From Diffusion model to Schrodinger bridge

Shikai Fang

2022/08

Group seminar

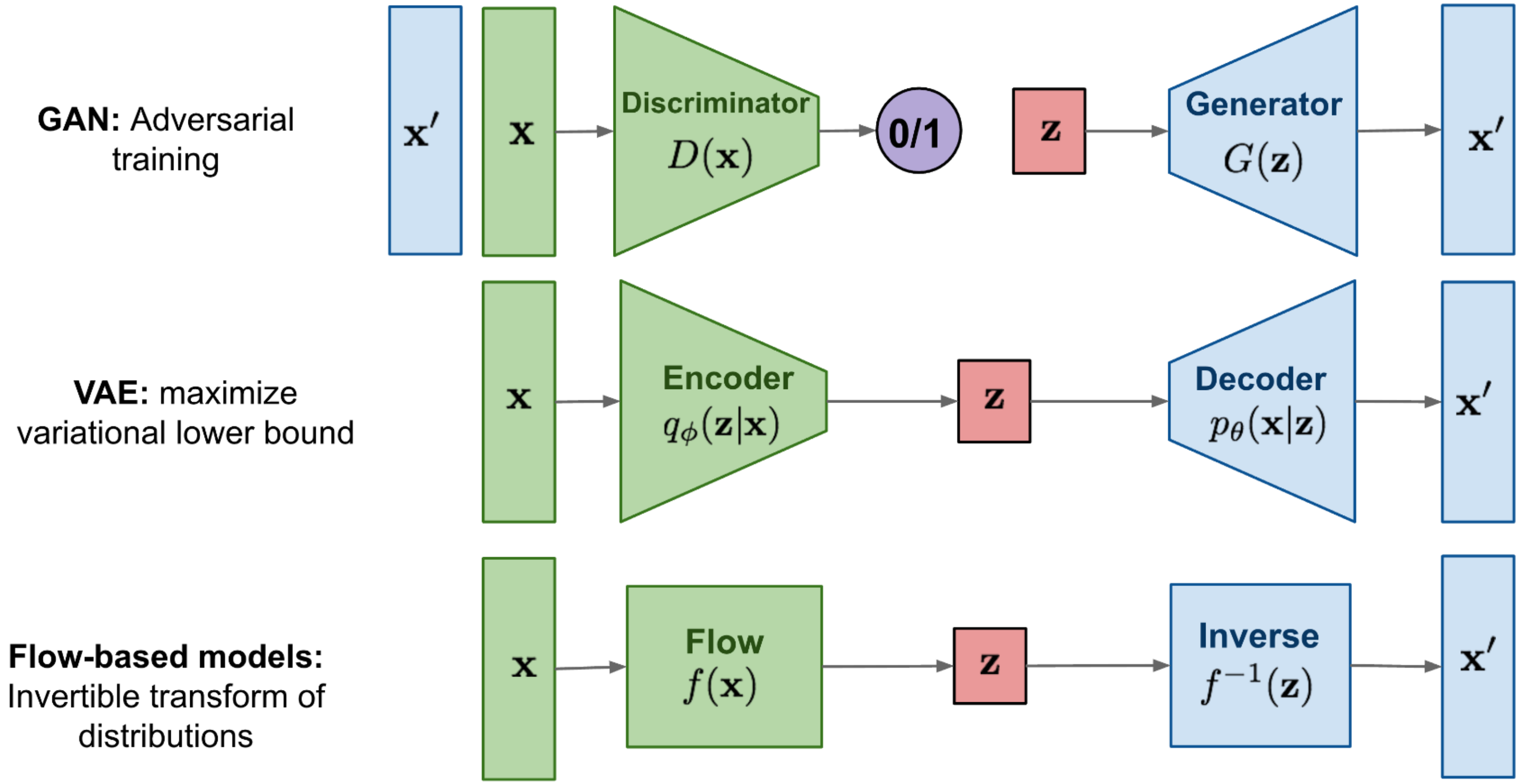
Material form books/paper from Simo Sarkka/Aron Solin, slides from xuecheng Li

Content

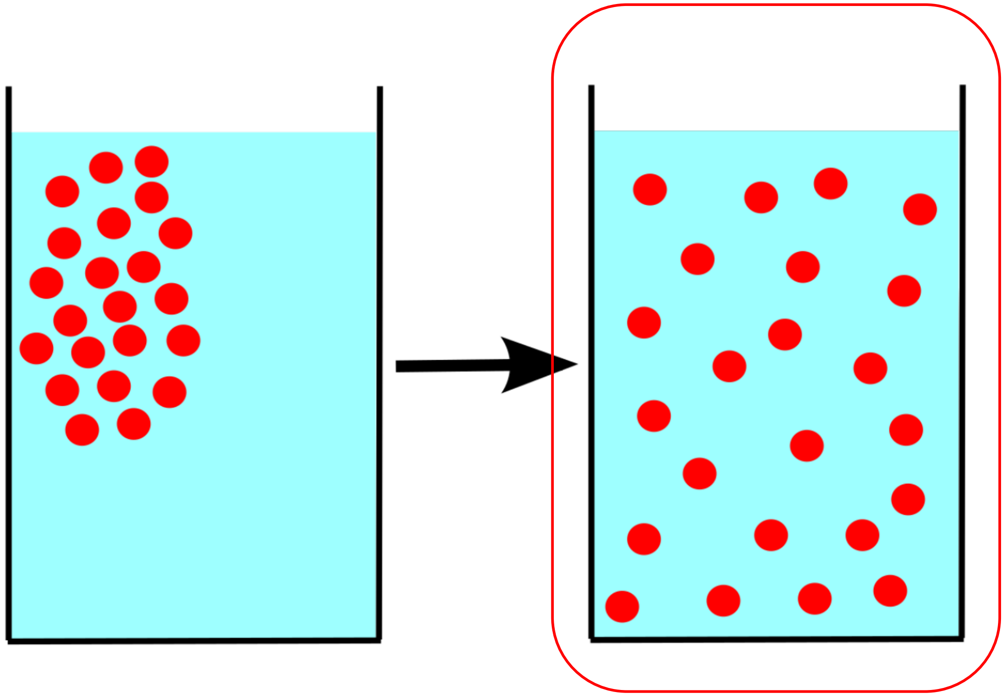
- Diffusion denoising model (DDPM)
- Score matching and Langevin dynamics (SMLD)
- Continues diffusion by SDE
- Schrodinger bridge
- Conditional case for supervised learning

Main-stream generative models:

Gaussian samples + well-trained models -> p(data) / p(x)



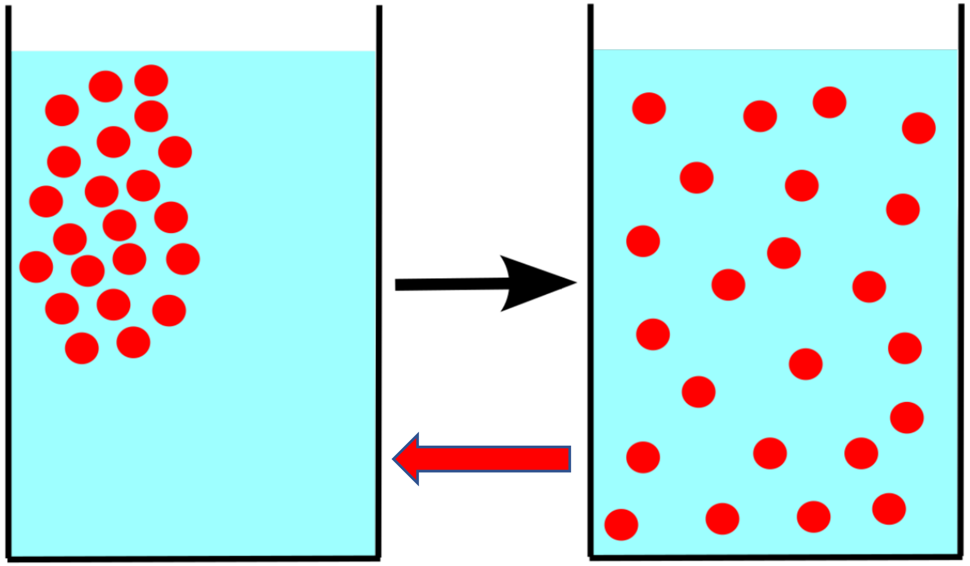
Diffusion



Gaussian Distribution!

Due to random motion, molecules of a high concentration will tend to flow towards a region in space where the concentration is lower.

Diffusion



Can we reverse it?

If we could, how to properly model it ?

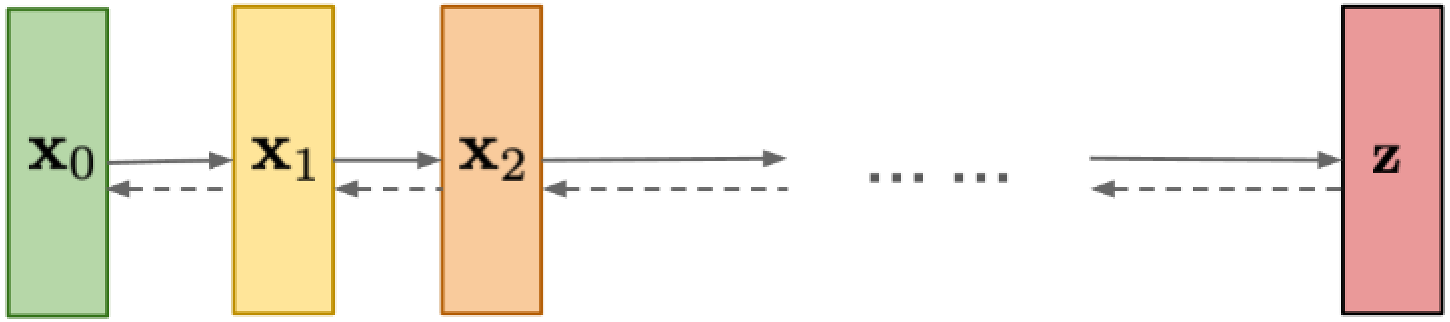
Due to random motion, molecules of a high concentration will tend to flow towards a region in space where the concentration is lower.

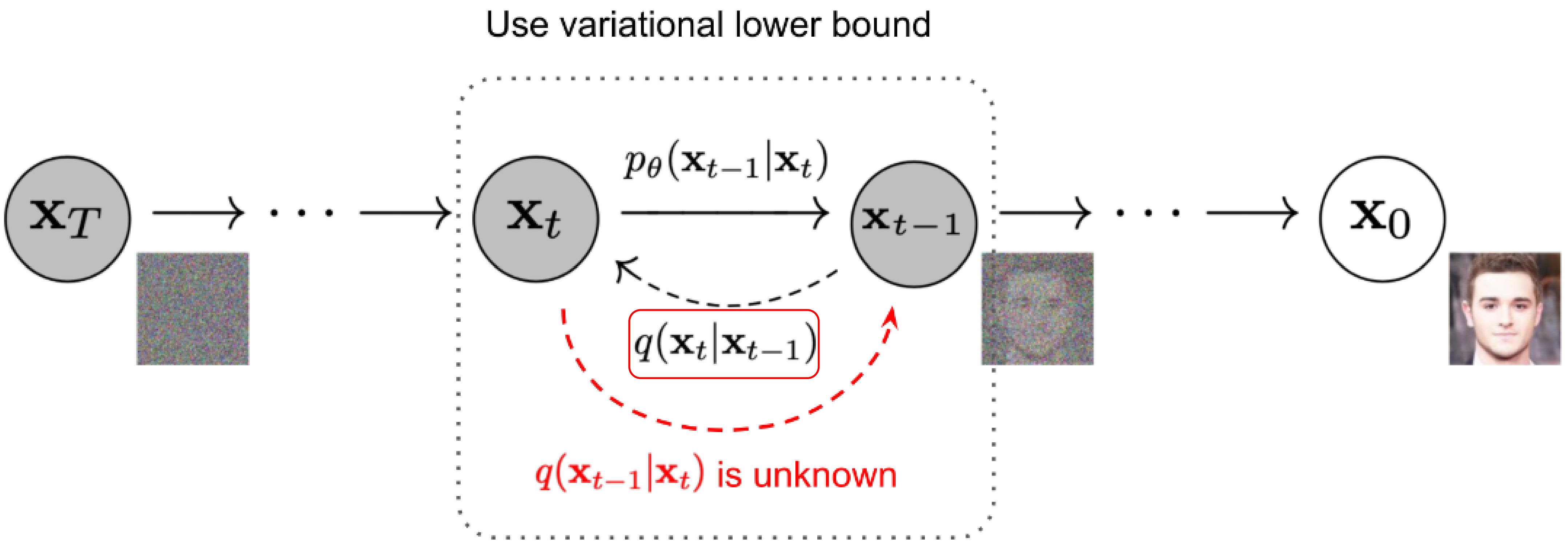
Diffusion model for generative tasks:

Model the RVs transition states with **bi-direct Markov chain**

Forward: **Gradually add noise**
Backward: **Gradually denoising**

Diffusion models:
Gradually add Gaussian noise and then reverse





Forward: **add noise** $\boxed{q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)}$ $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$

$\{\beta_t \in (0,1)\}_{t=1}^{T}$

Why with this form? To ensure $\boxed{T \to \infty, X_T \to \mathcal{N}(0,1)}$

How?

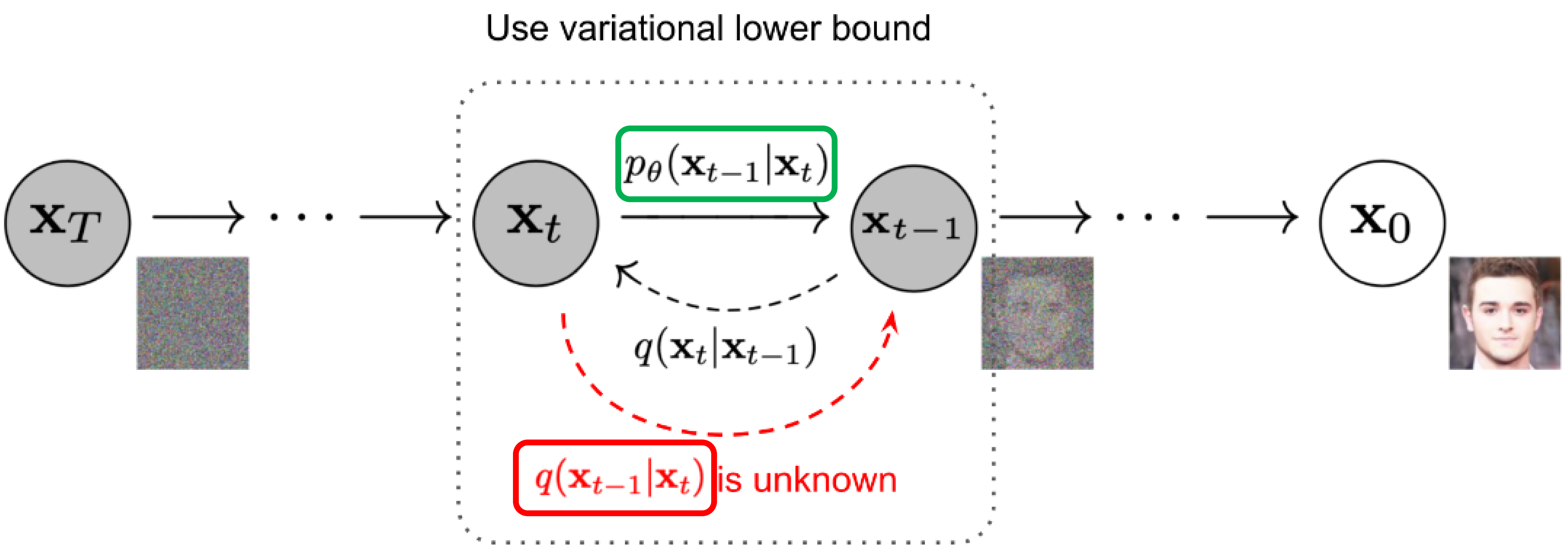
A nice property of the above process is that we can sample $\mathbf{x}_t$ at any arbitrary time step t in a closed form using reparameterization trick. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$:

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\mathbf{z}_{t-1} && \text{;where } \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
&= \sqrt{\alpha_t \alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t \alpha_{t-1}}\bar{\mathbf{z}}_{t-2} && \text{;where } \bar{\mathbf{z}}_{t-2} \text{ merges two Gaussians (*).} \\
&= \dots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{z}
\end{aligned}$$

$$\boxed{q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})}$$

(*) Recall that when we merge two Gaussians with different variance, $\mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I})$ and $\mathcal{N}(\mathbf{0}, \sigma_2^2\mathbf{I})$, the new distribution is $\mathcal{N}(\mathbf{0}, (\sigma_1^2 + \sigma_2^2)\mathbf{I})$. Here the merged standard deviation is $\sqrt{(1-\alpha_t) + \alpha_t(1-\alpha_{t-1})} = \sqrt{1-\alpha_t\alpha_{t-1}}$.

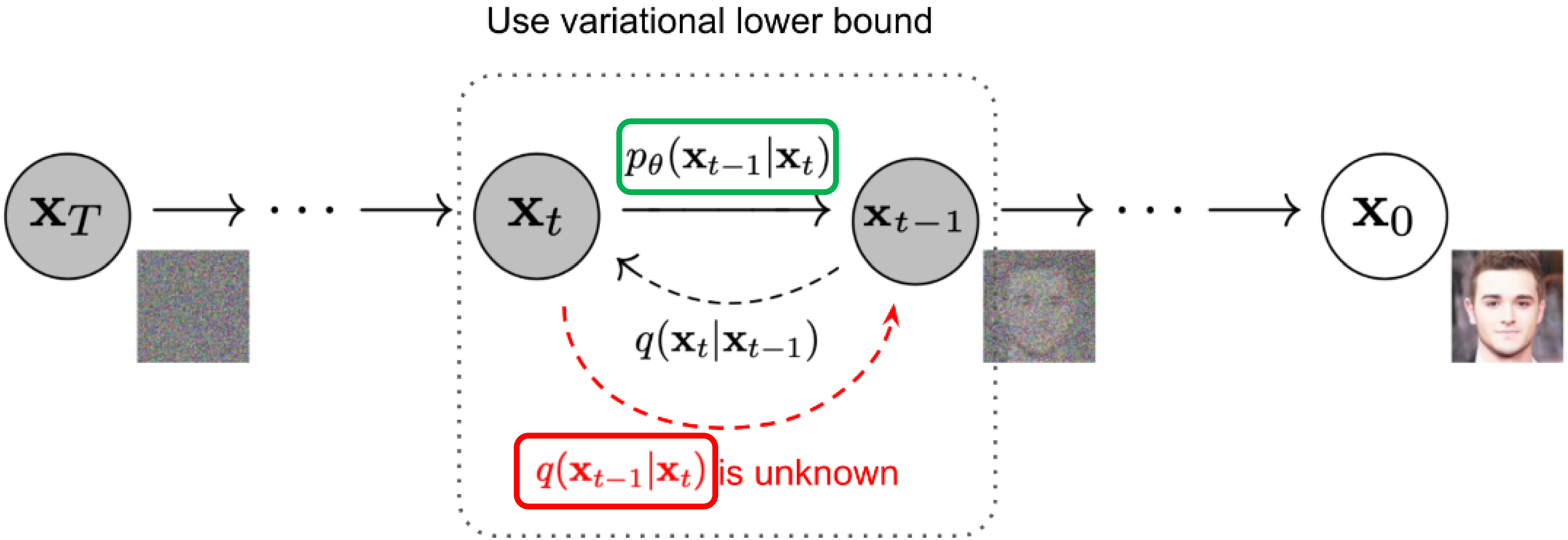
Usually, we can afford a larger update step when the sample gets noisier, so $\beta_1 < \beta_2 < \cdots < \beta_T$ and therefore $\bar{\alpha}_1 > \cdots > \bar{\alpha}_T$.



Forward: **add noise** $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$ $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$

Backward: **denoising**

- **analytic form is intractable (why? Write down the Bayes formula)**
- **build parameterized models to approx.**



Forward: **add noise** $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$ $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$

Backward: **denoising**

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \quad \boxed{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))}$$

The objective function: KL div-> the ELBO

$$
\begin{aligned}
\mathcal{L} &= -\mathbb{E}_{q(\mathbf{x}_0)} \log p_\theta(\mathbf{x}_0) \\
&= -\mathbb{E}_{q(\mathbf{x}_0)} \log \left(\int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \right) \\
&= -\mathbb{E}_{q(\mathbf{x}_0)} \log \left(\int q(\mathbf{x}_{1:T}|\mathbf{x}_0) \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \right) \\
&= -\mathbb{E}_{q(\mathbf{x}_0)} \log \left(\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right) \\
&\leq -\mathbb{E}_{q(\mathbf{x}_{0:T})} \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\
&\boxed{= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})} \right]} \quad \text{ELBO} \\
&= \cdots \\
&= \underbrace{\mathbb{E}_q[-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)]}_{L_0} + \sum_{t=2}^{T} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} + \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T))}_{L_T}
\end{aligned}
$$

The objective function: KL div-> the ELBO

$$\begin{aligned}
\mathcal{L} &= -\mathbb{E}_{q(\mathbf{x}_0)} \log p_\theta(\mathbf{x}_0) \\
&= -\mathbb{E}_{q(\mathbf{x}_0)} \log \left(\int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \right) \\
&= -\mathbb{E}_{q(\mathbf{x}_0)} \log \left(\int q(\mathbf{x}_{1:T}|\mathbf{x}_0) \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \right) \\
&= -\mathbb{E}_{q(\mathbf{x}_0)} \log \left(\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right) \\
&\leq -\mathbb{E}_{q(\mathbf{x}_{0:T})} \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\
&= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})} \right]
\end{aligned}$$

ELBO

$= \cdots$

$$= \underbrace{\mathbb{E}_q[-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)]}_{L_0} + \sum_{t=2}^{T} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} + \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T))}_{L_T}$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}$$

All Gaussian terms, analytic form: new Gaussian

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$$

$$\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \cdot \beta_t$$

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \mathbf{x}_0$$

$$\text{ELBO} = \mathbb{E}_q[\underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}] + \sum_{t=2}^{T} \boxed{\underbrace{D_{\mathrm{KL}}(\boxed{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \,\|\, \boxed{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)})}_{L_{t-1}}} + \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_T))}_{L_T}$$

Key fact:
D_KL div of two gaussian has the analytic form

$$\boxed{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})}$$

$$\boxed{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))}$$

$$\boxed{L_t = \mathbb{E}_q\left[\frac{1}{2\|\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\|_2^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2\right] + C}$$

$$\text{ELBO} = \mathbb{E}_q[\underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}] + \sum_{t=2}^{T} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} + \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T))}_{L_T}$$

Key fact:
D_KL div of two gaussian has the analytic form

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$$

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

$$\boxed{L_t = \mathbb{E}_q\left[\frac{1}{2\|\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\|_2^2} \|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2\right] + C}$$

Further simplify

set $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I}$ $(\sigma_t^2 = \tilde{\beta}_t \text{ or } \beta_t)$

Gaussian noise

Time-aware data2noise mapping

as $\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mathbf{z}_t\right)$, just set $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mathbf{z}_\theta(\mathbf{x}_t, t)\right)$

Further simplify

set $\Sigma_\theta(\mathbf{x}_t, t) = \sigma_t^2\mathbf{I}$ $(\sigma_t^2 = \tilde{\beta}_t \text{ or } \beta_t)$

Gaussian noise

Time-aware data2noise mapping

as $\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boxed{\mathbf{z}_t}\right)$, just set $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boxed{\mathbf{z}_\theta(\mathbf{x}_t, t)}\right)$

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\epsilon_t}\left[\left\|\epsilon_t - \mathbf{z}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\right\|^2\right]$$

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Breakthrough of DDPM: High-resolution generation

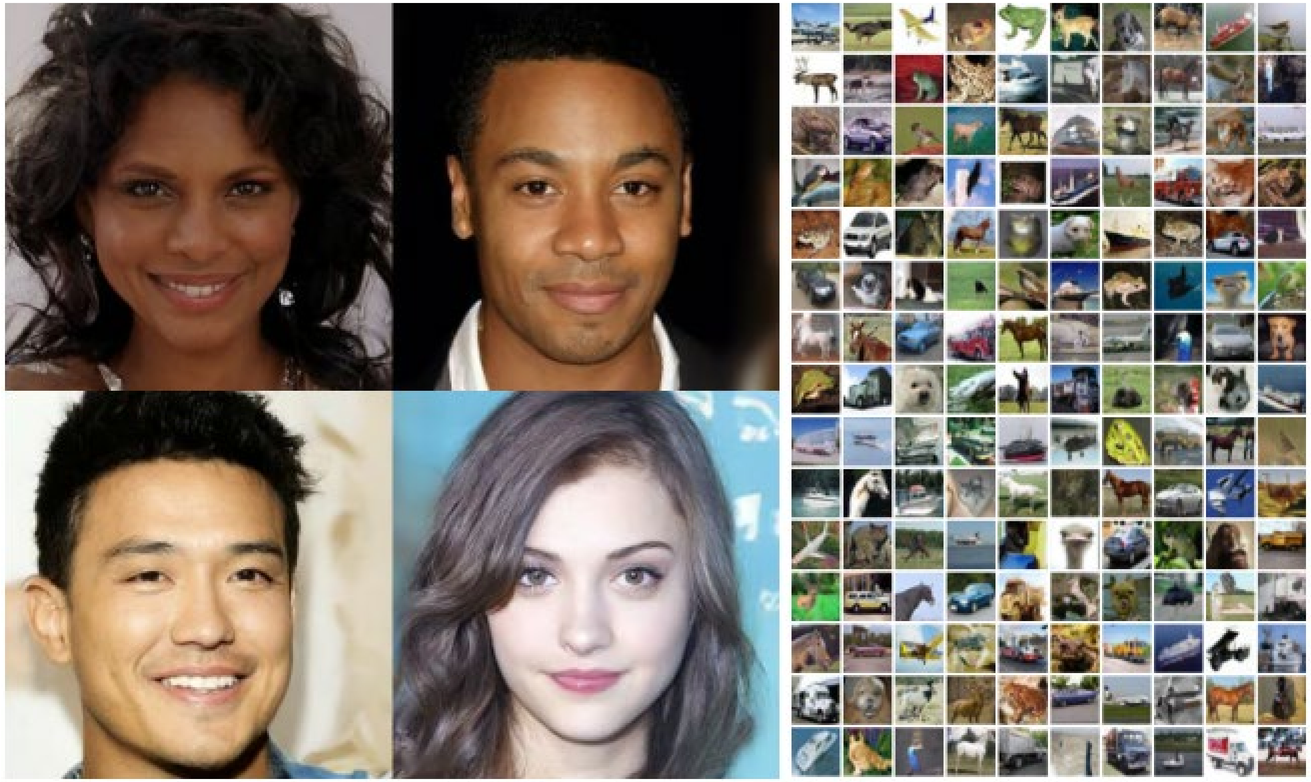


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

Rethink the loss, what does it learn???

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\epsilon_t}\left[\|\epsilon_t - \mathbf{z}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\|^2\right]$$

another perspective from

score-matching method and Langevin dynamic

World of Score function

- If we model a parameterized pdf like:

$$p_\theta(\mathbf{x}) = \frac{e^{-f_\theta(\mathbf{x})}}{Z_\theta}$$

Hard to handle from intractable Z

- Score-based solution, parameterize the **score function**

$$\mathbf{s}_\theta(\mathbf{x}) \approx \boxed{\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})} = -\nabla_{\mathbf{x}} f_\theta(\mathbf{x}) - \underbrace{\nabla_{\mathbf{x}} \log Z_\theta}_{=0} = -\nabla_{\mathbf{x}} f_\theta(\mathbf{x}).$$

  - Score-matching, family of methods to approx. score function by minimizing:

$$\mathbb{E}_{p(\mathbf{x})}\left[\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2\right]$$

Fisher divergence

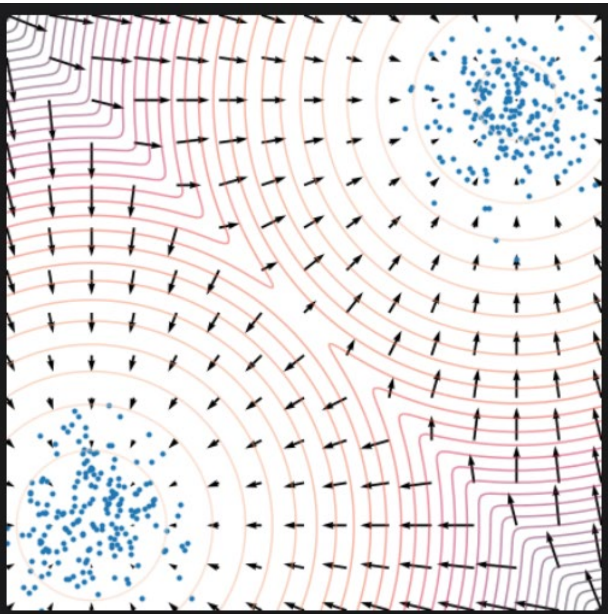
World of Score function

- Given the (approx.) **score function,** how to draw sample?
- **By Langevin dynamics**

Langevin dynamics provides an MCMC procedure to sample from a distribution $p(\mathbf{x})$ using only its score function $\nabla_\mathbf{x} \log p(\mathbf{x})$. Specifically, it initializes the chain from an arbitrary prior distribution $\mathbf{x}_0 \sim \pi(\mathbf{x})$, and then iterates the following

$$\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + \epsilon \nabla_\mathbf{x} \log p(\mathbf{x}) + \sqrt{2\epsilon}\, \mathbf{z}_i, \quad i = 0, 1, \cdots, K, \qquad ($$

where $\mathbf{z}_i \sim \mathcal{N}(0, I)$. When $\epsilon \to 0$ and $K \to \infty$, $\mathbf{x}_K$ obtained from the procedure in (6) converges to a sample from $p(\mathbf{x})$ under some regularity conditions. In practice, the error is negligible when ϵ is sufficiently small and K is sufficiently large.



Using Langevin dynamics to sample from a mixture of two Gaussians.

Rethink the loss, another perspective from

score-based method and Langevin dynamic

$$p_{\alpha_i}(\mathbf{x}_i \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_i; \sqrt{\alpha_i}\mathbf{x}_0, (1-\alpha_i)\mathbf{I}), \text{ where } \alpha_i := \prod_{j=1}^{i}(1-\beta_j).$$

Score-matching: $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} (1-\alpha_i)\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{p_{\alpha_i}(\tilde{\mathbf{x}}|\mathbf{x})}[\|\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, i) - \nabla_{\tilde{\mathbf{x}}} \log p_{\alpha_i}(\tilde{\mathbf{x}} \mid \mathbf{x})\|_2^2].$

Langevin dynamic: $\mathbf{x}_i^m = \mathbf{x}_i^{m-1} + \epsilon_i \mathbf{s}_{\boldsymbol{\theta}*}(\mathbf{x}_i^{m-1}, \sigma_i) + \sqrt{2\epsilon_i}\mathbf{z}_i^m, \quad m = 1, 2, \cdots, M,$

$\Downarrow$

Loss function of DDPM $L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\epsilon_t}\left[\left\|\epsilon_t - \mathbf{z}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\right\|^2\right]$

From discrete to continues: add noise by SDE

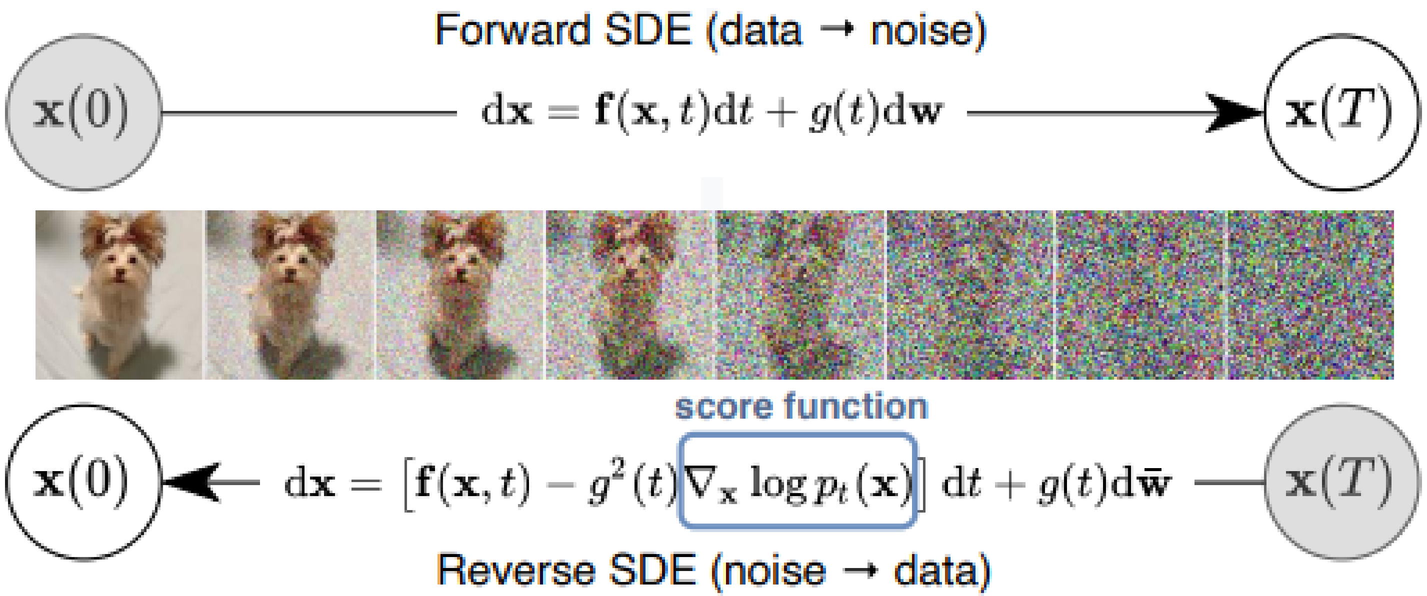


Figure 1: **Solving a reverse-time SDE yields a score-based generative model.** Transforming data to a simple noise distribution can be accomplished with a continuous-time SDE. This SDE can be reversed if we know the score of the distribution at each intermediate time step, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$.

score matching (Hyvärinen, 2005; Song et al., 2019a). To estimate $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, we can train a time-dependent score-based model $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ via a continuous generalization to Eqs. (1) and (3):

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\arg\min}\, \mathbb{E}_t \Big\{ \lambda(t) \mathbb{E}_{\mathbf{x}(0)} \mathbb{E}_{\mathbf{x}(t)|\mathbf{x}(0)} \Big[\big\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log p_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0)) \big\|_2^2 \Big] \Big\}. \quad (7)$$

Consistent with DDPM

Likewise for the perturbation kernels $\{p_{\alpha_i}(\mathbf{x} \mid \mathbf{x}_0)\}_{i=1}^N$ of DDPM, the discrete Markov chain is

$$\mathbf{x}_i = \sqrt{1-\beta_i}\mathbf{x}_{i-1} + \sqrt{\beta_i}\mathbf{z}_{i-1}, \quad i = 1, \cdots, N. \tag{10}$$

As $N \to \infty$, Eq. (10) converges to the following SDE,

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}\mathbf{w}. \tag{11}$$

Recall: the forward SDE is still given
the backward SDE is learned

One more step: also learn the forward SDE

We get **Schrodinger Bridge(SB)**

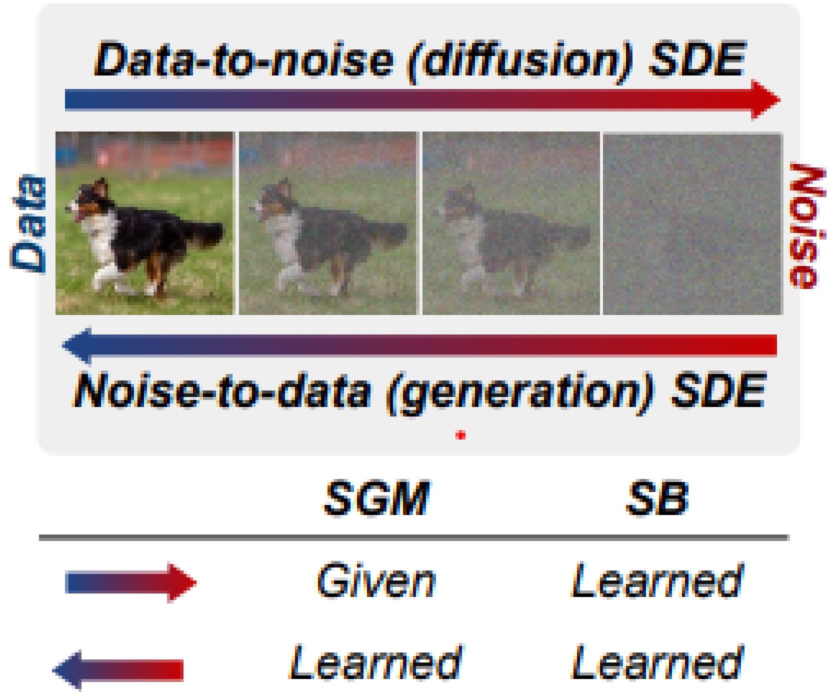


Figure 1: Both Score-based Generative Model (SGM) and Schrödinger Bridge (SB) transform between two distributions. While SGM requires pre-specifying the data-to-noise diffusion, SB instead *learns* the process.

Motivation for using SB in diffusion models:

Few step converges (SGM need large steps to be Gaussian)

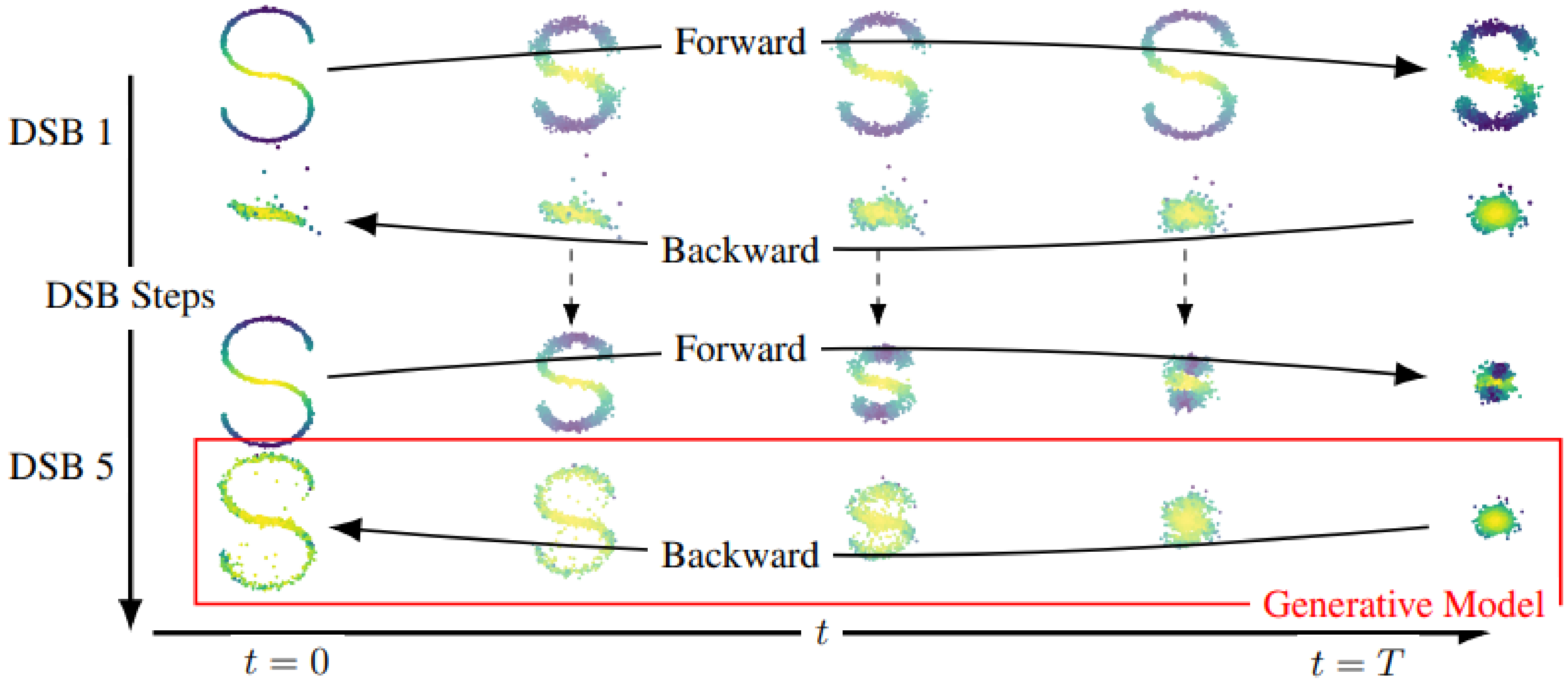


Figure 1: The reference forward diffusion initialized from the 2-dimensional data distribution fails to converge to the Gaussian prior in $T = 0.2$ diffusion-time ($N = 20$ discrete time steps), and the reverse diffusion initialized from the Gaussian prior does not converge to the data distribution. However, convergence does occur after 5 DSB iterations.

Formal formulation of SB:

2.2 Schrödinger Bridge (SB)

Following the dynamic expression of SB (Pavon & Wakolbinger, 1991; Dai Pra, 1991), consider

$$\min_{\mathbb{Q}\in\mathcal{P}(p_{\text{data}},p_{\text{prior}})} D_{\text{KL}}(\mathbb{Q}\,||\,\mathbb{P}), \tag{5}$$

where $\mathbb{Q} \in \mathcal{P}(p_{\text{data}}, p_{\text{prior}})$ belongs to a set of path measure with p_{data} and p_{prior} as its marginal densities at $t = 0$ and T. On the other hand, $\mathbb{P}$ denotes a reference measure, which we will set to the path measure of (1) for later convenience. The optimality condition to (5) is characterized by two PDEs that are coupled through their boundary conditions. We summarize the related result below.

Theorem 1 (SB optimality; Chen et al. (2021); Pavon & Wakolbinger (1991); Caluya & Halder (2021)). *Let $\Psi(t,\boldsymbol{x})$ and $\widehat{\Psi}(t,\boldsymbol{x})$ be the solutions to the following PDEs:*

$$\begin{cases} \frac{\partial \Psi}{\partial t} = -\nabla_{\boldsymbol{x}}\Psi^{\mathsf{T}} f - \frac{1}{2}\operatorname{Tr}(g^2\nabla_{\boldsymbol{x}}^2\Psi) \\ \frac{\partial \widehat{\Psi}}{\partial t} = -\nabla_{\boldsymbol{x}}\cdot(\widehat{\Psi} f) + \frac{1}{2}\operatorname{Tr}(g^2\nabla_{\boldsymbol{x}}^2\widehat{\Psi}) \end{cases} \quad \text{s.t. } \Psi(0,\cdot)\widehat{\Psi}(0,\cdot) = p_{\text{data}},\ \Psi(T,\cdot)\widehat{\Psi}(T,\cdot) = p_{\text{prior}} \tag{6}$$

Then, the solution to the optimization (5) can be expressed by the path measure of the following forward (7a), or equivalently backward (7b), SDE:

$$\mathrm{d}\mathbf{X}_t = [f + g^2\,\nabla_{\boldsymbol{x}}\log\Psi(t,\mathbf{X}_t)]\mathrm{d}t + g\,\mathrm{d}\mathbf{W}_t, \quad \mathbf{X}_0 \sim p_{\text{data}}, \tag{7a}$$

$$\mathrm{d}\mathbf{X}_t = [f - g^2\,\nabla_{\boldsymbol{x}}\log\widehat{\Psi}(t,\mathbf{X}_t)]\mathrm{d}t + g\,\mathrm{d}\mathbf{W}_t, \quad \mathbf{X}_T \sim p_{\text{prior}}, \tag{7b}$$

where $\nabla_{\boldsymbol{x}}\log\Psi(t,\mathbf{X}_t)$ and $\nabla_{\boldsymbol{x}}\log\Psi(t,\mathbf{X}_t)$ are the optimal forward and backward drifts for SB.

[2]Hereafter, we will sometimes drop $f \equiv f(t,\mathbf{X}_t)$ and $g \equiv g(t)$ for brevity.

Formal formulation of SB:

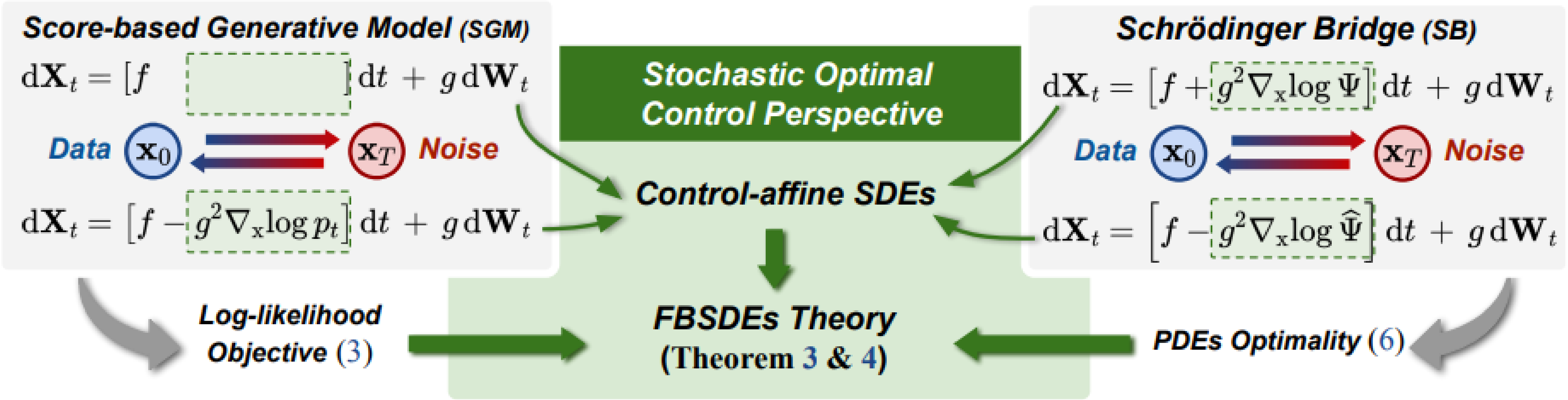


Figure 2: Schematic diagram of the our stochastic optimal control interpretation, and how it connects the objective of SGM (3) and optimality of SB (6) through Forward-Backward SDEs theory.

Formal formulation of SB:

Theorem 3 (FBSDEs to SB optimality (6)). *Consider the following set of coupled SDEs,*

$$\begin{cases} \mathrm{d}\mathbf{X}_t = (f + g\mathbf{Z}_t)\,\mathrm{d}t + g\mathrm{d}\mathbf{W}_t & \text{(13a)} \\ \mathrm{d}\mathbf{Y}_t = \frac{1}{2}\mathbf{Z}_t^\mathsf{T}\mathbf{Z}_t\mathrm{d}t + \mathbf{Z}_t^\mathsf{T}\mathrm{d}\mathbf{W}_t & \text{(13b)} \\ \mathrm{d}\widehat{\mathbf{Y}}_t = \left(\frac{1}{2}\widehat{\mathbf{Z}}_t^\mathsf{T}\widehat{\mathbf{Z}}_t + \nabla_{\boldsymbol{x}} \cdot (g\widehat{\mathbf{Z}}_t - f) + \widehat{\mathbf{Z}}_t^\mathsf{T}\mathbf{Z}_t\right)\mathrm{d}t + \widehat{\mathbf{Z}}_t^\mathsf{T}\mathrm{d}\mathbf{W}_t & \text{(13c)} \end{cases}$$

where f and g satisfy the same regularity conditions in Lemma 2 (see Footnote 4), and the boundary conditions are given by $\mathbf{X}(0) = \boldsymbol{x}_0$ and $\mathbf{Y}_T + \widehat{\mathbf{Y}}_T = \log p_{\text{prior}}(\mathbf{X}_T)$. Suppose $\Psi, \widehat{\Psi} \in C^{1,2}$, then the nonlinear Feynman-Kac relations between the FBSDEs (13) and PDEs (6) are given by

$$\begin{aligned} \mathbf{Y}_t &\equiv \mathbf{Y}(t, \mathbf{X}_t) = \log \Psi(t, \mathbf{X}_t), & \mathbf{Z}_t &\equiv \mathbf{Z}(t, \mathbf{X}_t) = g\nabla_{\boldsymbol{x}} \log \Psi(t, \mathbf{X}_t), \\ \widehat{\mathbf{Y}}_t &\equiv \widehat{\mathbf{Y}}(t, \mathbf{X}_t) = \log \widehat{\Psi}(t, \mathbf{X}_t), & \widehat{\mathbf{Z}}_t &\equiv \widehat{\mathbf{Z}}(t, \mathbf{X}_t) = g\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(t, \mathbf{X}_t). \end{aligned} \qquad (14)$$

Furthermore, $(\mathbf{Y}_t, \widehat{\mathbf{Y}}_t)$ obey the following relation:

$$\mathbf{Y}_t + \widehat{\mathbf{Y}}_t = \log p_t^{\text{SB}}(\mathbf{X}_t).$$

"Policy" which decide the forward and backward SDEs

Parameterized them to learn

Objective function

Joint training

Theorem 4 (Log-likelihood of SB model). *Given the solution satisfying the FBSDE system in (13), the log-likelihood of the SB model* $(\mathbf{Z}_t, \widehat{\mathbf{Z}}_t)$*, at a data point* $\boldsymbol{x}_0$*, can be expressed as*

$$\log p_0^{\text{SB}}(\boldsymbol{x}_0) = \mathbb{E}\left[\log p_T(\mathbf{X}_T)\right] - \int_0^T \mathbb{E}\left[\frac{1}{2}\|\mathbf{Z}_t\|^2 + \frac{1}{2}\|\widehat{\mathbf{Z}}_t - g\nabla_{\boldsymbol{x}} \log p_t^{\text{SB}} + \mathbf{Z}_t\|^2 - \frac{1}{2}\|g\nabla_{\boldsymbol{x}} \log p_t^{\text{SB}} - \mathbf{Z}_t\|^2 - \nabla_{\boldsymbol{x}} \cdot f\right] \mathrm{d}t \quad (15)$$

$$= \mathbb{E}\left[\log p_T(\mathbf{X}_T)\right] - \int_0^T \mathbb{E}\left[\frac{1}{2}\|\mathbf{Z}_t\|^2 + \frac{1}{2}\|\widehat{\mathbf{Z}}_t\|^2 + \nabla_{\boldsymbol{x}} \cdot (g\widehat{\mathbf{Z}}_t - f) + \widehat{\mathbf{Z}}_t^\top \mathbf{Z}_t\right] \mathrm{d}t, \quad (16)$$

where the expectation is taken over the forward SDE (13a) with the initial condition $\mathbf{X}_0 = \boldsymbol{x}_0$.

Similar to (3), Theorem 4 suggests a parameterized lower bound to the log-likelihoods, *i.e.* $\log p_0^{\text{SB}}(\boldsymbol{x}_0) \geq \mathcal{L}_{\text{SB}}(\boldsymbol{x}_0; \theta, \phi)$ where $\mathcal{L}_{\text{SB}}(\boldsymbol{x}_0; \theta, \phi)$ shares the same expression in (16) except that $\boxed{\mathbf{Z}_t \approx \mathbf{Z}(t, \boldsymbol{x}; \theta) \text{ and } \widehat{\mathbf{Z}}_t \approx \widehat{\mathbf{Z}}(t, \boldsymbol{x}; \phi)}$ are approximated with some parameterized models (*e.g.* DNNs). Note that $\nabla_{\boldsymbol{x}} \log p_t^{\text{SB}}$ is *intractable* in practice for any nontrivial $(\mathbf{Z}_t, \widehat{\mathbf{Z}}_t)$. Hence, we use the divergence-based objective in (16) as our training objective of both policies.

Alternative training

$$\widetilde{\mathcal{L}}_{\text{SB}}(\boldsymbol{x}_0; \phi) = -\int_0^T \mathbb{E}_{\mathbf{X}_t \sim (7\text{a})}\left[\frac{1}{2}\|\widehat{\mathbf{Z}}(t, \mathbf{X}_t; \phi)\|^2 + g\nabla_{\boldsymbol{x}} \cdot \widehat{\mathbf{Z}}(t, \mathbf{X}_t; \phi) + \mathbf{Z}_t^\top \widehat{\mathbf{Z}}(t, \mathbf{X}_t; \phi)\right] \mathrm{d}t, \quad (18)$$

$$\widetilde{\mathcal{L}}_{\text{SB}}(\boldsymbol{x}_T; \theta) = -\int_0^T \mathbb{E}_{\mathbf{X}_t \sim (7\text{b})}\left[\frac{1}{2}\|\mathbf{Z}(t, \mathbf{X}_t; \theta)\|^2 + g\nabla_{\boldsymbol{x}} \cdot \mathbf{Z}(t, \mathbf{X}_t; \theta) + \widehat{\mathbf{Z}}_t^\top \mathbf{Z}(t, \mathbf{X}_t; \theta)\right] \mathrm{d}t. \quad (19)$$

The flexibility of SB

2.2 Schrödinger Bridge (SB)

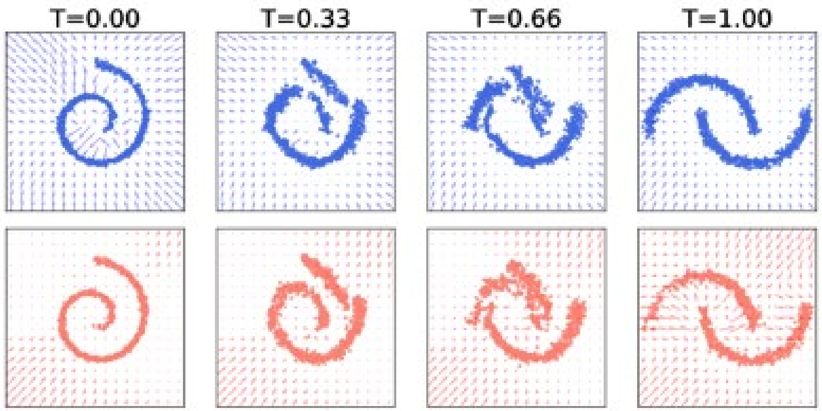
Following the dynamic expression of SB (Pavon & Wakolbinger, 1991; Dai Pra, 1991), consider

$$\min_{\mathbb{Q}\in\mathcal{P}(p_{\text{data}},p_{\text{prior}})} D_{\text{KL}}(\mathbb{Q}\,||\,\mathbb{P}), \tag{5}$$

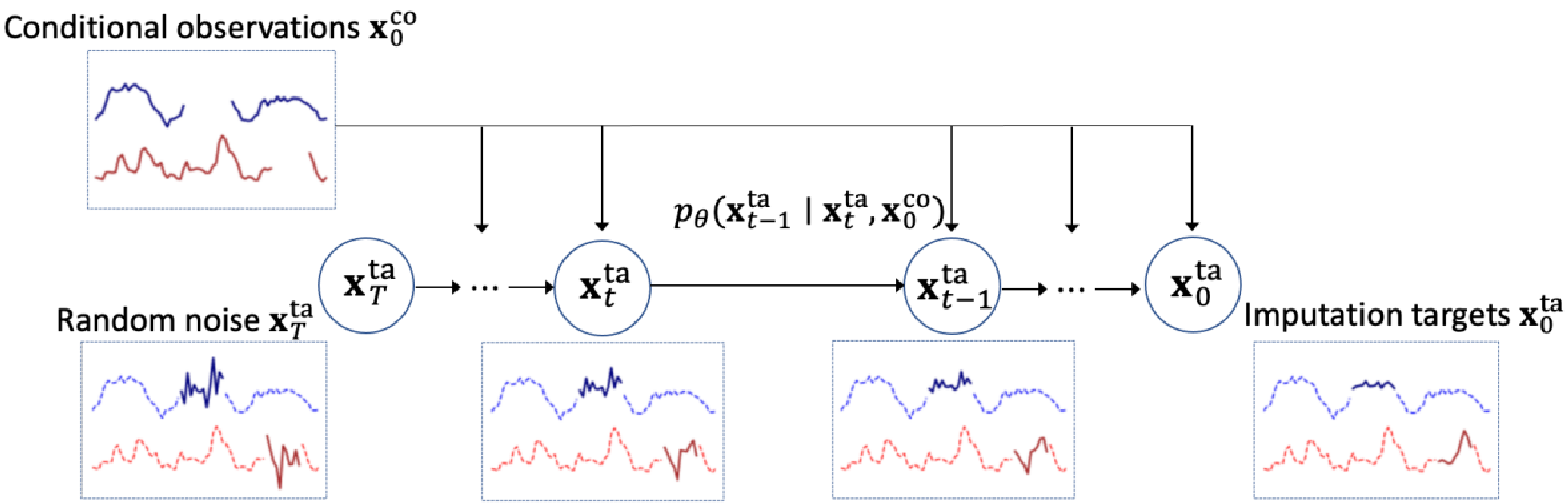
where $\mathbb{Q}\in\mathcal{P}(p_{\text{data}}\boxed{, p_{\text{prior}})}$ belongs to a set of path measure with p_{data} and p_{prior} as its marginal densities at $t=0$ and T. On the other hand, $\mathbb{P}$ denotes a reference measure, which we will set to the path measure of (1) for later convenience. The optimality condition to (5) is characterized by two

Can be arbitrary prior, not only Gaussian!

Spiral ⇆ Moon (`moon-to-spiral`)



Conditional diffusion/SB: more than unsupervised



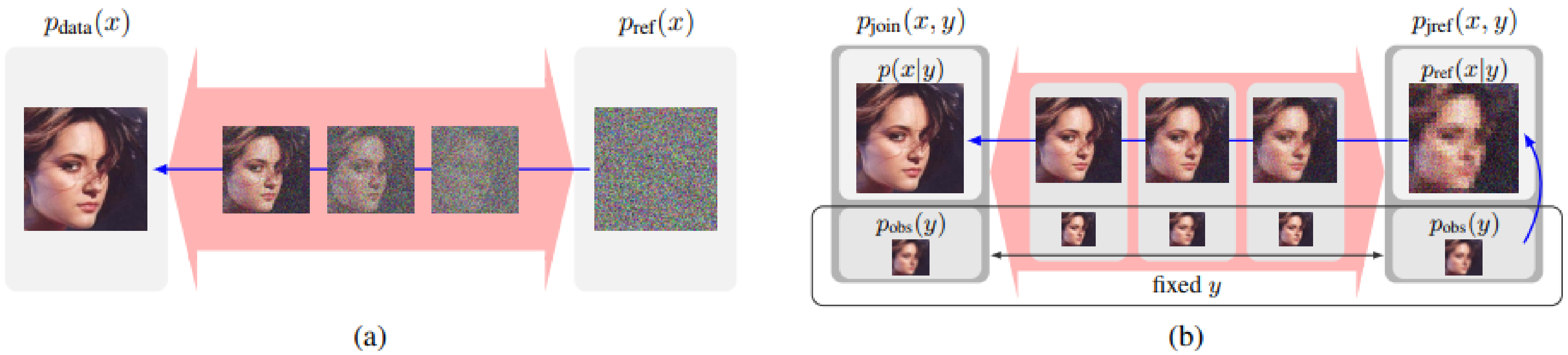


Figure 1: (a) An unconditional Schrödinger bridge (SB) between $p_{\mathrm{data}}(x)$ and $p_{\mathrm{ref}}(x)$; (b) our proposed conditional Schrödinger bridge (CSB) on the extended space between $p_{\mathrm{join}}(x, y)$ and $p_{\mathrm{jref}}(x, y)$. The blue arrows denote the direction of the generative procedure at simulation time.

Roadmap from diffusion model to SB