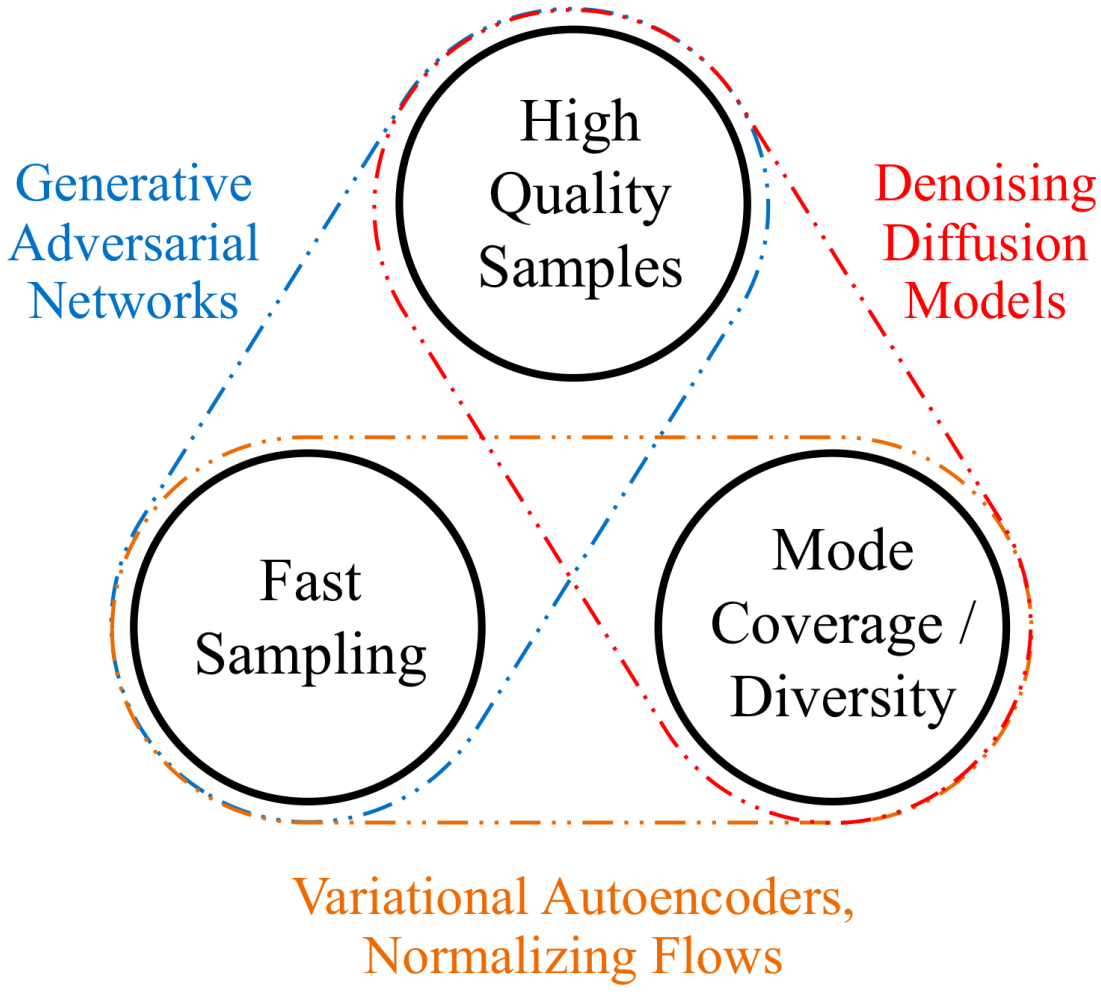
Diffusion models

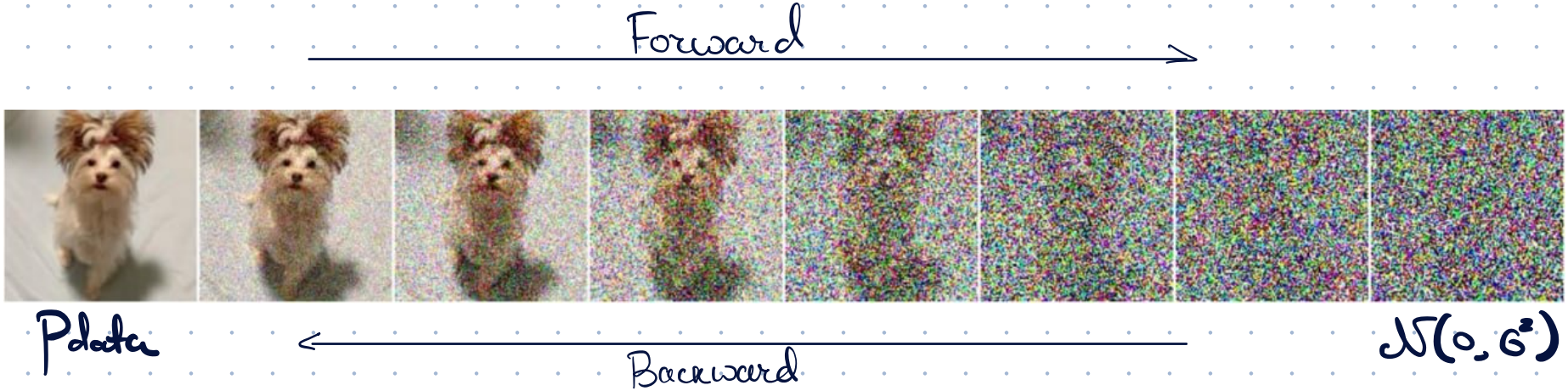
Denis Rakitin

Generative trilemma



https://arxiv.org/abs/2112.07804

Diffusion



Idea: define forward noising process and try to reverse it.

https://arxiv.org/abs/2011.13456

Discrete time

- Variance preserving (VP) process: $t = 1, \dots, T$

$$X_{t+1} = \sqrt{1-\beta_t}\, X_t + \sqrt{\beta_t}\, \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, I)$$

(If $\mathrm{Var}\, X_0 = I \Rightarrow \mathrm{Var}\, X_t = I$)

$$p_{t|0}(X_t | X_0) = \mathcal{N}(X_t \mid \alpha_t \cdot X_0, \sigma_t^2 I) \qquad \alpha_t \to 0, \quad \sigma_t^2 \to 1$$

$$p_T(X_T) \approx \mathcal{N}(X_T | 0, I)$$

https://arxiv.org/abs/2006.11239

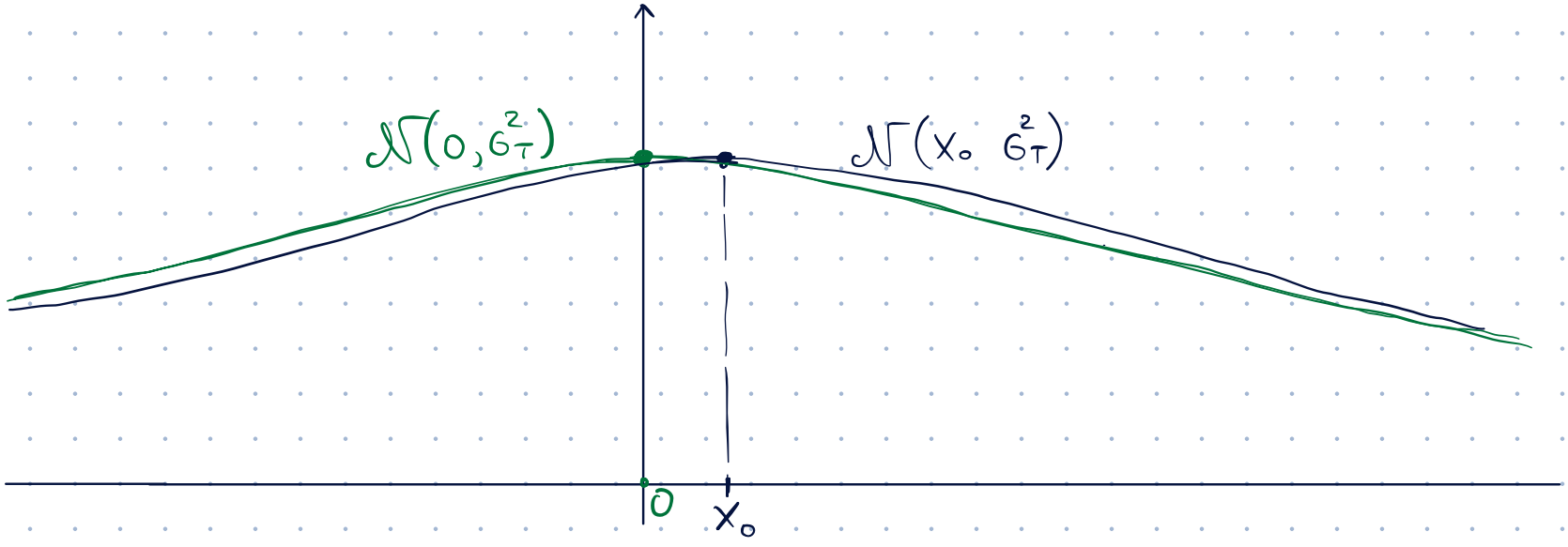
Discrete time

- Variance exploding (VE) process: $t = 1, \dots, T$

$$x_{t+1} = x_t + g(t) \cdot \varepsilon_t \,, \quad \varepsilon_t \sim \mathcal{N}(0, I)$$

$$p_{t|0}(x_t | x_0) = \mathcal{N}(x_t | x_0, \sigma_t^2 I) \,, \quad \sigma_t^2 = \sum_{s=t} g^2(s)$$

If σ_T^2 is big $\Rightarrow$ $p_T(x_T) \approx \mathcal{N}(0, \sigma_T^2 I)$



https://arxiv.org/abs/1907.05600

Basic idea (VE)

$$p_{t+1|t}(x_{t+1}|x_t) = \mathcal{N}(x_{t+1}|x_t, g^2(t)I)$$

$$p_T(x_T) \approx \mathcal{N}(x_T|0, \sigma_T^2), \quad \sigma_T^2 = \sum_{t=1}^{T} g^2(t).$$

Ideally:

- Sample $x_T \sim \mathcal{N}(0, \sigma_T^2)$
- Iterate $x_{t-1} \sim p_{t-1|t}(x_{t-1}|x_t) = \dfrac{p_{t|t-1}(x_t|x_{t-1})\, p_{t-1}(x_{t-1})}{p_t(x_t)}$

intractable ↓

intractable ↑

Basic idea (VE)

$$p_{t+1|t}(x_{t+1}|x_t) = \mathcal{N}(x_{t+1}|x_t, g^2(t)I)$$

$$p_T(x_T) \approx \mathcal{N}(x_T|0, \sigma_T^2), \quad \sigma_T^2 = \sum_{t=1}^{T} g^2(t).$$

Practical replacement:

- $$p_{t-1|t,0}(x_{t-1}|x_t, x_0) = \frac{p_{t|t-1}(x_t|x_{t-1})\, p_{t-1|0}(x_{t-1}|x_0)}{p_{t|0}(x_t|x_0)}$$

  ($p_{t|t-1}$: gaussian; $p_{t-1|0}$: gaussian; $p_{t|0}(x_t|x_0)$: gaussian)

- Make (maybe not ideal) prediction $\hat{x}_0 = D(x_t, t)$
- Sample $x_{t-1} \sim p_{t-1|t,0}(x_{t-1}|x_t, \hat{x}_0)$

Train NN: $$\sum_{t=1}^{T} \omega_t \mathbb{E} \| D_\theta(x_t, t) - x_0 \|^2 \to \min_\theta$$

Basic idea (summary)

Training: $\sum_{t=1}^{T} \omega_t \mathbb{E} \| D_\theta(X_t, t) - X_0 \|^2 \to \min_\theta$

Sampling:

- Claim: $P_{t-1|t,0}(X_{t-1} | X_t, \hat{X}_0) = \mathcal{N}\left(X_{t-1} \mid \mu_t, \Sigma_t\right)$

$$\mu_t = X_t \cdot \left(1 - \frac{g^2(t)}{\sigma_t^2}\right) + \hat{X}_0 \cdot \frac{g^2(t)}{\sigma_t^2}, \quad \Sigma_t = \boxed{\frac{\sigma_{t-1}^2}{\sigma_t^2}} g^2(t) \cdot I$$

- Sample $X_T \sim \mathcal{N}(0, \sigma_T^2)$
- Iterate

$$X_{t-1} = X_t \left(1 - \frac{g^2(t)}{\sigma_t^2}\right) + D_\theta(X_t, t) \cdot \frac{g^2(t)}{\sigma_t^2} + g(t)\varepsilon_t$$

Links

https://arxiv.org/abs/1907.05600 (VE process)

https://arxiv.org/abs/2006.11239 (VP process, method derivation)

ODEs, SDEs, PDEs

ODEs

$$dX_t = f(X_t, t)dt \quad / \quad \dot{X} = f(X, t)$$

Euler scheme: $X_{t+h} \approx X_t + h \cdot f(X_t, t)$

- Part of VP process: $X_{t+1} = \sqrt{1-\beta_t}\, X_t + \sqrt{\beta_t}\, \varepsilon_t$

- Cont. modification:

$$X_{t+h} = \sqrt{1-\beta_t h}\, X_t = \left(1 - h\frac{\beta_t}{2} + \bar{o}(h)\right) X_t$$

$$\frac{X_{t+h} - X_t}{h} = \frac{-h\frac{\beta_t}{2} + \bar{o}(h)}{h} \cdot X_t$$

$$\frac{X_{t+h} - X_t}{h} \to \dot{X}_t \qquad \frac{-h\frac{\beta_t}{2} + \bar{o}(h)}{h} \to -\frac{\beta_t}{2} X_t$$

Solution: $X_t = X_0 \cdot \exp\left(-\frac{1}{2}\int_0^t \beta_s\, ds\right)$

Wiener process

VE process with $g(t) = 1$:

$X_{t+1} = X_t + \varepsilon_t$, $\varepsilon_t \sim \mathcal{N}(0, I)$ ind. with X_t.

- Let $X_0 = 0$
- Independent increments:
  $t_1 < \dots < t_m \Rightarrow$ $X_{t_1}, X_{t_2} - X_{t_1}, \dots, X_{t_m} - X_{t_{m-1}}$ are indep.
- $X_t - X_s = \sum_{i=s+1}^{t} \varepsilon_i \sim \mathcal{N}(0, (t-s)I)$.

Def Wiener process is a continuous-time process

- $W_0 = 0$
- It has indep. increments
- $W_t - W_s \sim \mathcal{N}(0, (t-s)I)$
- It has continuous trajectories

SDEs

- VE process: $X_{t+1} = X_t + g(t)\,\varepsilon_t$
- VP process: $X_{t+1} = \sqrt{1-\beta_t}\,X_t + \sqrt{\beta_t}\,\varepsilon_t$

Continuous limits?

Def $dX_t = f(X_t,t)\,dt + G(X_t,t)\,d\omega_t$, if

$$X_{t+h} \approx X_t + f(X_t,t)\cdot h + G(X_t,t)\underbrace{(\omega_{t+h} - \omega_t)}_{\text{indep. with } X_t}$$

Euler scheme:

$$X_{t+h} \approx X_t + f(X_t,t)\cdot h + G(X_t,t)\cdot\sqrt{h}\cdot\varepsilon_t,$$

$$\varepsilon_t \sim \mathcal{N}(0,I) - \text{indep. with } X_t$$

SDEs

Discrete VE: $X_{t+1} = X_t + g(t)\varepsilon_t$

$\Leftrightarrow X_t + g(t)(\omega_{t+1} - \omega_t)$

Continuous VE: $X_{t+h} = X_t + g(t)(\omega_{t+h} - \omega_t)$

VE-SDE: $dX_t = g(t)d\omega_t$

$$X_t = X_0 + \int_0^t g(s)d\omega_s$$

$$\left(P_{X_t|X_0}(x_t|x_0) = \right.$$

$$\left. \right)$$

SDEs

Discrete VP: $X_{t+1} = \sqrt{1-\beta_t} X_t + \sqrt{\beta_t}\,\varepsilon_t$

Continuous VP: $X_{t+h} = \sqrt{1-\beta_t h}\, X_t + \sqrt{\beta_t h}\,\varepsilon_t$

$$\Leftrightarrow \left(1 - h\frac{\beta_t}{2} + o(h)\right) X_t + \sqrt{\beta_t}\,(W_{t+h} - W_t)$$

$$\boxed{\text{VP-SDE}: \; dX_t = -\frac{\beta_t}{2} X_t\, dt + \sqrt{\beta_t}\, dW_t}$$

$$X_t = \alpha_t X_0 + \alpha_t \int_0^t \alpha_s^{-\frac{1}{2}} \sqrt{\beta_s}\, dW_s, \qquad \alpha_t = \exp\left(-\int_0^t \beta_s\, ds\right)$$

$$\left(P_{X_t|X_0}(x_t \mid x_0) = \mathcal{N}\left(x_t \mid x_0 \cdot \sqrt{\alpha_t}, (1-\alpha_t) I\right)\right.$$

Backward process

VP/VE (simple)

Forward: $\begin{cases} dX_t = f(X_t, t)\,dt + g(t)\,dW_t \\ X_0 \sim p_{\text{data}} \end{cases}$ ← difficult

s.t. $p_1 := p_{X_1}$ is simple and known (s.t. $\mathcal{N}(0, \sigma^2 I)$).

difficult

Goal: define $X_t^{(b)}$: $\begin{cases} dX_t^{(b)} = f^{(b)}(X_t^{(b)}, t)\,dt + g^{(b)}(t)\,dW_t \\ X_0^{(b)} \sim p_1 \end{cases}$ ← simple

s.t. $q_t := p_{X_t^{(b)}}$ equals $p_{1-t} = p_{X_{1-t}}$.

Backward ODE is easy

Let y_t satisfy $\begin{cases} dy_t = v(y_t, t)dt \\ y_0 = y_0 \end{cases}$

Define $y_t^{(b)}$ as $\begin{cases} dy_t^{(b)} = -v(y_t^{(b)}, 1-t)dt \\ y_0^{(b)} = y_1 \end{cases}$

Then, $y_t^{(b)} = y_{1-t} \quad \forall t.$

Backward SDE scheme

$$\text{SDE} \xrightarrow{?} \text{ODE} \underset{v(\cdot,t) \leftrightarrow -v(\cdot,1-t)}{\longleftrightarrow} \text{b-ODE} \xrightarrow{?} \text{b-SDE}$$

Goal:

$$\begin{cases} dX_t = f(X_t, t)dt + g(t)\,dW_t \\ X_0 \sim p_0 \end{cases}$$

$$\Big\uparrow \Big\downarrow \quad \text{save marginals: } P_{X_t} = P_{Y_t}$$

$$\begin{cases} dY_t = v(Y_t, t)dt \\ Y_0 \sim p_0 \end{cases}$$

How? Study evolution of density.

Evolution of density

- $dX_t = f(X_t, t)\,dt$

$$\frac{\partial}{\partial t} p_t(x) = -\frac{\partial}{\partial x}\left(p_t(x)\, f(x,t)\right)$$

Continuity equation

- $dX_t = f(X_t, t)\,dt + g(t)\,dW_t$

$$\frac{\partial}{\partial t} p_t(x) = -\frac{\partial}{\partial x}\left(p_t(x)\, f(x,t)\right) + \frac{g^2(t)}{2}\frac{\partial^2}{\partial x^2} p_t(x)$$

Fokker-Planck (-Kolmogorov) equation

Interpretation

Conservation Laws (thanks to S.V. Shaposhnikov and T.I. Krasovitsky for their PDE course)

Let $u(x,t)$ be density of a substanse at time t.

Amount of substance in $[x_0, x_0+\Delta x]$ is

$$\int_{x_0}^{x_0+\Delta x} u(x,t)\,dx.$$

Change from time t_0 to $t_0+\Delta t$ is

$$\int_{x_0}^{x_0+\Delta x} u(x,t_0+\Delta t)\,dx - \int_{x_0}^{x_0+\Delta x} u(x,t_0)\,dx.$$

Interpretation

Let $F(x,t)$ be flow of the substance, i.e. the amount of flowed substance in the point x_0 from t_0 to $t_0+\Delta t$ is

$$\int_{t_0}^{t_0+\Delta t} F(x,t)\,dt.$$

Then,

$$\int_{x_0}^{x_0+\Delta x} u(x,t_0+\Delta t) - u(x,t_0)\,dx = \int_{t_0}^{t_0+\Delta t} F(x_0,t) - F(x_0+\Delta x,t)\,dt$$

$$\int_{x_0}^{x_0+\Delta x}\int_{t_0}^{t_0+\Delta t} \frac{\partial}{\partial t}u(x,t)\,dx\,dt = -\int_{x_0}^{x_0+\Delta x}\int_{t_0}^{t_0+\Delta t} \frac{\partial}{\partial x}F(x,t)\,dx\,dt$$

Conservation laws

$$\frac{\partial}{\partial t} u(x,t) = -\frac{\partial}{\partial x} F(x,t)$$

General FPE:

$$\frac{\partial}{\partial t} p_t(x) = -\frac{\partial}{\partial x}\left(p_t(x) f(x,t)\right) + \frac{g^2(t)}{2} \frac{\partial^2}{\partial x^2} p_t(x)$$

- $dX_t = f(X_t,t)dt$ $\quad F(x,t) = f(x,t) \cdot p_t(x)$
- $dX_t = g(t)dW_t$ $\quad F(x,t) = -\frac{g^2(t)}{2} \frac{\partial}{\partial x} p_t(x)$
- $dX_t = f(X_t,t)dt + g(t)dW_t$ $\quad F(x,t) = f(x,t) \cdot p_t(x) - \frac{g^2(t)}{2} \frac{\partial}{\partial x} p_t(x)$

Uniqueness

Thm Under regularity conditions, FPE

$$\begin{cases} \frac{\partial}{\partial t} p_t(x) = -\frac{\partial}{\partial x}\left(p_t(x) f(x,t)\right) + \frac{g^2(t)}{2} \frac{\partial^2}{\partial x^2} p_t(x) \\ p_0(x) = \hat{p}(x) \end{cases}$$

has unique solution, i.e. $\Leftrightarrow$ corresponds to evolution of density of SDE

$$\begin{cases} dX_t = f(X_t, t)dt + g(t)dW_t \\ X_0 \sim \hat{p} \end{cases}$$

SDE 2 ODE

$$dX_t = f(X_t, t)\,dt + g(t)\,dW_t$$

$$\downarrow$$

$$\frac{\partial}{\partial t} p_t(x) = -\frac{\partial}{\partial x}\left(p_t(x) f(x,t)\right) + \frac{g^2(t)}{2}\frac{\partial^2}{\partial x^2} p_t(x)$$

$$= -\frac{\partial}{\partial x}\left(p_t(x) f(x,t) - \frac{g^2(t)}{2}\frac{\partial}{\partial x} p_t(x)\right)$$

$$= -\frac{\partial}{\partial x}\left(p_t(x)\left[f(x,t) - \frac{g^2(t)}{2}\frac{\partial}{\partial x}\log p_t(x)\right]\right)$$

$$\uparrow$$

$$dY_t = \left(f(Y_t, t) - \frac{g^2(t)}{2}\frac{\partial}{\partial x}\log p_t(Y_t)\right)dt$$

— corresponding forward ODE.

ODE 2 b-ODE

$$d Y_t = v(Y_t, t)dt, \quad v(x,t) = f(x,t) - \frac{g^2(t)}{2} \frac{\partial}{\partial x} \log p_t(x)$$

$$\downarrow$$

$$d Y_t^{(b)} = -v(Y_t^{(b)}, 1-t)dt =$$

$$= \left(-f(Y_t^{(b)}, 1-t) + \frac{g^2(t)}{2} \frac{\partial}{\partial x} \log p_{1-t}(Y_t^{(b)}) \right) dt$$

– backward ODE

b-ODE 2 b-SDE

$$dY_t^{(b)} = -v(Y_t^{(b)}, 1-t)\,dt \qquad \left(q_t := p_{Y_t^{(b)}} = p_{Y_{1-t}} = p_{1-t}\right)$$

$$\downarrow$$

$$\frac{\partial}{\partial t} q_t(x) = -\frac{\partial}{\partial x}\left(q_t(x)\left(-v(x, 1-t)\right)\right)$$

$$= -\frac{\partial}{\partial x}\left(q_t(x)\left(-v(x,1-t) + \frac{g^2(1-t)}{2}\frac{\partial}{\partial x}\log q_t(x)\right)\right) + \frac{g^2(1-t)}{2}\frac{\partial^2}{\partial x^2} q_t(x)$$

$$\uparrow$$

$$dX_t^{(b)} = \left(-v\left(X_t^{(b)}, 1-t\right) + \frac{g^2(1-t)}{2}\frac{\partial}{\partial x}\log p_{1-t}\left(X_t^{(b)}\right)\right)dt + g(1-t)\,dW_t$$

b-ODE 2 b-SDE

$$dX_t^{(b)} = \left(-v\left(X_t^{(b)}, 1-t\right) + \frac{g^2(1-t)}{2} \frac{\partial}{\partial x} \log p_{1-t}\left(X_t^{(b)}\right) \right) dt + g(1-t)\, dW_t$$

Recall $v(x,t) = f(x,t) - \frac{g^2(t)}{2} \frac{\partial}{\partial x} \log p_t(x)$

$$dX_t^{(b)} = \left(-f\left(X_t^{(b)}, 1-t\right) + g^2(1-t) \frac{\partial}{\partial x} \log p_{1-t}\left(X_t^{(b)}\right) \right) dt + g(1-t)\, dW_t$$

Summary

$$dX_t = f(X_t, t)\,dt + g(t)\,dW_t$$

$$\downarrow$$

$$dY_t = \left(f(Y_t, t) - \frac{g^2(t)}{2} \frac{\partial}{\partial x} \log p_t(Y_t)\right) dt$$

$$\downarrow$$

$$dY_t^{(b)} = \left(-f(Y_t^{(b)}, 1-t) + \frac{g^2(1-t)}{2} \frac{\partial}{\partial x} \log p_{1-t}(Y_t^{(b)})\right) dt$$

$$\downarrow$$

$$dX_t^{(b)} = \left(-f(X_t^{(b)}, 1-t) + g^2(1-t) \frac{\partial}{\partial x} \log p_{1-t}(X_t^{(b)})\right) dt + g(1-t)\,dW_t$$

More precise result

$$dX_t = f(X_t, t)\,dt + g(t)\,dW_t$$

$$\downarrow$$

$$dX_t^{(b)} = \left(-f\left(X_t^{(b)}, 1-t\right) + g^2(1-t)\frac{\partial}{\partial x}\log p_{1-t}\left(X_t^{(b)}\right)\right)dt + g(1-t)\,dW_t$$

Thm (Anderson)

Under reg. conditions, if $X_0^{(b)} \sim p_{X_1^{(b)}}$, then for all measurable $B \in \mathcal{GL}\left(C[0,1]\right)$

$$P\left(\left(X_{1-t}, t\in[0,1]\right) \in B\right) = P\left(\left(X_t^{(b)}, t\in[0,1]\right) \in B\right).$$

VE-SDE

$$dX_t = g(t)\,dW_t$$

$$dY_t = -\frac{g^2(t)}{2}\frac{\partial}{\partial x}\log p_t(Y_t)\,dt$$

$$dY_t^{(b)} = \frac{g^2(1-t)}{2}\frac{\partial}{\partial x}\log p_{1-t}(Y_t^{(b)})\,dt$$

$$dX_t^{(b)} = g^2(1-t)\frac{\partial}{\partial x}\log p_{1-t}(Y_t^{(b)})\,dt + g(1-t)\,dW_t$$

Sampling

Both b-ODE $Y_t^{(b)}$ and b-SDE $X_t^{(b)}$ generate the reversed dynamics p_{1-t}, given $X_0^{(b)} \sim p_1$. ($Y_0^{(b)} \sim p_1$)

$$dY_t^{(b)} = \left(-f(Y_t^{(b)}, 1-t) + \frac{g^2(t)}{2} \frac{\partial}{\partial x} \log p_{1-t}(Y_t^{(b)}) \right) dt$$

$$dX_t^{(b)} = \left(-f(X_t^{(b)}, 1-t) + g^2(1-t) \frac{\partial}{\partial x} \log p_{1-t}(X_t^{(b)}) \right) dt + g(1-t)\, dW_t$$

Need to know $\frac{\partial}{\partial x} \log p_t(x)$ ← learn by neural net.

Summary

- Fokker-Planck equation

$$\frac{\partial}{\partial t} p_t(x) = -\frac{\partial}{\partial x}\left(p_t(x) f(x,t)\right) + \frac{g^2(t)}{2} \frac{\partial^2}{\partial x^2} p_t(x)$$

- SDE $\rightarrow$ ODE $\xrightarrow{v \to -v}$ ODE $\rightarrow$ SDE

  FPE $\rightarrow$ CE

  CE $\rightarrow$ FPE

- $$dX_t^{(b)} = \left(-f\left(X_t^{(b)}, 1-t\right) + g^2(1-t) \frac{\partial}{\partial x} \log p_{1-t}\left(X_t^{(b)}\right)\right) dt + g(1-t)\, dW_t$$

Links

http://wiki.cs.hse.ru/Уравнения_с_частными_производными_(2022-2023)

https://arxiv.org/abs/2011.13456 (SDE diffusion models)

Continuous-time diffusion models

Score function

$$p_t(x) = \int p_{t|0}(x|x_0)\, p_0(x_0)\, dx_0$$

$p_{t|0}(x|x_0)$: easy — $p_0(x_0)$: difficult ($= p_{data}$)

(VE-SDE: $p_{t|0}(x|x_0) = \mathcal{N}(x|x_0, \sigma_t^2 I)$)

$$\frac{\partial}{\partial x}\log p_t(x) = \frac{\partial/\partial x\, p_t(x)}{p_t(x)} = \frac{\partial/\partial x \int p_{t|0}(x|x_0)\, p_0(x_0)\, dx_0}{p_t(x)} =$$

$$= \int \left(\frac{\partial}{\partial x} p_{t|0}(x|x_0)\right) \frac{p_0(x_0)}{p_t(x)}\, dx_0 =$$

$$= \int \left(\frac{\partial}{\partial x} \log p_{t|0}(x|x_0)\right) \boxed{\frac{p_0(x_0)\, p_t(x|x_0)}{p_t(x)}}\, dx_0$$

$\leftarrow p_{0|t}(x_0|x)$

Score function

$$\frac{\partial}{\partial x}\log p_t(x) = \mathbb{E}\left[\frac{\partial}{\partial x_t}\log p_{t|0}(X_t|X_0)\,\middle|\,X_t = x\right]$$

easy ↓

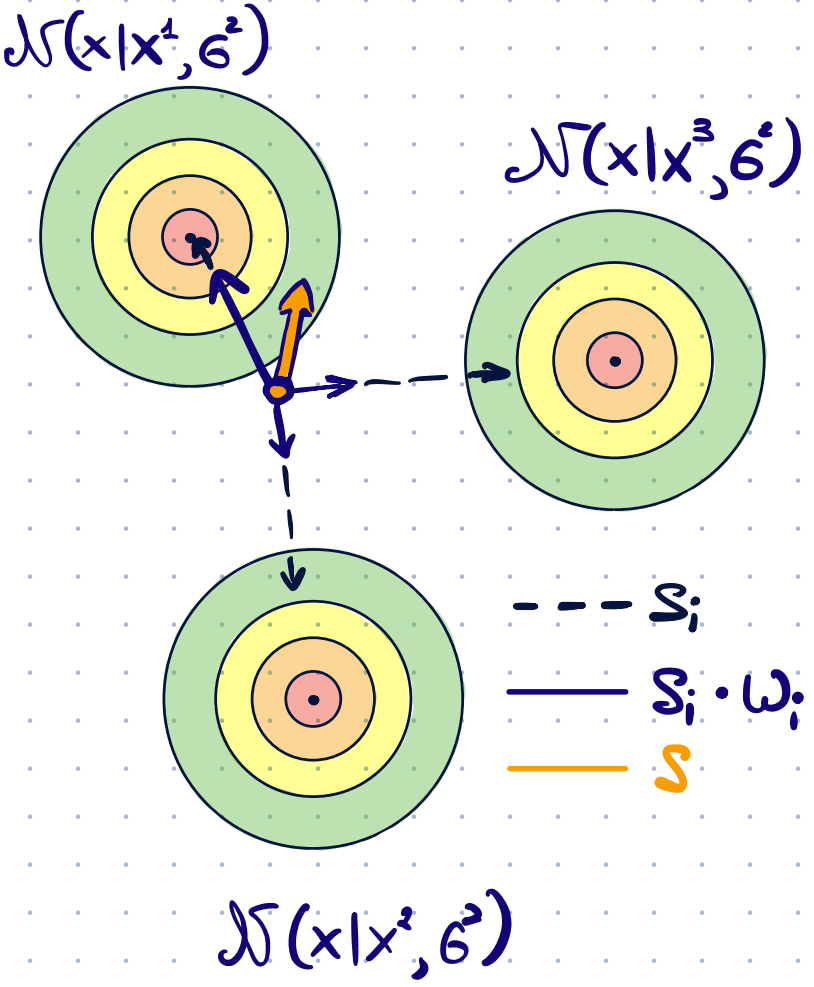
Interpretation:

Let $p_0(x) = \sum_{i=1}^{m} \omega_i \, \mathbb{I}\{x = x^i\}$

$$p_t(x) = \sum_{i=1}^{m} \omega_i \, \mathcal{N}(x|x^i, \sigma_t^2 I)$$

$$\underbrace{\frac{\partial}{\partial x}\log p_t(x)}_{s} =$$

$$= \sum_{i=1}^{m} \underbrace{\frac{\partial}{\partial x}\log p_t(x|x^i)}_{s_i} \cdot \underbrace{p_{0|t}(x^i|x)}_{\omega_i}$$



Learning score function

$$\frac{\partial}{\partial x}\log p_t(x) = \mathbb{E}\left[\frac{\partial}{\partial x_t}\log p_{t|0}(X_t \mid X_0) \,\middle|\, X_t = x\right]$$

Thm $g^*(x) = \mathbb{E}[Y \mid X = x] = \arg\min_g \mathbb{E}\| g(X) - Y \|^2$

In particular,

$$\mathbb{E}\| g(x) - \mathbb{E}[Y|X] \|^2 = \mathbb{E}\| g(x) - Y \|^2 + \text{const}$$

For score:

$$\mathbb{E}\| s_\theta(X_t, t) - \frac{\partial}{\partial x}\log p_t(X_t) \|^2 \to \min_\theta \iff$$

$$\iff \mathbb{E}\| s_\theta(X_t, t) - \frac{\partial}{\partial x}\log p_{t|0}(X_t \mid X_0) \|^2 \to \min_\theta$$

L2 projection proof

$$\mathbb{E}\|g(X) - \mathbb{E}[Y|X]\|^2 = \mathbb{E}\|g(X)\|^2 - 2\mathbb{E}\langle g(X), \mathbb{E}[Y|X]\rangle + \text{const}$$

$$\mathbb{E}\langle g(X), \mathbb{E}[Y|X]\rangle = \int \langle g(x), \int y \cdot p_{Y|X}(y|x)\,dy \rangle\, p_X(x)\,dx$$

$$= \iint \langle g(x), y\rangle\, p_{Y|X}(y|x)\, p_X(x)\,dx\,dy = \mathbb{E}\langle g(X), Y\rangle.$$

$$\mathbb{E}\|g(X) - \mathbb{E}[Y|X]\|^2 = \mathbb{E}\|g(X)\|^2 - 2\mathbb{E}\langle g(X), Y\rangle + \text{const}$$

$$= \mathbb{E}\|g(X) - Y\|^2 + \text{const}$$

□

Denoising Score Matching

Loss: $x_0 \sim p_{data}$, $\varepsilon \sim \mathcal{N}(0, I)$, $x_t = x_0 + \sigma_t \cdot \varepsilon$

$$\int_0^1 \mathbb{E} \| s_\theta(x_t, t) - \frac{\partial}{\partial x} \log p_t(x_t | x_0) \|^2 dt \to \min_\theta$$

- $p_{t|0}(x_t | x_0) = \mathcal{N}(x_t | x_0, \sigma_t^2 I) = C \cdot \exp\left(-\frac{1}{2\sigma_t^2} \|x_t - x_0\|^2\right)$

$$\frac{\partial}{\partial x_t} \log p_{t|0}(x_t | x_0) = -\frac{1}{\sigma_t^2}(x_t - x_0) = \frac{x_0}{\sigma_t^2} - \frac{x_t}{\sigma_t^2}$$

- Parameterize $s_\theta(x_t, t) = \frac{D_\theta(x_t, t)}{\sigma_t^2} - \frac{x_t}{\sigma_t^2}$

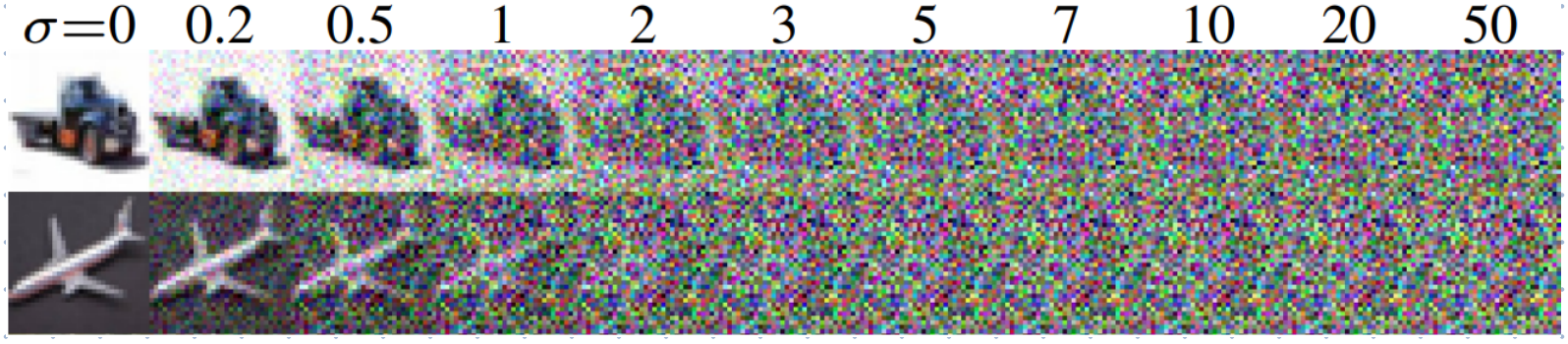
Loss: $\int_0^1 \omega_t \mathbb{E} \| D_\theta(x_t, t) - x_0 \|^2 dt \to \min_\theta$

↖ denoiser

Ideal denoiser

Theoretical optimum: $D^*(x_t, t) = \mathbb{E}[x_0 \mid x_t]$,

i.e. for $t=1$ $D^*(x_1, 1) = \mathbb{E}[x_0 \mid x_1] \approx \mathbb{E}x_0$.



$x_6 = x_0 + 6\varepsilon$

$D^*(x_6, 6)$

Sampling

$$dX_t^{(b)} = g^2(1-t)\, s_\theta\left(X_t^{(b)}, 1-t\right)dt + g(1-t)\, d\omega_t$$

$$X_T \sim \mathcal{N}\left(0, \sigma_T^2\right)$$

$$X_{t-h} \leftarrow X_t + h\, g^2(t)\, s_\theta(X_t, t) + \sqrt{h}\, g(t)\, \varepsilon_t$$

$$= X_t + h\, g^2(t)\, \frac{D_\theta(X_t, t) - X_t}{\sigma_t^2} + \sqrt{h}\, g(t)\, \varepsilon_t$$

$$= X_t\left(1 - \frac{h\, g^2(t)}{\sigma_t^2}\right) + D_\theta(X_t, t) \cdot \frac{h\, g^2(t)}{\sigma_t^2} + \sqrt{h}\, g(t)\, \varepsilon_t$$

Sampling

Remember this guy?

Sampling:

- Claim: $P_{t-1|t,0}(x_{t-1}|x_t, \hat{x}_0) = \mathcal{N}(x_{t-1}|\mu_t, \Sigma_t)$

$$\mu_t = x_t \cdot \left(1 - \frac{g^2(t)}{\sigma_t^2}\right) + \hat{x}_0 \cdot \frac{g^2(t)}{\sigma_t^2}, \quad \Sigma_t = \boxed{\frac{\sigma_{t-1}^2}{\sigma_t^2}} g^2(t) \cdot I$$

- Sample $x_T \sim \mathcal{N}(0, \sigma_T^2)$
- Iterate

$$x_{t-1} = x_t\left(1 - \frac{g^2(t)}{\sigma_t^2}\right) + D_\theta(x_t, t) \cdot \frac{g^2(t)}{\sigma_t^2} + g(t)\varepsilon_t$$

This is him now:

$$x_{t-h} = x_t\left(1 - \frac{h g^2(t)}{\sigma_t^2}\right) + D_\theta(x_t, t) \cdot \frac{h g^2(t)}{\sigma_t^2} + \sqrt{h}\, g(t)\varepsilon_t$$

ODE sampling

$$dY_t^{(b)} = \frac{g^2(1-t)}{2} S_\theta(Y_t^{(b)}, 1-t)\,dt$$

$$Y_{t-h} \leftarrow Y_t + h \cdot \frac{g^2(t)}{2} S_\theta(Y_t, t) =$$

$$= Y_t + h\,\frac{g^2(t)}{2}\,\frac{D_\theta(Y_t,t) - Y_t}{\sigma_t^2} =$$

$$= Y_t\left(1 - h\,\frac{g^2(t)}{2\sigma_t^2}\right) + h\,\frac{g^2(t)}{2\sigma_t^2} D_\theta(Y_t,t)$$

Example: $\sigma_t^2 = t^2 \quad g(t) = \sqrt{2t}$

$$Y_{t-h} \leftarrow Y_t\left(1 - \frac{h}{t}\right) + \frac{h}{t} \cdot D_\theta(Y_t,t)$$

Summary

- $\frac{\partial}{\partial x} \log p_t(x) = \mathbb{E}\left[\frac{\partial}{\partial x} \log p_{t|0}(X_t|X_0) \mid X_t = x\right]$
- $\int_0^t \omega_t \mathbb{E} \| S_\theta(X_t, t) - \frac{\partial}{\partial x} \log p_{t|0}(X_t|X_0) \|^2 dt \to \min_\theta$

<=>

$\int_0^t \omega_t \mathbb{E} \| D_\theta(X_t, t) - X_0 \|^2 dt \to \min_\theta$

- SDE sampling <=> $\hat{X}_0 = D(X_t, t)$; $X_{t-1} \sim p_{t-1|t,0}$
- ODE sampling : simple, interpretable

Links

https://arxiv.org/abs/1907.05600 (score matching)

https://arxiv.org/abs/2011.13456 (SDE diffusion models)

https://arxiv.org/abs/2206.00364 (Careful sampling design)