Score based model

jhhope1

July 3, 2023

Outline

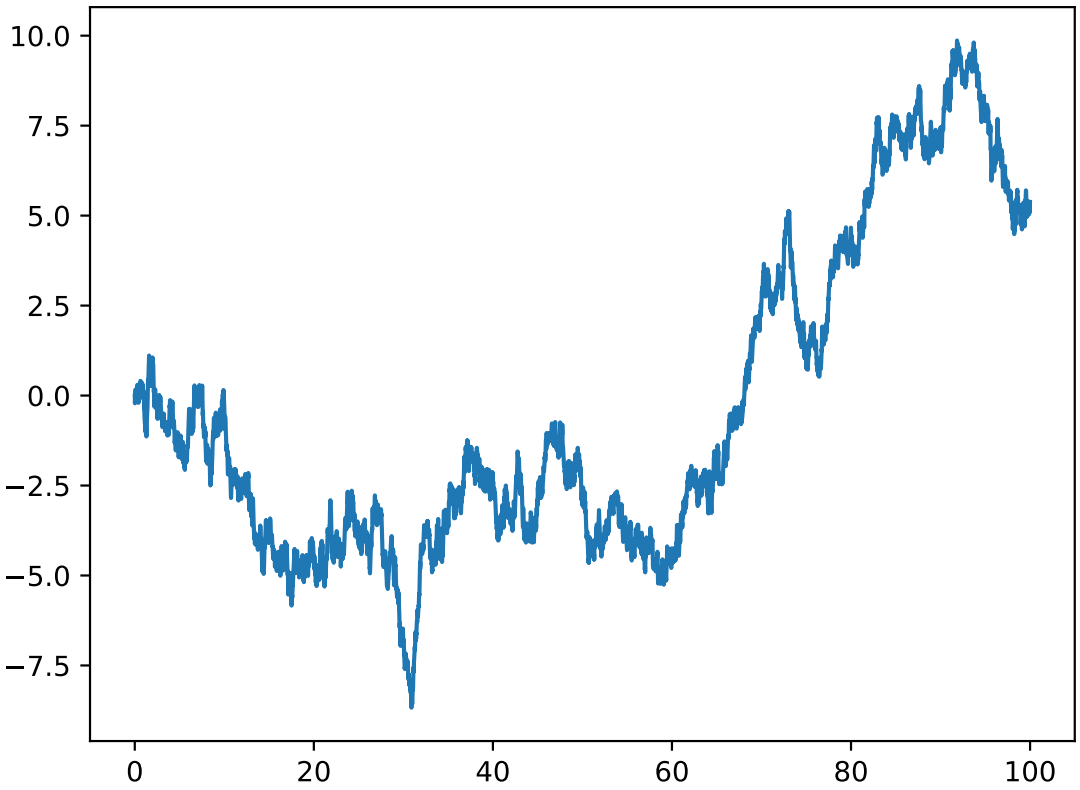
1. SDE
2. Generative models
3. References

Weiner process

A Wiener process W_t is a stochastic process which satisfies

1. $W_0 = 0$
2. W has independent increments: $\forall_{t>0}$, $W_{t+u} - W_t$, $u \geq 0$, are independent of the past values W_s
3. W has Gaussian increments: $W_{t+u} - W_t \sim \mathcal{N}(0, u)$
4. W has continous paths: W_t is continuous in t

Weiner process



Stochastic Differential Equation

A Stochastic Differential Equation(SDE) is a differential equation which contains a stochastic process.

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t$$

Ex. Weiner process

$X_0(x) \sim \delta(x)$, $dX_t = dW_t$

Then, $X_t = W_t$, $p_t(x) \sim \frac{1}{\sqrt{2\pi t}} e^{-x^2/(2t)}$

Time evolution of the probability density

Ex. diffusion process

$dX_t = dW_t$

Time evolution of the probability density

Ex. diffusion process

$dX_t = dW_t$

$$
\begin{aligned}
& p_{t+\Delta t}(x_{t+\Delta t}) \\
&= \int_{-\infty}^{\infty} G(\Delta x, \Delta t) p_t(x_t) dx_t \\
&= \int_{-\infty}^{\infty} G(\Delta x, \Delta t) p_t(x_{t+\Delta t} - \Delta x) d\Delta x \\
&= \int_{-\infty}^{\infty} G(\Delta x, \Delta t)(p_t(x_{t+\Delta t}) - \frac{\partial p_t}{\partial x}\Delta x + \frac{1}{2}\frac{\partial^2 p_t}{\partial x^2}(\Delta x)^2 + \mathcal{O}((\Delta x)^3)) d\Delta x \\
&= p_t(x_{t+\Delta t}) + \frac{1}{2}\frac{\partial^2 p_t}{\partial x^2}\mathbb{E}[(\Delta x)^2] = p_t(x_{t+\Delta t}) + \frac{1}{2}\frac{\partial^2 p_t}{\partial x^2}\Delta t
\end{aligned}
$$

Time evolution of the probability density

Ex. diffusion process

$dX_t = dW_t$

$$\therefore \frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2 p}{\partial x^2}$$

or,

$$\frac{\partial p}{\partial t} = \frac{1}{2}\nabla^2 p := D\nabla^2 p$$

D is called the diffusion coefficient.

Probability flow

Ex. Probability flow of the ODE

$$dX_t = \mu(X_t, t)dt$$

Probability flow

Ex. Probability flow of the ODE

$$dX_t = \mu(X_t, t)dt$$

$$\frac{\partial p}{\partial t} = -\frac{\partial(\mu p)}{\partial x}$$

or,

$$\frac{\partial p}{\partial t} = -\nabla \cdot (\mu p)$$

Fokker-Planck equation

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t$$

For a diffusion process, $p(x, t)$ follows the Fokker-Planck equation.

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}[\mu(x,t)p(x,t)] + \frac{\partial^2}{\partial x^2}[D(x,t)p(x,t)]$$

Where $D(x, t) = \frac{1}{2}\sigma^2(x, t)$

Ornstein–Uhlenbeck process

Ornstein–Uhlenbeck process

$$dX_t = -\theta X_t dt + \sigma dW_t$$

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}[\theta x p(x,t)] + \frac{\partial^2}{\partial x^2}[\frac{1}{2}\sigma^2 p(x,t)]$$

Then the stationary density is $p(x) = \sqrt{\frac{\theta}{\pi\sigma^2}} e^{-\theta x^2/\sigma^2}$.

Ornstein–Uhlenbeck process

Ornstein–Uhlenbeck process

$$dX_t = -\theta X_t dt + \sigma dW_t$$

For the same path $W_t(\omega)$,

$$\frac{d(X_t - Y_t)}{dt} = -\theta(X_t - Y_t)$$
$$X_t - Y_t = (X_0 - Y_0)e^{-\theta t}$$

$$\therefore \lim_{t\to\infty} p(x,t) = p(x) = \sqrt{\frac{\theta}{\pi\sigma^2}} e^{-\theta x^2/\sigma^2}$$

.

Langevin diffusion

Can we write a SDE satisfying

$$\lim_{t\to\infty} p(x,t) = p(x)$$

with arbitrary initial conditions?

Langevin diffusion

For the stationary process,

$$0 = \frac{\partial p(x)}{\partial t} = -\frac{\partial}{\partial x}[\mu(x)p(x)] + \frac{\partial^2}{\partial x^2}[D(x)p(x)]$$

$D(x) = 1$, $\mu(x) = \frac{1}{p(x)}\frac{\partial}{\partial x}p(x) = \frac{\partial}{\partial x}\log p(x)$ could be one answer.

Langevin diffusion

Langevin diffusion

$$dX_t = \nabla \log p(X_t)dt + \sqrt{2}dW_t$$

Then $p(x,t)$ converges to $p(x)$ as $t \to \infty$.

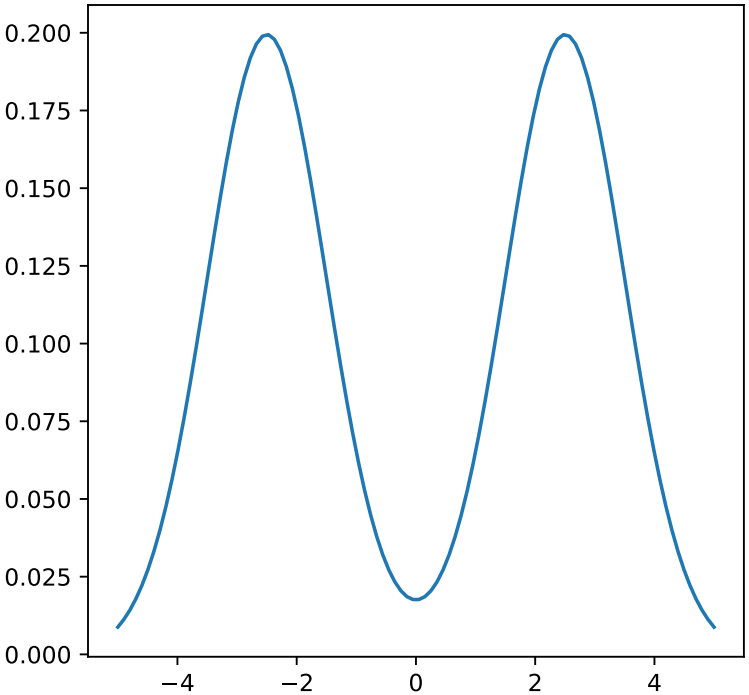
Langevin diffusion

Excercise

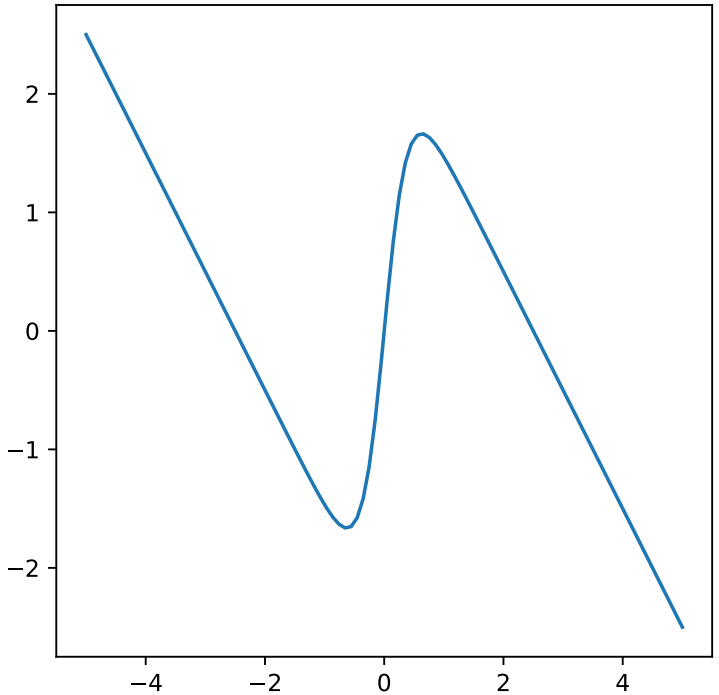
Let $\mathcal{F} = D_{kl}(p(x,t)||p(x)) = \mathbb{E}_{x\sim p(x)}[\log p(x) - \log p(x,t)]$.
By the Fokker-Planck equation,

$$\frac{\partial}{\partial t}\mathcal{F}(t) = ... = -\mathbb{E}[(\frac{\partial}{\partial x}\log p(x,t) - \frac{\partial}{\partial x}\log p(x))^2]$$

Langevin sampling

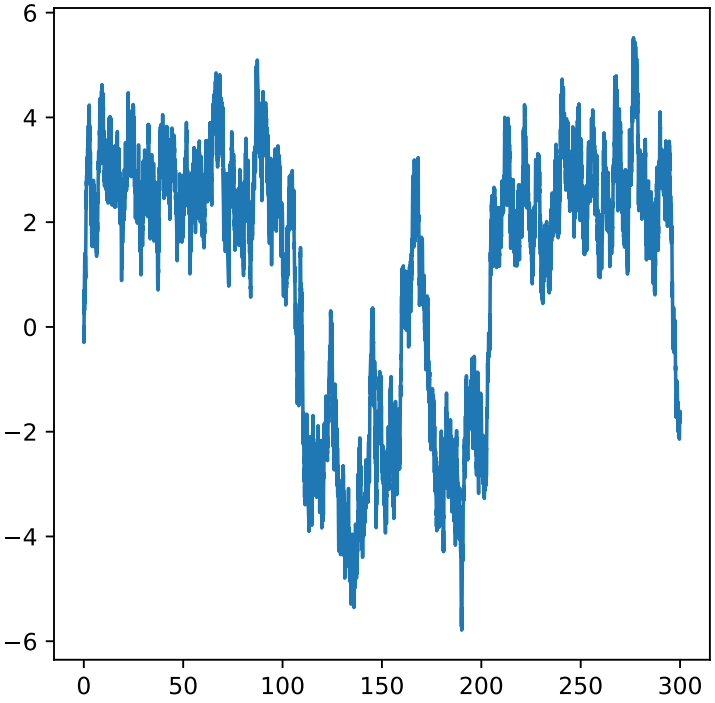


(a) Bimodal distribution

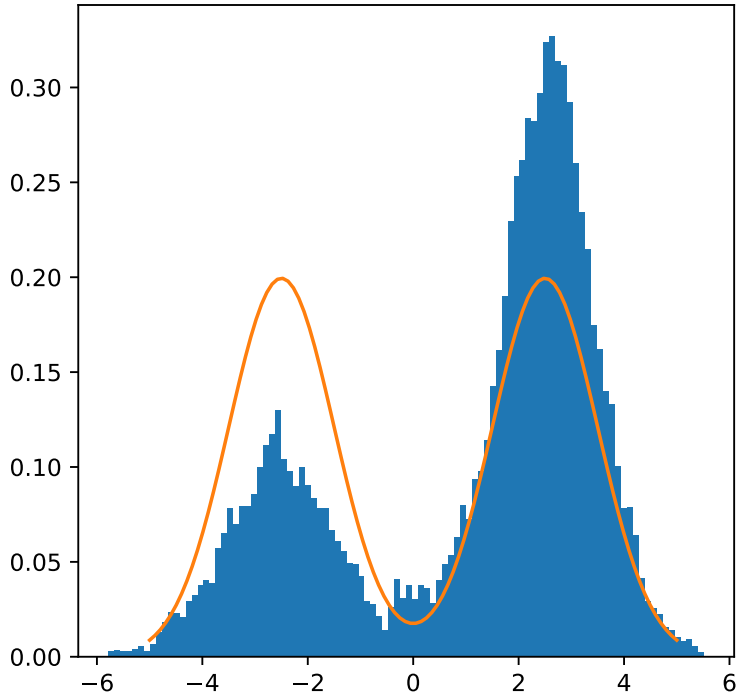


(b) Bimodal score

Langevin sampling

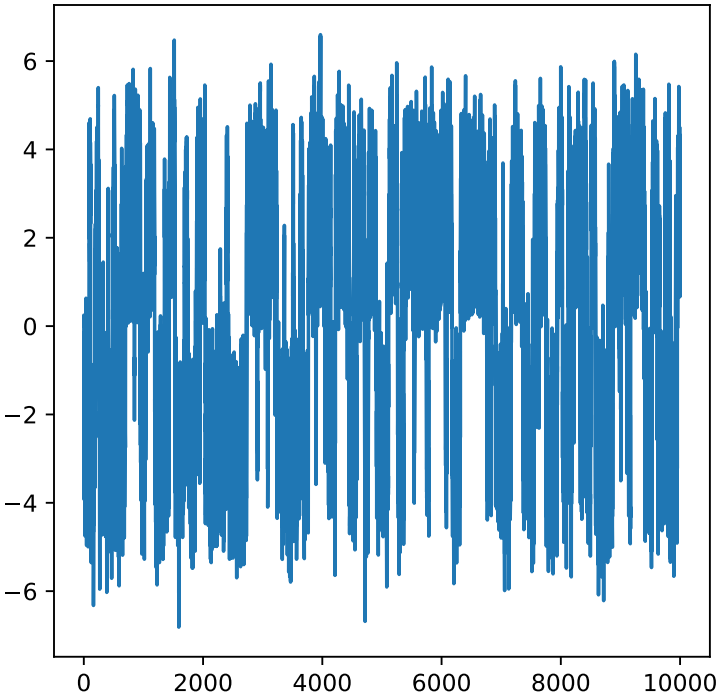


(c) Langevin sampling $\sim$ 300s

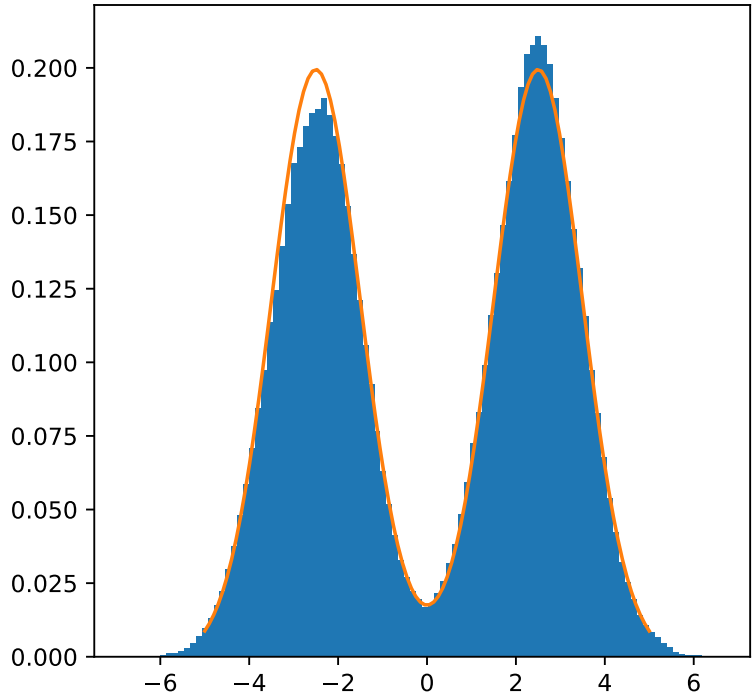


(d) Sample distribution

Langevin sampling



(e) Langevin sampling $\sim$ 10000s



(f) Sample distribution

Generative models

GM learns the real-world data distribution $p(x)$ and draws samples from it.

Energy based model

Energy based model learns a scalar energy function $E : X \to \mathbb{R}$ where

$$p_\theta(x) = e^{-E_\theta(x)}/Z(\theta)$$
$$Z(\theta) = \int e^{-E_\theta(x)} dx$$

Objective is the typical negative log-likelihood.

$$\mathcal{L} = \mathbb{E}_{x\sim p(x)}[-\log(p_\theta(x))] = \mathbb{E}_{x\sim p(x)}[E_\theta(x)] - Z(\theta)$$

Energy based model

Ex. Optimization without considering Z

$X \sim \mathcal{N}(0,1)$, $E_\theta(x) = -x^2/(2\theta)$

Estimation of the intractable term $Z(\theta)$ is needed. Requires sampling from $x \sim p_\theta(x)$ using Langevin MCMC or MALA at training time.
$\rightarrow$ Slow

Variational autoencoder

Variational autoencoder(VAE) models the true distribution $p(x)$ with combination of known distributions $p_\theta(x|z)$ (usually Gaussian).

$$p_\theta(x) = \int_{z \sim p(z)} p_\theta(x|z)p(z)dz$$

$\rightarrow$ restricts the model family.

Score based model

Recall) Energy based model objective

$$\mathcal{L} = \mathbb{E}_{x\sim p(x)}[-\log(p_\theta(x))] = \mathbb{E}_{x\sim p(x)}[E_\theta(x)] - Z(\theta)$$

To avoid the intractability of $Z(\theta)$, we focus on the "score".

$$s(x) := \nabla_x \log p(x)$$

Then, the intractable term vanishes(though the previous objective is invalid).

$$s_\theta(x) = \nabla_x \log(e^{-E_\theta(x)}/Z(\theta)) = -\nabla_x E_\theta(x)$$

Langevin diffusion

Langevin diffusion

$$dX_t = \frac{1}{2}\nabla \log p(X_t)dt + dW_t$$

We can draw a sample from $p(x, \infty) = p(x)$ by discretizing the SDE.

Euler–Maruyama method

$$dX_t = \frac{1}{2}\nabla \log p(X_t)dt + dW_t$$

$$\downarrow$$

$$X_{t+1} = X_t + \frac{\tau}{2}\nabla \log p(X_t)dt + \sqrt{\tau}d\xi_t$$

$$\cong X_t + \frac{\tau}{2}s_\theta(X_t)dt + \sqrt{\tau}d\xi_t$$

where $\xi_t \sim \mathcal{N}(0, 1)$

Score based model

We model a vector valued score function $s_\theta : \mathbb{R}^n \to T\mathbb{R}^n$

Fisher divergence

Fisher divergence is reduced to the simple expectation form through integration by parts.

$$
\begin{aligned}
\mathcal{F} &= \mathbb{E}_{x\sim p(x)}[\|s_\theta(x) - s(x)\|^2] \\
&= \mathbb{E}_{x\sim p(x)}[\|s_\theta(x)\|^2] - 2\sum_i \int s_{\theta i}(x)\partial_i p(x)dx + const \\
&= \mathbb{E}_{x\sim p(x)}[\|s_\theta(x)\|^2] + 2\sum_i \int \partial_i s_{\theta i}(x)p(x)dx + const \\
&= \mathbb{E}_{x\sim p(x)}[\|s_\theta(x)\|^2] + 2\sum_i \int \partial_i s_{\theta i}(x)p(x)dx + const
\end{aligned}
$$

Score matching

Our objective is

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim p(x)}[\frac{1}{2}\|s_\theta(x)\|^2 + tr(\nabla_x s_\theta(x))].$$

Under some assumptions, $\theta^* = \arg\min_\theta(\mathcal{L}(\theta))$ is consistent.

Score matching

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim p(x)}[\frac{1}{2}\|s_\theta(x)\|^2 + tr(\nabla_x s_\theta(x))].$$

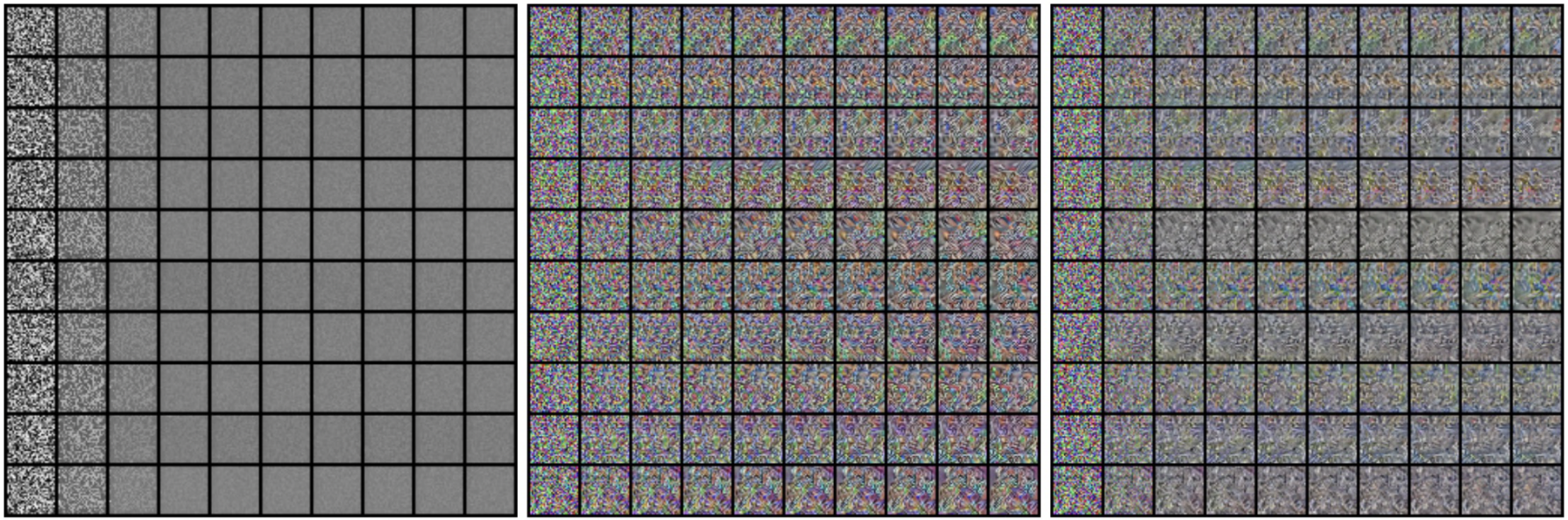
$\mathcal{L}$ requires derivatives of each $\partial_i s_\theta(x)$ for $i \in [1, n]$. $\rightarrow$ expensive!

Sliced score matching

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim p(x)}[\frac{1}{2}\|s_\theta(x)\|^2 + tr(\nabla_x s_\theta(x))]$$
$$= \mathbb{E}_{x\sim p(x)}[\frac{1}{2}\|s_\theta(x)\|^2] + \mathbb{E}_v[\mathbb{E}_{x\sim p(x)}[\nabla_v s_\theta(x)]]$$

Score based models

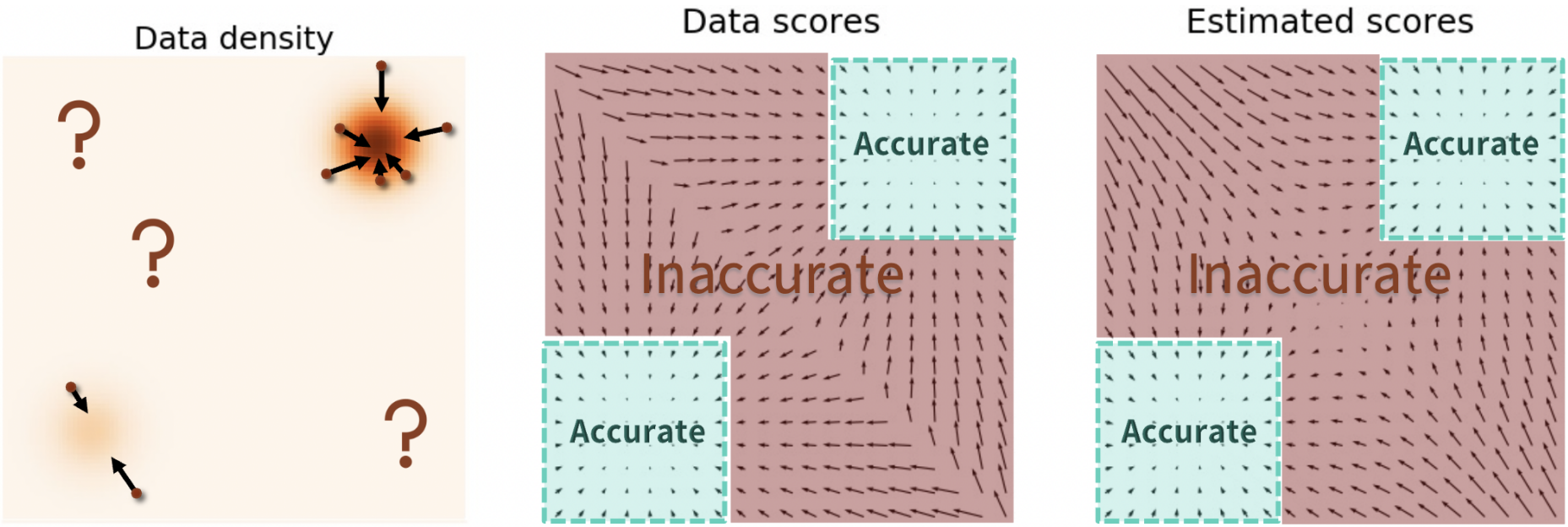
?



(a) MNIST (b) CelebA (c) CIFAR-10

Challenge in low density regions

Model cannot predict the score of outliers.



SDE revisit

$$dX_t = dW_t$$

This SDE gradually transforms the initial distribution $X_0 \sim p(x)$ into

$$\lim_{t \to \infty} X_t \sim \mathcal{N}(0, t)$$

.

SDE revisit

$$dX_t = dW_t$$

If we can reverse the SDE from $t = \infty$ to $t = 0$, we can sample $X_0 \sim p(x)$ from known distribution $\mathcal{N}(0, t)$

Time reversal diffusion

'Reverse-Time Diffusion Equation Model' by Brian D.O. Anderson (1982) gives the following surprising result.

Reverse-Time Diffusion Equation Model

Under nice conditions, the diffusion equation has a corresponding time-reversal diffusion equation.

From the above fact, let's think about the time-reversal diffusion equation.

Time reversal diffusion

First, consider the Weiner process $dX_t = dW_t$ where $X_0 \sim \delta(x)$.
The solution of the time-reversal equation must reach zero no matter where it starts.
So, The form of $dX_t = -(terms)\nabla p(x,t)dt + (terms)d\tilde{W}_t$ is highly suspected.

Time reversal diffusion

Let's find a time-reversal equation of the form
$dX_t = \tilde{\mu}(X_t, t)dt + \tilde{\sigma}(X_t, t)d\tilde{W}_t$

$$
\begin{aligned}
& p(x_t, t|x_{t+\Delta t}, t+\Delta t) \\
&= \frac{p(x_{t+\Delta t}, t+\Delta t|x_t, t)p(x_t, t)}{p(x_{t+\Delta t}, t+\Delta t)} \\
&= (const)exp[\frac{-(\Delta x - \mu(x_t, t)\Delta t)^2 - 2\sigma(x_t, t)^2\Delta t \log p(x_t, t)}{2\sigma(x_t, t)^2}\Delta t] \\
&\sim exp[\frac{-(\Delta x - \tilde{\mu}(x_{t+\Delta t}, t+\Delta t))^2}{2\tilde{\sigma}(x_{t+\Delta t})^2\Delta t}]
\end{aligned}
$$

Time reversal diffusion

By considering $x \sim \sqrt{\Delta t}$ and applying a Taylor expansion, we can derive the following time-reversal diffusion equation.

$$dX_t = [\mu(X_t, t) - \frac{\partial}{\partial x}(\sigma(X_t, t)^2 \log p(X_t, t))]dt + \sigma(X_t, t)d\tilde{W}_t$$

In our case,

$$dX_t = -\sigma(t)^2 \frac{\partial}{\partial x}(\log p(X_t, t))dt + \sigma(t)d\tilde{W}_t$$

Denoising score matching

Let q_σ be the perturbed distribution.

$$q_\sigma = \int q_\sigma(\tilde{x}|x)p(x)dx$$

Then score matching loss is

$$L_\sigma = \mathbb{E}_{x\sim p(x)}\mathbb{E}_{\tilde{x}\sim q(\tilde{x}|x)}\frac{1}{2}[\|s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)\|^2]$$

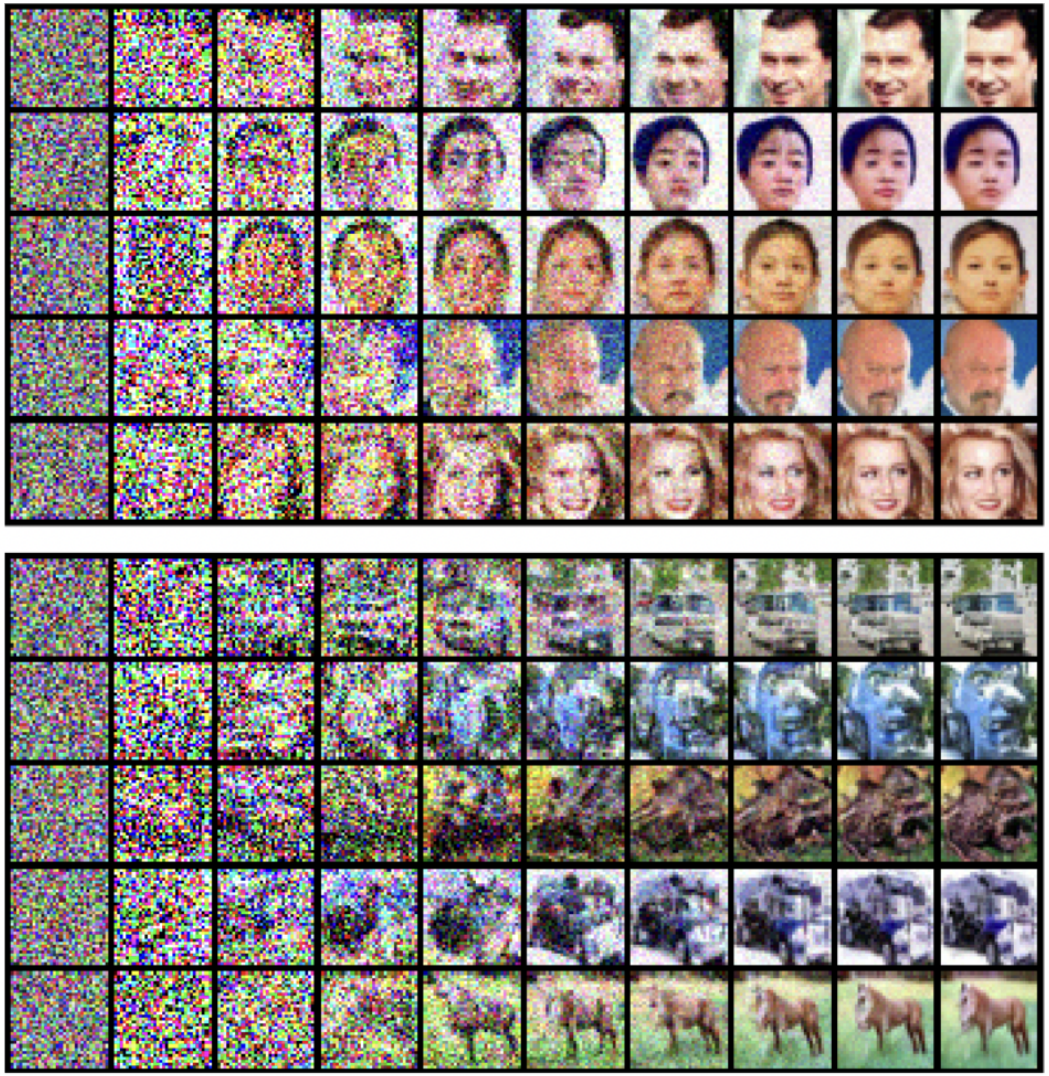
Denoising score matching

If we choose $q_\sigma(\tilde{x}|x) = G(x, \sigma^2)$,

$$L_\sigma = \mathbb{E}_{x\sim p(x)}\mathbb{E}_{\tilde{x}\sim q(\tilde{x}|x)}\frac{1}{2}[\|s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)\|^2]$$

$$= \mathbb{E}_{x\sim p(x)}\mathbb{E}_{\tilde{x}\sim q(\tilde{x}|x)}\frac{1}{2}[\|s_\theta(\tilde{x}) - (\tilde{x} - x)/\sigma^2\|^2]$$

Results



Fokker-Planck equation revisit

For the time-reverse SDE,

$$dX_t = \tilde{\mu}(X_t, t)dt + \tilde{\sigma}(X_t, t)d\tilde{W}_t$$

Let $\tau = -t$

$$dX_\tau = -\tilde{\mu}(X_t, t)d\tau + \tilde{\sigma}(X_t, t)dW_\tau$$

By Fokker-planck equation,

$$-\frac{\partial p(x,t)}{\partial t} = \frac{\partial p(x,t)}{\partial \tau} = -\frac{\partial}{\partial x}[-\tilde{\mu}(X_t, t)p(x,t)] + \frac{\partial^2}{\partial x^2}[\tilde{D}(x,t)p(x,t)]$$

Fokker-Planck equation revisit

$$
\begin{aligned}
&-\frac{\partial}{\partial x}[\mu(x,t)p(x,t)] + \frac{\partial^2}{\partial x^2}[D(x,t)p(x,t)] \\
&= \frac{\partial p(x,t)}{\partial t} \\
&= -\frac{\partial}{\partial x}[\tilde{\mu}(X_t,t)p(x,t)] - \frac{\partial^2}{\partial x^2}[\tilde{D}(x,t)p(x,t)]
\end{aligned}
$$

Our time-reversal solution $\tilde{D}(x,t) = D(x,t)$,

$\tilde{\mu}(x,t) = \mu(x,t) - \sigma(t)^2 \frac{\partial}{\partial x} \log p(x,t)$ satisfies the above equation.

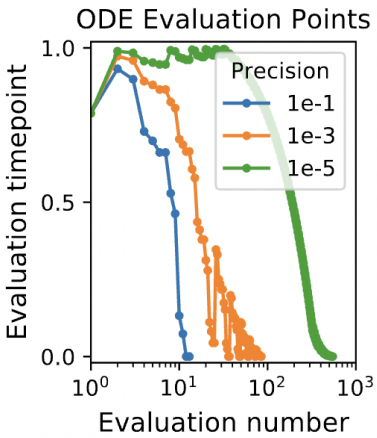
Probability flow revisit

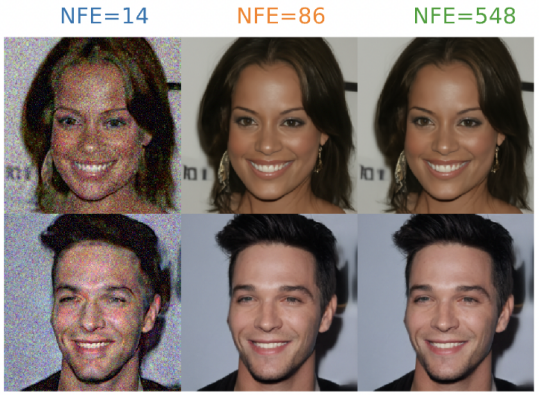
$$
\begin{aligned}
&- \frac{\partial}{\partial x}[\mu(x,t)p(x,t)] + \frac{\partial^2}{\partial x^2}[D(x,t)p(x,t)] \\
&= -\frac{\partial}{\partial x}[\tilde{\mu}(X_t,t)p(x,t)] - \frac{\partial^2}{\partial x^2}[\tilde{D}(x,t)p(x,t)]
\end{aligned}
$$

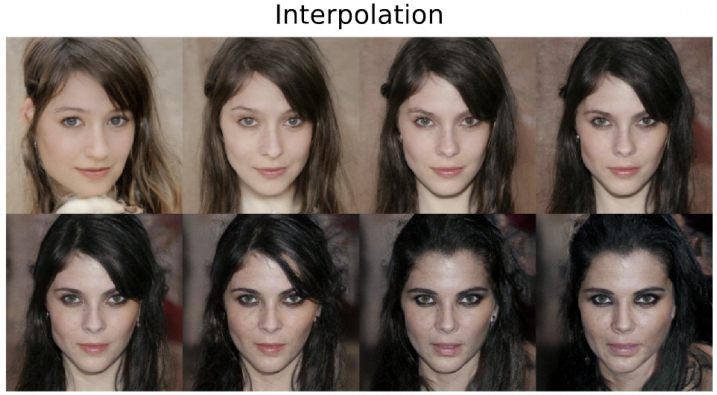
$\tilde{D}(x,t) = 0$, $\tilde{\mu}(x,t) = \mu(x,t) - \frac{1}{2}\sigma(t)^2 \frac{\partial}{\partial x} \log p(x,t)$ also satisfies the above equation.

$\rightarrow$ We can sample by solving a deterministic ODE!

Samples from solving Probability flow ODE







References

- Score-Based Generative Modeling through Stochastic Differential Equations.
- Generative Modeling by Estimating Gradients of the Data Distribution.
- Sliced Score Matching: A Scalable Approach to Density and Score Estimation.