Part 1: Generative Models & Differential Equations (Week 11)

1. ODE vs. SDE Formulation

Generative modeling can be viewed as transforming a simple distribution (prior) into a complex data distribution via a continuous time process.

- **Ordinary Differential Equation (ODE):** Deterministic flow.

$$\frac{dx_t}{dt} = u_t(x_t)$$

  - **Existence & Uniqueness:** Guaranteed by the **Picard-Lindelöf Theorem** if u_t is Lipschitz continuous.
  - **Numerical Solution (Euler Method):**

$$x_{t+h} = x_t + h \cdot u_t(x_t)$$

- **Stochastic Differential Equation (SDE):** Adds noise (diffusion).

$$dx_t = f(x_t, t)dt + g(t)dW_t$$

  - $f(x, t)$: Drift coefficient.
  - $g(t)$: Diffusion coefficient.
  - W_t: Wiener process (Brownian motion), where $W_{t+h} - W_t \sim \mathcal{N}(0, hI)$.
  - **Numerical Solution (Euler-Maruyama):**

$$x_{t+h} = x_t + f(x_t, t)h + g(t)\sqrt{h} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

2. Deriving the Fokker-Planck Equation

The Fokker-Planck equation describes the time evolution of the probability density function $p_t(x)$ of particles following an SDE.

Setup: Consider a test function $\phi(x)$ (smooth, compact support).
Using Taylor expansion for the SDE increment $\Delta x \approx f\Delta t + g\Delta W$:

$$\phi(x_{t+\Delta t}) \approx \phi(x_t) + \phi'(x_t)\Delta x + \frac{1}{2}\phi''(x_t)(\Delta x)^2$$

Taking expectations (Note: $E[\Delta W] = 0$ and $E[(\Delta W)^2] = \Delta t$):

$$\frac{E[\phi(x_{t+\Delta t})] - E[\phi(x_t)]}{\Delta t} \approx E\left[\phi'(x)f(x,t) + \frac{1}{2}\phi''(x)g^2(t)\right]$$

By definition of expectation $\int \phi(x)p_t(x)dx$ and integration by parts (assuming boundary terms vanish):

1. $\int \phi' f p dx = -\int \phi \nabla \cdot (fp) dx$
2. $\int \phi'' g^2 p dx = \int \phi \nabla^2 (g^2 p) dx$

Equating the time derivative $\partial_t p$:

$$\frac{\partial p_t(x)}{\partial t} = -\nabla \cdot [f(x,t)p_t(x)] + \frac{1}{2}g(t)^2 \nabla^2 p_t(x)$$

Part 2: The Reverse SDE & Score Matching (Week 12)

1. The Reverse SDE

To generate data, we must reverse the diffusion process (from noise to data).
For a forward SDE $dx = f(x,t)dt + g(t)dW_t$, the **Reverse SDE** flows backward in time (from T to 0):

$$d\bar{x}_t = \left[f(\bar{x}_t, t) - g(t)^2 \nabla_x \log p_t(\bar{x}_t)\right] dt + g(t)d\bar{W}_t$$

- dt here represents a negative time step.
- $\nabla_x \log p_t(x)$ is the **Score Function**.

2. Score Matching Objectives

Since $p_t(x)$ is unknown, we train a neural network $s_\theta(x,t)$ to approximate the score $\nabla \log p_t(x)$.

A. Explicit Score Matching (Fisher Divergence):

$$\mathcal{L}(\theta) = E_{p_{data}}\left[\frac{1}{2}||s_\theta(x)||^2 + \mathrm{Tr}(\nabla_x s_\theta(x))\right]$$

- *Issue:* Calculating $\text{Tr}(\nabla s_\theta)$ (Jacobian trace) is computationally expensive ($O(d^2)$).
- *Hutchinson's Trace Estimator:* Approximate trace using random vectors $v \sim \mathcal{N}(0, I)$:

$$\text{Tr}(A) = E_v[v^T A v]$$

B. Denoising Score Matching (DSM):
Instead of matching the intractable $\nabla \log p_t(x)$, we match the conditional score $\nabla \log p_{t|0}(x_t|x_0)$, which is analytically known (Gaussian).

$$\mathcal{L}_{DSM}(\theta) = E_{x_0, x_t}\left[||s_\theta(x_t, t) - \nabla_{x_t} \log p_{t|0}(x_t|x_0)||^2\right]$$

For a Gaussian perturbation $x_t|x_0 \sim \mathcal{N}(\mu_t, \sigma_t^2 I)$:

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

Part 3: Tweedie's Formula & Signal Recovery (Week 13)

1. Tweedie's Formula Derivation

Problem: Recover the clean signal X from a noisy observation $Y = X + \delta Z$, where $Z \sim \mathcal{N}(0, I)$ and δ is noise level.

The marginal density $p_Y(y)$ is the convolution of p_X and the noise kernel ϕ_δ:

$$p_Y(y) = \int p_X(x) \frac{1}{(2\pi\delta^2)^{d/2}} \exp\left(-\frac{||y - x||^2}{2\delta^2}\right) dx$$

Differentiating $\log p_Y(y)$ with respect to y:

$$\nabla \log p_Y(y) = \frac{\nabla p_Y(y)}{p_Y(y)} = \frac{1}{p_Y(y)} \int p_X(x) \frac{\partial}{\partial y} \phi_\delta(y - x) dx$$

Using $\frac{\partial}{\partial y}\phi_\delta(y - x) = -\frac{y-x}{\delta^2}\phi_\delta(y - x)$:

$$\nabla \log p_Y(y) = \frac{1}{p_Y(y)} \int -\frac{y - x}{\delta^2} p_X(x) \phi_\delta(y - x) dx$$

$$\nabla \log p_Y(y) = -\frac{1}{\delta^2}\left(y - E[X|Y=y]\right)$$

Result (Tweedie's Formula):

$$E[X|Y=y] = y + \delta^2 \nabla \log p_Y(y)$$

This confirms that the score function points towards the mean of the clean data posterior.

2. Bayesian Interpretation of Reconstruction

The posterior $p(x|y)$ combines the prior and likelihood:

$$p(x|y) \propto p(y|x)p(x) \approx \mathcal{N}(y, \delta^2 I) \cdot p(x)$$

Using Taylor expansion on the prior $\log p_x(x)$ around y:

$$\log p(x|y) \approx -\frac{||y-x||^2}{2\delta^2} + \langle \nabla \log p_x(y), x - y \rangle + C$$

Completing the square reveals the posterior is approximately Gaussian with mean:

$$\mu_{post} = y + \delta^2 \nabla \log p_x(y)$$

Part 4: The Ornstein-Uhlenbeck (OU) Process (Week 11 & 14)

1. SDE Formulation

$$dX_t = -\beta X_t dt + \delta dW_t$$

- Mean reversion term: $-\beta X_t$.
- Diffusion term: δ.

2. Analytical Solution

Use the integrating factor $e^{\beta t}$. Apply Ito's Product Rule to $Y_t = e^{\beta t} X_t$:

$$d(e^{\beta t} X_t) = (\beta e^{\beta t} X_t)dt + e^{\beta t}(-\beta X_t dt + \delta dW_t)$$

$$d(Y_t) = \delta e^{\beta t} dW_t$$

Integrate from 0 to t:

$$e^{\beta t} X_t - X_0 = \int_0^t \delta e^{\beta s} dW_s$$

$$X_t = e^{-\beta t} X_0 + \delta \int_0^t e^{-\beta(t-s)} dW_s$$

3. Moments (Mean & Variance)

- **Mean:** $E[X_t] = e^{-\beta t} x_0$ (since expectation of Ito integral is 0).
- **Variance:** Using Ito Isometry $E[(\int f(t) dW_t)^2] = \int f(t)^2 dt$:

$$\mathrm{Var}(X_t) = \delta^2 e^{-2\beta t} \int_0^t e^{2\beta s} ds$$

$$\mathrm{Var}(X_t) = \delta^2 e^{-2\beta t} \left[\frac{e^{2\beta t} - 1}{2\beta} \right] = \frac{\delta^2}{2\beta}(1 - e^{-2\beta t})$$

- **Stationary Distribution ($t \to \infty$):**

$$X_\infty \sim \mathcal{N}\left(0, \frac{\delta^2}{2\beta} I\right)$$

Part 5: DDPM (Denoising Diffusion

Probabilistic Models) (Week 14)

1. Discrete Forward Process

$$x_t = \sqrt{1-\beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon_t$$

Using notation $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

2. Training Objective

The Evidence Lower Bound (ELBO) leads to a simplified MSE loss between the actual noise ϵ and the predicted noise ϵ_θ:

$$\mathcal{L}_{\text{simple}} = E_{t, x_0, \epsilon} \left[||\epsilon - \epsilon_\theta(x_t, t)||^2 \right]$$

Connection to Score Matching:
From the forward pass parameterization:

$$\nabla_{x_t} \log p_{t|0}(x_t|x_0) = -\frac{\epsilon}{\sqrt{1-\bar{\alpha}_t}}$$

Thus, the network predicts the (scaled) score:

$$\epsilon_\theta(x_t, t) \approx -\sqrt{1-\bar{\alpha}_t} \nabla_{x_t} \log p_t(x_t)$$

3. Discrete Reverse Step

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z$$

- This update rule is derived by matching the posterior $q(x_{t-1}|x_t, x_0)$.

4. SDE Limit (Variance Preserving)

As steps $T \to \infty$, the discrete DDPM converges to the following SDE:

- **Forward:**

$$dx_t = -\frac{1}{2}\beta(t)x_t dt + \sqrt{\beta(t)}dW_t$$

- **Reverse:**

$$d\bar{x}_t = \left[-\frac{1}{2}\beta(t)\bar{x}_t - \beta(t)\nabla \log p_t(\bar{x}_t)\right] dt + \sqrt{\beta(t)}d\bar{W}_t$$