Diffusion Models

Instructor: John Thickstun

Discussion Board: Available on Ed

Zoom Link: Available on Canvas

Instructor Contact: thickstn@cs.washington.edu

Course Webpage: https://courses.cs.washington.edu/courses/cse599i/20au/

Score Function Sampling

- Learn a score function $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ to approximate $s(x) = \nabla_x \log p(x)$.

- Sample via Langevin dynamics:

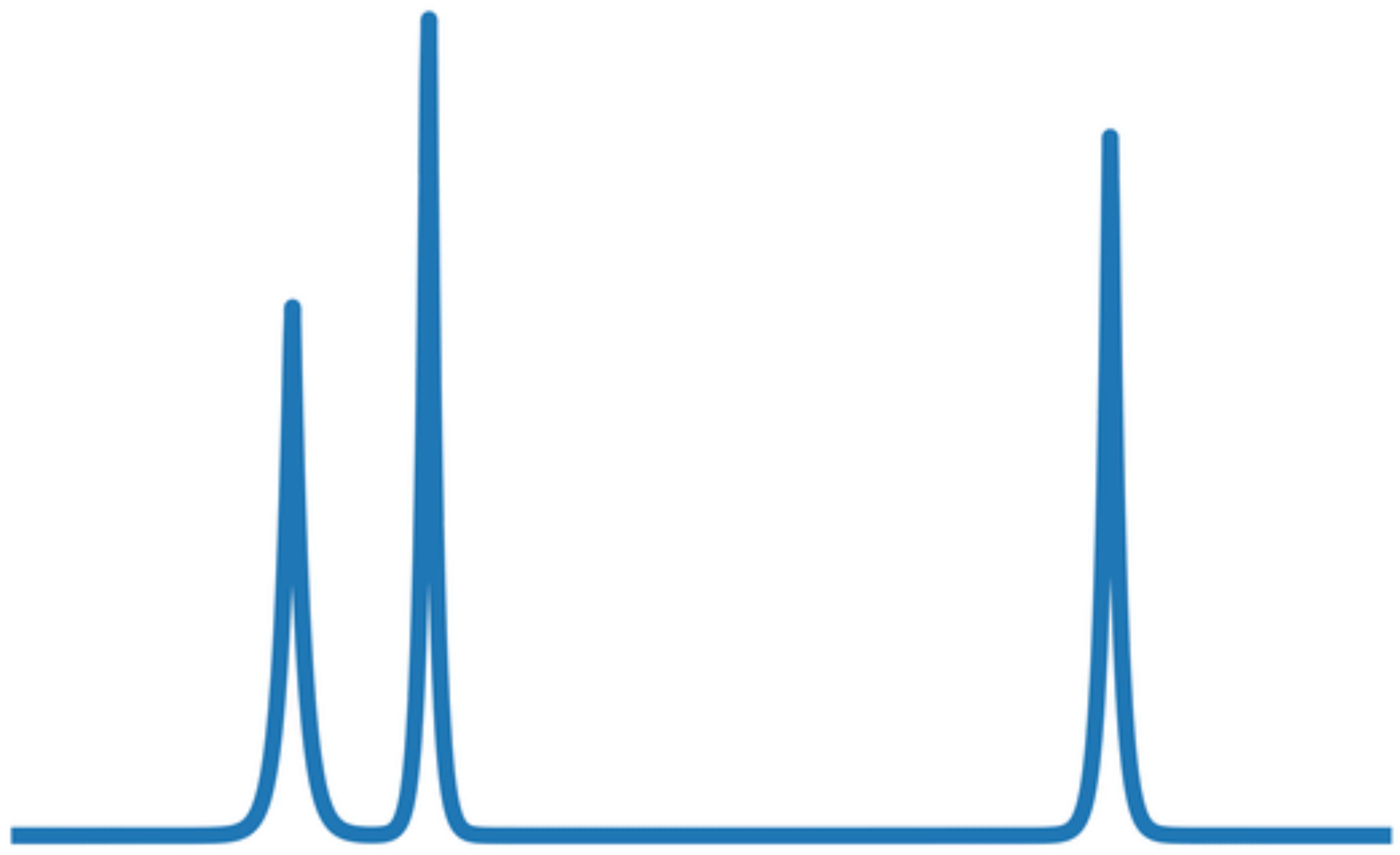
$$
\begin{aligned}
x_{t+1} &= x_t - \eta \nabla_x \log p_\theta(x_t) + \sqrt{2\eta}\varepsilon_t \\
&= x_t + \eta s_\theta(x_t) + \sqrt{2\eta}\varepsilon_t .
\end{aligned}
$$

- In continuous limit as $\eta \to 0,\ D(x_t \parallel p_\theta) \to 0 \text{ as } t \to \infty$.

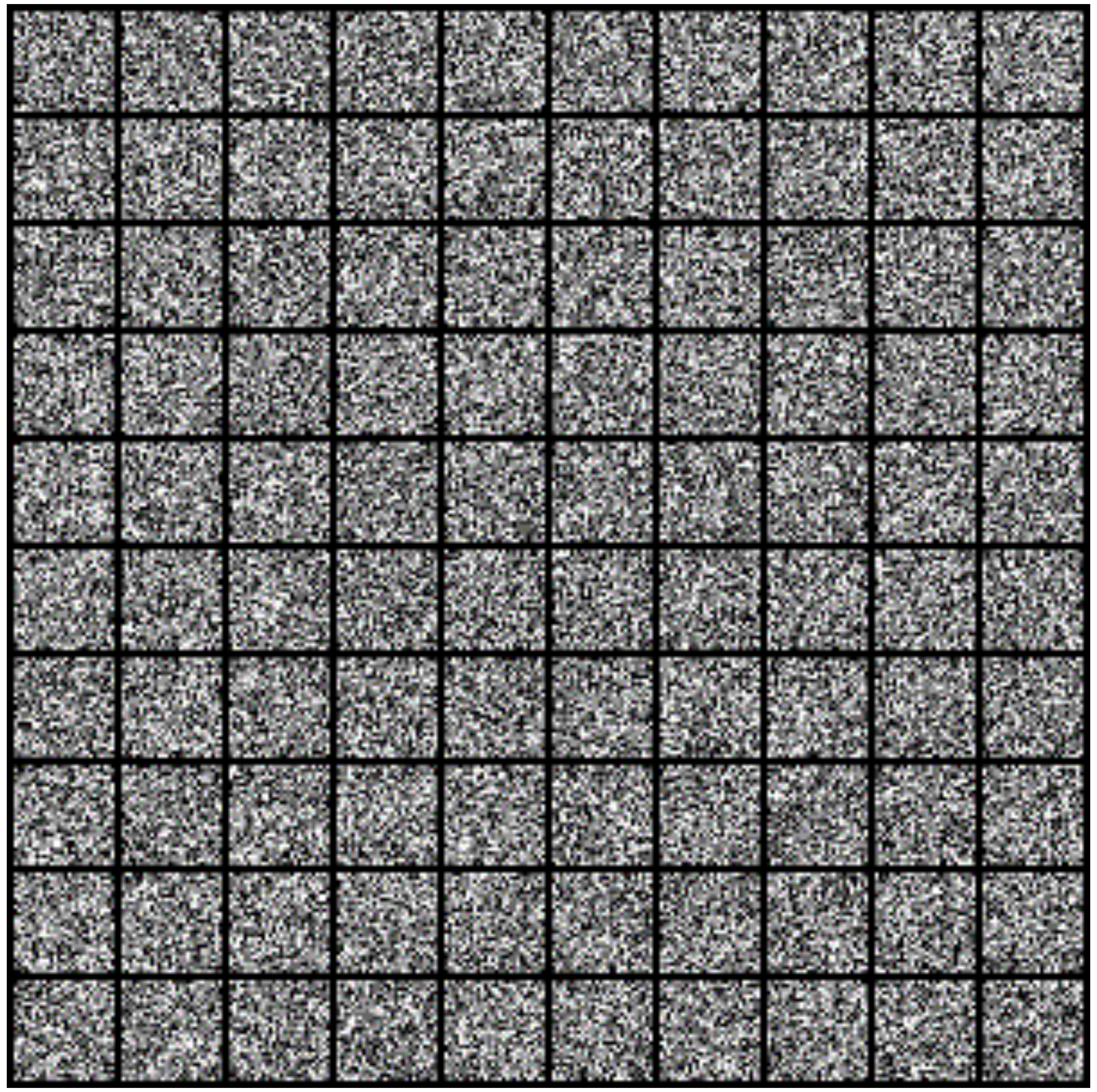
- How long will I need to run this Markov chain? A long long time.

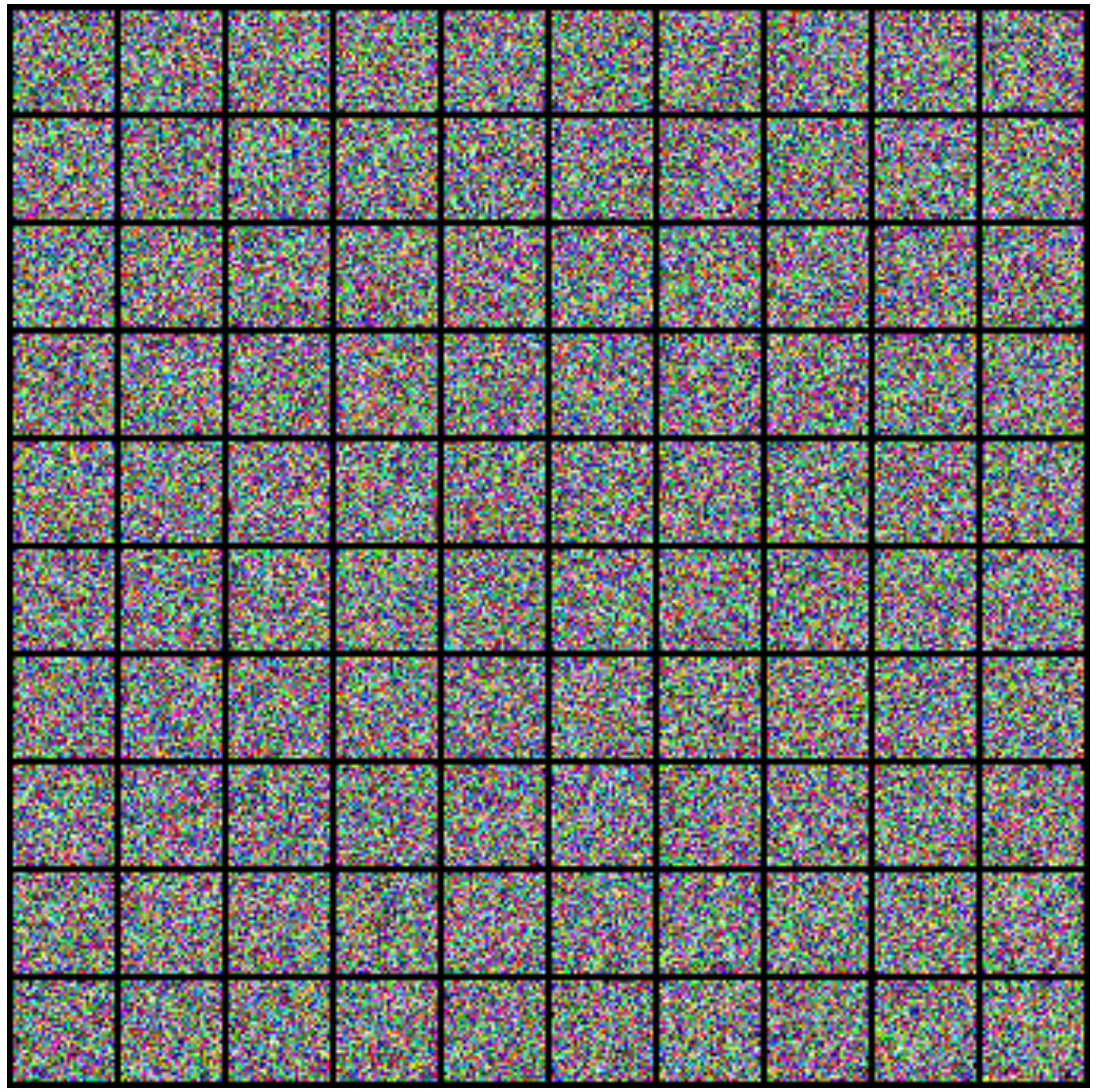
Accelerated Mixing

- Loss surface is highly non-Lipschitz.
- Let $p_\sigma(\mathbf{x})$ be the distribution of $\mathbf{x} + \epsilon_\sigma$.
- Where $\mathbf{x} \sim p, \text{ and } \epsilon_\sigma \sim \mathcal{N}(0, \sigma^2 I).$
- Simulated annealing: smooth out the likelihood surface.
- Gradually un-smooth, while slowing down the learning rate.



Langevin Dynamics in Practice





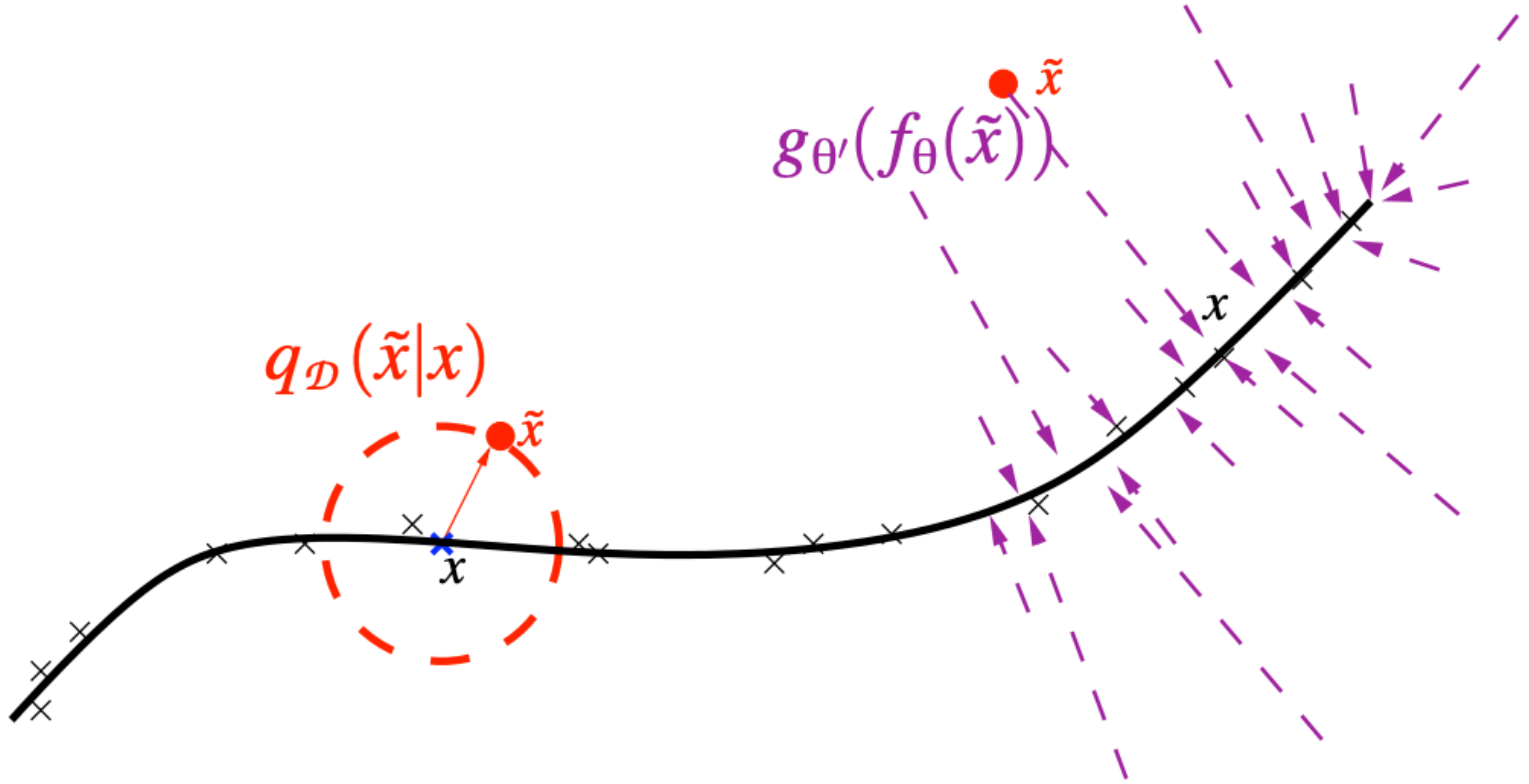
Sampling from MNIST (left) and CIFAR-10 (right) EBMs Using Langevin Dynamics

Song and Ermon, Neurips 2019

Denoising Autoencoders

- Learn to denoise noise-corrupted data: recover x given $\tilde{x} = x + \varepsilon$.
- Suppose data lives on a low-dimensional manifold.
- Learn to project noise-corrupted data back onto the manifold
- Optimize a reconstruction objective:

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \mathop{\mathbb{E}}_{(x,\tilde{x})\sim p} \|x - g_\theta(f_\phi(\tilde{x}))\|^2.$$

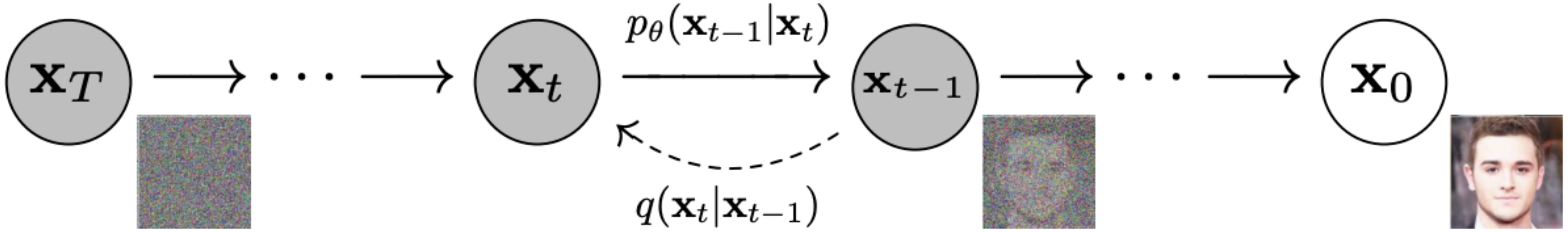


Vincent et. al., JMLR 2010

Generative Denoising AE

- Can we turn denoising autoencoders into a generative model?

- Idea: construct a Markov chain of progressively less noisy samples:



Ho, Jain, and Abbeel, Neurips 2020

- What if each transition $p_\theta(x_{t-1}|x_t)$ were given by a denoising autoencoder?

Diffusion Models

- Want to model the distribution of $\mathbf{x} \sim p$, where $\mathbf{x} \in \mathbb{R}^d$.
- Construct a Markov chain $\mathbf{x}_0, \ldots, \mathbf{x}_T \in \mathbb{R}^d$.
- Learn a transition model, e.g. $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}|\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)\right)$.
- Base case: $p(\mathbf{x}_T) = \mathcal{N}(0, I)$.
- Marginal distribution over $\mathbf{x}_0$ is $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_0, \ldots, \mathbf{x}_T)\, d\mathbf{x}_1 \ldots d\mathbf{x}_T$.
- Learn the parameters so that $p(\mathbf{x}_0) \approx p_\theta(\mathbf{x}_0)$.

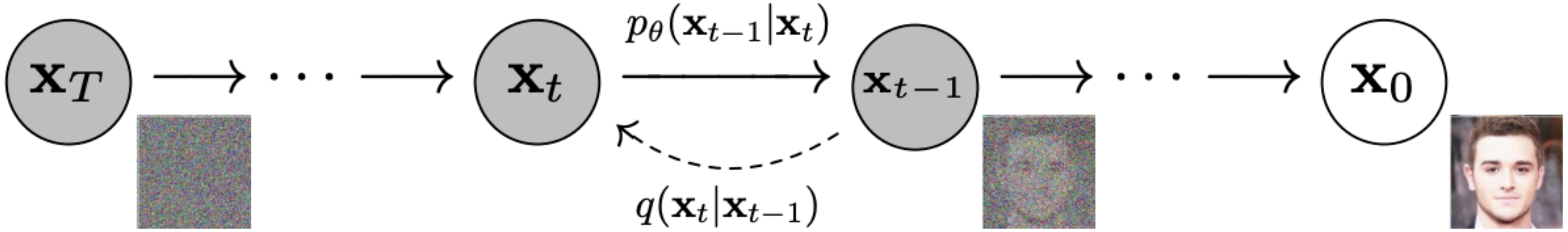
Maximize the Likelihood?

- Want to model the distribution of $\mathbf{x} \sim p,$ where $\mathbf{x} \in \mathbb{R}^d$.
- Marginal distribution over $\mathbf{x}_0$ is $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_0, \ldots, \mathbf{x}_T) \, d\mathbf{x}_1 \ldots d\mathbf{x}_T$.
- Learn the parameters so that $p(\mathbf{x}_0) \approx p_\theta(\mathbf{x}_0)$.
- Likelihood is intractable:

$$\arg\max_\theta \mathop{\mathbb{E}}_{\mathbf{x}_0 \sim p} [\log p_\theta(\mathbf{x}_0)] = \mathop{\mathbb{E}}_{\mathbf{x}_0 \sim p} \left[\log \int p_\theta(\mathbf{x}_0, \ldots, \mathbf{x}_T) \, d\mathbf{x}_1 \ldots d\mathbf{x}_T \right].$$

- Use the variational approximation!

Posterior Approximation



- Want to model the distribution of $\mathbf{x} \sim p,$ where $\mathbf{x} \in \mathbb{R}^d$.
- Marginal distribution over $\mathbf{x}_0$ is $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_0, \ldots, \mathbf{x}_T)\, d\mathbf{x}_1 \ldots d\mathbf{x}_T$.
- Fix $q_t(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I)$ (hyper-parameters β_t).
- The evidence lower-bound (Jensen):

$$\arg\max_\theta \mathop{\mathbb{E}}_{\mathbf{x}_0\sim p}[\log p_\theta(\mathbf{x}_0)] \geq \mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p \\ \mathbf{x}_{1:T}\sim q}}\left[\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right].$$

Unpacking the ELBO

- The forward process: $q_t(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I)$.
- The reverse process: $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}|\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)\right)$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \mathbf{x}_{1:T}\sim q}}\left[\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] &= \mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \mathbf{x}_{1:T}\sim q}}\left[-\log p(\mathbf{x}_T) - \sum_{t>0}\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)}\right] \\
&= \mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \mathbf{x}_{1:T}\sim q}}\left[-\log p(\mathbf{x}_T) - \sum_{t>0}\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}\frac{q(\mathbf{x}_{t-1})|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}\right] \\
&= \mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \mathbf{x}_{1:T}\sim q}}\left[-\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} - \sum_{t>1}\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right].
\end{aligned}$$

Closed Form Conditionals

- The forward process: $q_t(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I)$.
- The ELBO: $\underset{\substack{\mathbf{x}_0 \sim p \\ \mathbf{x}_{1:T} \sim q}}{\mathbb{E}} \left[-\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} - \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right].$
- Conditionals have closed form: $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t I\right).$
- Where: $\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\tilde{\alpha}_t}\beta_t}{1-\tilde{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t}\mathbf{x}_t.$
- And: $\alpha_t = 1 - \beta_t, \quad \tilde{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \quad \tilde{\beta}_t = \frac{1-\tilde{\alpha}_{t-1}}{1-\tilde{\alpha}_t}\beta_t.$

A Reconstruction Objective

- The reverse process: $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}|\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)\right).$
- The ELBO: $\mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \mathbf{x}_{1:T}\sim q}}\left[-\log\frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} - \sum_{t>1}\log\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right].$
- Optimization of the parameters decomposes term-wise:

$$\mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \mathbf{x}_{1:T}\sim q}}\left[-\sum_{t>1}\log\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}\right] = \mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \mathbf{x}_t\sim q_t}}\left[D(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0 \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))\right]$$

$$= \mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \mathbf{x}_t\sim q_t}}\left[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)\|^2\right] + C.$$

Reparameterization

- We can directly compute $q_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\tilde{\alpha}_t}\mathbf{x}_0, (1-\tilde{\alpha}_t)I)$.
- Define $x_t(\mathbf{x}_0, \varepsilon) = \sqrt{\tilde{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\tilde{\alpha}_t}\varepsilon$.
- Re-parameterize $\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta}{\sqrt{1-\tilde{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right)$.

$$\mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \mathbf{x}_{1:T}\sim q}}\left[-\sum_{t>1}\log\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\right] = \mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \varepsilon\sim\mathcal{N}(0,I)}}\left[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0)-\mu_\theta(\mathbf{x}_t,t)\|^2\right]+C$$

$$= \mathop{\mathbb{E}}_{\substack{\mathbf{x}_0\sim p\\ \varepsilon\sim\mathcal{N}(0,I)}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\tilde{\alpha}_t)}\|\varepsilon-\varepsilon_\theta\left(x_t(\mathbf{x}_0,\varepsilon),t\right)\|^2\right]+C.$$

Analogies

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
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

(Like Denoising Score Matching)

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **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** $\mathbf{x}_0$

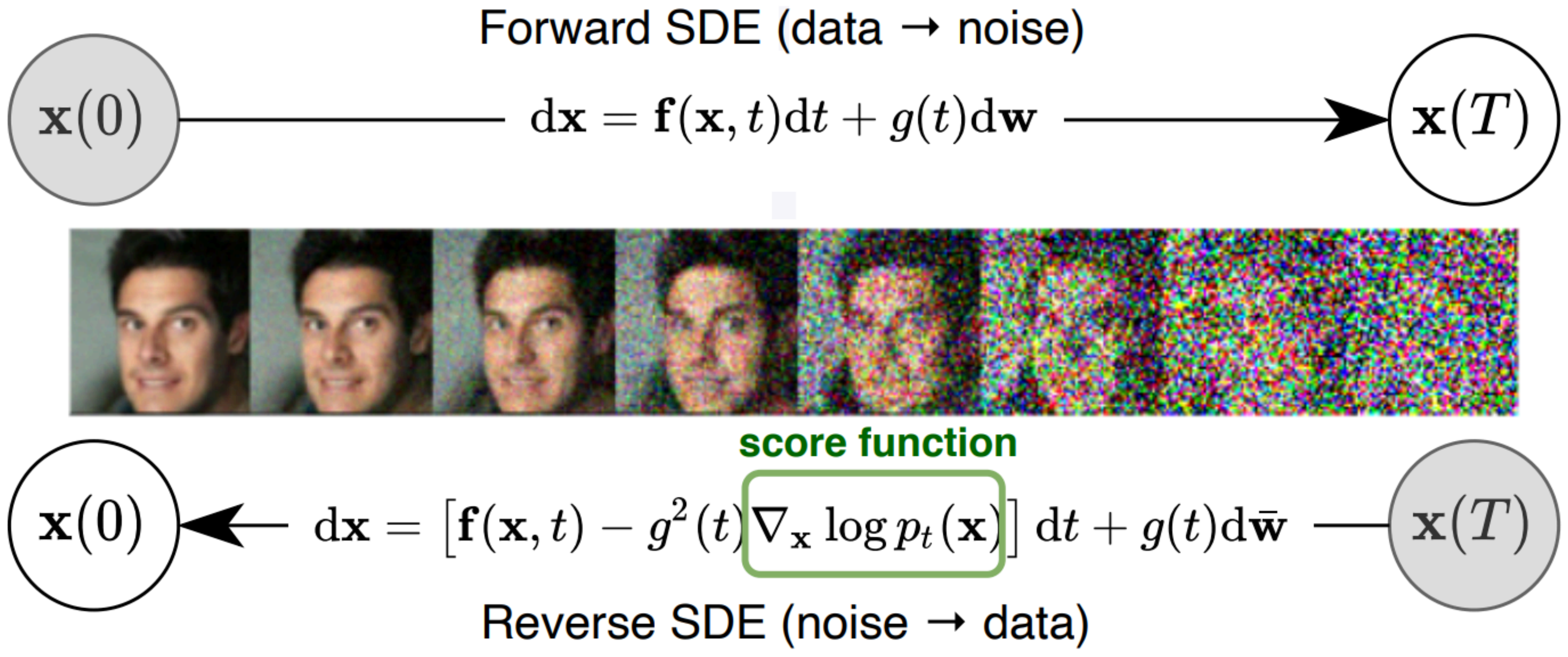
(Like Annealed Langevin Dynamics)

Algorithm 1 Annealed Langevin dynamics.

Require: $\{\sigma_i\}_{i=1}^L, \epsilon, T$.
1: Initialize $\tilde{\mathbf{x}}_0$
2: **for** $i \leftarrow 1$ to L **do**
3: $\alpha_i \leftarrow \epsilon \cdot \sigma_i^2/\sigma_L^2$ ▷ α_i is the step size.
4: **for** $t \leftarrow 1$ to T **do**
5: Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$
6: $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\, \mathbf{z}_t$
7: **end for**
8: $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
9: **end for**
return $\tilde{\mathbf{x}}_T$

- Training is like Denoising Score Matching.
- Sampling is like Annealed Langevin Dynamics.
- Can we think of a denoising diffusion model as a model trained to optimally step through the annealing levels of the Langevin sampling procedure?

Stochastic Differential Equations



Song et. al., Preprint 2020

- Generalize the Markov chain perspective to continuous SDE's.
- Analogous to the Neural ODE perspective.
- Score Matching and Denoising Diffusions can be viewed as discretization of two different continuous SDE dynamics.

Thank You