Diffusion Models for Generative Learning

1 Probability from approximation point of view; conditional expectation

1.1 Random variable (RV)

Definition. Let $(\Omega, \mathcal{F}, P)$ be a probability space ($P(\Omega) = 1$). A random variable is a measurable function from Ω to $\mathbb{R}$, namely

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}),$$

where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$, i.e., for all $B \in \mathcal{B}$,

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

Expectation. Let $X \in L^1(\Omega, \mathcal{F}, P)$, i.e.

$$\int_\Omega |X(\omega)|\, p(d\omega) < \infty.$$

Then

$$\mathbb{E}[X] \stackrel{\text{def}}{=} \int_\Omega X(\omega)\, p(d\omega).$$

Claim. $\mathbb{E}X$ is the best approximation in the L^2 norm, i.e.

$$\|X - \mathbb{E}X\|_{L^2(\Omega)} = \min_{a \in \mathbb{R}} \|X - a\|.$$

Proof.

$$\begin{aligned}
\|X - a\|^2 &= \|X - \mathbb{E}X + \mathbb{E}X - a\|^2 \\
&= \|X - \mathbb{E}X\|^2 + 2\langle X - \mathbb{E}X, \mathbb{E}X - a\rangle + |\mathbb{E}X - a|^2 \\
&= \|X - \mathbb{E}X\|^2 + |\mathbb{E}X - a|^2 + 2\,\mathbb{E}[X - \mathbb{E}X]\,(\mathbb{E}X - a) \\
&= \|X - \mathbb{E}X\|^2 + |\mathbb{E}X - a|^2 \geq \|X - \mathbb{E}X\|^2.
\end{aligned}$$

Hence $\mathbb{E}X$ is the orthogonal projection of X onto $\mathbb{R}$.

A second proof uses

$$\begin{aligned}
\langle X - \mathbb{E}X, a\rangle &= \int_\Omega (X - \mathbb{E}X) a\, p(d\omega) \\
&= a \int_\Omega X\, p(d\omega) - a \int_\Omega \mathbb{E}X\, p(d\omega) \\
&= a\,\mathbb{E}X - a\,\mathbb{E}X = 0.
\end{aligned}$$

Thus $\mathbb{E}X$ is the projection of X into $\mathbb{R}$.

1.2 Variance, correlation and covariance matrices

Let RV $X : \Omega \to \mathbb{R}$. Then the variance of X is

$$V(X) \stackrel{\text{def}}{=} \mathbb{E}|X - \mathbb{E}X|^2.$$

Thus $\sqrt{V(X)}$ is the L^2 error of approximating X by $\mathbb{E}X$.
Let $X : \Omega \to \mathbb{R}^m$ and $Y : \Omega \to \mathbb{R}^n$. The correlation matrix is defined by

$$\operatorname{cor}(X, Y) \stackrel{\text{def}}{=} \mathbb{E}\big[(X - \mathbb{E}X) \otimes (Y - \mathbb{E}Y)\big],$$

where

$$a \otimes b \stackrel{\text{def}}{=} [a_i b_j]_{m \times n}.$$

The covariance matrix of X is

$$\operatorname{cov}(X) \stackrel{\text{def}}{=} \operatorname{cor}(X, X) = \mathbb{E}\big[(X - \mathbb{E}X) \otimes (X - \mathbb{E}X)\big].$$

Equivalently,

$$\operatorname{cov}(X) = \big(\mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)]\big)_{i,j=1}^m.$$

1.3 Independent random variables

(1) **Conditional probability.** Let $A, B \in \mathcal{F}$. Then

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

(2) **Independent events.** A and B are independent if

$$P(B \mid A) = P(B), \qquad \text{or equivalently} \qquad P(A \cap B) = P(A)P(B).$$

(3) **Independent random variables.** X and Y are independent if

$$\{X \in A\} \text{ and } \{Y \in B\}$$

are independent for all Borel sets $A, B \in \mathcal{B}$.

If X and Y are independent, then

$$\mathbb{E}(X, Y) = (\mathbb{E}X, \mathbb{E}Y).$$

In particular, if X_A and X_B are indicator variables,

$$\mathbb{E}[X_A X_B] = \mathbb{E}[X_{A \cap B}] = P(A \cap B) = P(A)P(B) = \mathbb{E}[X_A]\mathbb{E}[X_B].$$

If X and Y are independent, then

$$\operatorname{cor}(X, Y) = 0,$$

i.e. they are uncorrelated.

2 Conditional expectation

Definition. Given probability spaces $(\Omega, \mathcal{F}, P)$ and $(\Omega, \mathcal{G}, P)$ with $\mathcal{G} \subset \mathcal{F}$, the conditional expectation of a random variable in $(\Omega, \mathcal{F}, P)$ with respect to $\mathcal{G}$ is defined as a random variable in $(\Omega, \mathcal{G}, P)$ such that

$$\int_A \mathbb{E}(X \mid \mathcal{G})\, p(d\omega) = \int_A X\, p(d\omega) \qquad \text{for all random variables } 1_A \in (\Omega, \mathcal{G}, P).$$

Equivalently, for all $A \in \mathcal{G}$,

$$\int_A \mathbb{E}(X \mid \mathcal{G})\, p(d\omega) = \int_A X\, p(d\omega).$$

Claim. Let $X \in L^2(\Omega, \mathcal{F}, P)$. Then $\mathbb{E}(X \mid \mathcal{G})$ is the projection of X in $L^2(\Omega, \mathcal{G}, P)$, i.e.

$$\mathbb{E}(X \mid \mathcal{G}) = \arg \min_{Y \in L^2(\Omega, \mathcal{G}, P)} \|X - Y\|_{L^2(\Omega, \mathcal{F}, P)}.$$

Proof. WLOG, assume scalar-valued random variables. Need to show

$$\langle X - \mathbb{E}(X \mid \mathcal{G}), Y \rangle_{L^2(\Omega, \mathcal{G}, P)} = 0, \qquad \forall Y \in L^2(\Omega, \mathcal{G}, P).$$

(1) If Y is a characteristic function, $Y = \chi_A$ with $A \in \mathcal{G}$, then

$$\begin{aligned} \langle X - \mathbb{E}(X \mid \mathcal{G}), \chi_A \rangle &= \int_\Omega X \chi_A\, p(d\omega) - \int_\Omega \mathbb{E}(X \mid \mathcal{G}) \chi_A\, p(d\omega) \\ &= \int_A X\, p(d\omega) - \int_A \mathbb{E}(X \mid \mathcal{G})\, p(d\omega) = 0. \end{aligned}$$

(2) If Y is simple, $Y = \sum_{i=1}^n \alpha_i \chi_{A_i}$ with $A_i \in \mathcal{G}$, the result follows by linearity.

(3) If $Y \in L^2(\Omega, \mathcal{G}, P)$, there exists a sequence of simple functions $Y_i \in L^2(\Omega, \mathcal{G}, P)$ such that $Y_i \to Y$ in $L^2(\Omega, \mathcal{G}, P)$. Then

$$0 = \langle X - \mathbb{E}(X \mid \mathcal{G}), Y_i \rangle \to \langle X - \mathbb{E}(X \mid \mathcal{G}), Y \rangle.$$

Properties of conditional expectation

(1)

$$\mathbb{E}[\mathbb{E}(X \mid \mathcal{G})] = \mathbb{E}X.$$

Indeed, since $\Omega \in \mathcal{G}$,

$$\int_\Omega X(\omega)\, p(d\omega) = \int_\Omega \mathbb{E}(X \mid \mathcal{G})\, p(d\omega).$$

Hence

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}(X \mid \mathcal{G})].$$

(2) If X is independent of $\mathcal{G}$, i.e. X is independent of any $\mathcal{G}$-measurable random variable, then

$$\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X].$$

Proof: need to show that $\mathbb{E}[X]$ is the projection of X onto $L^2(\Omega, \mathcal{G}, P)$. For any $Y \in L^2(\Omega, \mathcal{G}, P)$,

$$\langle X - \mathbb{E}X, Y \rangle_{L^2} = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0,$$

since X and Y are independent.

(3) If $\mathcal{G}$ and $\mathcal{H}$ are σ-algebras such that $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$, then

$$\mathbb{E}[\mathbb{E}(X \mid \mathcal{H}) \mid \mathcal{G}] = \mathbb{E}(X \mid \mathcal{G}).$$

Proof: for any $Y \in L^2(\Omega, \mathcal{G}, P)$, also $Y \in L^2(\Omega, \mathcal{H}, P)$, so

$$\langle \mathbb{E}(X \mid \mathcal{H}) - \mathbb{E}(X \mid \mathcal{G}), Y \rangle_{L^2} = \langle X, Y \rangle - \langle X, Y \rangle = 0.$$

Remark:

$$\mathbb{E}[\mathbb{E}(X \mid \{\Omega, \varnothing\})] = \mathbb{E}[X \mid \{\Omega, \varnothing\}] = \mathbb{E}X.$$

(4) If Y is $\mathcal{G}$-measurable, then

$$\mathbb{E}[YX \mid \mathcal{G}] = Y\,\mathbb{E}[X \mid \mathcal{G}].$$

(If Y is $\mathcal{G}$-measurable then it behaves like a constant under conditioning.)

(5) If X is independent of $\mathcal{G}$ and Y is $\mathcal{G}$-measurable, then

$$\mathbb{E}[f(X, Y) \mid \mathcal{G}] = \big(\mathbb{E}[f(x, y)]\big)_{y=Y}$$

for appropriate f. In particular,

$$\mathbb{E}[YX \mid \mathcal{G}] = Y\,\mathbb{E}[X \mid \mathcal{G}] = Y\,\mathbb{E}[X].$$

3 PDF, Brownian motion, Gaussian random variables

3.1 Probability density function (PDF), Gaussian RVs

Definition. Given an RV X in $(\Omega, \mathcal{F})$, it induces a probability measure μ_X on $(\mathbb{R}^n, \mathcal{B})$ by

$$\mu_X(B) \stackrel{\text{def}}{=} P(X \in B), \qquad \forall B \in \mathcal{B}.$$

Let λ be Lebesgue measure and assume $\mu_X \ll \lambda$. Then there exists a density $p \in L^1(\mathbb{R}^n)$ such that

$$\mu_X(B) = \int_B p(x)\,dx.$$

Thus

$$p = \frac{d\mu_X}{d\lambda}$$

in the Radon–Nikodym sense. p is called the PDF of X.

PDF of a Gaussian random variable. If

$$X \sim \mathcal{N}(\mu, \Sigma), \qquad \mu \in \mathbb{R}^n, \quad \Sigma \in \mathbb{R}^{n \times n},$$

then

$$p(x) = \frac{1}{(\sqrt{2\pi})^n |\det \Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right).$$

For $n = 1$,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad X \sim \mathcal{N}(\mu, \sigma^2).$$

Also,

$$\mu = \mathbb{E}[X], \qquad \Sigma = \operatorname{cov}(X).$$

The standard Gaussian has $\mu = 0$ and $\Sigma = I_n$.

PDF of $Z = f(X)$, $\mathbb{R}^n \to \mathbb{R}^n$. Let f be a one-to-one and differentiable function. If $X \sim p_X(x)$, then the PDF of Z is

$$p_Z(z) = p_X(f^{-1}(z)) \left| \det \frac{\partial f^{-1}}{\partial z} \right| = p_X(f^{-1}(z)) \left| \det \frac{\partial z}{\partial x} \right|^{-1}.$$

PDF of an independent shift. Let

$$Z = X + Y,$$

where $X \sim p_X(x)$ and Y is independent of X. Then

$$p_{Z|Y}(z \mid Y) = p_X(z - Y).$$

Important property. Assume X and Y are Gaussian. Then

$$X \text{ and } Y \text{ are independent} \iff X \text{ and } Y \text{ are uncorrelated},$$

that is,

$$\operatorname{cor}(X, Y) = 0.$$

3.2 Stochastic process

A stochastic process is a set of random variables parameterized by time:

$$X(t), \qquad t \geq 0.$$

3.3 Brownian motion (Wiener process)

$W(t)$ is a stochastic process such that:

(1) $W_0 = 0$ a.s. and $W_t(\omega)$ is continuous in t a.s. $\omega \in \Omega$.

(2) $W_t - W_s \sim \mathcal{N}(0, (t-s)I)$; in particular $W_t \sim \mathcal{N}(0, tI)$.

(3) W_t has independent increments: for $s < t_1 < t_2 < t_3 < t_4$,

$$W(t_2) - W(t_1) \quad \text{and} \quad W(t_4) - W(t_3)$$

are independent.

Fact 1. The PDF of $W(t)$ is the fundamental solution of the diffusion equation

$$\begin{cases} \dfrac{\partial p}{\partial t} = \dfrac{1}{2}\Delta p, & (t, x) \in (0, \infty) \times \mathbb{R}^n, \\ p(x, 0) = \delta(x), & x \in \mathbb{R}^n. \end{cases}$$

Namely,

$$p(t, x) = \frac{1}{(\sqrt{2\pi t})^n} e^{-\frac{|x|^2}{2t}}.$$

For $n = 1$,

$$\begin{cases} \dfrac{\partial p}{\partial t} = \dfrac{1}{2}\dfrac{\partial^2 p}{\partial x^2}, \\ p(x,0) = \delta(x). \end{cases}$$

Fact 2.

$$(dW(t))^2 = dt.$$

Heuristic explanation: $dW(t) := W(t+dt) - W(t)$, and

$$\mathbb{E}[(dW(t))^2] = \mathbb{E}[(W(t+dt) - W(t))^2] = \mathrm{Var}[W(t+dt) - W(t)] = dt.$$

Rigorous explanation: let $0 = t_0 < t_1 < \cdots < t_n = t$. Then

$$\int_0^t (dW(s))^2 := \lim_{\max |t_{i+1} - t_i| \to 0} \sum_{i=0}^{n-1} (W(t_{i+1}) - W(t_i))^2 = t$$

in mean square. Hence one writes $(dW(t))^2 = dt$.

4 Introduction to stochastic calculus

4.1 Filtration

Given $(\Omega, \mathcal{F}, P)$:

(i) **Definition.** $\{\mathcal{F}_t\}_{t \geq 0}$ is said to be a filtration if it is right-continuous and

$$\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}, \qquad s \leq t.$$

(ii) **Adapted process / stochastic process.** X_t is said to be adapted to $\{\mathcal{F}_t\}_{t \geq 0}$ if X_t is measurable with respect to $\mathcal{F}_t$. For example,

$$\mathcal{F}_t = \sigma\{W(s), 0 \leq s \leq t\},$$

then W_t is adapted to $\mathcal{F}_t$.

(iii) **Progressively measurable process.** $W(t, \omega)$ is measurable with respect to the product measure $\mathcal{B}([0,t]) \times \mathcal{F}_t$. Then W_t is progressively measurable.

4.2 Itô integral

Let

$$L^2_{\mathcal{F}}(0,T) = \left\{ X(t) : X(t) \text{ is progressively measurable, } \mathbb{E}\int_0^T |X_t|^2 dt < \infty \right\}.$$

If X_t is a simple process, i.e.

$$X_t = X_{t_i}, \qquad t_i \leq t < t_{i+1}, \quad 0 = t_0 < t_1 < \cdots < t_n = T,$$

then X_{t_i} is $\mathcal{F}_{t_i}$-measurable and we define

$$\int_0^T X_t \, dW_t \overset{\text{def}}{=} \sum_{i=0}^{n-1} X_{t_i} \big(W(t_{i+1}) - W(t_i) \big).$$

Properties.

(1)
$$\mathbb{E}\int_0^T X_t\,dW_t = 0.$$
Indeed, each X_{t_i} is measurable with respect to $\mathcal{F}_{t_i}$, while $W(t_{i+1}) - W(t_i)$ is independent of $\mathcal{F}_{t_i}$ and has mean zero.

(2) **Itô isometry.**
$$\mathbb{E}\left\|\int_0^T X_t\,dW_t\right\|^2 = \mathbb{E}\int_0^T |X_t|^2 dt.$$
Proof for simple processes:
$$\begin{aligned}\mathbb{E}\left|\sum_{i=0}^{n-1} X_{t_i}(W(t_{i+1}) - W(t_i))\right|^2 &= \sum_{i=0}^{n-1}\mathbb{E}|X_{t_i}|^2\,\mathbb{E}|W(t_{i+1}) - W(t_i)|^2\\ &= \sum_{i=0}^{n-1}\mathbb{E}|X_{t_i}|^2\,(t_{i+1} - t_i)\\ &= \mathbb{E}\int_0^T |X_t|^2 dt,\end{aligned}$$
where the cross terms vanish by independence.

For general $X \in L^2_{\mathcal{F}}(0,T)$, there exists a sequence of simple functions X_n such that
$$\mathbb{E}\int_0^T |X_n - X|^2 dt \to 0.$$
Then
$$\mathbb{E}\left|\int_0^T (X_n - X_m)\,dW_t\right|^2 \le \mathbb{E}\int_0^T |X_n - X_m|^2 dt \to 0,$$
so $\int_0^T X_n\,dW_t$ converges in $L^2(\Omega)$. Define
$$\int_0^T X_t\,dW_t := \lim_{n\to\infty}\int_0^T X_n(t)\,dW_t.$$

4.3 Itô's formula: chain rule in stochastic calculus

Itô differential. Let
$$X(t) = \int_0^t f(s)\,ds + \int_0^t g(s)\,dW(s),$$
where $f \in L^1(0,T)$ and $g \in L^2(0,T)$. Then the Itô differential is defined by
$$dX(t) = f(t)dt + g(t)dW(t).$$
Recall that
$$(dt)^\alpha = 0 \ \ (\alpha > 1), \qquad (ds)^\alpha = 0, \qquad \left(\int_0^t \phi(s)dW(s)\right)^2 \sim dt, \qquad (dW(t))^2 = dt.$$

Thus

$$\begin{aligned}(dX(t))^2 &= f(t)^2(dt)^2 + 2f(t)g(t)dt\,dW(t) + g(t)^2(dW(t))^2\\ &= g(t)^2dt \neq 0.\end{aligned}$$

Theorem (Itô's formula). Let $F \in C^2((0,T)\times\mathbb{R})$. Then

$$d(F(t,X_t)) = \left(\frac{\partial F}{\partial t} + \frac{\partial F}{\partial x}f + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}g^2\right)dt + \frac{\partial F}{\partial x}g\,dW_t.$$

Proof.

$$\begin{aligned}dF(t,X_t) &= \frac{\partial F}{\partial t}dt + \frac{\partial F}{\partial x}dX_t + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}(dX_t)^2\\ &= \frac{\partial F}{\partial t}dt + \frac{\partial F}{\partial x}(fdt + gdW_t) + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}g^2dt.\end{aligned}$$

Example. Evaluate $\int_0^T W(s)\,dW(s)$. Let $X(t) = W(t)$ and $F(x) = x^2$. Then

$$d(W(t)^2) = 2W(t)dW(t) + dt.$$

Hence

$$\int_0^T W(t)\,dW(t) = \frac{1}{2}\int_0^T d(W(t)^2) - \int_0^T dt = \frac{1}{2}W(T)^2 - T.$$

(As written in the notes, the final line keeps the same normalization convention from the displayed identity.)

Itô's formula in higher dimensions. Let

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

where $X_t, f \in \mathbb{R}^n$, $\sigma \in \mathbb{R}^{n\times m}$ and $W \in \mathbb{R}^m$. Then

$$dF(t,X_t) = \frac{\partial F}{\partial t} + \langle\nabla_x F, f\rangle\,dt + \frac{1}{2}\langle\nabla_x^2 F\,dX_t, dX_t\rangle + \nabla_x F\cdot\sigma\,dW_t,$$

that is,

$$dF(t,X_t) = \left(\frac{\partial F}{\partial t} + \langle\nabla_x F, f\rangle + \frac{1}{2}\nabla_x^2 F : \sigma\sigma^T\right)dt + \nabla_x F\cdot\sigma\,dW_t.$$

5 Stochastic differential equations (SDEs)

Definition. $X(t)$ is said to be a solution of the SDE

$$\begin{cases}dX(t) = f(t,X(t))dt + g(t,X(t))dW(t), & t > 0,\\ X(0) = X_0,\end{cases}$$

if X satisfies

$$X(t) = X_0 + \int_0^t f(s,X(s))\,ds + \int_0^t g(s,X(s))\,dW(s).$$

Example 1: geometric Brownian motion. Let μ, σ be constants. Solve

$$\begin{cases} dX(t) = \mu X(t)dt + \sigma X(t)dW(t), \quad t > 0, \\ X(0) = 1. \end{cases}$$

Then

$$\frac{1}{X(t)} dX(t) = \mu dt + \sigma dW(t).$$

Apply Itô's formula to $\ln X(t)$:

$$\begin{aligned} d(\ln X(t)) &= \frac{1}{X(t)} dX(t) - \frac{1}{2X(t)^2} (dX(t))^2 \\ &= \mu dt + \sigma dW(t) - \frac{1}{2}\sigma^2 dt \\ &= \left(\mu - \frac{1}{2}\sigma^2\right) dt + \sigma dW(t). \end{aligned}$$

Integrating from 0 to t gives

$$\ln \frac{X(t)}{X_0} = \left(\mu - \frac{1}{2}\sigma^2\right) t + \sigma W(t).$$

Hence

$$X(t) = X_0 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right) t + \sigma W(t)\right).$$

Example 2: Langevin / Ornstein–Uhlenbeck equation. Solve

$$\begin{cases} dX = -bX dt + \sigma dW, \\ X(0) = X_0, \end{cases}$$

where X_0 is independent of $W(t)$. Then

$$X(t) = e^{-bt} X_0 + \sigma \int_0^t e^{-b(t-s)}\, dW(s).$$

Write $X(t) = Y(t) + Z(t)$, where

$$Y(t) = e^{-bt} X_0, \qquad Z(t) = \sigma \int_0^t e^{-b(t-s)}\, dW(s).$$

Since $Z(t)$ is Gaussian with mean zero and

$$\begin{aligned} \mathbb{E}[Z(t)^2] &= \sigma^2 \int_0^t e^{-2b(t-s)}\, ds \\ &= \sigma^2 \frac{1 - e^{-2bt}}{2b}, \end{aligned}$$

we have

$$Z(t) \sim \mathcal{N}\left(0, \sigma^2 \frac{1 - e^{-2bt}}{2b}\right).$$

Therefore

$$X(t) \mid X_0 \sim \mathcal{N}\left(e^{-bt} X_0,\, \sigma^2 \frac{1 - e^{-2bt}}{2b}\right).$$

6 Diffusion models for generative learning

6.1 Generative learning and normalizing flow method

Given samples $\{x_i\}_{i=1}^n$ of a random variable $X \in \mathbb{R}^n$:

(1) Generate an arbitrary number of samples.

(2) Generate (or learn) the PDF of X.

Normalizing flow method. Learn a map F such that

$$X = F(Z),$$

where

$$Z \sim \mathcal{N}(0, I_n)$$

is the standard Gaussian.

Once F is known, one can perform generative learning on X:

(1) Generate samples for X:

$$x_i = F(z_i), \qquad i = 1, \dots, M,$$

where z_i are samples of Z.

(2) Generate the PDF of X:

$$p_X(x) = p_Z(F^{-1}(x)) \left| \det \left(\frac{\partial F}{\partial x} \right)^{-1} \right|.$$

Diffusion model. Use a continuous method to obtain F.

6.2 Liouville's theorem, score function

Liouville's theorem. Consider the transport equation in terms of PDEs:

$$\begin{cases} \dfrac{\partial p}{\partial t} = -\nabla \cdot (\mu p), & \text{in } [0, T] \times \mathbb{R}^n, \\ p(x, 0) = p_0(x), & x \in \mathbb{R}^n. \end{cases}$$

Example: if $T = \infty$, $\Omega = (-\infty, \infty)$ and μ is constant, then

$$\frac{\partial p}{\partial t} = -\mu \frac{\partial p}{\partial x} \qquad \Rightarrow \qquad p(x, t) = p_0(x - \mu t).$$

The transport equation in terms of probability-flow ODE is

$$\begin{cases} \dfrac{dZ_t}{dt} = \mu(Z_t, t), & t > 0, \\ Z_0 = X \sim p_0(x), & x \in \mathbb{R}^n, \end{cases}$$

where X is a random variable with PDF p_0.

Example: $n = 1$, μ constant. Then

$$Z_t = X + \mu t = f(X).$$

The PDF of Z_t is

$$p(x,t) = p(f^{-1}(z_t))\,(f^{-1})'(z_t) = p_0(x-\mu t)\cdot 1 = p_0(x-\mu t),$$

which is the same as the PDE solution.

Liouville theorem. Let $p = p(x,t)$ be the PDF of the solution Z_t of the probability-flow ODE. Then p is the solution of the transport PDE.

Proof. Let $\varphi \in C_0^\infty(\mathbb{R}^n)$ be a test function. Since

$$\mathbb{E}[\varphi(Z_t)] = \int_{\mathbb{R}^n} \varphi(x)p(x,t)\,dx = \langle \varphi, p(\cdot,t)\rangle,$$

we have

$$\frac{d}{dt}\mathbb{E}[\varphi(Z_t)] = \left\langle \varphi, \frac{\partial p}{\partial t}\right\rangle.$$

On the other hand,

$$\begin{aligned}\frac{d}{dt}\mathbb{E}[\varphi(Z_t)] &= \mathbb{E}\left[\nabla\varphi(Z_t)\cdot\frac{dZ_t}{dt}\right] = \mathbb{E}[\nabla\varphi(Z_t)\cdot\mu(Z_t,t)] \\ &= \int_{\mathbb{R}^n} \nabla\varphi(x)\cdot\mu(x,t)p(x,t)\,dx \\ &= -\int_{\mathbb{R}^n} \varphi(x)\nabla\cdot(\mu p)(x,t)\,dx = \langle\varphi, -\nabla\cdot(\mu p)\rangle.\end{aligned}$$

Thus

$$\left\langle \varphi, \frac{\partial p}{\partial t}\right\rangle = \langle\varphi, -\nabla\cdot(\mu p)\rangle, \qquad \forall \varphi \in C_0^\infty(\mathbb{R}^n),$$

which implies

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

Diffusion as a transport equation; score function. Consider the heat equation

$$\begin{cases}\dfrac{\partial p}{\partial t} = \Delta p, & \text{in } \mathbb{R}^n\times(0,\infty),\\ p(x,0) = p_0(x), & x\in\mathbb{R}^n.\end{cases}$$

Note that

$$\Delta p = \nabla\cdot(\nabla p) = \nabla\cdot\left(\frac{\nabla p}{p}p\right) = -\nabla\cdot(\mu p),$$

where

$$\mu = -\frac{\nabla p}{p} = -\nabla \ln p(x,t) \overset{\text{def}}{=} -S(x,t).$$

Here

$$S = \nabla \ln p$$

is called the score function.

Interpretation from the notes:

- The score function is in the steepest ascent direction of p.
- Connection to entropy:

$$\frac{d}{dt}H(p_t) = \int_{\mathbb{R}^n} p_t \, |S(x,t)|^2 dx = \mathbb{E}|S(\cdot,t)|^2.$$

The corresponding probability-flow ODE is

$$\begin{cases} \dfrac{dZ_t}{dt} = -S(Z_t, t), \\ Z_0 = X \sim p_0(x). \end{cases}$$

The heat equation cannot be solved backward in time, but the PF-ODE can be solved backward to retrieve X.

Diffusion model idea. Starting at T (chosen so that $Z_T \sim \mathcal{N}(0, I_d)$), solve the PF-ODE backward to find X.

6.3 Fokker–Planck equation

Assume $p = p(x,t)$ is the PDF of the solution Z_t of the SDE

$$\begin{cases} dZ_t = \mu(Z_t, t)dt + \sigma(Z_t, t)dW_t, \\ Z_0 = X \sim p_0(x). \end{cases}$$

Then p is the solution of the Fokker–Planck equation

$$\begin{cases} \dfrac{\partial p}{\partial t} = -\nabla \cdot (\mu p) + \dfrac{1}{2}\nabla \cdot \nabla \cdot (\sigma\sigma^T p), \\ p(x,0) = p_0. \end{cases}$$

Proof. Let $\varphi \in C_0^\infty(\mathbb{R}^d)$. Then

$$\frac{d}{dt}\mathbb{E}[\varphi(Z_t)] = \frac{d}{dt}\langle \varphi, p\rangle = \left\langle \varphi, \frac{\partial p}{\partial t} \right\rangle.$$

On the other hand, by Itô's formula,

$$\begin{aligned} d\mathbb{E}[\varphi(Z_t)] &= \mathbb{E}[d\varphi(Z_t)] \\ &= \mathbb{E}\left[\nabla\varphi(Z_t)\cdot dZ_t + \frac{1}{2}\nabla^2\varphi(Z_t) : dZ_t \otimes dZ_t\right] \\ &= \mathbb{E}\left[\nabla\varphi(Z_t)\cdot \mu\, dt + \nabla\varphi(Z_t)\cdot \sigma\, dW_t + \frac{1}{2}\nabla^2\varphi(Z_t) : \sigma\sigma^T dt\right]. \end{aligned}$$

Taking expectation removes the stochastic integral, hence

$$\begin{aligned} \frac{d}{dt}\mathbb{E}[\varphi(Z_t)] &= \mathbb{E}[\nabla\varphi(Z_t)\cdot\mu] + \frac{1}{2}\mathbb{E}[\nabla^2\varphi(Z_t) : \sigma\sigma^T] \\ &= \langle \nabla\varphi, \mu p\rangle + \frac{1}{2}\langle \nabla^2\varphi, \sigma\sigma^T p\rangle \\ &= -\langle \varphi, \nabla\cdot(\mu p)\rangle + \frac{1}{2}\langle \varphi, \nabla\cdot\nabla\cdot(\sigma\sigma^T p)\rangle. \end{aligned}$$

Therefore

$$\frac{\partial p}{\partial t} = -\nabla\cdot(\mu p) + \frac{1}{2}\nabla\cdot\nabla\cdot(\sigma\sigma^T p).$$

Corresponding Probability Flow ODE (general form). Consider the Itô SDE in $\mathbb{R}^d$:

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

where $\sigma(x,t) \in \mathbb{R}^{d\times m}$ and define the diffusion tensor

$$a(x,t) = \sigma(x,t)\sigma(x,t)^\top.$$

The associated Fokker–Planck equation is

$$\partial_t p_t = -\nabla \cdot (b\, p_t) + \frac{1}{2}\sum_{i,j} \partial_{ij}(a_{ij}(x,t)\, p_t)\,.$$

The corresponding *probability flow ODE* (PFODE) is given by

$$\frac{dX_t}{dt} = b(X_t,t) - \frac{1}{2}\left[a(X_t,t)\,\nabla \log p_t(X_t) + (\nabla \cdot a)(X_t,t)\right],$$

where $(\nabla \cdot a)_i = \sum_j \partial_j a_{ij}$.
This deterministic flow induces the same marginal density evolution $p_t(x)$ as the SDE, i.e.,

$$\partial_t p_t + \nabla \cdot \big(v\, p_t\big) = 0, \quad v = b - \frac{1}{2}\left[a\nabla \log p_t + \nabla \cdot a\right].$$

Hence, the stochastic SDE and the deterministic PFODE are equivalent at the level of probability densities.

6.4 Diffusion model for generative learning

Recall the Liouville theorem. Assume that $p = p(x,t)$ is the solution of the PF-ODE

$$\begin{cases} \dfrac{dZ_t}{dt} = \mu(Z_t, t), & t > 0, \\ Z_0 = X \sim p_0(x), & x \in \mathbb{R}^n. \end{cases}$$

Then p is the solution of the transport PDE

$$\begin{cases} \dfrac{\partial p}{\partial t} = -\nabla \cdot (\mu p), & \text{in } \mathbb{R}^n \times (0,\infty), \\ p(x,0) = p_0(x), & x \in \mathbb{R}^n. \end{cases}$$

Goal: "generatively learn RV X."

Step 1. Construct an SDE

$$\begin{cases} dZ_t = \mu(Z_t, t)dt + \sigma(Z_t, t)dW_t, \\ Z_0 = X, \end{cases}$$

such that

$$Z_T \sim \mathcal{N}(0, I_d).$$

Example 1: Langevin equation.

$$\begin{cases} dZ_t = -Z_t dt + \sqrt{2}\, dW_t, \\ Z_0 = X \sim p_0(x), \qquad x \in \mathbb{R}^d. \end{cases}$$

Then

$$Z_t = e^{-t} X + \sqrt{2} \int_0^t e^{-(t-s)}\, dW(s),$$

and therefore

$$Z_t \mid X \sim \mathcal{N}\big(Xe^{-t}, (1 - e^{-2t}) I_d\big).$$

For large T,

$$Z_T \mid X \sim \mathcal{N}(Xe^{-T}, (1 - e^{-T}) I_d) \approx \mathcal{N}(0, I_d).$$

(The notes write the terminal approximation in this informal way.)

Example 2. Take

$$\mu(z,t) = \frac{1}{t-1} z, \qquad \sigma(z,t) = \sqrt{\frac{1+t}{1-t}}\, I_d.$$

Then the SDE is

$$dZ_t = \frac{1}{t-1} Z_t dt + \sqrt{\frac{1+t}{1-t}}\, dW_t.$$

Using the integrating factor,

$$d\left(\frac{1}{1-t} Z_t\right) = \frac{1}{1-t} \sqrt{\frac{1+t}{1-t}}\, dW_t,$$

so

$$\frac{1}{1-t} Z_t - \frac{1}{1-0} X = \int_0^t \frac{1}{1-s} \sqrt{\frac{1+s}{1-s}}\, dW_s.$$

Hence

$$Z_t = (1-t)X + (1-t) \int_0^t \frac{1}{1-s} \sqrt{\frac{1+s}{1-s}}\, dW_s.$$

Therefore

$$Z_t \mid X \sim \mathcal{N}\left((1-t)X,\, \mathbb{E}\left[(1-t) \int_0^t \frac{1}{1-s} \sqrt{\frac{1+s}{1-s}}\, dW_s \right]^2 I_d \right).$$

Compute the variance:

$$\begin{aligned} \mathbb{E}\left[(1-t) \int_0^t \frac{1}{1-s} \sqrt{\frac{1+s}{1-s}}\, dW_s \right]^2 &= (1-t)^2 \int_0^t \frac{1+s}{(1-s)^3}\, ds \\ &= (1-t)^2 \int_0^t \frac{1+s}{(1-s)^3}\, ds \\ &= t. \end{aligned}$$

So

$$Z_t \mid X \sim \mathcal{N}\big((1-t)X, tI_d\big).$$

In particular,

$$Z_0 = X, \qquad Z_1 \mid X \sim \mathcal{N}(0, I_d).$$

Step 2.

Step 1: Forward SDE and probability flow representation. Consider the general Itô SDE

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

with diffusion tensor

$$a(x,t) := \sigma(x,t)\sigma(x,t)^\top.$$

The corresponding probability flow ODE (PFODE) is

$$\frac{dX_t}{dt} = b(X_t, t) - \frac{1}{2}\left[a(X_t, t)\,\nabla \log p_t(X_t) + \nabla \cdot a(X_t, t)\right].$$

Both the SDE and the PFODE induce the same marginal density evolution $p_t(x)$. Write p as the PDF of Z_t. Then the PDF of Z_t satisfies the Fokker–Planck equation. Write the Fokker–Planck equation in the form of a transport PDE:

$$\frac{\partial p}{\partial t} = -\nabla \cdot (\mu p) + \frac{1}{2}\frac{1+t}{1-t}\Delta p.$$

Using

$$\Delta p = \nabla \cdot \big(S(p,t)p\big), \qquad S(p,t) = \nabla \log p(\cdot, t),$$

this becomes

$$\frac{\partial p}{\partial t} = -\nabla \cdot \left(\frac{1}{t-1} x\, p - \frac{1}{2}\frac{1+t}{1-t} S(p,t)p\right).$$

Hence, by Liouville's theorem, p is also the PDF of the PF-ODE solution $\widetilde{Z}_t$ satisfying

$$\begin{cases} \dfrac{d\widetilde{Z}_t}{dt} = \dfrac{1}{t-1}\widetilde{Z}_t - \dfrac{1}{2}\dfrac{1+t}{1-t} S(\widetilde{Z}_t, t), \\ \widetilde{Z}_0 = X. \end{cases}$$

Step 3: Identification of the velocity field. Comparing the Fokker–Planck equation with the continuity equation, we identify the velocity field

$$v(x,t) = b(x,t) - \frac{1}{2}\left[a(x,t)\,\nabla \log p_t(x) + \nabla \cdot a(x,t)\right].$$

Thus, the PFODE provides an explicit realization of the PDODE.

In particular, the dynamics depend on the score function $\nabla \log p_t(x)$, which plays a central role in generative diffusion models. **Step 2 (as written at the end of the notes).** Evaluate the score function based on the given sample of X.

Step 3. NN approximation (Monte Carlo / kernel approximation are mentioned in the notes), and then generate the samples of X by solving the PF-ODE.