

CS886: Theory and Applications of Generative Models

Lec 02: Diffusion

Yaoliang Yu



UNIVERSITY OF
WATERLOO

FACULTY OF MATHEMATICS
**DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE**

September 21, 2026

Notation for divergence

$$\langle \nabla, \mathbf{b} \rangle := \sum_i \nabla_i b_i, \quad \text{in particular} \quad \langle \nabla, \nabla p \rangle = \sum_i \nabla_i (\nabla_i p) = \sum_i \nabla_i^2 p =: \Delta p.$$

- $\mathbf{b} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and $p : \mathbb{R}^d \rightarrow \mathbb{R}$
- Not to confuse the scalar $\langle \nabla, \nabla p \rangle$ with the matrix $\nabla^2 p$

$$\text{// } \int b(x) \cdot d p(x) = b(x) p(x) \Big| - \int p(x) d b(x)$$

$$\boxed{\int \langle \nabla p(\mathbf{x}), \mathbf{b}(\mathbf{x}) \rangle d\mathbf{x} = - \int p(\mathbf{x}) \langle \nabla, \mathbf{b}(\mathbf{x}) \rangle d\mathbf{x}}$$

- Assuming each $p b_i$ vanishes at the boundary
- “Just” pull the scale-valued function p out of the inner product and negate the sign

$$\langle \nabla^2, A \rangle = \sum_{i,j} \nabla_{ij}^2 A_{ij}, \quad \text{in particular} \quad \langle \nabla^2, pI \rangle = \Delta p.$$

- $A : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ and $p : \mathbb{R}^d \rightarrow \mathbb{R}$
- Again, not to confuse the scalar $\langle \nabla^2, A \rangle$ with the tensor $\nabla^2 A$
- We omit arguments of functions whenever no confusion will arise
 - for instance, we will rewrite integration by parts simply as

$$\int \langle \nabla p, \mathbf{b} \rangle = - \int p \langle \nabla, \mathbf{b} \rangle \iff \int \langle \nabla p + p \nabla, \mathbf{b} \rangle = \int \langle \nabla, p \mathbf{b} \rangle = 0$$

- it should be clear that p and $\mathbf{b}$ are functions, whose arguments are being integrated over

Neural Ordinary Differential Equations (ODE)

$$\underline{y = T(x)}$$

$$\mathbf{x}_{t+1} \approx \mathbf{x}_t + \eta_t \cdot \mathbf{v}_t(\mathbf{x}_t) =: \mathbf{T}_t(\mathbf{x}_t)$$

$$d\mathbf{x}_t = \mathbf{v}_t(\mathbf{x}_t) dt$$

$$p(y) dy = p(x) dx$$

$$p(y) = p(x) \cdot \frac{1}{\left(\frac{dy}{dx}\right)}$$

$$= p(x) \cdot \frac{1}{T'(x)}$$

$\nwarrow$ $\tilde{p}(y)$ $\searrow$ $\tilde{T}$

- Suppose $\mathbf{x}_t \sim p_t$
- Apply the change-of-variable formula we know $\mathbf{x}_{t+1} \sim p_{t+1}$, where

$$\begin{aligned} \log p_{t+1}(\mathbf{x}_{t+1}) &= \log p_t(\mathbf{x}_t) - \log |\det \nabla_{\mathbf{x}} \mathbf{T}_t(\mathbf{x}_t)| \\ &= \log p_t(\mathbf{x}_t) - \log |\det [\text{Id} + \eta_t \cdot \nabla_{\mathbf{x}} \mathbf{v}_t(\mathbf{x}_t)]| \\ &\approx \log p_t(\mathbf{x}_t) - \eta_t \cdot \langle \nabla_{\mathbf{x}}, \mathbf{v}_t(\mathbf{x}_t) \rangle \end{aligned}$$

- Continuous change-of-variable formula:

$$\log(1+x) \approx x$$

$$\boxed{\frac{d \log p_t(\mathbf{x}_t)}{dt} = - \langle \nabla_{\mathbf{x}}, \mathbf{v}_t(\mathbf{x}_t) \rangle}$$

$$\frac{d \log p_t(\mathbf{x}_t)}{dt} = \frac{1}{p_t} [\partial_t p_t + \langle \nabla_{\mathbf{x}} p_t; \mathbf{v}_t \rangle] = \frac{1}{p_t} [-\langle \nabla_{\mathbf{x}}; p_t \mathbf{v}_t \rangle + \langle \nabla_{\mathbf{x}} p_t; \mathbf{v}_t \rangle] = -\langle \nabla_{\mathbf{x}}; \mathbf{v}_t(\mathbf{x}_t) \rangle$$

- Can compute the log-likelihood along the trajectory of $\mathbf{x}_t$:

$$\begin{cases} \dot{\mathbf{x}}_t &= \mathbf{v}_t(\mathbf{x}_t, \mathbf{u}) \\ \dot{l}_t &= -\langle \nabla_{\mathbf{x}}; \mathbf{v}_t(\mathbf{x}_t, \mathbf{u}) \rangle \end{cases}, \text{ with initial condition } \mathbf{x}_0 \text{ and } l_0. \quad (1)$$

- More importantly, can solve (1) in reverse, from end condition $\mathbf{x}_1$ and l_1 , so that

$$\boxed{\log p_1(\mathbf{x}_1) - \log p_0(\mathbf{x}_0) = l_1 - l_0}$$

MLE for Neural ODE

$$\dot{X}_t = \mathbf{v}_t(X_t; \mathbf{w}), \quad X_0 \sim p_{\text{noise}}, \quad \text{e.g., } \mathcal{N}(\mathbf{0}, I)$$

- Learn the vector field $\mathbf{v}_t(\cdot; \mathbf{w})$ so that $X_1 \sim p_{\text{data}}$

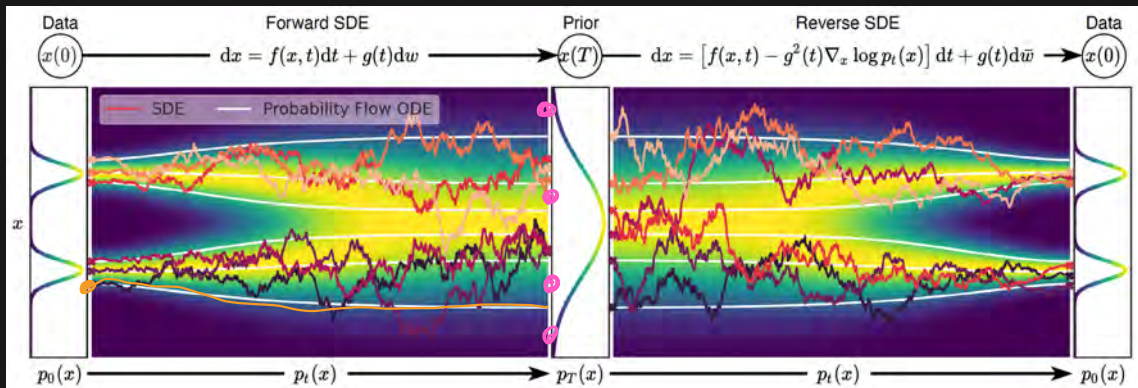
Representation

Can every p_{data} be obtained through some vector field $\mathbf{v}_t$?

- Maximum likelihood estimation:

$$\min \text{KL}(p_{\text{data}} \| p_1) = \text{KL}(q_0 \| p_{\text{noise}})$$

- p_1 is obtained by running ODE forward with initialization p_{noise}
- q_0 is obtained by running ODE backward with initialization p_{data}



Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.

- $\text{KL}(p_{\text{data}} \| p_1) \equiv -\mathbb{E}_{p_{\text{data}}} \log p_1(\mathbf{X})$
 - have a sample from p_{data}
 - can compute $\log p_1(\mathbf{X}_1)$ using change-of-variable: $\frac{d \log p_t(\mathbf{x}_t)}{dt} = -\langle \nabla_{\mathbf{x}}, \mathbf{v}_t(\mathbf{x}_t) \rangle$
 - but $\mathbf{X}_1 \sim p_1 \neq p_{\text{data}}$!
- $\text{KL}(q_0 \| p_{\text{noise}}) \equiv -\mathbb{E}_{q_0} [\log p_{\text{noise}}(\mathbf{X}_0) - \log q_0(\mathbf{X}_0)]$
 - can sample from q_0 : $\dot{\mathbf{X}}_t = \mathbf{v}_t(\mathbf{X}_t; \mathbf{w})$, $\mathbf{X}_1 \sim p_{\text{data}}$
 - can compute $\log q_0(\mathbf{X}_0)$ using change-of-variable: $\frac{d \log p_t(\mathbf{x}_t)}{dt} = -\langle \nabla_{\mathbf{x}}, \mathbf{v}_t(\mathbf{x}_t) \rangle$
 - of course, can compute $\log p_{\text{noise}}(\mathbf{X}_0)$

A Little Bit of Math

Markov Process

$$X_t = X_0 + \int_0^t \mathbf{b}_s(X_s) ds + \int_0^t G_s(X_s) \underline{dB_s}$$
$$dX_t = \underbrace{\mathbf{b}_t(X_t)}_{dt} + G_t(X_t) dB_t, \text{ given } X_0$$

For any bounded and continuous function f :

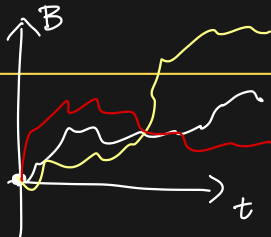
$$\forall t \geq s, \quad \mathbb{E}^{\mathbf{x}}[f(X_t) | \mathcal{F}_s] = \mathbb{E}^{\mathbf{x}}[f(X_t) | X_s]$$

- $\mathbb{E}^{\mathbf{x}}$: when we start with $X_0 = \mathbf{x}$
- Given the past, future only depends on the present

Brownian Motion

A stochastic process $\{B_t : t \geq 0\}$ is called **Brownian motion** if

- Initialization: $B_0 \equiv 0$
- Independent increment: $\forall n, \forall t_0 \leq t_1 \leq \dots \leq t_n, B_{t_1} - B_{t_0} \perp \dots \perp B_{t_n} - B_{t_{n-1}}$
- Stationary increment: $\forall s \leq t, B_t - B_s \simeq B_{t-s} - B_0$
- Gaussian: $B_t \simeq \mathcal{N}(0, t) = \sqrt{t} \cdot \mathcal{N}(0, 1)$
- Continuous sample path: for (almost) all ω , $t \mapsto B_t(\omega)$ is continuous



Brownian motion is a (continuous) Gaussian process with covariance kernel

$$\kappa(s, t) := \mathbb{E}(B_s B_t) = s \wedge t$$

Transition and Semigroup

$k(x, A)$



- Let $p(s, \mathbf{x}, t, d\mathbf{y})$ be a Markov transition kernel, satisfying Chapman-Kolmogorov:

$$\forall s \leq \tau \leq t, \quad p(s, \mathbf{x}, t, d\mathbf{y}) = \int_{\mathbf{z}} p(s, \mathbf{x}, \tau, d\mathbf{z}) p(\tau, \mathbf{z}, t, d\mathbf{y})$$

$p(t, y | s, x) = \int p(t, y | \tau, z) p(\tau, z | s, x) dz$

- Two bounded (in fact, contractive) and **positive** linear operators:

$$\mathcal{P}_{s,t} : b\Sigma \rightarrow b\Sigma, \quad \mathcal{P}_{s,t} f(\mathbf{x}) = (\mathcal{P}_{s,t} f)(\mathbf{x}) := \int p(s, \mathbf{x}, t, d\mathbf{y}) f(\mathbf{y})$$

$$\mathcal{P}_{s,t}^* : \mathcal{M}(\Sigma) \rightarrow \mathcal{M}(\Sigma), \quad \mu \mathcal{P}_{s,t}^*(d\mathbf{y}) = (\mu \mathcal{P}_{s,t}^*)(d\mathbf{y}) := \int p(s, \mathbf{x}, t, d\mathbf{y}) \mu(d\mathbf{x})$$

$= \mathbb{E}[f(X_t) | X_s = x]$

- From Chapman-Kolmogorov we derive a semigroup:

$$\mathcal{P}_{s,t} = \mathcal{P}_{s,\tau} \mathcal{P}_{\tau,t}, \quad \mathcal{P}_{t,t} = \text{Id},$$

$$\mathcal{P}_{s,t} \mathbf{1} = \mathbf{1}. \quad \int p(s, x, t, y) dy$$

Adjoint and Generator

$$\langle f; \mu \rangle := \int f(\mathbf{x}) \mu(d\mathbf{x}) = \mathbb{E}f(\mathbf{X}), \quad \text{where } \mathbf{X} \sim \mu$$

- Adjoint, as “transpose”:

$$\langle \mathcal{P}_{s,t} f; \mu \rangle = \langle f; \mu \mathcal{P}_{s,t}^* \rangle$$

- Generator, as infinitesimal change:

$$\mathcal{L}_t := \lim_{h \downarrow 0} \frac{\mathcal{P}_{t,t+h} - \text{Id}}{h}, \quad \mathcal{L}_t^* := \lim_{h \downarrow 0} \frac{\mathcal{P}_{t,t+h}^* - \text{Id}}{h}. \quad [\text{recall } \mathcal{P}_{t,t} = \text{Id}]$$

- Formally, $\frac{d\mathcal{P}_{s,t}}{dt} = \mathcal{P}_{s,t} \mathcal{L}_t$ and $\frac{d\mathcal{P}_{s,t}^*}{dt} = \mathcal{P}_{s,t}^* \mathcal{L}_t^*$

$$\frac{\mathcal{P}_{s,t+h} - \mathcal{P}_{s,t}}{h} = \frac{\mathcal{P}_{s,t} \cdot \mathcal{P}_{t,t+h} - \mathcal{P}_{s,t}}{h} = \mathcal{P}_{s,t} \cdot \frac{\mathcal{P}_{t,t+h} - \text{Id}}{h} = \mathcal{P}_{s,t} \cdot \mathcal{L}_t$$

Diffusion Process

Under mild conditions, a Markov process has generator

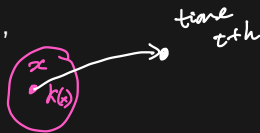
$$\mathcal{L}_t f(\mathbf{x}) = \frac{1}{2} \langle A(t, \mathbf{x}), \nabla^2 f(\mathbf{x}) \rangle + \langle \mathbf{b}(t, \mathbf{x}), \nabla f(\mathbf{x}) \rangle, \quad \text{where } A(t, \mathbf{x}) \succeq 0$$

iff the following holds for all t and $\mathbf{x}$: for all compact nhood $K(\mathbf{x})$,

$$\lim_{h \downarrow 0} \frac{1}{h} \int_{\mathbb{X} \setminus K(\mathbf{x})} p(t, \mathbf{x}, t+h, d\mathbf{y}) = 0,$$

$$\lim_{h \downarrow 0} \frac{1}{h} \int_{K(\mathbf{x})} (\mathbf{y} - \mathbf{x}) \cdot p(t, \mathbf{x}, t+h, d\mathbf{y}) = \mathbf{b}(t, \mathbf{x}),$$

$$\lim_{h \downarrow 0} \frac{1}{h} \int_{K(\mathbf{x})} (\mathbf{y} - \mathbf{x}) \otimes (\mathbf{y} - \mathbf{x}) \cdot p(t, \mathbf{x}, t+h, d\mathbf{y}) = A(t, \mathbf{x}).$$



$$p\mathcal{L}_t^* = \frac{1}{2} \langle \nabla^2, pA \rangle - \langle \nabla, p\mathbf{b} \rangle$$

Tracking the Mean and Covariance

 $\mathbf{x}_t$
 $\mathbb{E}\mathbf{x}_t$

$$\lim_{h \downarrow 0} \frac{\mathbb{E}[\mathbf{x}_{t+h} | \mathbf{x}_t] - \mathbf{x}_t}{h} = \mathbf{b}(t, \mathbf{x}_t), \quad \lim_{h \downarrow 0} \frac{\mathbb{E}[(\mathbf{x}_{t+h} - \mathbf{x}_t) \otimes (\mathbf{x}_{t+h} - \mathbf{x}_t) | \mathbf{x}_t]}{h} = A(t, \mathbf{x}_t).$$

 $\mathbf{b}_t(\mathbf{x}_t) = \mathbf{x}_t$

$$\underline{d\mathbf{m}_t} := d\mathbb{E}[\mathbf{x}_t] = \lim_{h \downarrow 0} \frac{\mathbb{E}[\mathbf{x}_{t+h} - \mathbf{x}_t]}{h} = \mathbb{E} \left[\lim_{h \downarrow 0} \frac{\mathbb{E}[\mathbf{x}_{t+h} - \mathbf{x}_t | \mathbf{x}_t]}{h} \right] = \mathbb{E}[\mathbf{b}_t(\mathbf{x}_t)]$$

 $= \mathbb{E}\mathbf{x}_t = \underline{m_t}$

$$d\Sigma_t := d\mathbb{E}[(\mathbf{x}_t - \mathbf{m}_t) \otimes (\mathbf{x}_t - \mathbf{m}_t)]$$

$$= \lim_{h \downarrow 0} \frac{\mathbb{E}[(\mathbf{x}_{t+h} - \mathbf{m}_{t+h}) \otimes (\mathbf{x}_{t+h} - \mathbf{m}_{t+h})] - \mathbb{E}[(\mathbf{x}_t - \mathbf{m}_t) \otimes (\mathbf{x}_t - \mathbf{m}_t)]}{h},$$

 $C \cdot \text{explicit}$

$$= \lim_{h \downarrow 0} \frac{\mathbb{E}[(\mathbf{x}_{t+h} - \mathbf{x}_t) \otimes (\mathbf{x}_{t+h} - \mathbf{x}_t)] + \mathbb{E}[(\mathbf{x}_t - \mathbf{m}_t) \otimes (\mathbf{x}_{t+h} - \mathbf{x}_t)] + \mathbb{E}[(\mathbf{x}_{t+h} - \mathbf{x}_t) \otimes (\mathbf{x}_t - \mathbf{m}_t)]}{h}$$

$$= \mathbb{E}[A_t(\mathbf{x}_t) + (\mathbf{x}_t - \mathbf{m}_t) \otimes \mathbf{b}_t(\mathbf{x}_t) + \mathbf{b}_t(\mathbf{x}_t) \otimes (\mathbf{x}_t - \mathbf{m}_t)]$$

Ornstein-Uhlenbeck Process: Affine Drift and Diffusion

When the drift $\mathbf{b}_t(\mathbf{X}_t)$ and diffusion $A_t(\mathbf{X}_t)$ are affine in $\mathbf{X}_t$:

$$\begin{aligned}d\mathbf{m}_t &= \mathbf{b}_t(\mathbf{m}_t) =: C_t \mathbf{m}_t + \mathbf{a}_t \\d\Sigma_t &= A_t(\mathbf{m}_t) + \Sigma_t C_t^\top + C_t \Sigma_t.\end{aligned}$$

Introducing the flow $\Phi(t, s)$ (fundamental matrix, solution of the homogeneous ODE):

$$\begin{aligned}\mathbf{m}_t &:= \Phi(t, s) \mathbf{m}_s + \int_s^t \Phi(t, \tau) \mathbf{a}_\tau d\tau \\ \Sigma_t &:= \Phi(t, s) \Sigma_s \Phi^\top(t, s) + \int_s^t \Phi(t, \tau) A_\tau(\mathbf{m}_\tau) \Phi^\top(t, \tau) d\tau.\end{aligned}$$

When $C_t = c_t \cdot C$, we simply have $\Phi(t, s) = \exp(\int_s^t c_\tau d\tau \cdot C)$.

Fokker-Planck-Kolmogorov (FPK) Equation

- Recall that formally, $\frac{d\mathcal{P}_{s,t}}{dt} = \mathcal{P}_{s,t}\mathcal{L}_t$, hence

$$\begin{aligned}\langle f; \partial_t p(s, \mathbf{x}, t, \cdot) \rangle &= \partial_t \langle f; \underbrace{p(s, \mathbf{x}, t, \cdot)}_{\mathcal{P}_{s,t} f} \rangle = \mathcal{P}_{s,t} \mathcal{L}_t f(\mathbf{x}) = \langle \mathcal{L}_t f; p(s, \mathbf{x}, t, \cdot) \rangle \\ &= \langle f; \mathcal{L}_t^* p(s, \mathbf{x}, t, \cdot) \rangle\end{aligned}$$

- For a diffusion process:

$$\partial_t p = \mathcal{L}_t^* p = - \langle \nabla, p \mathbf{b} \rangle + \frac{1}{2} \langle \nabla^2, p A \rangle \quad (\text{forward})$$

$$-\partial_s p = \mathcal{L}_s p = \langle \mathbf{b}, \nabla p \rangle + \frac{1}{2} \langle A, \nabla^2 p \rangle \quad (\text{backward})$$

FPK for SDE

Consider the stochastic differential equation

$$d\mathbf{X}_t = \mathbf{b}_t(\mathbf{X}_t) dt + \cancel{G_t(\mathbf{X}_t)} d\mathbf{B}_t, \quad \mathcal{L}_t f(\mathbf{x}) := \langle \mathbf{b}_t(\mathbf{x}), \nabla f(\mathbf{x}) \rangle + \frac{1}{2} \langle G_t(\mathbf{x}) G_t(\mathbf{x})^\top, \nabla^2 f(\mathbf{x}) \rangle$$

Apply Itô's formula to obtain

$$df(\mathbf{X}_t) = \langle \mathbf{b}_t(\mathbf{X}_t), \nabla f(\mathbf{X}_t) \rangle + \frac{1}{2} \langle G_t(\mathbf{X}_t) G_t^\top(\mathbf{X}_t), \nabla^2 f(\mathbf{X}_t) \rangle dt + \nabla f(\mathbf{X}_t) \cdot G_t(\mathbf{X}_t) d\mathbf{B}_t.$$

Take expectation from s at position $\mathbf{x}$ to t on both sides:

$$\begin{aligned} d\mathbb{E}[f(\mathbf{X}_t)] &= \mathbb{E} \left[\langle \mathbf{b}_t(\mathbf{X}_t), \nabla f(\mathbf{X}_t) \rangle + \frac{1}{2} \langle G_t(\mathbf{X}_t) G_t^\top(\mathbf{X}_t), \nabla^2 f(\mathbf{X}_t) \rangle \right] dt \\ \partial_t \langle f; p(s, \mathbf{x}, t, \mathbf{y}) \rangle &= \langle f; \partial_t p(s, \mathbf{x}, t, \mathbf{y}) \rangle = \langle \mathcal{L}_t f; p(s, \mathbf{x}, t, \mathbf{y}) \rangle = \langle f; \mathcal{L}_t^* p(s, \mathbf{x}, t, \mathbf{y}) \rangle \end{aligned}$$

$$\partial_t p = \mathcal{L}_t^* p = -\langle \nabla, p \mathbf{b} \rangle + \frac{1}{2} \langle \nabla^2, p G G^\top \rangle$$

$$\partial_s p = \mathcal{L}_s p = \langle \mathbf{b}, \nabla p \rangle + \frac{1}{2} \langle G G^\top, \nabla^2 p \rangle$$

- Continuity equation when $G = 0$

$$\frac{1}{2} \langle \nabla^2, p \cdot \sqrt{2} \mathbf{I} \cdot \sqrt{2} \mathbf{I} \rangle$$

$$= \langle \nabla^2, p \mathbf{I} \rangle$$

$$= \begin{bmatrix} \nabla_{ij}^2 \end{bmatrix} \cdot \begin{bmatrix} p & 0 \\ 0 & \dots & 0 \end{bmatrix}$$

Heat Equation

The heat equation

$$= \sum_i \nabla_{x_i}^2 P$$

$$\partial_t p(t, \mathbf{x}) = \Delta_{\mathbf{x}} p(t, \mathbf{x})$$

corresponds to the (trivial) SDE

$$dX_t = \sqrt{2} dB_t.$$

$$\underbrace{\sqrt{2} \cdot I \cdot dB_t}$$

- Note that $G_t = \sqrt{2}I$, not merely $\sqrt{2}$
- With $X_0 = 0$ we have $X_t = \sqrt{2}B_t \sim \mathcal{N}(0, 2t)$
- Since $p(t, \mathbf{x})$ is the density of X_t , the Gaussian $\mathcal{N}(0, 2t)$ solves the heat equation
 - can also be verified directly through tedious calculation

$$\begin{aligned} X_t &= \int_0^t dX_s = \int_0^t \sqrt{2} dB_s \\ &= \sqrt{2} (B_t - B_0) = \sqrt{2} B_t \end{aligned}$$

Itô Process

$$dX(t, \omega) = F(t, \omega) dt + G(t, \omega) dB(t, \omega)$$

$$X(t, \omega) = X(0, \omega) + \int_0^t F(s, \omega) ds + \int_0^t G(s, \omega) dB(s, \omega)$$

- The 1st integral is the familiar Riemann or Lebesgue integral
- The 2nd integral is Itô's stochastic integral

Itô's formula:

$$df(X_t, V_t) = f_x(X_t, V_t) dX_t + f_y(X_t, V_t) dV_t + \frac{1}{2} f_{xx}(X_t, V_t) d[X]_t$$

- $[X]_t := \lim_{n \rightarrow \infty} \sum_{k=0}^n (X_{t_{k+1}} - X_{t_k})^2 = G^\top G \cdot [B]_t, \quad [B]_t = t$

Handwritten notes:

$$\begin{aligned} & \cancel{(dt)^2} \quad \cancel{dt dB_t} \\ & (dB_t)^2 \\ & (dt + dB_t)^2 \\ & (dX_t)^2 \end{aligned}$$

Itô's Stochastic Differential Equation (SDE)

$$dX_t = b(t, X_t) dt + G(t, X_t) dB_t$$

$$X_t = X_0 + \int_0^t b_s(X_s) ds + \int_0^t G_s(X_s) dB_s$$

- $t \mapsto X_t$ is continuous
- We assume b and G are continuous
- The 1st integral is the usual Riemann integral
- The 2nd integral is Itô's stochastic integral

But, does such X_t actually exist?

Geometric Brownian Motion

$$\frac{dS_t}{S_t} = b dt + \sigma dB_t, \text{ or equivalently } \boxed{dS_t = bS_t dt + \sigma S_t dB_t}$$

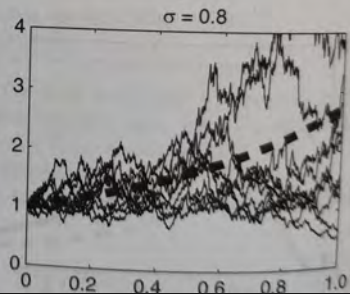
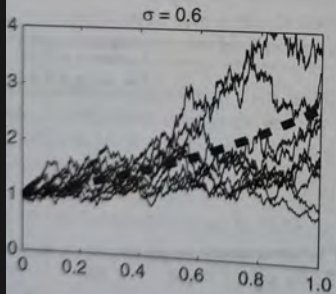
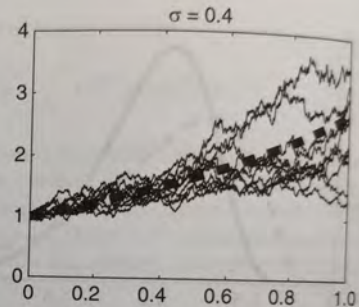
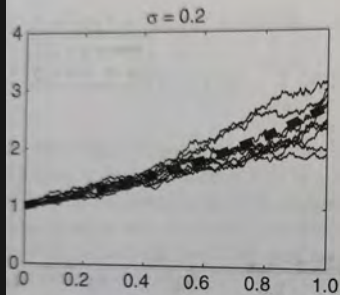
- $b(S_t) = bS_t$
- $G(S_t) = \sigma S_t$
- Apply Itô's formula to $f(S_t) = \ln(S_t)$:

$$d \ln(S_t) = \frac{1}{S_t} dS_t - \frac{1}{2S_t^2} d[S]_t = \frac{1}{S_t} dS_t - \frac{\sigma^2}{2} dt = (b - \frac{\sigma^2}{2}) dt + \sigma dB_t$$

$$\boxed{S_t = S_0 \cdot \exp \left[(b - \frac{\sigma^2}{2})t + \sigma B_t \right]}$$

- Take expectation:

$$d\mathbb{E}[S_t] = b\mathbb{E}[S_t] \cdot dt \implies \mathbb{E}[S_t] = \mathbb{E}[S_0] \cdot \exp(bt)$$



Langevin's Equation

$$dX_t = -bX_t dt + \sigma dB_t$$

- X_t is the velocity of a particle
- b is the friction coefficient
- σ models random perturbation

$$X_t = \exp(-bt) \cdot X_0 + \sigma \int_0^t \exp[-b(t-s)] dB_s$$

$$\mathbb{E}[X_t] = \exp(-bt) \cdot \mathbb{E}[X_0], \quad \mathbb{E}[X_t^2] = \exp(-2bt) \cdot \mathbb{E}[X_0^2] + \frac{\sigma^2}{2b} [1 - \exp(-2bt)]$$

- As $t \rightarrow \infty$, $X_t \rightarrow \mathcal{N}(0, \frac{\sigma^2}{2b})$

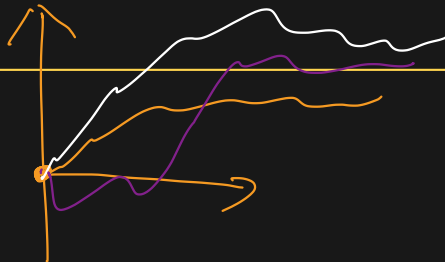
Brownian Motion on a Riemannian Manifold

$$\mathbf{X}(t) = (\cos B_t, \sin B_t)$$

$$\begin{cases} dX_1 &= -\frac{1}{2}X_1 dt - X_2 dB_t \\ dX_2 &= -\frac{1}{2}X_2 dt + X_1 dB_t \end{cases} \iff d\mathbf{X} = -\frac{1}{2}\mathbf{X} dt + J\mathbf{X} dB_t, \quad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$d\mathbb{E}[\mathbf{X}] = -\frac{1}{2}\mathbb{E}[\mathbf{X}] dt \implies \mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{X}(0)] \cdot \exp(-\frac{1}{2}t)$$

Brownian Bridge



$$dX_t = \frac{-X_t}{1-t} dt + dB_t, \quad 0 \leq t < 1, \quad X_0 = 0$$

$$X_t = \int_0^t \frac{1-s}{1-t} dB_s, \quad \mathbb{E}[X_t] \equiv 0$$



Existence and Uniqueness

$$X_t = X_0 + \int_0^t \mathbf{b}_s(X_s) \, ds + \int_0^t G_s(X_s) \, dB_s$$

There exists a **unique continuous** solution of SDE, if

- b and G are **locally** Lipschitz continuous (e.g., continuously differentiable)
 - i.e., $|b(\mathbf{x}) - b(\mathbf{y})| \leq L \cdot |\mathbf{x} - \mathbf{y}|$ for all $\mathbf{x}, \mathbf{y} \in B(0, r)$
- b and G have **linear growth**
 - i.e., $|b(\mathbf{x})|^2 \leq K \cdot (1 + |\mathbf{x}|^2)$
 - trivially holds if b is linear in $\mathbf{x}$

$$\begin{aligned} X_t &= X_s + \int_s^t \mathbf{b}_\tau(X_\tau) d\tau + \int_s^t G_\tau(X_\tau) dB_\tau \\ &\approx X_s + \mathbf{b}_s(X_s) \cdot [t - s] + G_s(X_s)[B_t - B_s] \end{aligned}$$

- Divide $0 := t_0 < t_1 < \dots < t_n < t_{n+1} = t$
- For $k = 1, \dots, n$, compute

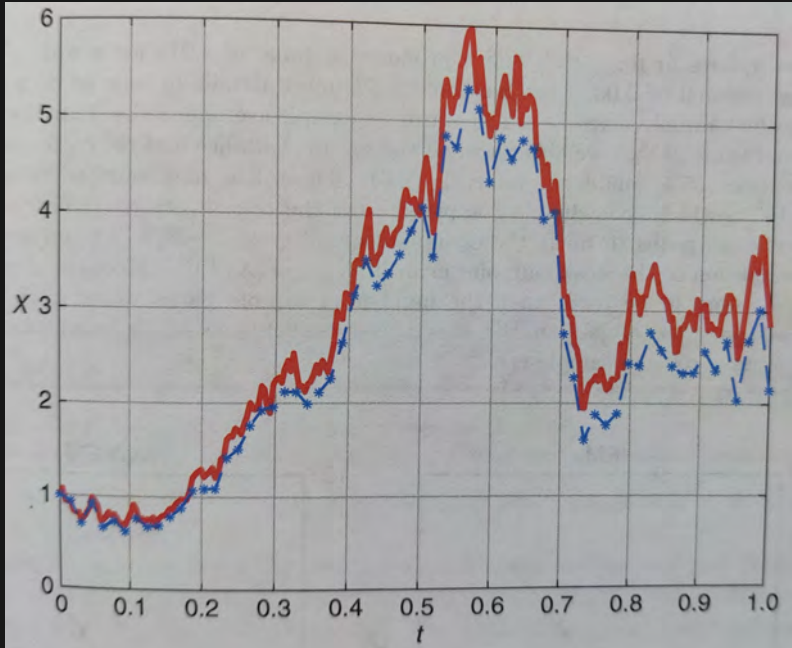
$$X_{t_{k+1}} = X_{t_k} + b_{t_k}(X_{t_k}) \cdot \Delta t_k + G_{t_k}(X_{t_k}) \cdot \Delta B_{t_k}$$

$$- \Delta B_{t_k} \stackrel{i.i.d.}{\cong} \mathcal{N}(0, \Delta t_k)$$

- Interpolate

$$- \hat{X}_t := \sum_k X_{t_k} \cdot \mathbb{I}[t_k \leq t < t_{k+1}]$$

$$- \bar{X}_t := \mathcal{T}(\hat{X}_t)$$



"I did not hear any reaction from other mathematicians about my paper until long after the end of the war, when [Gisiro Maruyama](#) told me in person that he had been drafted into the army and had read my paper in the barracks in 1942. Perhaps he and I were the only researchers then interested in the problem of sample paths."

— [Kiyosi Itô](#)

Langevin again

$$d\mathbf{X}_t = -b(\mathbf{X}_t) dt + \sigma_t dB_t$$

- Euler-Maruyama:

$$\mathbf{X}_{k+1} = \mathbf{X}_k - b(\mathbf{X}_k)\Delta t_k + \mathcal{N}(0, \sigma_k^2 \Delta t_k)$$

- Let $\Delta t_k =: \eta_k$ and $b(\mathbf{X}_k) =: \nabla f(\mathbf{X}_k)$

$$\mathbf{X}_{k+1} = \mathbf{X}_k - \eta_k [\nabla f(\mathbf{X}_k) + \mathcal{N}(0, \sigma_k^2 / \eta_k)]$$

- In typical SGD, $\eta_k = O(1/\sqrt{k})$ and $\sigma_k^2 / \eta_k = O(1)$, meaning $\sigma_k = k^{-1/4}$
- In Langevin gradient, $\sigma_k \equiv 1$

Complexity

It can be shown that the Euler-Muruyama scheme has

- weak convergence of order $O(\Delta t)$
- strong convergence of order $O(\sqrt{\Delta t})$

More generally, for a scheme with convergence order $O(\Delta t)^p$, to achieve ϵ accuracy:

- Need $\Delta t = \epsilon^{1/p}$
- Need to evaluate b and G roughly $1/\Delta t$ times
- For $p = 1/2$, obtain the familiar rate $1/\epsilon^2$ for SGD

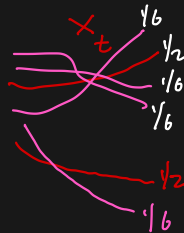
Reverse-time SDE

$$(x_1, x_2, x_3, \dots, x_n) \quad P(x_1, x_2, \dots, x_n)$$

$$(x_n, x_{n-1}, \dots, x_1)$$

$$dX_t = b_t(X_t) dt + G_t(X_t) dB_t \quad (\text{forward-SDE})$$

$$d\bar{X}_t = \bar{b}_t(\bar{X}_t) dt + \bar{G}_t(\bar{X}_t) d\bar{B}_t \quad (\text{reverse-SDE})$$



- FPK to reverse-SDE (negation due to time reversal: $\bar{X}_t = X_{1-t}$):

$$-\partial_s \bar{p}(s, \mathbf{x}, t, \mathbf{y}) = -\langle \nabla, \bar{p} \bar{\mathbf{b}} \rangle + \frac{1}{2} \langle \nabla^2, \bar{p} \bar{A} \rangle, \quad \text{where} \quad \bar{A} := \bar{G} \bar{G}^\top$$

- FPK to forward-SDE for $p(s, \mathbf{x}, t, \mathbf{y})$ and $q(s, \mathbf{x})$:

$$-\partial_s \log p = \frac{\langle \mathbf{b}, \nabla p \rangle + \frac{1}{2} \langle A, \nabla^2 p \rangle}{p}, \quad -\partial_s \log q = \frac{\langle \nabla, q \mathbf{b} \rangle - \frac{1}{2} \langle \nabla^2, q A \rangle}{q}, \quad A := G G^\top$$

Let $r = pq$ (the joint density of X_s and X_t). We add the above two equations:

$$\begin{aligned}
 -\partial_s \log r &= \frac{\langle \mathbf{b}, \nabla p \rangle + \frac{1}{2} \langle A, \nabla^2 p \rangle}{p} + \frac{\langle \nabla, q \mathbf{b} \rangle - \frac{1}{2} \langle \nabla^2, q A \rangle}{q} \\
 &= \langle \mathbf{b}, \nabla \log p \rangle + \frac{1}{2p} \langle A, \nabla^2 p \rangle + \langle \nabla \log q, \mathbf{b} \rangle + \langle \nabla, \mathbf{b} \rangle - \frac{1}{2q} \langle \nabla^2, q A \rangle \\
 &= \frac{1}{r} [\langle \nabla, r \mathbf{b} \rangle + \frac{q}{2} \langle A, \nabla^2 p \rangle - \frac{p}{2} \langle \nabla^2, q A \rangle] \\
 &= \frac{1}{r} [\langle \nabla, r \mathbf{b} \rangle + \frac{1}{2} \langle q A, \nabla^2 p \rangle - \frac{1}{2} \langle p \nabla, \nabla \cdot (q A) \rangle] \\
 &= \frac{1}{r} [\langle \nabla, r \mathbf{b} \rangle + \frac{1}{2} \langle q A, \nabla^2 p \rangle + \frac{1}{2} \langle \nabla p, \nabla \cdot (q A) \rangle - \frac{1}{2} \langle \nabla p, \nabla \cdot (q A) \rangle - \\
 &\quad - \frac{1}{2} \langle p \nabla, \nabla \cdot (q A) \rangle] \\
 &= \frac{1}{r} [\langle \nabla, r \mathbf{b} \rangle + \frac{1}{2} \langle \nabla, (\nabla p) \cdot (q A) \rangle - \frac{1}{2} \langle \nabla, p \nabla \cdot (q A) \rangle] \\
 &= \frac{1}{r} [\langle \nabla, r \mathbf{b} \rangle + \frac{1}{2} \langle \nabla, \nabla \cdot (pq A) \rangle - \langle \nabla, p \nabla \cdot (q A) \rangle] \\
 &= \frac{1}{r} \left[\left\langle \nabla, r \left(\mathbf{b} - \frac{1}{q} \nabla \cdot (q A) \right) \right\rangle + \frac{1}{2} \langle \nabla^2, r A \rangle \right]
 \end{aligned}$$

$$\text{KL}(p \parallel q) \geq \text{KL}(\tau_{\#} p \parallel \tau_{\#} q)$$

$$-\partial_s r = \left\langle \nabla, r(\mathbf{b} - \frac{1}{q} \nabla \cdot (qA)) \right\rangle + \frac{1}{2} \langle \nabla^2, rA \rangle$$

- Dividing both sides by $q(t, \mathbf{y})$ (and noting that ∇ and ∇^2 are w.r.t. $\mathbf{x}$):

$$\begin{aligned} -\partial_s \bar{p}(s, \mathbf{x}, t, \mathbf{y}) &= \left\langle \nabla, \bar{p}(\mathbf{b} - \frac{1}{q} \nabla \cdot (qA)) \right\rangle + \frac{1}{2} \left\langle \nabla^2, \bar{p}A \right\rangle \\ &= \left\langle \nabla, \bar{p}(\mathbf{b} - A\mathbf{s}_q + \frac{1}{2}A\mathbf{s}_{\bar{p}} - \frac{1}{2}\nabla \cdot A) \right\rangle \end{aligned}$$

- Comparing with the FPK for reverse-SDE, we may identify

$$\bar{G}_{1-t} = G_t, \quad \bar{\mathbf{b}}_{1-t} = -\mathbf{b}_t + \frac{1}{q} \nabla \cdot (qG_t G_t^\top), \quad \text{or,}$$

$$\bar{G}_{1-t} = 0, \quad \bar{\mathbf{b}}_{1-t} = -\mathbf{b}_t + G_t G_t^\top \mathbf{s}_q - \frac{1}{2} G_t G_t^\top \mathbf{s}_{\bar{p}} + \frac{1}{2} \nabla \cdot (G_t G_t^\top)$$

Doob's h -transform

Let X_t be a Markov process with transition kernel $p(s, \mathbf{x}, t, d\mathbf{y})$. Fix an end time T .

- Define

$$h(t, \mathbf{y}) := p(t, \mathbf{y}, T, A) = \Pr(X_T \in A | X_t = \mathbf{y})$$

- Conditioned on the event $X_T \in A$, we have a new Markov kernel:

$$\begin{aligned} \tilde{p}(s, \mathbf{x}, t, B) &:= \Pr(X_t \in B | X_s = \mathbf{x}, X_T \in A) = \frac{\Pr(X_t \in B, X_T \in A | X_s = \mathbf{x})}{\Pr(X_T \in A | X_s = \mathbf{x})} \\ &= \int_B h(t, \mathbf{y}) p(s, \mathbf{x}, t, d\mathbf{y}) / h(s, \mathbf{x}), \end{aligned}$$

or written more succinctly in terms of the balance equation:

$$h(s, \mathbf{x}) \tilde{p}(s, \mathbf{x}, t, d\mathbf{y}) = p(s, \mathbf{x}, t, d\mathbf{y}) h(t, \mathbf{y}), \quad i.e., \quad \tilde{p} = h^{-1} p h.$$

$$\tilde{p}(t, \mathbf{x}, t, d\mathbf{y}) = h(t, \mathbf{y})p(t, \mathbf{x}, t, d\mathbf{y})/h(t, \mathbf{x}) = \delta_{\mathbf{x}}(d\mathbf{y})$$

$$\tilde{p}(s, \mathbf{x}, t, d\mathbf{y}) = h(t, \mathbf{y})p(s, \mathbf{x}, t, d\mathbf{y})/h(s, \mathbf{x})$$

$$= h(t, \mathbf{y}) \int_{\mathbf{z}} p(s, \mathbf{x}, \tau, d\mathbf{z})p(\tau, \mathbf{z}, t, d\mathbf{y})/h(s, \mathbf{x})$$

$$= \int_{\mathbf{z}} h(\tau, \mathbf{z})p(s, \mathbf{x}, \tau, d\mathbf{z})/h(s, \mathbf{x}) \cdot h(t, \mathbf{y})p(\tau, \mathbf{z}, t, d\mathbf{y})/h(\tau, \mathbf{z})$$

$$= \int_{\mathbf{z}} \tilde{p}(s, \mathbf{x}, \tau, d\mathbf{z})\tilde{p}(\tau, \mathbf{z}, t, d\mathbf{y})$$

$$\tilde{p}(s, \mathbf{x}, T, A) = 1 \implies \Pr(\tilde{\mathbf{X}}_T \in A) = 1$$

More Generally

- Everything remains true for any **positive h that is consistent**:

$$\mathbb{E}[h(t, \mathbf{X}_t) | \mathbf{X}_s] = h(s, \mathbf{X}_s), \quad i.e., \quad h = \mathcal{P}_{s,t}h.$$

- In particular, for Markov densities we may set

$$h(t, \mathbf{y}) := p(t, \mathbf{y}, T, \mathbf{x}_T), \quad h(T, \mathbf{y}) = p(T, \mathbf{y}, T, \mathbf{x}_T) = \delta_{\mathbf{x}_T}(\mathbf{y}),$$

where $\mathbf{x}_T$ is a fixed value that the process ends at.

h -transform for Diffusions

Consider the diffusion process

$$d\mathbf{X}_t = \mathbf{b}_t(\mathbf{X}_t) dt + G_t(\mathbf{X}_t) dB_t, \quad \mathcal{L}f = \langle \nabla f, \mathbf{b} \rangle + \frac{1}{2} \langle \nabla^2 f, A \rangle, \quad A = G_t G_t^\top$$

The generator of the conditional Markov kernel $\tilde{p}$ is

$$\tilde{\mathcal{L}}_s f = \frac{1}{h} (\partial_s + \mathcal{L}_s)(hf) = \langle \nabla \log h \otimes \nabla f, A \rangle + \mathcal{L}_s f = \langle \nabla f, A \nabla \log h \rangle + \mathcal{L}_s f,$$

corresponding to the SDE:

$$d\mathbf{X}_t = [\mathbf{b}_t(\mathbf{X}_t) + G_t(\mathbf{X}_t) G_t(\mathbf{X}_t)^\top \nabla \log h(t, \mathbf{X}_t)] dt + G_t(\mathbf{X}_t) dB_t.$$

$$\begin{aligned}
\tilde{\mathcal{L}}_s f &= \lim_{\delta \downarrow 0} \frac{\mathbb{E}[f(\tilde{\mathbf{X}}_{s+\delta}) - f(\mathbf{X}_s) | \mathbf{X}_s]}{\delta} = \lim_{\delta \downarrow 0} \frac{\mathbb{E}[f(\mathbf{X}_{s+\delta}) \frac{h(s+\delta, \mathbf{X}_{s+\delta})}{h(s, \mathbf{X}_s)} - f(\mathbf{X}_s) | \mathbf{X}_s]}{\delta} \\
&= \frac{1}{h(s, \mathbf{X}_s)} \lim_{\delta \downarrow 0} \frac{\mathbb{E}[f(\mathbf{X}_{s+\delta}) (h(s, \mathbf{X}_{s+\delta}) + \partial_s h(s, \mathbf{X}_{s+\delta}) \delta + o(\delta)) - f(\mathbf{X}_s) h(s, \mathbf{X}_s) | \mathbf{X}_s]}{\delta} \\
&= \frac{1}{h} (\partial_s + \mathcal{L}_s)(hf).
\end{aligned}$$

- Setting $f \equiv 1$ we know $\frac{1}{h}(\partial_s + \mathcal{L}_s)h = 0$
- Simplify

Brownian Bridge

Consider a Brownian motion with transition density p :

$$\forall t \in [0, T], \quad dX_t = dB_t, \quad X_0 = 0$$

- Define

$$h(t, \mathbf{x}) := p(t, \mathbf{x}, T, \mathbf{0}) \propto \exp \left(-\frac{\|\mathbf{0} - \mathbf{x}\|^2}{2(T-t)} \right)$$

- Thus, $\nabla \log h(t, \mathbf{x}) = -\frac{\mathbf{x}}{T-t}$
- The conditional process corresponds to the SDE:

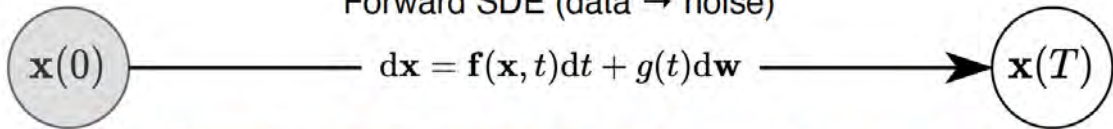
$$dX_t = -\frac{X_t}{T-t} dt + dB_t,$$

namely the Brownian bridge.

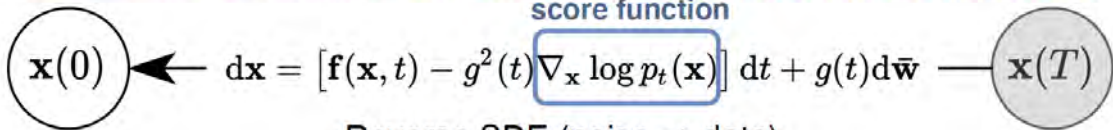
Back to Machine Learning

Diffusion Models

Forward SDE (data $\rightarrow$ noise)



score function



Reverse SDE (noise $\rightarrow$ data)

Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.

Stochastic Differential Equations (SDE)

$$d\mathbf{X}_t = \mathbf{v}_t(\mathbf{X}_t) dt + \sigma_t dB_t$$

$$\mathbf{X}_{t+1} \approx \mathbf{X}_t + \eta_t \cdot \mathbf{v}_t(\mathbf{X}_t) + \sigma_t \sqrt{\eta_t} \cdot \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- Intuitively, $\mathbf{X}_{t+1}$ is now a deformed, **noisy** version of $\mathbf{X}_t$
 - drift term $\mathbf{v}_t$: the infinitesimal change of $\mathbf{E}[\mathbf{X}_u | \mathbf{X}_t]$ when $u \searrow t$
 - diffusion term σ_t : the infinitesimal change of $\mathbf{E}[(\mathbf{X}_u - \mathbf{X}_t)(\mathbf{X}_u - \mathbf{X}_t)^\top | \mathbf{X}_t]$ when $u \searrow t$
- Kolmogorov forward equation (a.k.a. Fokker-Planck equation):

$$\boxed{\nabla_t p_t = -\langle \nabla_{\mathbf{x}}, p_t \mathbf{v}_t \rangle + \frac{1}{2} \sigma_t^2 \langle \nabla_{\mathbf{x}} \nabla_{\mathbf{x}}^\top, p_t \rangle}, \quad \mathbf{x}_t \sim p_t$$

- Kolmogorov backward equation (with fixed end time $t > s$):

$$-\partial_s p_s = \langle \mathbf{v}_s, \nabla_{\mathbf{x}} p_s \rangle + \frac{1}{2} \sigma_s^2 \nabla_{\mathbf{x}} \nabla_{\mathbf{x}}^\top p_s$$

ODE $\Leftrightarrow$ SDE

$$d\mathbf{x}_t = \mathbf{v}_t(\mathbf{x}_t) dt$$

$$d\mathbf{X}_t = \mathbf{v}_t(\mathbf{X}_t) dt + \sigma_t d\mathbf{B}_t$$

- Any ODE is a (trivial) SDE with $\sigma_t \equiv 0$
- Conversely, any SDE is “equivalent” (meaning marginal preserving) to an ODE:

$$\mathbf{v}_t \leftarrow \mathbf{v}_t - \frac{1}{2}\sigma_t^2 \nabla_{\mathbf{x}} \log p_t$$

- The **score function** plays an important role:

$$\mathbf{s}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x}) := \nabla_{\mathbf{x}} \log p(\mathbf{x})$$

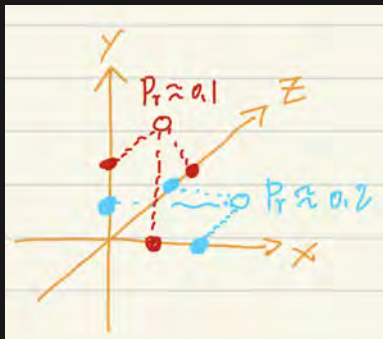
A Naive Approach

$$d\mathbf{X}_t = \mathbf{v}_t(\mathbf{X}_t) dt + \sigma_t d\mathbf{B}_t, \quad \mathbf{X}_0 \sim p_{\text{noise}}$$

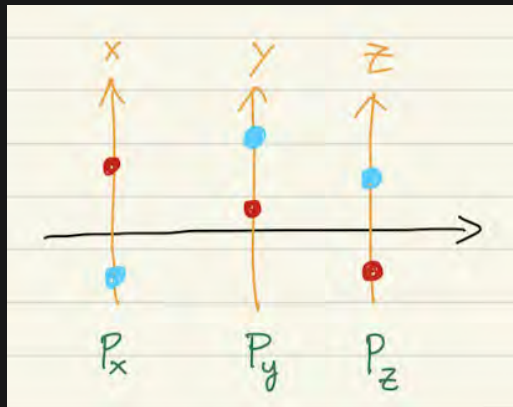
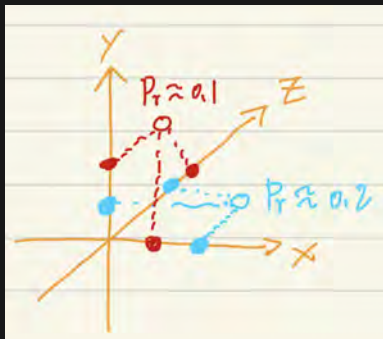
$$\mathbf{X}_{t+1} \approx \mathbf{X}_t + \eta_t \cdot \mathbf{v}_t(\mathbf{X}_t) + \sigma_t \sqrt{\eta_t} \cdot \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- $\text{KL}(p_{\text{data}} \| p_1) \equiv -\mathbb{E}_{p_{\text{data}}} \log p_1(\mathbf{X})$
 - have a sample from p_{data}
 - can compute $\log p_1(\mathbf{X}_1)$ using Fokker-Planck $\partial_t p_t = -\langle \partial_{\mathbf{x}}, p_t \mathbf{v}_t \rangle + \frac{1}{2} \sigma_t^2 \langle \partial_{\mathbf{x}} \partial_{\mathbf{x}}^\top, p_t \rangle$??
 - but $\mathbf{X}_1 \sim p_1 \neq p_{\text{data}}$!
- $\text{KL}(q_0 \| p_{\text{noise}}) \equiv -\mathbb{E}_{q_0} [\log p_{\text{noise}}(\mathbf{X}_0) - \log q_0(\mathbf{X}_0)]$
 - can we even run an SDE backwards to get q_0 ?

Sampling from a Distribution on $\mathbb{R}^d$



Sampling from a Distribution on $\mathbb{R}^d$



Path Space and Path Measure



- $\mathbf{x} = (\mathbf{x}_t)_{t \in [0,1]}$ is a path
- Disintegration (fancy name for joint = marginal $\times$ conditional):

$$p(\mathbf{x}) = p_0(\mathbf{x}_0) \cdot p(\mathbf{x}|\mathbf{x}_0) = p_1(\mathbf{x}_1) \cdot p(\mathbf{x}|\mathbf{x}_1)$$

Time Reversal



$$dX_t = \mathbf{v}_t(X_t) dt + \sigma_t dB_t$$

$$d\bar{X}_t = [\mathbf{v}_t(\bar{X}_t) - \sigma_t^2 \nabla \log p_t(\bar{X}_t)] dt + \sigma_t d\bar{B}_t$$

H. Föllmer. "An entropy approach to the time reversal of diffusion processes". In: *Stochastic Differential Systems Filtering and Control*. Springer, 1985, pp. 156–163.

Reverse-time SDE

$$\begin{aligned}dX_t &= \mathbf{v}_t(X_t) dt + \sigma_t dB_t \\d\bar{X}_t &= \bar{\mathbf{v}}_t(\bar{X}_t) dt + \sigma_t d\bar{B}_t, \quad \text{where}\end{aligned}$$

$$\bar{\mathbf{v}}_t = \mathbf{v}_t - \sigma_t^2 \partial_{\mathbf{x}} \log p_t$$

- Time flows backwards for the bar quantities
- Forward SDE: diffuses data into noise
- Reverse SDE: molds noise into data
- $\mathbf{v}_t$ and σ_t together specify $\bar{\mathbf{v}}_t$: key is to estimate the score $\partial_{\mathbf{x}} \log p_t$

$$dX_t = \mathbf{u}_t(X_t) dt + \sigma_t(X_t) dB_t$$

$$dZ_t = \mathbf{v}_t(Z_t) dt + \sigma_t(Z_t) dB_t$$

Theorem: Girsanov's Change of Variable

$$\frac{dX}{dZ}(X) = \exp \left[\int \left(\frac{\mathbf{u}_t(X_t) - \mathbf{v}_t(X_t)}{\sigma_t(X_t)} \right) dB_t + \frac{1}{2} \int \left\| \frac{\mathbf{u}_t(X_t) - \mathbf{v}_t(X_t)}{\sigma_t(X_t)} \right\|^2 dt \right]$$

Under mild conditions:

$$KL(X \parallel Z) := \mathbb{E}_X \left[\log \frac{dX}{dZ} \right] = KL(X_0 \parallel Z_0) + \frac{1}{2} \int \mathbb{E} \left\| \frac{\mathbf{u}_t(X_t) - \mathbf{v}_t(X_t)}{\sigma_t(X_t)} \right\|^2 dt$$

Diffusion Model

$$dX_t = \mathbf{0} dt + \sigma_t dB_t$$

$$d\bar{X}_t = -\sigma_t^2 \nabla \log p_t(\bar{X}_t) dt + \sigma_t d\bar{B}_t$$

$$d\bar{Z}_t = -\sigma_t^2 \mathbf{s}_t(\bar{Z}_t; \boldsymbol{\theta}) dt + \sigma_t d\bar{B}_t$$

$$X_t = X_0 + \int_0^t \sigma_s dB_s$$

- Applying Girsanov's change of variable:

$$\text{KL}(\bar{X} \parallel \bar{Z}) = \cancel{\text{KL}(\bar{X}_1 \parallel \bar{Z}_1)} + \frac{1}{2} \int \mathbb{E}[\sigma_t^2 \|\nabla \log p_t(\bar{X}_t) - \mathbf{s}_t(\bar{X}_t; \boldsymbol{\theta})\|^2] dt$$

– this is nothing but (weighted) score matching!

- Can sample $\bar{X}_t = X_{1-t}$ by simulating the forward SDE:

$$X_0 \sim p_{\text{data}}, \quad dX_t = \mathbf{0} dt + \sigma_t dB_t$$

- But, how to avoid the true score $\nabla \log p_t$ in the objective?

Denoising Score Matching

Theorem: Tweedie's Formula

Let $X = Y + Z$ where $Y \perp Z \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Then,

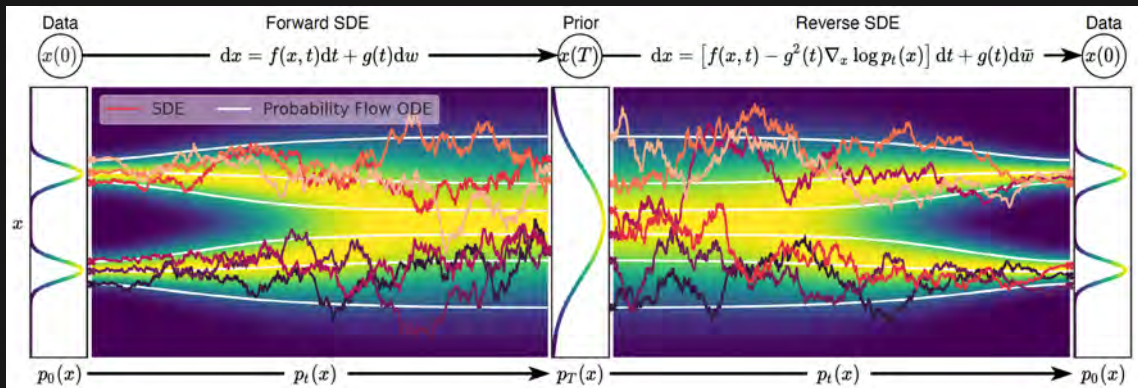
$$\nabla \log p_{\bullet}(\mathbf{x}) = -\Sigma^{-1} \cdot \mathbb{E}[Z|X = \mathbf{x}]$$

$$dX_t = \sigma_t dB_t \implies \boxed{X_t = X_0 + \epsilon_t}, \quad \text{where } \epsilon_t \sim \mathcal{N}(\mathbf{0}; \varsigma_t^2 I), \quad \varsigma_t^2 = \int_0^t \sigma_s^2 ds$$

$$\implies \nabla \log p_t(\mathbf{x}_t) = \mathbb{E} \left[\frac{\mathbf{x}_0 - \mathbf{x}_t}{\varsigma_t^2} \mid X_t = \mathbf{x}_t \right]$$

$$2\text{KL}(X \| Z) = \int \mathbb{E}[\sigma_t^2 \|\nabla \log p_t(X_t) - \mathbf{s}_t(X_t; \theta)\|^2] dt$$

$$= \int \mathbb{E}[\sigma_t^2 \|\mathbb{E} \left[\frac{\mathbf{x}_0 - \mathbf{x}_t}{\varsigma_t^2} \mid X_t \right] - \mathbf{s}_t(X_t; \theta)\|^2] dt \equiv \boxed{\mathbb{E}_{t, X_0, X_t} \left[\sigma_t^2 \left\| \frac{\mathbf{x}_0 - \mathbf{x}_t}{\varsigma_t^2} - \mathbf{s}_t(X_t; \theta) \right\|^2 \right]}$$



Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.

Score-based Diffusion Generative Models

$$d\mathbf{X}_t = \mathbf{v}_t(\mathbf{X}_t) dt + \sigma_t d\mathbf{B}_t$$

$$\mathbf{X}_{t+1} \approx \mathbf{X}_t + \eta_t \cdot \mathbf{v}_t(\mathbf{X}_t) + \sigma_t \sqrt{\eta_t} \cdot \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- Key is to estimate the score $\mathbf{s}_t(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_t$
- Apply denoising score matching: $\mathbb{E}[\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{z}) \mid \mathbf{x}] = \nabla_{\mathbf{x}} \log p(\mathbf{x})$

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t \sim \mu, (\mathbf{X}_t, \mathbf{X}_0) \sim p(\mathbf{x}_t, \mathbf{x}_0)} \lambda_t \|\mathbf{s}_t(\mathbf{X}_t; \boldsymbol{\theta}) - \nabla_{\mathbf{x}} \log p(\mathbf{X}_t | \mathbf{X}_0)\|_2^2$$

- $\mathbf{X}_0 \sim p_{\text{data}}(\mathbf{x})$, the data density
- $p(\mathbf{x}_t | \mathbf{x}_0)$ can be derived from the forward SDE, in closed-form if $\mathbf{v}_t$ is affine

Inference After Learning

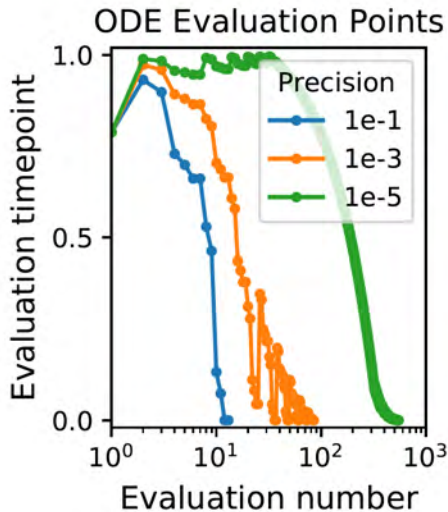
$$d\mathbf{X}_t = \mathbf{v}_t(\mathbf{X}_t) dt + \sigma_t d\mathbf{B}_t$$

$$d\bar{\mathbf{X}}_t = \left[\mathbf{v}_t(\bar{\mathbf{X}}_t) - \sigma_t^2 \mathbf{s}_t(\bar{\mathbf{X}}_t; \boldsymbol{\theta}) \right] dt + \sigma_t d\bar{\mathbf{B}}_t$$

$$d\bar{\mathbf{x}}_t = \left[\mathbf{v}_t(\bar{\mathbf{x}}_t) - \frac{1}{2} \sigma_t^2 \mathbf{s}_t(\bar{\mathbf{x}}_t; \boldsymbol{\theta}) \right] dt$$

- Run the reverse SDE or the equivalent ODE
 - sample $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \text{Id})$
 - apply numerical SDE or ODE solver (e.g., [Euler-Maruyama](#))

D. J. Higham. "An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations". *SIAM Review*, vol. 43, no. 3 (2001), pp. 525–546.



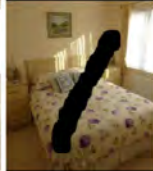
Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.



Interpolation

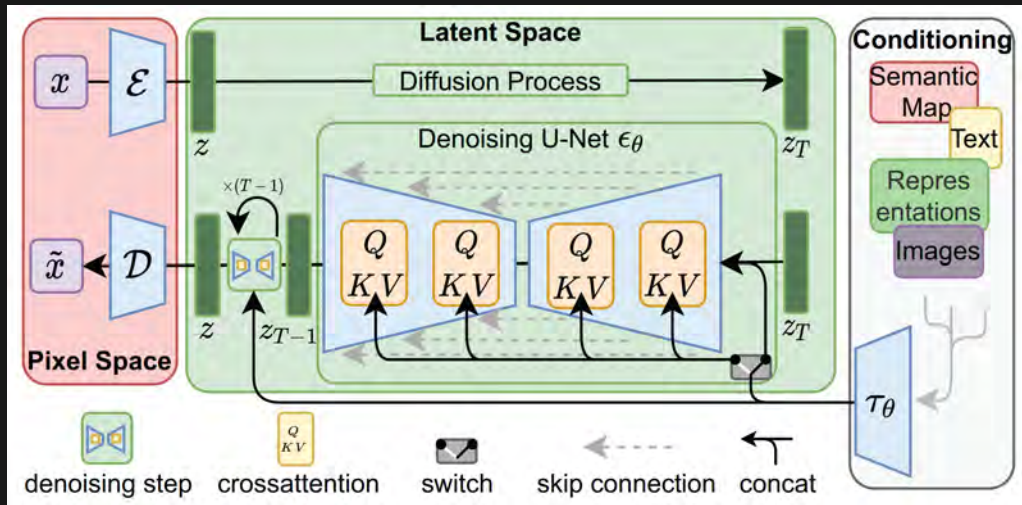


Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.





Stable Diffusion



R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10674–10685.



Simple or Sophisticated?

Consider two stochastic processes and their reverse processes:

$$\begin{aligned}d\mathbf{X}_t &= d\mathbf{B}_t, \quad \mathbf{X}_0 \sim p_0, & d\bar{\mathbf{X}}_t &= -\nabla \log p_t(\bar{\mathbf{X}}_t) + d\mathbf{B}_t, \\d\mathbf{Z}_t &= d\mathbf{B}_t, \quad \mathbf{Z}_0 \sim q_0, & d\bar{\mathbf{Z}}_t &= -\nabla \log q_t(\bar{\mathbf{Z}}_t) + d\mathbf{B}_t.\end{aligned}$$

Then, applying Girsanov's theorem twice we obtain

$$\text{KL}(\mathbf{X} \parallel \mathbf{Z}) = \text{KL}(p_0 \parallel q_0) + 0 = \text{KL}(\bar{\mathbf{X}} \parallel \bar{\mathbf{Z}}) = \text{KL}(\mathbf{X}_T \parallel \mathbf{Z}_T) + \frac{1}{2} \int_0^T \mathbb{E}[\|\nabla \log p_t(\mathbf{X}_t) - \nabla \log q_t(\mathbf{X}_t)\|_2^2] dt,$$

Thus,

$$\lim_{T \rightarrow \infty} \text{KL}(\mathbf{X}_T \parallel \mathbf{Z}_T) \rightarrow 0 \implies \text{KL}(p_0 \parallel q_0) = \frac{1}{2} \int_0^\infty \mathbb{E}[\|\nabla \log p_t(\mathbf{X}_t) - \nabla \log q_t(\mathbf{X}_t)\|_2^2] dt.$$