CS 59300 – Algorithms for Data Science

Classical and Quantum approaches

Lecture 13 (10/23)

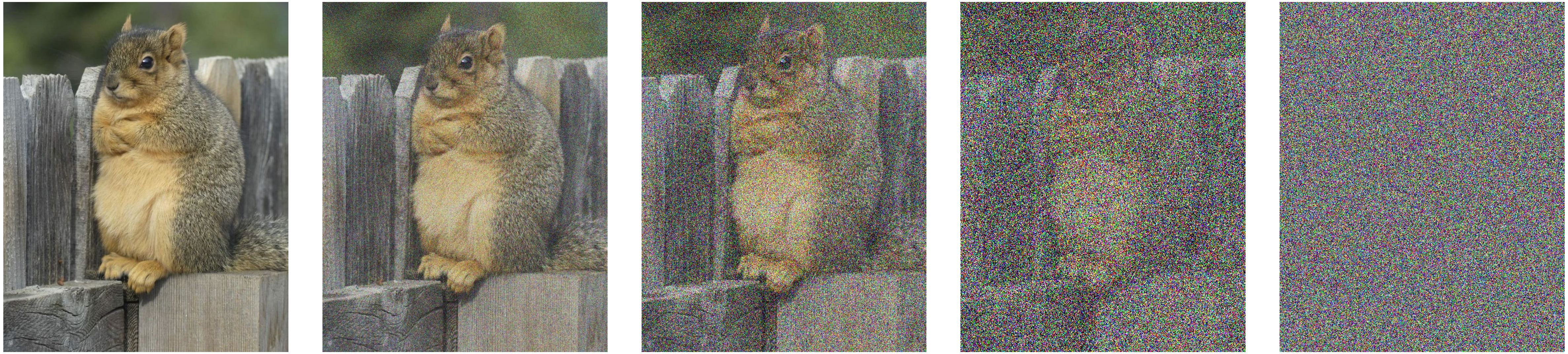
Diffusion Models

https://ruizhezhang.com/course_fall_2025.html

Slides are borrowed from Sitan Chen's course

Diffusion models

Forward process: Ornstein-Uhlenbeck



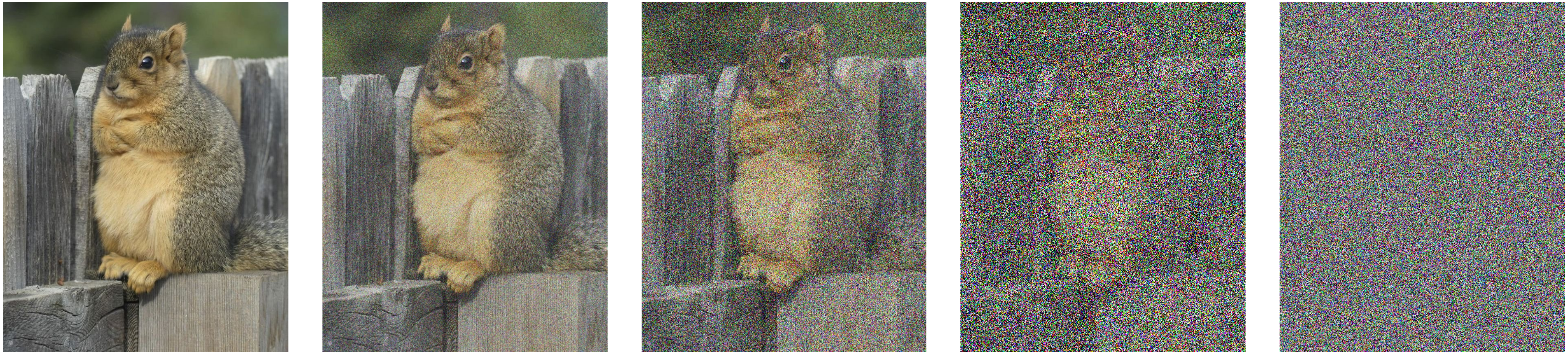
$t = 0$ $\longrightarrow$ $t = \infty$

$$\mathrm{d}X_t = -X_t + \sqrt{2}\mathrm{d}B_t$$

$X_0 \sim q$ (data distribution)

Diffusion models

Forward process: Ornstein-Uhlenbeck



$t = 0$ $\longrightarrow$ $t = \infty$

$$X_t = e^{-t} \cdot X_0 + \sqrt{1 - e^{-2t}} \cdot g \quad \text{for } g \sim N(0, I)$$

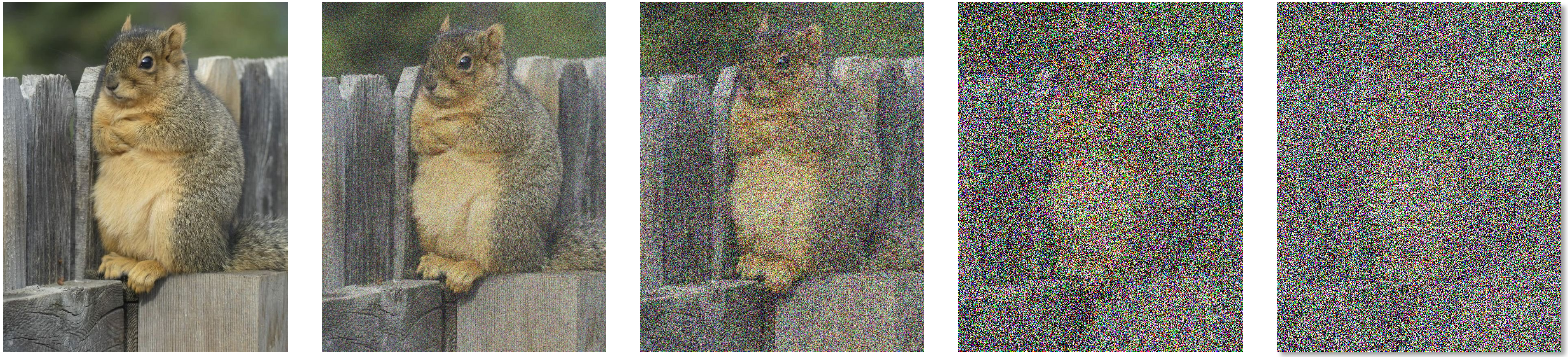
$X_0 \sim q$ (data distribution)

(prove this latter)

Diffusion models

Forward process:

$q_0 = q$ $\qquad\qquad\qquad\qquad$ q_t



$t = 0$ $\longrightarrow$ $t = T$

$$\mathrm{d}X_t = -X_t + \sqrt{2}\mathrm{d}B_t$$

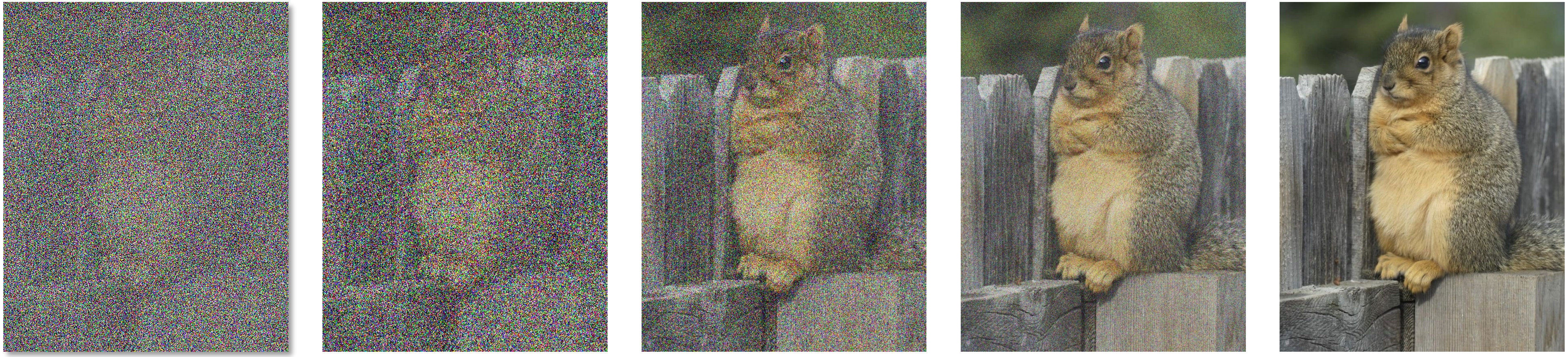
$X_0 \sim q$ (data distribution)

in practice,
can only run
for finite time

Diffusion models

Reverse process:

q_T $\quad q_{T-1}$ $\quad q_{T-t}$ $\quad q_0 = q$



$t = 0$ $\longrightarrow$ $t = T$

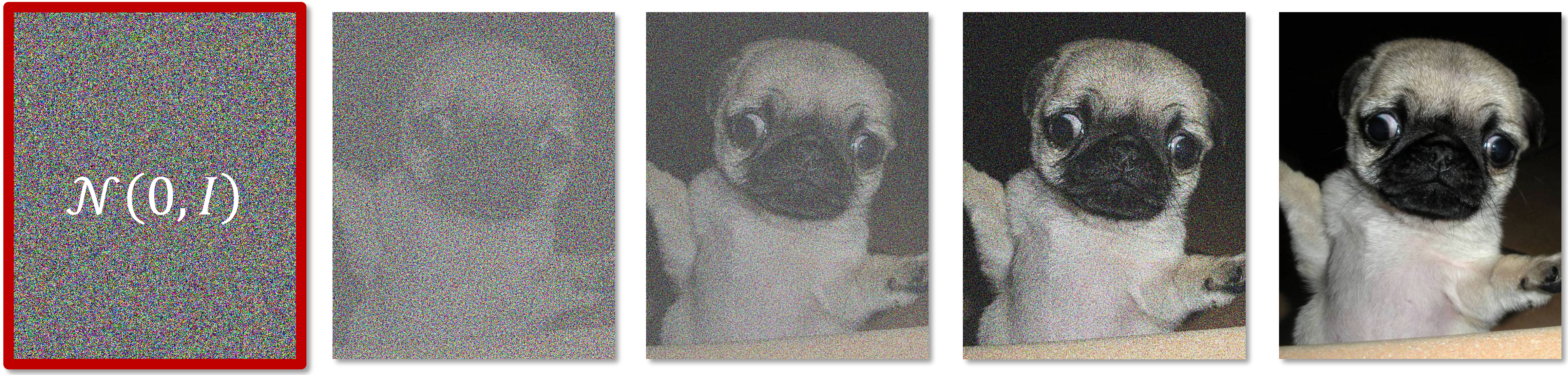
$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\nabla \ln q_{T-t}(X_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

(prove this latter)

$X_0^{\leftarrow} \sim q_T$ (forward process at time T)

Diffusion models

To sample fresh images, run reverse process with Gaussian initialization



$t = 0$ $\longrightarrow$ $t = T$

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + \boxed{2\nabla \ln q_{T-t}(X_t^{\leftarrow})}\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

$X_0^{\leftarrow} \sim q_T$ (forward process at time T)

"score function"

How to estimate it?

Score matching

Tweedie's formula. Given $\tilde{x} = x + e$ for $x \sim p$ and $e \sim \mathcal{N}(0, \sigma^2 I)$,

$$\mathbb{E}[x \mid \tilde{x}] = \tilde{x} + \sigma^2 \cdot \nabla \ln \tilde{p}(\tilde{x})$$

where $\tilde{p}$ is the density for $\tilde{x}$ $\quad \Longleftrightarrow \quad \hat{e}_{\text{Bayes}} = -\sigma^2 \nabla \ln \tilde{p}(\tilde{x})$

Song-Ermon '19: reduce estimating the score function to a supervised learning task:

Given noisy image X_t, predict noise γ that was added

$$X_t = \underbrace{e^{-t} \cdot X_0}_{x} + \underbrace{\sqrt{1 - e^{-2t}} \cdot g}_{e \sim \mathcal{N}(0, (1 - e^{-2t}) I)}$$

Score matching

Tweedie's formula. Given $\tilde{x} = x + e$ for $x \sim p$ and $e \sim \mathcal{N}(0, \sigma^2 I)$,

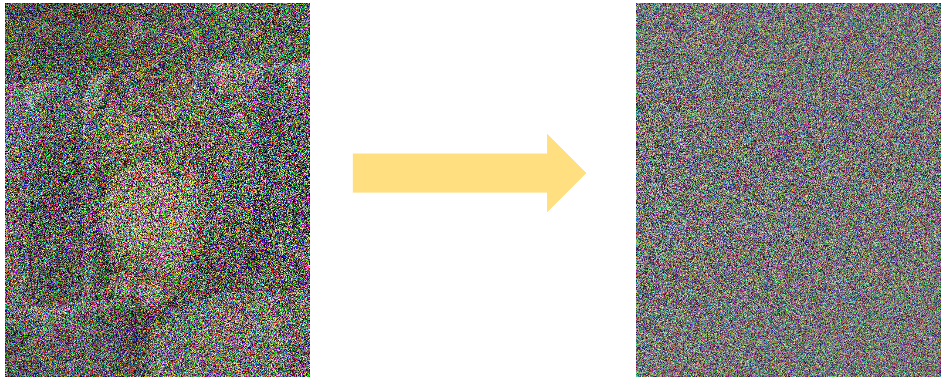
$$\mathbb{E}[x \mid \tilde{x}] = \tilde{x} + \sigma^2 \cdot \nabla \ln \tilde{p}(\tilde{x})$$

where $\tilde{p}$ is the density for $\tilde{x}$ $\quad\iff\quad \hat{e}_{\text{Bayes}} = -\sigma^2 \nabla \ln \tilde{p}(\tilde{x})$

Song-Ermon '19: reduce estimating the score function to a supervised learning task:

Given noisy image X_t, predict noise γ that was added

Fit a neural net to training examples to drawn from q_t



Score matching

Tweedie's formula. Given $\tilde{x} = x + e$ for $x \sim p$ and $e \sim \mathcal{N}(0, \sigma^2 I)$,

$$\mathbb{E}[x \mid \tilde{x}] = \tilde{x} + \sigma^2 \cdot \nabla \ln \tilde{p}(\tilde{x})$$

where $\tilde{p}$ is the density for $\tilde{x}$ $\quad\Longleftrightarrow\quad \hat{e}_{\mathrm{Bayes}} = -\sigma^2 \nabla \ln \tilde{p}(\tilde{x})$

Song-Ermon '19: reduce estimating the score function to a supervised learning task:

Given noisy image X_t, predict noise γ that was added

$g^{(i)} \sim \mathcal{N}(0, I)$

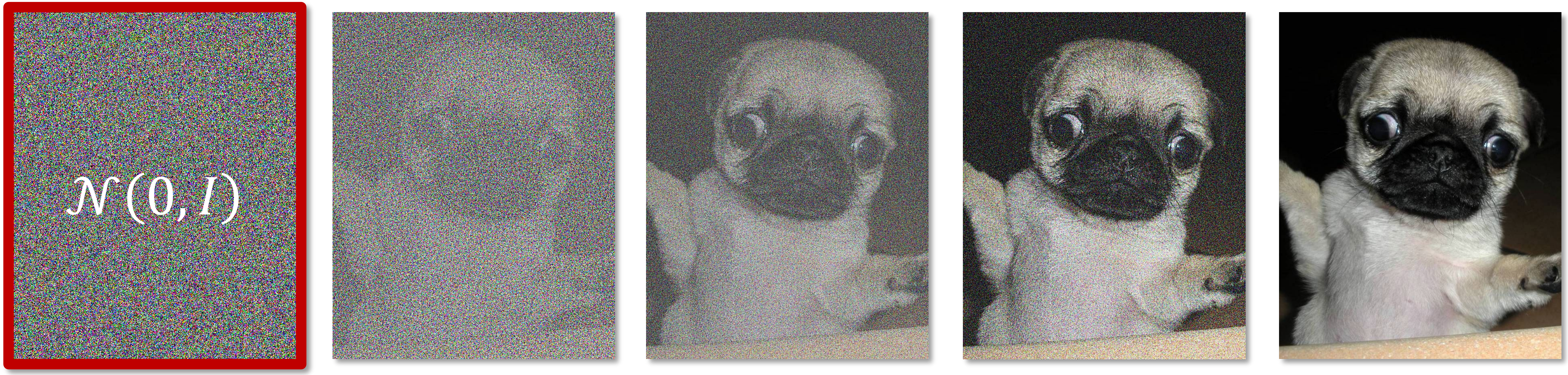
Fit a neural net to training examples to drawn from q_t

$$s_t \coloneqq \arg\min_{\mathrm{NN} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left\| \frac{1}{\sqrt{1 - e^{-2t}}} g^{(i)} + \mathrm{NN}\left(e^{-t} X^{(i)} + \sqrt{1 - e^{-2t}} g^{(i)}\right) \right\|^2$$

Assumption: $\mathbb{E}_{q_t}[\|s_t(X_t) - \nabla \ln q_t(X_t)\|^2] \le \epsilon_{\mathrm{sc}}^2 \quad \forall\, t$

Diffusion models

To sample fresh images, run reverse process with Gaussian initialization



$t = 0$ $\longrightarrow$ $t = T$

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + \boxed{2\nabla \ln q_{T-t}(X_t^{\leftarrow})}\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

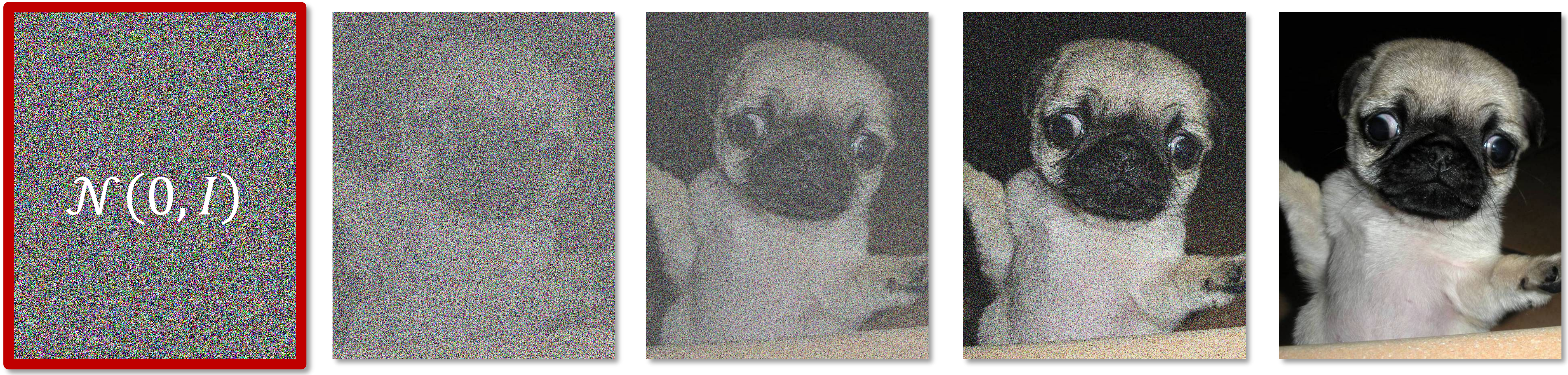
$X_0^{\leftarrow} \sim q_T$ (forward process at time T)

"score function"

How to estimate it?

Diffusion models

To sample fresh images, run reverse process with Gaussian initialization



$t = 0$ $\longrightarrow$ $t = T$

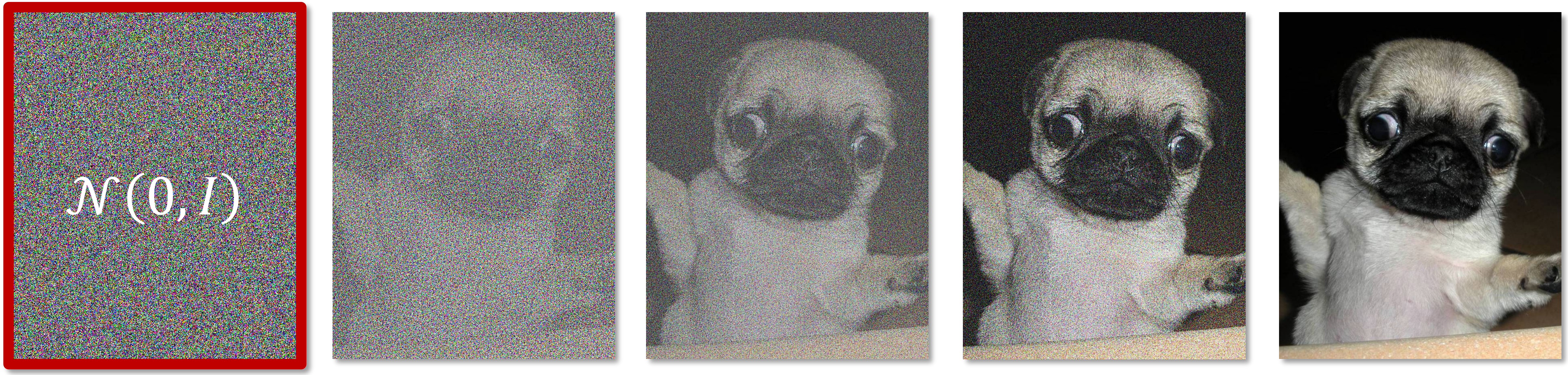
$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2s_{T-t}(X_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

$X_0^{\leftarrow} \sim q_T$ (forward process at time T)

can't run in continuous time, need to **discretize**

Diffusion models

To sample fresh images, run reverse process with Gaussian initialization



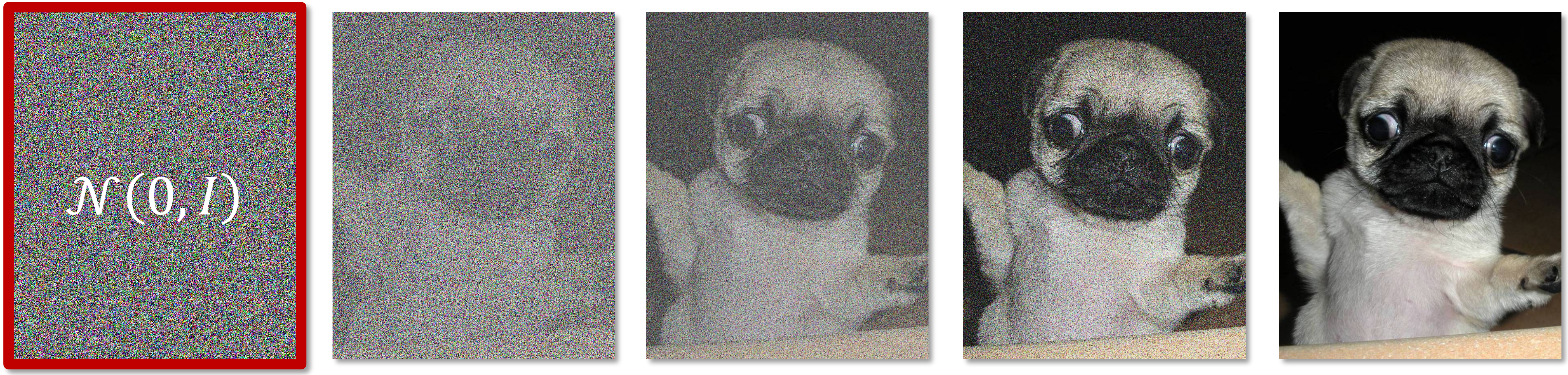
$t = 0$ $\quad t = h, 2h, 3h, \ldots \quad$ $t = T$

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2 s_{T-kh}(X_{kh}^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \qquad \text{for } t \in [kh, (k+1)h)$$

$X_0^{\leftarrow} \sim q_T$ (forward process at time T)

Diffusion models

To sample fresh images, run reverse process with Gaussian initialization



$t = 0$ $\quad$ $t = h, 2h, 3h, \dots$ $\quad$ $t = T$

$$X^{\leftarrow}_{(k+1)h} = e^{h}\, X^{\leftarrow}_{kh} + \left(e^{h} - 1\right) 2 s_{T-kh}(X^{\leftarrow}_{kh}) + \mathcal{N}\left(0, \left(e^{2h} - 1\right) I\right)$$

$X^{\leftarrow}_{0} \sim q_T$ (forward process at time T)

Deferred proofs

1. Integral solution for the Ornstein-Uhlenbeck process

$$\mathrm{d}X_t = -X_t + \sqrt{2}\mathrm{d}B_t \quad \Longrightarrow \quad X_t = e^{-t} \cdot X_0 + \mathcal{N}(0, (1 - e^{-2t})I)$$

2. Tweedie's formula

$$\mathbb{E}[x \mid \tilde{x}] = \tilde{x} + \sigma^2 \cdot \nabla \ln \tilde{p}(\tilde{x})$$

3. Integral solution for the discretized reverse process:

$$X^{\leftarrow}_{(k+1)h} = e^h \, X^{\leftarrow}_{kh} + \left(e^h - 1\right) 2 s_{T-kh}(X^{\leftarrow}_{kh}) + \mathcal{N}\left(0, (e^{2h} - 1)I\right)$$

4. Reverse-time SDE

$$\mathrm{d}X^{\leftarrow}_t = \{X^{\leftarrow}_t + 2\nabla \ln q_{T-t}(X^{\leftarrow}_t)\} \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t$$

Solve linear SDE

$$\mathrm{d}X_t = -X_t + \sqrt{2}\mathrm{d}B_t$$

- $e^t(X_t + \mathrm{d}X_t) = \sqrt{2}e^t\mathrm{d}B_t$, which implies that

$$\mathrm{d}(e^t X_t) = \sqrt{2}e^t\mathrm{d}B_t$$

- Integrate from 0 to t on both sides:

$$e^t X_t - X_0 = \sqrt{2}\int_0^t e^s\,\mathrm{d}B_s \quad \Longrightarrow \quad X_t = e^{-t}X_0 + \boxed{\sqrt{2}\int_0^t e^{s-t}\,\mathrm{d}B_s}$$

- By Itô's isometry,

$$\sqrt{2}\int_0^t e^{s-t}\,\mathrm{d}B_s \equiv \mathcal{N}\left(0{,}2\int_0^t (e^{s-t})^2\mathrm{d}s\right) = \mathcal{N}(0,(1-e^{-2t})I)$$

Proof of Tweedie's formula

$\tilde{x} = x + e$ for $x \sim p$ and $e \sim \mathcal{N}(0, \sigma^2 I)$

By Bayes' rule,

$$\mathbb{P}[\, x \mid \tilde{x} \,] = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\tilde{x} - x)^2}{2\sigma^2}\right) \cdot p(x)}{\tilde{p}(\tilde{x})}$$

so

$$\mathbb{E}\left[\frac{x - \tilde{x}}{\sigma^2} \,\middle|\, \tilde{x}\right] = \tilde{p}(\tilde{x})^{-1} \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \int \exp\left(-\frac{(\tilde{x} - x)^2}{2\sigma^2}\right) \cdot \frac{x - \tilde{x}}{\sigma^2} \cdot p(x)\, \mathrm{d}x}_{\nabla \tilde{p}(\tilde{x})}$$

Observe that

$$\tilde{p}(\tilde{x}) = \int \exp\left(-\frac{(\tilde{x} - x)^2}{2\sigma^2}\right) \cdot p(x) \mathrm{d}x$$

Proof of Tweedie's formula

$$\tilde{x} = x + e \text{ for } x \sim p \text{ and } e \sim \mathcal{N}(0, \sigma^2 I)$$

By Bayes' rule,

$$\mathbb{P}[\, x \mid \tilde{x} \,] = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\tilde{x} - x)^2}{2\sigma^2}\right) \cdot p(x)}{\tilde{p}(\tilde{x})}$$

so

$$\mathbb{E}\left[\frac{x - \tilde{x}}{\sigma^2} \,\middle|\, \tilde{x}\right] = \frac{\nabla \tilde{p}(\tilde{x})}{\tilde{p}(\tilde{x})} = \nabla \ln \tilde{p}(\tilde{x})$$

$\blacksquare$

Reverse-time SDE

- Forward process: $\mathrm{d}X_t = -X_t + \sqrt{2}\mathrm{d}B_t$

$$\frac{\partial}{\partial t} q_t(\mathbf{x}) = \nabla \cdot \big(\mathbf{x} q_t(\mathbf{x})\big) + \Delta q_t$$

- Reverse process: $\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\nabla \ln q_{T-t}(X_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$

$$\frac{\partial}{\partial t} q_t^{\leftarrow}(\mathbf{x}) = -\nabla \cdot \Big((\mathbf{x} + 2\nabla \ln q_t^{\leftarrow}(\mathbf{x}))q_t^{\leftarrow}(\mathbf{x})\Big) + \Delta q_t^{\leftarrow}$$

Fokker-Planck equation. Let $\{x_t\}_{t\geq 0}$ follows $\mathrm{d}\mathbf{x}_t = \boldsymbol{\mu}_t(\boldsymbol{x}_t)\mathrm{d}t + \boldsymbol{\sigma}_t \mathrm{d}\boldsymbol{B}_t$ and $\mathbf{x}_0 \sim \pi_0$. Then for all $t \geq 0$, denoting the law of x_t by π_t, we have

$$\frac{\partial}{\partial t} \pi_t(\mathbf{x}) = -\nabla \cdot \big(\boldsymbol{\mu}_t(\mathbf{x})\pi_t(\mathbf{x})\big) + \frac{1}{2} \sum_{i,j\in[d]} \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j} \Big(\boldsymbol{\sigma}_t \boldsymbol{\sigma}_t^{\top}(\mathbf{x})_{ij} \pi_t(\mathbf{x})\Big)$$

Reverse-time SDE

- Forward process: $\mathrm{d}X_t = -X_t + \sqrt{2}\mathrm{d}B_t$

$$\frac{\partial}{\partial t} q_t(\mathbf{x}) = \nabla \cdot \left(\mathbf{x} q_t(\mathbf{x})\right) + \Delta q_t$$

- Reverse process: $\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\nabla \ln q_{T-t}(X_t^{\leftarrow})\}\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t$

$$\begin{aligned}\frac{\partial}{\partial t} q_t^{\leftarrow}(\mathbf{x}) &= -\nabla \cdot \left((\mathbf{x} + 2\nabla \ln q_t^{\leftarrow}(\mathbf{x}))q_t^{\leftarrow}(\mathbf{x})\right) + \Delta q_t^{\leftarrow} \\ &= -\nabla \cdot \left(\mathbf{x} q_t^{\leftarrow}(\mathbf{x})\right) - \Delta q_t^{\leftarrow}\end{aligned}$$

$$\frac{\partial}{\partial t} q_t(\mathbf{x}) = \nabla \cdot \left(\mathbf{x} q_t(\mathbf{x})\right) + \Delta q_t \quad \xleftrightarrow{t \rightleftharpoons T-t} \quad \frac{\partial}{\partial t} q_t^{\leftarrow}(\mathbf{x}) = -\nabla \cdot \left(\mathbf{x} q_t^{\leftarrow}(\mathbf{x})\right) - \Delta q_t^{\leftarrow}$$

Discretization analysis

- Let p_T be the law of the output of the algorithm after $N = T/h$ steps
- Our goal is to show that $\mathrm{TV}(p_T, q)$ is small

We need to bound three sources of error:

1) the initialization of the algorithm at pure Gaussian noise rather than at q_T
2) the estimation of the score function
3) the discretization of the SDE with step size $h > 0$

Assumption I (L^2-accurate score estimate): For all $t = 0, h, 2h, \dots, T$, the score estimate $s_t(\cdot)$ satisfies

$$\mathbb{E}_{q_t}[\|s_t(X_t) - \nabla \ln q_t(X_t)\|^2] \le \epsilon_{\mathrm{sc}}^2$$

Assumption II (Smoothness): For all $t \ge 0$, $\nabla \ln q_t(\cdot)$ is L-Lipschitz

Assumption III (Bounded second moment): $\mathfrak{m}_2^2 := \mathbb{E}_q[\|x\|^2] < \infty$

Convergence guarantee

Theorem (Chen et al. '23). Under Assumptions I-III, if p_T is the law of the output of the algorithm after $N = T/h$ iterations with step size h,

$$\mathrm{TV}(p_T, q) \lesssim \underbrace{\sqrt{\mathrm{KL}(q\|\gamma)} \cdot \exp(-T)}_{\substack{\text{error from initializing at} \\ \gamma = \mathcal{N}(0, I) \text{ instead of } q_T}} + \underbrace{\left(L\sqrt{dh} + L\mathfrak{m}_2 h\right)\sqrt{T}}_{\text{discretization error}} + \underbrace{\epsilon_{\mathrm{sc}}\sqrt{T}}_{\substack{\text{score} \\ \text{error}}}$$

- Suppose $\mathrm{KL}(q\|\gamma) \le \mathrm{poly}(d)$ and $\mathfrak{m}_2^2 \lesssim d$. Choose $T = \log(\mathrm{KL}(q\|\gamma)/\epsilon)$ and $h = \frac{\epsilon^2}{L^2 d}$ gives
$$\mathrm{TV}(p_T, q) = \tilde{O}(\epsilon + \epsilon_{\mathrm{sc}})$$
with $N = \tilde{O}(L^2 d/\epsilon^2)$ steps
- To get ϵ-close in TV, it suffices to estimate the score function to within $\epsilon_{sc} \le \tilde{O}(\epsilon)$ accuracy

Bound initialization error

1. Forward process converges exponentially quickly to Gaussian

$$\mathrm{KL}(q_T \| \gamma) \leq \exp(-2T)\,\mathrm{KL}(q_0 \| \gamma) = \exp(-2T)\,\mathrm{KL}(q \| \gamma)$$

2. Running reverse SDE starting from Gaussian vs. from q_T cannot increase distance between them (data processing inequality)

$$\mathrm{KL}\big(q \| p_T\big) = \mathrm{KL}\big(\mathrm{reverse}(q_T) \| \mathrm{reverse}(\gamma)\big) \leq \mathrm{KL}(q_T \| \gamma)$$

3. Pinsker's inequality: $\mathrm{TV}(p_T, q) \leq \sqrt{\frac{1}{2}\mathrm{KL}(q \| p_T)}$

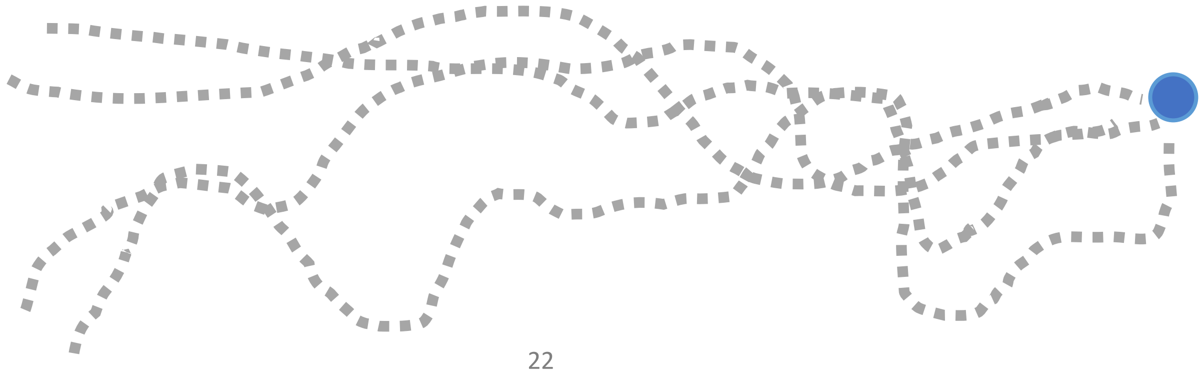
Discretization argument

Consider the ideal reverse process (continuous, perfect score) and algorithm (discrete, estimated score), with **both initialized at q_T**

- $\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\nabla \ln q_{T-t}(X_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$
- $\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2s_{T-kh}(X_{kh}^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$ $\quad X_{(k+1)h}^{\leftarrow} = e^h X_{kh}^{\leftarrow} + \left(e^h - 1\right) 2s_{T-kh}(X_{kh}^{\leftarrow}) + \mathcal{N}\left(0, e^{2h} - 1\right)$

To control the distance between these processes, we use Girsanov's theorem, a powerful tool for bounding distance between laws of processes driven by the same Brownian motion

Intuition: distribution over last iterate of reverse process is hard to characterize, but distribution over trajectory just given by a bunch of Gaussian samples



Girsanov's theorem

Consider the SDE's

$$\mathrm{d}X_t = b_t \, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t$$

$$\mathrm{d}X_t = b_t' \, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t$$

with the same initial distribution.

Let Q_T, P_T denote the laws of the **trajectories** over time T, i.e. $\mathcal{C}\left([0,T];\mathbb{R}^d\right)$. Then

$$\frac{\mathrm{d}Q_T}{\mathrm{d}P_T} = \exp\left(\frac{1}{\sqrt{2}}\int_0^T (b_t - b_t')\,\mathrm{d}B_t - \frac{1}{4}\int_0^T \|b_t - b_t'\|^2\,\mathrm{d}t\right)$$

where B_t is a standard Brownian motion w.r.t. P_T

Heuristic proof of Girsanov's theorem

For the SDE's:

$$\mathrm{d}x_t = b_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \quad \text{(in } Q\text{-world)}$$

$$\mathrm{d}x_t = b'_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \quad \text{(in } P\text{-world)}$$

consider the discrete-time approximation:

$$\hat{x}_{(k+1)h} = \hat{x}_{kh} + h b_{kh}(\hat{x}_{kh}) + \sqrt{2h} g_{kh} \quad \text{(in } Q\text{-world)}$$

$$\hat{x}_{(k+1)h} = \hat{x}_{kh} + h b'_{kh}(\hat{x}_{kh}) + \sqrt{2h} g_{kh} \quad \text{(in } P\text{-world)}$$

For a trajectory $(\hat{x}_0, \hat{x}_h, \hat{x}_{2h}, \ldots, \hat{x}_{Nh})$, what are the likelihoods in Q-world and P-world?

- Q-world: $L_Q \propto \prod_{k=0}^{N-1} \exp\left(-\frac{1}{4h}\left\|\hat{x}_{(k+1)h} - \hat{x}_{kh} - h b_{kh}(\hat{x}_{kh})\right\|^2\right)$
- P-world: $L_P \propto \prod_{k=0}^{N-1} \exp\left(-\frac{1}{4h}\left\|\hat{x}_{(k+1)h} - \hat{x}_{kh} - h b'_{kh}(\hat{x}_{kh})\right\|^2\right)$

Heuristic proof of Girsanov's theorem

For a trajectory $(\hat{x}_0, \hat{x}_h, \hat{x}_{2h}, \dots, \hat{x}_{Nh})$, what are the likelihoods in Q-world and P-world?

$$\frac{L_Q}{L_P} = \frac{\prod_{k=0}^{N-1} \exp\left(-\frac{1}{4h}\left\|\hat{x}_{(k+1)h} - \hat{x}_{kh} - h b_{kh}(\hat{x}_{kh})\right\|^2\right)}{\prod_{k=0}^{N-1} \exp\left(-\frac{1}{4h}\left\|\hat{x}_{(k+1)h} - \hat{x}_{kh} - h b'_{kh}(\hat{x}_{kh})\right\|^2\right)}$$

$$= \prod_{k=0}^{N-1} \exp\left(-\frac{1}{4h}\left(h^2\|b_{kh}(\hat{x}_{kh})\|^2 - h^2\|b'_{kh}(\hat{x}_{kh})\|^2 - 2h\left\langle \overbrace{\hat{x}_{(k+1)h} - \hat{x}_{kh}}^{hb'_{kh}(\hat{x}_{kh}) + \sqrt{2h}g_{kh} \text{ in } P\text{-world}}, b_{kh}(\hat{x}_{kh}) - b'_{kh}(\hat{x}_{kh})\right\rangle\right)\right)$$

$$= \prod_{k=0}^{N-1} \exp\left(-\frac{1}{4h}\left(h^2\|b_{kh}(\hat{x}_{kh}) - b'_{kh}(\hat{x}_{kh})\|^2 - 2\sqrt{2}h\left\langle \sqrt{h}g_{kh}, b_{kh}(\hat{x}_{kh}) - b'_{kh}(\hat{x}_{kh})\right\rangle\right)\right)$$

$$= \exp\left(-\frac{1}{4}\sum_{k=0}^{N-1} h\|b_{kh}(\hat{x}_{kh}) - b'_{kh}(\hat{x}_{kh})\|^2 + \frac{1}{\sqrt{2}}\sum_{k=0}^{N-1}\left\langle \sqrt{h}g_{kh}, b_{kh}(\hat{x}_{kh}) - b'_{kh}(\hat{x}_{kh})\right\rangle\right)$$

Heuristic proof of Girsanov's theorem

For a trajectory $(\hat{x}_0, \hat{x}_h, \hat{x}_{2h}, \dots, \hat{x}_{Nh})$, what are the likelihoods in Q-world and P-world?

$$\frac{L_Q}{L_P} = \exp\left(-\frac{1}{4}\sum_{k=0}^{N-1} h\|b_{kh}(\hat{x}_{kh}) - b'_{kh}(\hat{x}_{kh})\|^2 + \frac{1}{\sqrt{2}}\sum_{k=0}^{N-1} \langle \underbrace{\sqrt{h} g_{kh}}_{\text{“}\mathrm{d}B_{kh}\text{”}}, b_{kh}(\hat{x}_{kh}) - b'_{kh}(\hat{x}_{kh}) \rangle \right)$$

$$\xrightarrow{(h\to 0)} \exp\left(-\frac{1}{4}\int_0^T \|b_t - b'_t\|^2 \mathrm{d}t + \underbrace{\frac{1}{\sqrt{2}}\int_0^T (b_t - b'_t)\mathrm{d}B_t}_{\text{martingale in } P\text{-world}} \right) = \frac{\mathrm{d}Q_T}{\mathrm{d}P_T}$$

KL divergence bound from Girsanov's theorem

BM in P-world

$$\mathrm{KL}\left(Q_T \| P_T\right) = \mathbb{E}_Q\left[-\frac{1}{4}\int_0^T \|b_t - b_t'\|^2 \mathrm{d}t + \frac{1}{\sqrt{2}}\int_0^T (b_t - b_t')\mathrm{d}B_t\right]$$

By Girsanov's theorem, we can relate the Brownian motion in P-world to Q-world:

$$b_t \mathrm{d}t + \sqrt{2}\mathrm{d}\tilde{B}_\mathrm{t} = b_t' \mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \Longrightarrow \quad \mathrm{d}\tilde{B}_t = \frac{1}{\sqrt{2}}(b_t' - b_t)\mathrm{d}t + \mathrm{d}B_t$$

a new BM in Q-world

$$\begin{aligned}\mathrm{KL}\left(Q_T \| P_T\right) &= \mathbb{E}_Q\left[-\frac{1}{4}\int_0^T \|b_t - b_t'\|^2 \mathrm{d}t + \frac{1}{\sqrt{2}}\int_0^T \left\langle b_t - b_t', \frac{1}{\sqrt{2}}(b_t - b_t')\mathrm{d}t + \mathrm{d}\tilde{B}_t\right\rangle\right] \\ &= \frac{1}{4}\mathbb{E}_Q\left[\int_0^T \|b_t - b_t'\|^2 \mathrm{d}t\right] + \frac{1}{\sqrt{2}}\cancel{\mathbb{E}_Q[(b_t - b_t')\mathrm{d}\tilde{B}_t]}\end{aligned}$$

Warmup: Girsanov analysis for Langevin dynamics

Suppose the target distribution q is strongly log-concave: $\alpha \cdot I \preccurlyeq -\nabla^2 \ln q \preccurlyeq L \cdot I$

ULA: $\mathrm{d}\hat{x}_t = -\nabla \ln q(\hat{x}_{kh})\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad$ for $t \in [kh, (k+1)h]$

LD: $\mathrm{d}x_t = -\nabla \ln q(x_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$

$$\begin{aligned}
\mathrm{KL}(\mathrm{LD}\|\mathrm{ULA}) &= \frac{1}{4}\mathbb{E}_{\mathrm{LD}}\left[\sum_{k=0}^{T/h-1}\int_{kh}^{(k+1)h}\|\nabla \ln q(x_t) - \nabla \ln q(x_{kh})\|^2 \mathrm{d}t\right] \\
&\le \frac{L^2}{4}\mathbb{E}_{\mathrm{LD}}\left[\sum_{k=0}^{T/h-1}\int_{kh}^{(k+1)h}\|x_t - x_{kh}\|^2 \mathrm{d}t\right]
\end{aligned}$$

$$\begin{aligned}
\|x_t - x_{kh}\|^2 &= \left\|\int_0^{t-kh}\nabla \ln q(x_{kh+s})\mathrm{d}s + \sqrt{2}(B_t - B_{kh})\right\|^2 \\
&\le 2h\int_0^{t-kh}\|\nabla \ln q(x_{kh+s})\|^2 \mathrm{d}s + 4\|B_t - B_{kh}\|^2
\end{aligned}$$

Warmup: Girsanov analysis for Langevin dynamics

Suppose the target distribution q is strongly log-concave: $\alpha \cdot I \preccurlyeq -\nabla^2 \ln q \preccurlyeq L \cdot I$

ULA: $\mathrm{d}\hat{x}_t = -\nabla \ln q(\hat{x}_{kh})\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$ for $t \in [kh, (k+1)h]$

LD: $\mathrm{d}x_t = -\nabla \ln q(x_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$

$$\begin{aligned}
\mathrm{KL}(\mathrm{LD}\|\mathrm{ULA}) &= \frac{1}{4}\mathbb{E}_{\mathrm{LD}}\left[\sum_{k=0}^{T/h-1}\int_{kh}^{(k+1)h}\|\nabla \ln q(x_t) - \nabla \ln q(x_{kh})\|^2 \mathrm{d}t\right] \\
&\leq \frac{L^2}{4}\mathbb{E}_{\mathrm{LD}}\left[\sum_{k=0}^{T/h-1}\int_{kh}^{(k+1)h}\|x_t - x_{kh}\|^2 \mathrm{d}t\right] \\
&\lesssim L^2 h^2 \int_0^T \mathbb{E}_{\mathrm{LD}}[\|\nabla \ln q(x_t)\|^2]\mathrm{d}t + L^2 dhT
\end{aligned}$$

$$\mathbb{E}_{\mathrm{LD}}[\|\nabla \ln q(x_t)\|^2] \lesssim Ld + \frac{L^2}{\alpha}\mathrm{KL}(\mathrm{Law}(x_0)\|q)$$

Warmup: Girsanov analysis for Langevin dynamics

Suppose the target distribution q is strongly log-concave: $\alpha \cdot I \preccurlyeq -\nabla^2 \ln q \preccurlyeq L \cdot I$

ULA: $\mathrm{d}\hat{x}_t = -\nabla \ln q(\hat{x}_{kh})\, \mathrm{d}t + \sqrt{2}\mathrm{d}B_t$ for $t \in [kh, (k+1)h]$

LD: $\mathrm{d}x_t = -\nabla \ln q(x_t)\, \mathrm{d}t + \sqrt{2}\mathrm{d}B_t$

$$\mathrm{KL}(\mathrm{LD}\|\mathrm{ULA}) \lesssim L^2 h^2 T \left(Ld + \frac{L^2}{\alpha} \mathrm{KL}(\mathrm{Law}(x_0)\|q) \right) + L^2 dhT$$

- For sufficiently small ϵ, if we take $h = \frac{\epsilon^2}{L^2 dT}$, then

$$\mathrm{KL}(\mathrm{Law}(x_T)\|\mathrm{Law}(\hat{x}_T)) \le \mathrm{KL}(\mathrm{LD}\|\mathrm{ULA}) \lesssim \epsilon^2$$

- Since q satisfies α-LSI, if we take $T = \frac{1}{\alpha}\log(\mathrm{KL}(\mathrm{Law}(x_0)\|q)/\epsilon)$, then $\mathrm{KL}(\mathrm{Law}(x_T)\|q) \le \epsilon^2$

- Finally, by Pinsker and triangle inequality for TV distance, we get that

$$\mathrm{TV}(\mathrm{Law}(\hat{x}_T), q) \lesssim \sqrt{\mathrm{KL}(\mathrm{Law}(x_T)\|\mathrm{Law}(\hat{x}_T))} + \sqrt{\mathrm{KL}(\mathrm{Law}(x_T)\|q)} \le \epsilon$$

Warmup: Girsanov analysis for Langevin dynamics: bound $\mathbb{E}_{\mathrm{LD}}[\|\nabla \ln q(x_t)\|^2]$

Suppose the target distribution q is strongly log-concave: $\alpha \cdot I \preccurlyeq -\nabla^2 \ln q \preccurlyeq L \cdot I$

By the definition of W_2,

$$\mathbb{E}_{\mathrm{LD}}[\|\nabla \ln q(x_t)\|^2] \lesssim \mathbb{E}_q[\|\nabla \ln q(x)\|^2] + L^2 W_2^2(\mathrm{Law}(x_t), q)$$

- For the first term,

$$\mathbb{E}_q[\|\nabla \ln q(x)\|^2] = \int \langle \nabla q, \nabla \ln q \rangle \,\mathrm{d}x \overset{\text{i.b.p.}}{=} -\int (\Delta \ln q) q(x) \mathrm{d}x \leq Ld$$

- For the second term,

$$\underbrace{L^2 W_2^2(\mathrm{Law}(x_t), q) \leq \frac{L^2}{\alpha} \mathrm{KL}(\mathrm{Law}(x_t) \| q)}_{\text{Talagrand's } \mathrm{T}_2 \text{ inequality}} \underbrace{\leq \frac{L^2}{\alpha} \mathrm{KL}(\mathrm{Law}(x_0) \| q)}_{\text{data processing inequality}}$$

Girsanov analysis for diffusion model

ALG: $\mathrm{d}X_t = \left(\hat{X}_t + 2s_{T-kh}\left(\hat{X}_{kh}\right)\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$ for $t \in [kh, (k+1)h]$

TRUE: $\mathrm{d}X_t = (X_t + 2\nabla \ln q_{T-t}(X_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$

Following the same argument,

$$\mathrm{KL}(\mathrm{TRUE} \| \mathrm{ALG}) = \mathbb{E}_{\mathrm{TRUE}}\left[\sum_{k=0}^{T/h-1} \int_{kh}^{(k+1)h} \|\nabla \ln q_{T-t}(X_t) - s_{T-kh}(X_{kh})\|^2 \mathrm{d}t\right]$$

Girsanov analysis for diffusion model

We can decompose $\mathbb{E}_{\text{TRUE}}[\|\nabla \ln q_{T-t}(X_t) - \nabla \ln s_{T-kh}(X_{kh})\|^2]$ into:

1. $\mathbb{E}[\|\nabla \ln q_{T-kh}(X_{kh}) - s_{T-kh}(X_{kh})\|^2] \leq \epsilon_{\text{sc}}^2$

   score estimation error, bounded by assumption

2. $\mathbb{E}[\|\nabla \ln q_{T-t}(X_{kh}) - \nabla \ln q_{T-kh}(X_{kh})\|^2]$

   bounded because score doesn't change much over short period

3. $\mathbb{E}[\|\nabla \ln q_{T-t}(X_t) - \nabla \ln q_{T-t}(X_{kh})\|^2]$

   bounded because score is Lipschitz, and process doesn't move too much over short time

Space discretization

- By Lipschitzness of the score function,

$$\mathbb{E}[\|\nabla \ln q_{T-t}(X_t) - \nabla \ln q_{T-t}\,(X_{kh})\|^2] \lesssim L^2\mathbb{E}[\|X_t - X_{kh}\,\|^2]$$

- The joint distribution of $(X_t, X_{kh}) \equiv \left(X_t, e^{-(t-kh)}X_t + \mathcal{N}\left(0, \left(1 - e^{-2(t-kh)}\right)I\right)\right)$

$$\mathbb{E}[\|X_t - X_{kh}\,\|^2] \lesssim \left(1 - e^{-(t-kh)}\right)^2\mathbb{E}[\|X_t\|^2] + \left(1 - e^{-2(t-kh)}\right)d$$

- $X_t \equiv e^{-(T-t)}X + \mathcal{N}\left(0, \left(1 - e^{-2(T-t)}\right)I\right)$ for $X \sim q$

$$\mathbb{E}[\|X_t\|^2] \lesssim e^{-2(T-t)}\mathbb{E}_q[\|X\|^2] + \left(1 - e^{-2(T-t)}\right)d \le \mathfrak{m}_2^2 + d$$

- Therefore,

$$\mathbb{E}[\|\nabla \ln q_{T-t}(X_t) - \nabla \ln q_{T-t}\,(X_{kh})\|^2] \le L^2h^2\mathfrak{m}_2^2 + L^2hd$$

Time discretization

$$\mathbb{E}[\|\nabla \ln q_{T-t}(X_{kh}) - \nabla \ln q_{T-kh}(X_{kh})\|^2]$$

- Consider $p \propto \exp(-V)$, $\nabla^2 V \succcurlyeq L \cdot I$, and $p' = p \star \mathcal{N}(0, \sigma^2 I)$, i.e. Gaussian convolution
- Our goal: $\mathbb{E}_{p'}[\|\nabla \ln p - \nabla \ln p'\|^2]$
- Notice that

$$\nabla \ln p'(x) = -\mathbb{E}_{p_{x,\sigma}}[\nabla V(y)] \quad \text{where } p_{x,\sigma} = \mathrm{Law}(y|y + \sigma g = x) \text{ for } y \sim p$$

- $\mathbb{E}_{x\sim p'}[\|\nabla \ln p - \nabla \ln p'\|^2] = \mathbb{E}_{x\sim p'}\left[\left\|\mathbb{E}_{y\sim p_{x,\sigma}}[\nabla V(y) - \nabla V(x)]\right\|^2\right] \leq L^2 \mathbb{E}_{x\sim p'} \mathbb{E}_{y\sim p_{x,\sigma}}[\|x - y\|^2]$
- $\mathrm{Law}(x, y) \equiv \mathrm{Law}(\tilde{y} + \sigma g, \tilde{y})$ for $\tilde{y} \sim p$
- Thus, $\mathbb{E}_{x\sim p'}[\|\nabla \ln p - \nabla \ln p'\|^2] \leq L^2 \sigma^2 d$

See Lemma C.12 in (Lee-Lu-Tan ’22) for the full proof

Convergence guarantee

Assumption I (L^2-accurate score estimate): For all $t = 0, h, 2h, \dots, T$, the score estimate $s_t(\cdot)$ satisfies

$$\mathbb{E}_{q_t}[\|s_t(X_t) - \nabla \ln q_t(X_t)\|^2] \le \epsilon_{\mathrm{sc}}^2$$

~~**Assumption II** (Smoothness): For all $t \ge 0$, $\nabla \ln q_t(\cdot)$ is L-Lipschitz~~

"early stopping" (Chen-Lee-Lu '23)

Assumption III (Bounded second moment): $\mathfrak{m}_2^2 \coloneqq \mathbb{E}_q[\|x\|^2] < \infty$

Theorem (Chen et al. '23). Under Assumptions I-III, if p_T is the law of the output of the algorithm after $N = T/h$ iterations with step size h,

$$\mathrm{TV}(p_T, q) \lesssim \underbrace{\sqrt{\mathrm{KL}(q\|\gamma)} \cdot \exp(-T)}_{\text{initialization error}} + \underbrace{\left(L\sqrt{dh} + L\mathfrak{m}_2 h\right)\sqrt{T}}_{\text{discretization error}} + \underbrace{\epsilon_{\mathrm{sc}}\sqrt{T}}_{\text{score error}}$$

- Choose $T = \log(\mathrm{KL}(q\|\gamma)/\epsilon)$ and $h = \frac{\epsilon^2}{L^2 d}$ gives $\mathrm{TV}(p_T, q) = \tilde{O}(\epsilon + \epsilon_{\mathrm{sc}})$ with $N = \tilde{O}(L^2 d/\epsilon^2)$ steps

Using ODE flow + Langevin corrector can achieve $\sqrt{d}$ steps (Chen-Chewi-Lee-Li-Lu-Salim '23)

SOTA DDPM convergence bound: $N = d/\epsilon$ (Li-Yan '25)

Stochastic localization

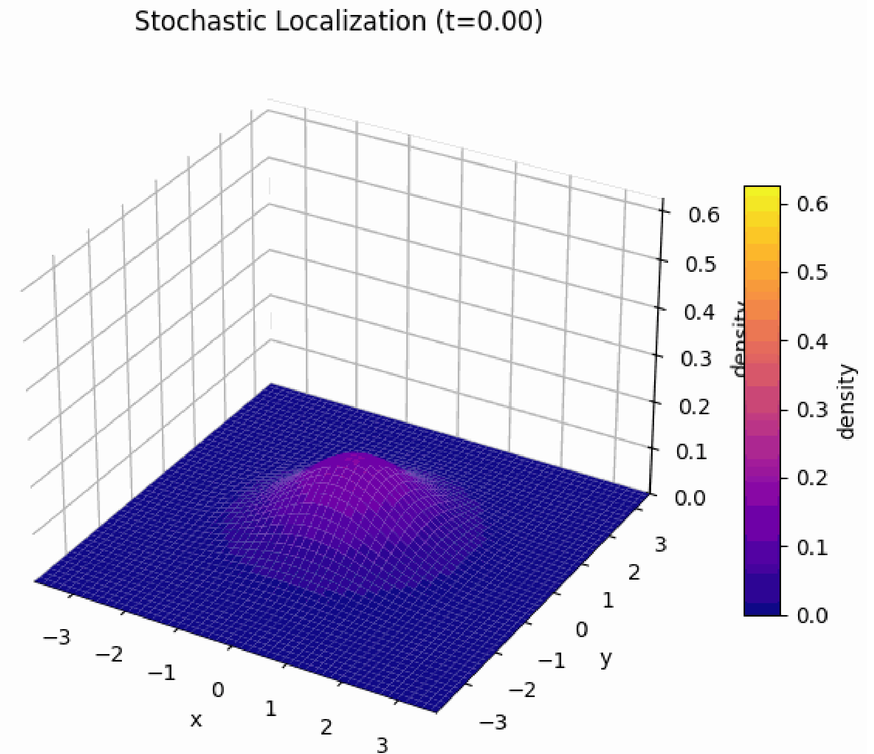
Let π_0 be a probability measure over $\mathbb{R}^d$ with mean $\mathbf{m}_0$

Define a stochastic process $\{\mathbf{c}_t\}_{t\geq 0}$ as follows:

$$\mathbf{c}_0 = 0, \qquad \mathrm{d}\mathbf{c}_t = \mathbf{m}_t \mathrm{d}t + \mathrm{d}B_t$$

where $\mathbf{m}_t := \mathbb{E}_{\pi_t}[\mathbf{x}]$ and $\pi_t(\mathbf{x}) \propto \exp\left(\langle \mathbf{c}_t, \mathbf{x}\rangle - \frac{t}{2}\|\mathbf{x}\|^2\right)\pi_0(\mathbf{x})$

We call the random induced measures $\{\pi_t\}_{t\geq 0}$ the stochastic localizatio



- As $t \to \infty$, $\pi_t \to \delta_{\mathbf{x}}$ i.e. π_t localizes towards a delta-measure at son
- $\{\pi_t\}_{t\geq 0}$ is measure-valued martingale i.e. $\mathbb{E}[\pi_t(\mathbf{x})] = \pi_0(x)$ for all

SL $\Longleftrightarrow$ Diffusion model

$t \leftrightarrow \frac{1}{2}\log\frac{t+1}{t}$

Consider the reverse process of diffusion model (re-parameterized):

$$\mathrm{d}X_t^{\leftarrow} = \left\{\frac{X_t^{\leftarrow}}{2t(t+1)} + \frac{1}{t(t+1)}\nabla \ln \pi_t^{\leftarrow}(X_t^{\leftarrow})\right\} \mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\, \mathrm{d}B_t$$

with $X_0^{\leftarrow} \sim \mathcal{N}(0, I)$

Then the processes $\{\mathbf{c}_t\}_{t\geq 0}$ and $\{X_t^{\leftarrow}\}_{t\geq 0}$ satisfy $\sqrt{t(t+1)}X_t^{\leftarrow} = \mathbf{c}_t$

- $\mathrm{d}\mathbf{c}_t = \mathbf{m}_t \mathrm{d}t + \mathrm{d}B_t$ with the change of variable $\sqrt{t(t+1)}X_t^{\leftarrow} = \mathbf{c}_t$ gives

$$\mathrm{d}X_t^{\leftarrow} = -\frac{2t+1}{2t(t+1)}X_t^{\leftarrow}\mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\mathbf{m}_t\mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\mathrm{d}B_t$$

SL $\Longleftrightarrow$ Diffusion model

Consider the reverse process of diffusion model (re-parameterized):

$t \leftrightarrow \frac{1}{2}\log\frac{t+1}{t}$

$$\mathrm{d}X_t^{\leftarrow} = \left\{\frac{X_t^{\leftarrow}}{2t(t+1)} + \frac{1}{t(t+1)}\nabla \ln \pi_t^{\leftarrow}(X_t^{\leftarrow})\right\}\mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\,\mathrm{d}B_t$$

with $X_0^{\leftarrow} \sim \mathcal{N}(0, I)$

Then the processes $\{\mathbf{c}_t\}_{t\geq 0}$ and $\{X_t^{\leftarrow}\}_{t\geq 0}$ satisfy $\sqrt{t(t+1)}X_t^{\leftarrow} = \mathbf{c}_t$

- $X_t^{\leftarrow} = X_{t'}$ for $t' = \frac{1}{2}\log\frac{t+1}{t}$ in the forward process, which distributed as $e^{-t'}X_0 + \left(1 - e^{-2t'}\right)g$
- By Tweedie's formula, $\nabla \ln \pi_t^{\leftarrow}(X_t^{\leftarrow}) = \sqrt{t(t+1)}\mathbb{E}_{\pi_t^{\leftarrow}}[X_0] - (t+1)X_t^{\leftarrow}$

$$\mathrm{d}X_t^{\leftarrow} = \left\{-\frac{2t+1}{2t(t+1)}X_t^{\leftarrow} + \frac{1}{\sqrt{t(t+1)}}\mathbb{E}_{\pi_t^{\leftarrow}}[X_0]\right\}\mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\,\mathrm{d}B_t$$

SL $\Longleftrightarrow$ Diffusion model

$t \leftrightarrow \frac{1}{2}\log\frac{t+1}{t}$

Consider the reverse process of diffusion model (re-parameterized):

$$\mathrm{d}X_t^{\leftarrow} = \left\{\frac{X_t^{\leftarrow}}{2t(t+1)} + \frac{1}{t(t+1)}\nabla \ln \pi_t^{\leftarrow}(X_t^{\leftarrow})\right\}\mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\mathrm{d}B_t$$

with $X_0^{\leftarrow} \sim \mathcal{N}(0, I)$

Then the processes $\{\mathbf{c}_t\}_{t\geq 0}$ and $\{X_t^{\leftarrow}\}_{t\geq 0}$ satisfy $\sqrt{t(t+1)}X_t^{\leftarrow} = \mathbf{c}_t$

$$\mathrm{d}X_t^{\leftarrow} = -\frac{2t+1}{2t(t+1)}X_t^{\leftarrow}\mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\mathbf{m}_t\mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\mathrm{d}B_t$$

$$\mathrm{d}X_t^{\leftarrow} = \left\{-\frac{2t+1}{2t(t+1)}X_t^{\leftarrow} + \frac{1}{\sqrt{t(t+1)}}\mathbb{E}_{\pi_t^{\leftarrow}}[X_0]\right\}\mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\mathrm{d}B_t$$

SL $\Longleftrightarrow$ Diffusion model

Consider the reverse process of diffusion model (re-parameterized):

$t \leftrightarrow \frac{1}{2}\log\frac{t+1}{t}$

$$\mathrm{d}X_t^{\leftarrow} = \left\{\frac{X_t^{\leftarrow}}{2t(t+1)} + \frac{1}{t(t+1)}\nabla \ln \pi_t^{\leftarrow}(X_t^{\leftarrow})\right\}\mathrm{d}t + \frac{1}{\sqrt{t(t+1)}}\,\mathrm{d}B_t$$

with $X_0^{\leftarrow} \sim \mathcal{N}(0, I)$

Then the processes $\{\mathbf{c}_t\}_{t\geq 0}$ and $\{X_t^{\leftarrow}\}_{t\geq 0}$ satisfy $\sqrt{t(t+1)}X_t^{\leftarrow} = \mathbf{c}_t$

- $\pi_t^{\leftarrow}$ is the density of $e^{-t'}X_0 + \left(1 - e^{-2t'}\right)g = \sqrt{\frac{t}{t+1}}X_0 + \frac{1}{\sqrt{t+1}}g$

$$\pi_t^{\leftarrow}(X_0|X_t^{\leftarrow}) \propto \exp\left(-\frac{t+1}{2}\left\|\sqrt{t/t+1}\,X_0 - X_t^{\leftarrow}\right\|^2\right)\pi_0(X_0)$$

$$\propto \exp\left(\sqrt{t(t+1)}\langle X_t^{\leftarrow}, X_0\rangle - \frac{t}{2}\|X_0\|^2\right)\pi_0(X_0) = \overbrace{\exp\left(\langle \mathbf{c}_t, X_0\rangle - \frac{t}{2}\|X_0\|^2\right)\pi_0(X_0)}^{\pi_t(X_0)}$$

Provable score estimation

- **El Alaoui-Montanari-Sellke '22; Celentano '24:** Sampling from the Sherrington–Kirkpatrick model where
$$\pi_W(x) \propto \exp(-\beta x^\top W x)$$
$W \in \mathbb{R}^{d\times d}$ is a random matrix with i.i.d. Gaussian entries

- **El Alaoui-Montanari-Sellke '23; Huang-Montanari-Pham '24; Huang-Mohanty-Rajaraman-Wu '24:** Sampling from the p-spin spherical spin glass

- **Montanari-Wu '23; Montanari-Wu '24:** Bayesian posterior sampling: observing $A = \frac{\beta}{d}\theta^\top\theta + W$ where $\theta \sim_u \{-1,1\}^d$, sample from the posterior distribution
$$P(\theta|A) \propto \exp\left(-\frac{\beta}{2}\theta^\top A\theta\right)$$

- **Shah-Chen-Klivans '23; Chen-Kontonis-Shah '24; Gatmiry-Kelner-Lee '24:** learning mixtures of Gaussians

- **Chewi-Kalavasis-Mehrotra-Montasser '25:** Statistical theory for score estimation (and a comprehensive literature review on diffusion model theory)