Y. How to find the joint density function $p \in \mathcal{M}(\pi_X, \pi_Y)$ that minimizes $\mathrm{KL}(p, q)$? This is sometimes known as the static Schrödinger Bridge problem. It turns out that the solution has a simple characterization. For the proof of the following theorem, see, e.g., [12]

Theorem 13.1. *Let $q(x, y)$ denote the joint Lebesgue density function of (X, Y). If*

$$\inf\{\mathrm{KL}(p, q) \colon p \in \mathcal{M}(\pi_X, \pi_Y)\} < \infty, \tag{1}$$

then there exists a unique $p^* \in \mathcal{M}(\pi_X, \pi_Y)$ achieving the minimum KL divergence in (1), and it can be expressed as

$$p^*(x, y) = a(x) b(y) q_{Y|X}(y \,|\, x), \tag{2}$$

for some functions $a, b \geq 0$.

Remark 13.1. We can also express p^* by $p^*(x, y) = \tilde{a}(x) b(y) q(x, y)$ with $\tilde{a}(x) = a(x)/q_X(x)$.

Hence, to solve the static SB problem, it only remains to find the functions a, b in (2), which is equivalent to solving the so-called Schrödinger system (or Schrödinger equations):

$$\int a(x) b(y) q_{Y|X}(y \,|\, x) \mathrm{d}y = \pi_X(x), \tag{3}$$

$$\int a(x) b(y) q_{Y|X}(y \,|\, x) \mathrm{d}x = \pi_Y(y). \tag{4}$$

Iterative proportional fitting (IPF) is an iterative algorithm for solving this system, which has a long history in statistics and was first used for estimating cell probabilities of contingency tables [4]. We begin by considering the joint distribution $p^{(0)}(x, y) = \pi_X(x) q_{Y|X}(y \,|\, x)$, which in general does not have marginal π_Y. So we update this distribution by matching the marginal π_Y, which leads to $p^{(1)}(x, y) = \pi_Y(y) p^{(0)}_{X|Y}(x \,|\, y)$. Repeating this procedure yields a sequence of joint distributions $p^{(n)}$ that converge to p^*. See [11] for the proof of the convergence.

Algorithm 13.1 (Iterative Proportional Fitting)**.** Let $a_0(x) = \pi(x), b_0(y) = 1$ and $p^{(0)}(x, y) = a_0(x) b_0(y) q_{Y|X}(y \,|\, x)$. For $k = 1, 2, \ldots,$

(i) Set $p^{(2k)}(x, y) = \pi_Y(y) p^{(2k-1)}_{X|Y}(x \,|\, y) = a_{2k}(x) b_{2k}(y) q_{Y|X}(y \,|\, x)$ where

$$a_{2k}(x) = a_{2k-1}(x), \quad b_{2k}(y) = \frac{\pi_Y(y)}{\int a_{2k-1}(x) q_{Y|X}(y \,|\, x) \mathrm{d}x}.$$

(ii) Set $p^{(2k+1)}(x,y) = \pi_X(x)p^{(2k)}_{Y|X}(y\,|\,x) = a_{2k+1}(x)b_{2k+1}(y)q_{Y|X}(y\,|\,x)$ where

$$a_{2k+1}(x) = \frac{\pi_X(x)}{\int b_{2k}(y)q_{Y|X}(y\,|\,x)\mathrm{d}y}, \quad b_{2k+1}(y) = b_{2k}(y).$$

13.3 Discrete-time Schrödinger Bridge

We can generalize the static SB problem and IPF algorithm by considering a stochastic process $(X_0, X_1, \ldots, X_N)$ with joint density $q(x_0, x_1, \ldots, x_N)$. Let $\mathcal{M}(\pi_0, \pi_N)$ denote the collection of all joint densities $p(x_0, x_1, \ldots, x_N)$ such that the marginal distributions of the first and last component equal π_0 and π_N respectively. The dynamic SB problem searches for the optimal $p \in \mathcal{M}(\pi_0, \pi_N)$ that minimizes $\mathrm{KL}(p, q)$.

A simple argument shows that the dynamic SB problem can be reduced to the static one. Let $q_{0,N}(x_0, x_N)$ denote the marginal distribution of (x_0, x_N) and $q_{\cdot\,|\,0,N}(x_1, \ldots, x_{N-1}\,|\,x_0, x_N)$ denote the conditional distribution of $(x_1, \ldots, x_{N-1})$ given (x_0, x_N). It can be shown that

$$\mathrm{KL}(p, q) = \mathrm{KL}(p_{0,N}, q_{0,N}) + \mathbb{E}\left[\mathrm{KL}(p_{\cdot\,|\,0,N}(\cdot\,|\,X_0, X_N),\, q_{\cdot\,|\,0,N}(\cdot\,|\,X_0, X_N))\right] \tag{5}$$

where the expectation is taken over the distribution $p_{0,N}$. Since KL divergence is non-negative, it is clear that the optimal p^* should take the form

$$p^*(x_0, x_1, \ldots, x_N) = p^*_{0,N}(x_0, x_N)q_{\cdot\,|\,0,N}(x_1, \ldots, x_{N-1}\,|\,x_0, x_N),$$

where $p^*_{0,N}$ is the solution to the static SB problem.

Now let's further assume that $(X_0, \ldots, X_N)$ is Markovian. In this case, we obtain an interesting generalization of the IFP iterations [3]. In the n-th iteration (assuming n is even), we update our joint density estimate forward in time by

$$p^{(n)}(x_0, \ldots, x_N) = \pi_0(x_0)p^{(n-1)}_{\cdot\,|\,0}(x_1, \ldots, x_N\,|\,x_0) = \pi_0(x_0)\prod_{j=1}^{N} p^{(n-1)}_{j\,|\,j-1}(x_j\,|\,x_{j-1}),$$

and in the $(n+1)$-th iteration, we perform the update backward in time by

$$p^{(n+1)}(x_0, \ldots, x_N) = \pi_N(x_N)p^{(n)}_{\cdot\,|\,N}(x_0, x_1, \ldots, x_{N-1}\,|\,x_N) = \pi_N(x_N)\prod_{j=1}^{N} p^{(n)}_{j-1\,|\,j}(x_{j-1}\,|\,x_j).$$

So instead of working with the conditional density $p_{N\,|\,0}(x_N\,|\,x_0)$, we split it into multiple time steps and update the transition density at each time step separately.

Example 13.1. We present a simple numerical example with $N = 4$ and $X_j \in \mathbb{R}^2$. Let π_0 be the bivariate normal distribution $N(0, I)$ and π_3 be the bivariate normal distribution with mean $(1, 1)$ and covariance matrix $0.25 * \begin{bmatrix} 1 & -0.99 \\ -0.99 & 1 \end{bmatrix}$. Let q be the density such that $X_j\,|\,X_{j-1} \sim N(X_{j-1}, 0.05I)$. We show the first four iterations (2 forward and 2 backward) in Figure 1, where the red dots indicate the mean.

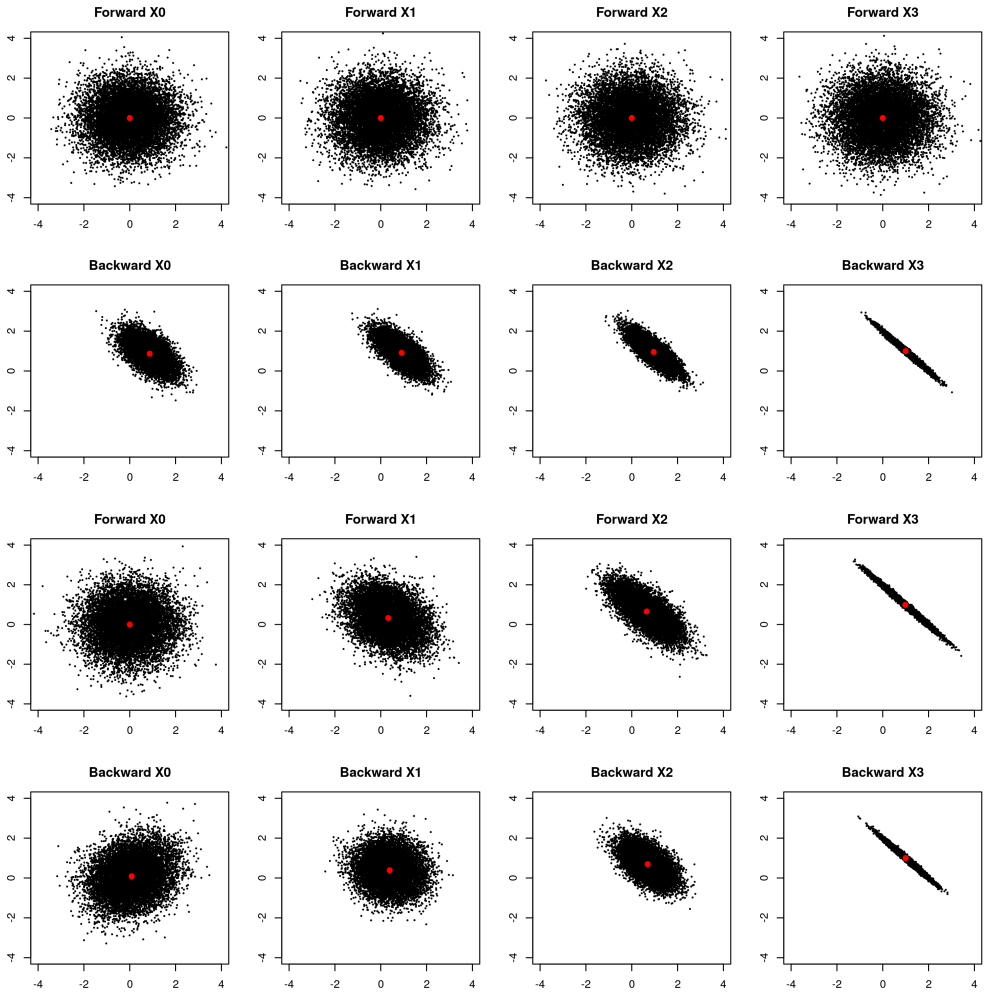

Figure 1: IPF for the stochastic process (X_0, X_1, X_2, X_3) in Example 13.1. We generate 10^4 replicates to visualize the distributions.

Remark 13.2. The Euler-Maruyama discretization of a diffusion process yields a (possibly non-homogeneous) Markov chain $(X_{t_0}, X_{t_1}, \dots, X_{t_N})$ with $0 = t_0 < t_1 < \cdots < t_N = T$. Assuming that the time step is sufficiently small, one can argue that the conditional distributions $p^{(n)}_{j-1\,|\,j}$ and $p^{(n)}_{j\,|\,j-1}$ in IPF are both approximately normal (recall Remark 12.1). So in this case, essentially $p^{(n)}$ is obtained by calculating the reverse-time SDE of $p^{(n-1)}$, which can be done by score matching in generative modeling. This leads to an iterative algorithm that trains a diffusion process that evolves from an arbitrary initial distribution to the desired terminal distribution. Efficient algorithms based on this idea have been proposed in [13, 3].

Exercise 13.4. Prove (5).

13.4 Continuous-time Schrödinger Bridge

Now we present the SB problem for continuous-time diffusion processes. Consider a diffusion X_t on $\mathbb{R}^d$ over the time interval $[0, T]$, evolving by

$$\mathrm{d}X_t = b(X_t, t)\mathrm{d}t + \sigma \mathrm{d}B_t, \quad t \in [0, T],$$

with initial distribution $\mathrm{Law}(X_0) = \mu_0$. Denote the distribution of $(X_t)_{0\le t\le T}$ by $\mathbb{P}_b$ (which is a probability measure on the space of all continuous functions on $[0, T]$). When $b \equiv 0$, the conditional distribution of X_T given $X_0 = x$ has density

$$q_{T|0}(y \,|\, x) = \phi_{\sigma\sqrt{T}}(y - x), \tag{6}$$

where the marginal distribution of X_T is $\mu_0 * \phi_{\sigma\sqrt{T}}$ (where $*$ denotes convolution). Denote the distribution of $(X_t)_{0\le t\le T}$ with $b \equiv 0$ by $\mathbb{P}_0$, which will be treated as the reference measure. The SB problem asks how to find the "optimal" b such that $\mathrm{Law}(X_T) = \pi$ that achieves minimal KL divergence between $\mathbb{P}_b$ and $\mathbb{P}_0$. Solutions have been derived via different techniques [7, 1, 10, 6, 5]. Below we present a result from [2].

Theorem 13.2. *Let $q_{T|0}$ be given by (6). For every continuous distribution π on $\mathbb{R}^d$, there exists a unique (up to scaling) pair of σ-finite measures, (ν_0, ν_T), such that*

$$\frac{\mathrm{d}\mu_0}{\mathrm{d}\nu_0}(x) = \int q_{T|0}(y \,|\, x)\nu_T(\mathrm{d}y), \text{ for } \mu_0\text{-almost every } x, \tag{7}$$

$$\frac{\mathrm{d}\pi}{\mathrm{d}\nu_T}(y) = \int q_{T|0}(y \,|\, x)\nu_0(\mathrm{d}x), \text{ for every } y \in \mathbb{R}^d. \tag{8}$$

If $\int x^2\mu_0(\mathrm{d}x) < \infty, \int(\mathrm{d}\mu_0/\mathrm{d}\nu_0)\mathrm{d}\mu_0 < \infty$, and $\mathrm{KL}(\pi, \nu_0 * \phi_{\sigma\sqrt{T}}) < \infty$ where $\phi_{\sigma\sqrt{T}}$ denotes the normal distribution $N(0, \sigma^2 T I_d)$, then the SB problem has a solution given by

$$b^*(x, t) = \sigma^2 \nabla_x \log h(x, t), \text{ where } h(x, t) = \int \phi_{\sigma\sqrt{T-t}}(y - x)\nu_T(\mathrm{d}y).$$

That is, $b^* = \arg\min_{b\in\mathcal{C}(\pi)} \mathrm{KL}(\mathbb{P}_b, \mathbb{P}_0)$ where

$$\mathcal{C}(\pi) = \{b\colon \text{the marginal distribution of } X_T \text{ under } \mathbb{P}_b \text{ equals } \pi\}.$$

Proof. See [2]. □

Remark 13.3. Equations (7) and (8) are essentially the Schrödinger system given in (3) and (4). Equations (7) and (8) are more general since here we do not assume that μ_0 is a continuous distribution.

Example 13.2. Let μ_0 be the Dirac measure at x_0. Then, the solution to the Schrödinger system is given by $\nu_0 = \mu_0$, and

$$\nu_T(\mathrm{d}y) = \frac{\pi(y)}{q_{T|0}(y \,|\, x_0)}\mathrm{d}y = \frac{\pi(y)}{\phi_{\sigma\sqrt{T}}(y - x_0)}\mathrm{d}y.$$

In this case,

$$h(x,t) = \int \frac{\pi(y)}{\phi_{\sigma\sqrt{T}}(y - x_0)}\phi_{\sigma\sqrt{T-t}}(x - y)\mathrm{d}y.$$

This has been used in [9, 14] for devising sampling algorithms for generative modeling or other purposes. In Figure 2, we give an example from [8] illustrating the dynamics of this optimal diffusion process where $d = 2$ and π is the mixture of four normal distributions.

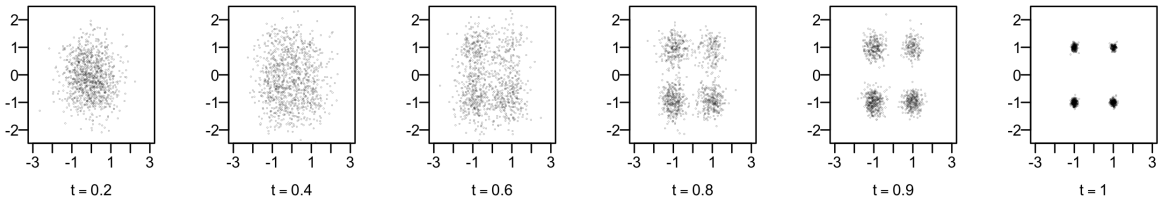

Figure 2: Distribution of $1,000$ replicates.

Example 13.3. We can now also derive Theorem 12.1 as a special case of the SB problem. Let $\mu_0 = \pi * \phi_{\sigma\sqrt{T}}$. The solution to the Schrödinger system is given by $\nu_0(x) = 1$ and $\nu_T(y) = \pi(y)$. Hence, $h(t,x) = \int \pi(y)\phi_{\sigma\sqrt{T-t}}(x - y)\mathrm{d}y$.

References

[1] Arne Beurling. An automorphism of product measures. *Annals of Mathematics*, pages 189–200, 1960.

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

[3] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

[4] 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.

[5] Wendell H Fleming. Logarithmic transformations and stochastic control. In *Advances in Filtering and Optimal Stochastic Control: Proceedings of the IFIP-WG 7/1 Working Conference Cocoyoc, Mexico, February 1–6, 1982*, pages 131–141. Springer, 2005.

[6] H Föllmer. Random fields and diffusion processes. *Ecole d'Ete de Probabilites de Saint-Flour XV-XVII, 1985-87*, 1988.

[7] Robert Fortet. Résolution d'un système d'équations de m. Schrödinger. *Journal de Mathématiques Pures et Appliquées*, 19(1-4):83–105, 1940.

[8] Jhanvi Garg, Xianyang Zhang, and Quan Zhou. Soft-constrained schrödinger bridge: a stochastic control approach. In *International Conference on Artificial Intelligence and Statistics*, pages 4429–4437. PMLR, 2024.

[9] Jian Huang, Yuling Jiao, Lican Kang, Xu Liao, Jin Liu, and Yanyan Liu. Schrödinger-Föllmer sampler: sampling without ergodicity. *arXiv preprint arXiv:2106.10880*, 2021.

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

[11] Ludger Rueschendorf. Convergence of the iterative proportional fitting procedure. *Annals of Statistics*, 23, 08 1995.

[12] Ludger Rüschendorf and Wolfgang Thomsen. Note on the schrödinger equation and I-projections. *Statistics & probability letters*, 17(5):369–375, 1993.

[13] Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving Schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021.

[14] Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep generative learning via Schrdinger bridge. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10794–10804. PMLR, 18–24 Jul 2021.