1. Generative Models via Differential Equations

Concept: Generative modeling can be viewed as transforming a simple prior distribution (e.g., Gaussian) into a complex data distribution through a continuous time process.

A. ODE Flow Models

We can define a deterministic path for data generation using an Ordinary Differential Equation (ODE):

$$\frac{dX_t}{dt} = u_t(X_t)$$

- **Generation:** Solve the ODE starting from $X_0 \sim p_{\text{prior}}$ to obtain $X_T \sim p_{\text{data}}$.
- **Existence:** By the **Picard-Lindelöf Theorem**, if the velocity field u_t is Lipschitz continuous, a unique solution exists.
- **Simulation (Euler Method):**

$$X_{t+h} = X_t + hu_t(X_t)$$

B. Stochastic Differential Equations (SDE)

To add randomness, we introduce a diffusion term (Brownian motion).

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

- $f(X_t, t)$: **Drift** coefficient (deterministic force).
- $g(t)$: **Diffusion** coefficient (noise scale).
- W_t: Wiener process (Brownian motion), where $W_{t+h} - W_t \sim \mathcal{N}(0, hI)$.

Simulation (Euler-Maruyama Method):
Unlike standard calculus, stochastic calculus requires specific discretization.

$$X_{t+h} = X_t + hf(X_t, t) + g(t)\sqrt{h}Z, \quad \text{where } Z \sim \mathcal{N}(0, I)$$

2. The Fokker-Planck Equation (FPE)

The FPE describes the time evolution of the probability density function $p_t(x)$ of a particle moving according to an SDE.

Theorem:
For the SDE $dX_t = f(X_t, t)dt + g(t)dW_t$, the density $p_t(x)$ satisfies:

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

Proof Sketch (1D case):

1. Consider the time evolution of the expectation of a test function $\phi(x)$:

$$\partial_t \mathbb{E}[\phi(X_t)] = \lim_{\Delta t \to 0} \frac{\mathbb{E}[\phi(X_{t+\Delta t})] - \mathbb{E}[\phi(X_t)]}{\Delta t}$$

2. Using Taylor expansion on $\phi(X_{t+\Delta t})$ and applying Ito's rules ($\mathbb{E}[\Delta W] = 0, \mathbb{E}[\Delta W^2] = \Delta t$):

$$\partial_t \mathbb{E}[\phi(X_t)] = \mathbb{E}\left[\phi'(X_t)f(X_t,t) + \frac{1}{2}\phi''(X_t)g(t)^2\right]$$

3. Express expectations as integrals against $p_t(x)$:

$$\int \phi(x)\partial_t p_t dx = \int \left(\phi'(x)fp_t + \frac{1}{2}\phi''(x)g^2 p_t\right) dx$$

4. Apply integration by parts (assuming boundary terms vanish) to move derivatives from ϕ to p_t:

$$\int \phi(x)\partial_t p_t dx = \int \phi(x) \left(-\partial_x(fp_t) + \frac{1}{2}g^2\partial_x^2 p_t\right) dx$$

5. Since this holds for any ϕ, the terms inside the integral must be equal.

3. The Ornstein-Uhlenbeck (OU) Process

A fundamental SDE used in diffusion models to corrupt data into noise.

Equation:

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

- Drift: Pulls X_t toward 0 (mean reversion).
- Diffusion: Adds constant noise.

Exact Solution:
Using the integrating factor $e^{\beta t}$:

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

Distributional Properties:
Given X_0, the conditional distribution is Gaussian:

$$X_t | X_0 \sim \mathcal{N}\left(X_0 e^{-\beta t}, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t}) I \right)$$

Derivation of Variance (Ito Isometry):

$$\mathrm{Var}(X_t) = \mathrm{Var}\left(\sigma e^{-\beta t} \int_0^t e^{\beta s} dW_s \right)$$

Using Ito Isometry $\mathbb{E}[(\int f(t) dW_t)^2] = \mathbb{E}[\int f(t)^2 dt]$:

$$= \sigma^2 e^{-2\beta t} \int_0^t e^{2\beta s} ds = \sigma^2 e^{-2\beta t} \left[\frac{e^{2\beta s}}{2\beta} \right]_0^t = \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})$$

- As $t \to \infty$, the distribution converges to the stationary distribution $\mathcal{N}(0, \frac{\sigma^2}{2\beta} I)$.

4. Reverse Time SDE and Score Matching

To generate data, we must reverse the diffusion process.

A. Reverse SDE Formulation

If the forward process is $dX_t = f(X_t, t)dt + g(t)dW_t$, the **reverse time SDE** (running from T to 0) is given by **Anderson's Theorem**:

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

- **Key Insight:** To simulate backward, we need the **Score Function**: $\nabla_x \log p_t(x)$.
- We replace the unknown score with a neural network approximation $s_\theta(x, t) \approx \nabla \log p_t(x)$.

B. Score Matching Objectives

How do we train $s_\theta(x, t)$?

1. Explicit Score Matching (Intractable):
Minimizing $\mathbb{E}[\|s_\theta(x) - \nabla \log p(x)\|^2]$ requires knowing $\nabla \log p(x)$, which is unknown.

2. Denoising Score Matching (DSM):
Instead of the true score, we match the conditional score given the clean data X_0:

$$\mathcal{L}_{DSM}(\theta) = \mathbb{E}_{X_0, X_t}\left[\|s_\theta(X_t, t) - \nabla_{X_t} \log p(X_t|X_0)\|^2\right]$$

- Since $p(X_t|X_0)$ is Gaussian (e.g., in OU process), $\nabla \log p(X_t|X_0)$ is easily calculable:

$$\nabla_{X_t} \log p(X_t|X_0) = -\frac{X_t - \mu_t(X_0)}{\Sigma_t}$$

3. Sliced Score Matching (SSM):
Used when $p(X_t|X_0)$ is unknown. It uses integration by parts to avoid the true score.

$$\mathcal{L}_{SSM}(\theta) = \mathbb{E}_{X_t, v}\left[v^T \nabla_x s_\theta(X_t, t) v + \frac{1}{2}\|s_\theta(X_t, t)\|^2\right]$$

- Involves the Jacobian trace $\text{Tr}(\nabla_x s_\theta)$, estimated efficiently using random projection vectors v (Hutchinson's trick).

5. Tweedie's Formula

This formula connects **denoising** (estimating the clean signal) to **score matching**.

Setup:
Observe a noisy signal $Y = X + \delta Z$, where $X \sim p_X$ (clean data) and $Z \sim \mathcal{N}(0, I)$ (noise).

The Formula:
The posterior mean (optimal denoised estimate) is:

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

Derivation Sketch:

1. Write $p_Y(y) = \int p_{Y|X}(y|x) p_X(x) dx$.
2. Use the Gaussian property: $\nabla_y p_{Y|X}(y|x) = -\frac{y-x}{\delta^2} p_{Y|X}(y|x)$.
3. Compute $\nabla p_Y(y)$ using differentiation under the integral.
4. Rearrange to find that $\frac{\nabla p_Y(y)}{p_Y(y)} = \frac{1}{\delta^2}(\mathbb{E}[X|Y = y] - y)$.

Implication for Diffusion:

- **Score = Denoising Error.** The score $\nabla \log p_Y(y)$ is proportional to the residual $(X - Y)$.
- This explains why predicting the noise ϵ in DDPM is equivalent to learning the score function.

6. Denoising Diffusion Probabilistic Models (DDPM)

DDPMs are discrete-time approximations of the underlying SDEs.

A. Forward Process

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

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

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

B. Reverse Process

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

The mean μ_θ is parameterized to predict the noise ϵ_θ:

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

C. Training Objective

The Variational Lower Bound simplifies to a Mean Squared Error (MSE) on the noise vectors:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x_0, \epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right]$$

D. Connection to SDEs

- **Forward Limit:** As the number of steps $T \to \infty$, the discrete DDPM process converges to the continuous Variance Preserving (VP) SDE:

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

- **Reverse Limit:** The reverse update step corresponds to the Reverse SDE solver.
- **Tweedie's Relation:** The score network s_θ and noise network ϵ_θ are related by:

$$s_\theta(x_t, t) \approx -\frac{\epsilon_\theta(x_t, t)}{\sqrt{1 - \bar{\alpha}_t}}$$