Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling[1]

–

Computational Optimal Transport

Hugo Simon

06/02/2022

[1] Valentin De Bortoli et al. (2021). "Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling". In: *arXiv preprint arXiv:2106.01357*.

- In generative modeling we are interested into designing algorithms which transform a given distribution p_{prior} into a given data distribution p_{data} .
- We have access to the data distribution only through samples.
- There is a flurry of frameworks which aim at solving this problem such as Energy-Based models, Generative Adversarial Networks, normalizing flows, Variational Auto-encoders or diffusion score-matching techniques.

Score-based generative modeling (SGM)

- We have access to a *noising process*[2] (ergodic) perturbing the data distribution p_{data} into p_{prior} e.g. Ornstein-Ulhenbeck process

$$\mathrm{d}\mathbf{X}_t = -\alpha \mathbf{X}_t \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\mathbf{B}_t, \quad \mathbf{X}_0 \sim p_{\text{data}}, \quad \alpha > 0$$

- The backward dynamic is retrieved by injecting forward dynamic information[3] as marginals scores $\nabla \log p_{T-t}$ where p_t is the density of $\mathbf{X}_t$

$$\mathrm{d}\mathbf{Y}_t = (\alpha \mathbf{Y}_t + 2\nabla \log p_{T-t}(\mathbf{Y}_t)) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\mathbf{B}_t, \quad \mathbf{Y}_0 \sim p_{\text{prior}}$$

- A generative model can be obtained using an Euler-Maruyama discretization of the previous diffusion process with stepsize γ

$$Y^\star_{k+1} = Y^\star_k + \gamma \left\{ \alpha Y^\star_k + 2\nabla \log p_{T-t_k}(Y^\star_k) \right\} + \sqrt{2\gamma} \mathbf{Z}_k$$

- **EXCEPT** we often do not analytically know the score.
- In score-based generative modeling techniques this score is approximated by a neural network $s_{\theta^\star}$.

[2] Jascha Sohl-Dickstein et al. (2015). *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. arXiv: 1503.03585 [cs.LG].

[3] Yang Song et al. (2021). *Score-Based Generative Modeling through Stochastic Differential Equations*. arXiv: 2011.13456 [cs.LG].

Score-based generative modeling with SDE

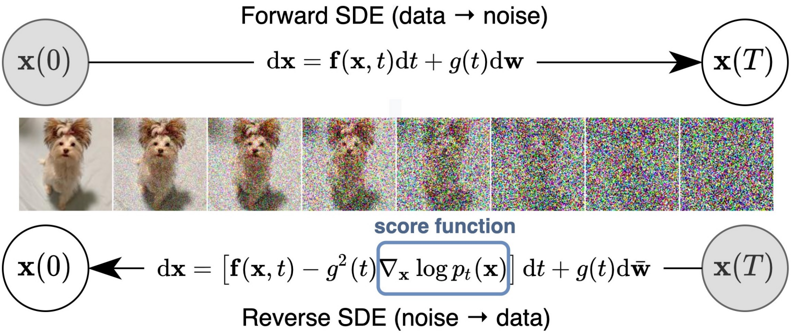
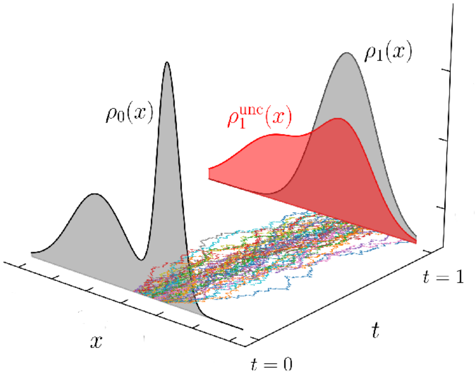


Figure 1: Solving a reverse SDE yields a score-based generative model.

One of the main limitation of score-based generative modeling is that they require a large number of step sizes so that the initial forward dynamics is close to the distribution p_{prior} and small enough step sizes so that the neural network approximation holds.

Table of content

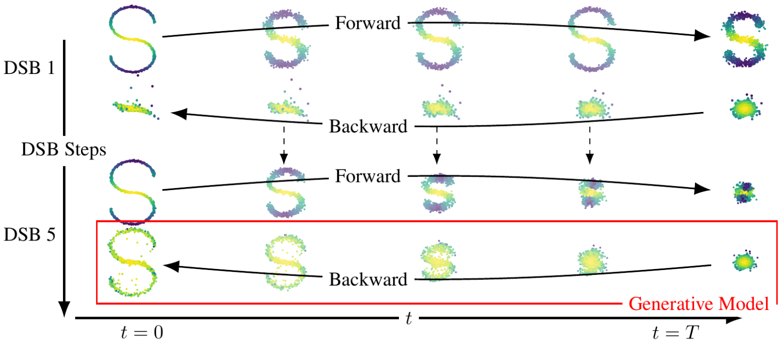
$$\pi^\star = \operatorname{argmin}\left\{\mathrm{KL}(\pi \mid \mathbb{P}), \pi_0 = p_{\text{data}}\,, \pi_N = p_{\text{prior}}\,\right\}$$



Classical problem[4] in physics, optimal control and probability. Knowing departure distribution and reference dynamic, what is the actual dynamic knowing also the arrival?

[4] E. Schrödinger (1932). "Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique". fr. In: *Annales de l'institut Henri Poincaré* 2.4, pp. 269–310. URL: http://www.numdam.org/item/AIHP_1932__2_4_269_0/.

Iterative Proportional Fitting (IPF)



In order to find the solution of the SB problem IPF operates iteratively by successively solving half-bridge problems. Starting with the reference dynamic $\pi^0 = \mathbb{P}$, then project alternatively

$$\pi^{2n+1} = \operatorname{argmin}\left\{\mathrm{KL}\left(\pi \mid \pi^{2n}\right), \pi_N = p_{\text{prior}}\right\}$$
$$\pi^{2n+2} = \operatorname{argmin}\left\{\mathrm{KL}\left(\pi \mid \pi^{2n+1}\right), \pi_0 = p_{\text{data}}\right\}$$

Algorithm 1 Implementation of Diffusion Schrödinger Bridge

Input: Samples from p_{data} and p_{prior}, stepsize γ, number of timesteps N.

1. Sample forward $X_0 \sim p_{\text{data}}$ and $X_{k+1} = F_k^{\alpha}(X_k) + \sqrt{2\gamma} Z_{k+1}$ (Ornstein-Ulhenbeck process as initialization)
2. Compute backward loss on batches and update weights
   - $\hat{\ell}^b(\beta) = \textbf{Mean}\left(\|B_{k+1}^{\beta}(X_{k+1}) - (X_{k+1} + F_k^{\alpha}(X_{k+1}) - F_k^{\alpha}(X_k))\|^2\right)$
   - $\beta \leftarrow$ Gradient Step$(\hat{\ell}^b(\beta))$
3. Sample backward $X_N \sim p_{\text{prior}}$ and $X_{k-1} = B_k^{\beta}(X_k) + \sqrt{2\gamma}\tilde{Z}_k$
4. Compute forward loss on batches and update weights
   - $\hat{\ell}^f(\alpha) = \textbf{Mean}\left(\|F_k^{\alpha}(X_k) - (X_k + B_{k+1}^{\beta}(X_{k+1}) - B_{k+1}^{\beta}(X_k))\|^2\right)$
   - $\alpha \leftarrow$ Gradient Step$(\hat{\ell}^f(\alpha))$
5. Iterate until convergence

Output: Sampling function for p_{data} from p_{prior} (and *vice-versa*)

Theorem (Approximation bound for Ornstein–Uhlenbeck based SGM)

Assume $\exists \mathrm{M} \geq 0$ *such that* $\forall (t, x) \in [0, T] \times \mathbb{R}^d$

$$\|s_{\theta^\star}(t, x) - \nabla \log p_t(x)\| \leq \mathrm{M},$$

Assume some mild other bounds on the score. Then for any $\forall \alpha > 0$, $\exists B_\alpha, C_\alpha, D_\alpha \geq 0$ *such that* $\forall N \in \mathbb{N}$, *the following bounds on the total variation distance hold:*

$$\|\mathcal{L}(X_0) - p_{\mathit{data}}\|_{\mathrm{TV}} \leq C_\alpha \left(\mathrm{M} + \gamma^{1/2}\right) \exp\left[D_\alpha T\right] + B_\alpha \exp\left[-\alpha^{1/2} T\right]$$

where $\mathcal{L}(X_0)$ *is the distribution of* X_0 *(backsampled from* X_N*).*

Proposition (Convergence rate of IPF)

It is known[a][b] *that IPF converges at geometric rate in the case where* p_{data} *and* p_{prior} *are compactly supported. Moreover, convergence rate is* $o(n)$ *in the non-compact setting.*

[a] Gabriel Peyré et al. (2020). *Computational Optimal Transport*. arXiv: 1803.00567 [stat.ML].

[b] Yongxin Chen et al. (2015). *Entropic and displacement interpolation: a computational approach using the Hilbert metric*. arXiv: 1506.04255 [math.OC].

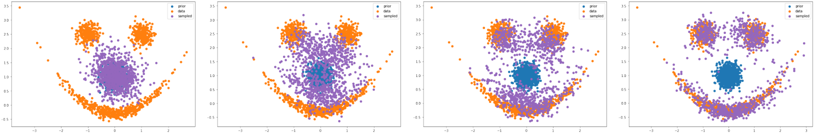


Figure 2: Backward-sampling a mixture of densities from a gaussian prior after 5 DSB steps. Intermediate states constitute data interpolation.

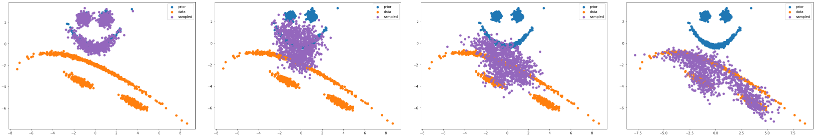


Figure 3: Backward sampling a mixture of densities from a affine-transformed prior after 5 DSB steps. Intermediate states constitute data interpolation.

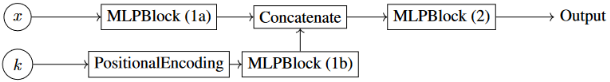


Figure 4: Multi Layer Perceptron with Positional Encoding

Observations

- less timesteps and iterations needed than existing SGM.
- not very robust, must choose architectures and parameters with care.
- simpler losses **Mean** $\left(\|F_k^\alpha(X_k) - X_{k+1}\|^2\right)$ and **Mean** $\left(\|B_{k+1}^\beta(X_{k+1}) - X_k\|^2\right)$ work as well.

Table of content

Strengths

- can be used on top of existing algorithms.
- contrary to existing SGM, does not require reference process to converge to p_{prior} .
- does not require p_{prior} to be Gaussian. In particular, able to perform dataset interpolation.

Weaknesses

- non dedicated architecture. Requires careful selection of the parameters of DSB.
- as all sequential architectures (SGM, LSTM...), non parallelized.
- no approximation results for DSB.

Perspectives

- dedicated architectures, approximation results.

Chen, Yongxin et al. (2015). *Entropic and displacement interpolation: a computational approach using the Hilbert metric*. arXiv: 1506.04255 [math.OC].

De Bortoli, Valentin et al. (2021). "Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling". In: *arXiv preprint arXiv:2106.01357*.

Peyré, Gabriel et al. (2020). *Computational Optimal Transport*. arXiv: 1803.00567 [stat.ML].

Schrödinger, E. (1932). "Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique". fr. In: *Annales de l'institut Henri Poincaré* 2.4, pp. 269–310. URL: http://www.numdam.org/item/AIHP_1932__2_4_269_0/.

Sohl-Dickstein, Jascha et al. (2015). *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. arXiv: 1503.03585 [cs.LG].

Song, Yang et al. (2021). *Score-Based Generative Modeling through Stochastic Differential Equations*. arXiv: 2011.13456 [cs.LG].

Thank you!