Lecture 01 - Introduction

① Main focus:

The Principles of diffusion models

1. forward process:

   data (x) → intermediate distributions → noise (z)

2. reverse process: (main goal)

   noise (z) → intermediate distributions → data (x)

Three ways: (regression problem)

① Variational View (VAE) chp2 → Fokker-Planck Equ.

② Score-based View (EBMs) chp3

③ Flow-based View (Normalizing Flows) chp.5

- A learned time-dependent velocity field whose flow transports a simple prior to the data.
- Sampling = Solving a differential equation.

Part A: Introduction to DGMs (chp. 1)

Part B: Foundations of DMs (chp. 2-7)

Part C: Sampling of DMs (chp 8-9)

Part D: Learning fast generators (chp. 10-11)

②. Deep Generative Modeling

learn a probability distribution p_ϕ.

$p_\phi \approx p_{data}$ via $D_{tr} := \{x_i\}_{i=1}^{N} \sim p_{data}$

measure : $D(p_{data}, p_\phi)$ using D_{tr}.

- $D_{tr} := \{x_1, x_2, \cdots, x_N\}$: i.i.d p_{data}
- p_{data} is intractable
- $p_\phi \approx p_{data}$ via

  $\phi^* \in \arg\min_\theta D(p_{data}, p_\phi)$ s.t.

  $p_{\phi^*}(x) \approx p_{data}(x)$

choices of D :

(1). KL-Divergence

$$D_{KL}(p_{data} \| p_\phi) = \int p_{data}(x) \cdot \log \frac{p_{data}(x)}{p_\phi(x)} dx$$

$$= \mathbb{E}_{x \sim p_{data}}[\log p_{data}(x) - \log p_\phi(x)]$$

- $D_{KL}(p \| q) \neq D_{KL}(q, p)$

$p = (0.5, 0.5)$. $q = (0.9, 0.1)$

$$\begin{cases} KL(p\|q) = 0.511 \\ KL(q\|p) = 0.367 \end{cases}$$

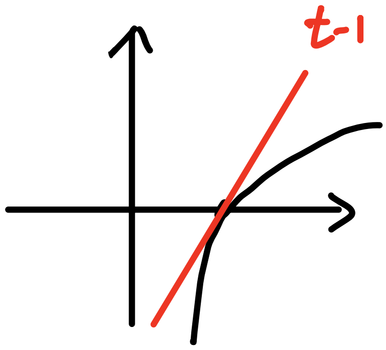
- $D_{KL}(p \| q) \geq 0$
- $D_{KL}(p \| q) = 0$ iff $p = q$.

Gibb's Inequality:

$$\begin{cases} D_{KL}(p\|q) \geq 0 \\ D_{KL}(p\|q) = 0 \text{ iff } p(x) = q(x) \text{ for all } x. \end{cases}$$

proof: $\forall t \geq 0,\ \log t \leq t-1$

$\log t = t-1$ iff $t=1$



$t = \frac{q(x)}{p(x)}$

$$D_{KL} = \sum_x p(x) \cdot \left(-\log \frac{q(x)}{p(x)}\right)$$

$$-\log t \geq 1-t \iff -\log \frac{q(x)}{p(x)} \geq 1 - \frac{q(x)}{p(x)}$$

$$\Rightarrow p(x) \cdot \left(-\log \frac{q(x)}{p(x)}\right) \geq p(x) - q(x)$$

$$\Rightarrow \sum_x p(x) \cdot \left(-\log \frac{q(x)}{p(x)}\right) \geq \sum_x \left(p(x) - q(x)\right)$$

$$= \sum_x p(x) - \sum_x q(x)$$

$$= 0$$

$\frac{q(x)}{p(x)} = 1 \iff p(x) = q(x). \iff D_{KL}(p\|q) = 0.$ ∎

(2). Fisher-Divergence:

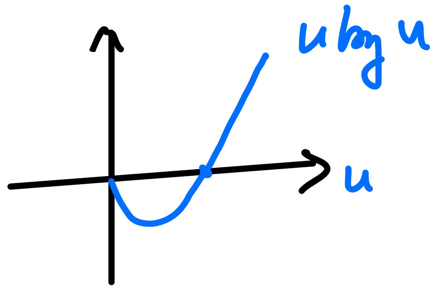
$$D_F(p_{data} \| p_\phi) = \mathbb{E}_{x \sim p_{data}}\left[\| \underbrace{\nabla_x \log p_{data}(x)}_{\text{score fun.}} - \underbrace{\nabla_x \log p_\phi(x)}_{\text{score fun.}} \|_2^2\right]$$

· score matching

[3]. f-divergence:

$$D_f(p \| q) = \int q(x)\, f\left(\frac{p(x)}{q(x)}\right) dx \,, \quad f(1) = 0.$$

① $f(u) = u \log u$, $D_f = D_{KL}$



② Jensen-Shannon divergence:

$$D_{JS}(p \| q) = \frac{1}{2} D_{KL}(p \| m) + \frac{1}{2} D_{KL}(q \| m)$$

$m = \frac{1}{2}(p+q)$

- $0 \le D_{JS} \le 1$: $\log_2$
- $D_{JS}(p \| q) \le \frac{1}{2} \| p - q \|_1$

③ $f(u) = \frac{1}{2}|u-1|$, total variational distance

About $p_\phi(x)$:

① $p_\phi(x) \geq 0$

② $\int p_\phi(x)\, dx = 1$.

$E_\phi(x)$: neural network

$x \in \mathbb{R}^d$, $E_\phi(x)$

$\Rightarrow E_\phi(x)$: $\exp(E_\phi(x)) := \tilde{p}_\phi(x)$

$$Z(\phi) = \int \exp(E_\phi(x'))\, dx'$$

$$\Rightarrow p_\phi(x) = \frac{\tilde{p}_\phi(x)}{Z(\phi)}$$

③ Examples of DGMs: trade-offs
1. tractability
2. expressiveness
3. training efficiency

(1). EBMs: $E_\phi(x): \mathbb{R}^d \to \mathbb{R}$

$$p_\phi(x) = \frac{\exp(-E_\phi(x))}{Z(\phi)}. \quad Z(\phi) = \int \exp(-E_\phi(x))\, dx$$

Score-base models:

$$\begin{cases} p(x) = e^{f(x)} & \int p(x)\,dx = 1 \\ q(x) = e^{g(x)} & \int q(x)\,dx = 1 \end{cases}$$

if $\nabla f = \nabla g \Rightarrow f \equiv g$.

Proof: $f(x) = g(x) + c$

$$e^{f(x)} = e^{g(x)+c} = e^c e^{g(x)}$$

$$1 = \int e^{f(x)} dx = \int e^c e^{g(x)} dx = e^c \int e^{g(x)} dx$$
$$= e^c \cdot 1 = e^c$$

We must have $c = 0$.

Score function: $s(x) = \nabla_x \log p(x)$ $\quad s: \mathbb{R}^d \to \mathbb{R}^d$.

$$p(x) = \frac{\tilde{p}(x)}{Z(\phi)}$$

$$\nabla_x \log p(x) = \nabla_x \log \tilde{p}(x) - \underbrace{\nabla_x \log Z}_{=0}$$

$$= \nabla_x \log \tilde{p}(x)$$

(2) AR-model:

$$P_\phi(x) = \prod_{i=1}^{D} P_\phi(x_i \mid x_{<i})$$

- $x = (x_1, x_2, \cdots, x_D)$ sequence data
- generation is slow
- struggle with high-resolution images
- weak in denoising tasks
- worse for image / video generation.
- Enforce a fixed ordering

But, • works great for text data.

- no need to compute $Z(\phi)$.

(3) VAE : learn : P_ϕ by decomposing into two parts.

- encoder (inference network) :
  $q_\theta(z|x)$ $\quad x \xrightarrow{} \underset{q_\theta(z|x)}{\text{encoder}} \xrightarrow{} z$
  (original : $P_\phi(z|x)$)
- Decoder (generator) :
  $P_\phi(x|z)$ $\quad z \xrightarrow{} \underset{P_\phi(x|z)}{\text{decoder}} \xrightarrow{} x'$

$\Rightarrow$ core idea : maximize a lower bound of $\log P_\phi(x)$.

Evidence Lower Bound (ELBO) :

$$P_\phi(x) = \int P_\phi(x|z) \cdot p(z)\, dz$$

intractable

posterior : $$P_\phi(z|x) = \frac{P_\phi(x|z) \cdot p(z)}{P_\phi(x)} = \frac{P_\phi(x|z) \cdot p(z)}{\boxed{\int P_\phi(x, z')\, dz'}}$$

$$q_\theta(z|x) \approx p_\phi(z|x)$$

$$\log p_\phi(x) = \log \int p_\phi(x,z)\,dz$$

$$= \log \int q_\theta(z|x) \cdot \frac{p_\phi(x,z)}{q_\theta(z|x)}\,dz$$

$$= \log \mathbb{E}_{z\sim q_\theta(z|x)}\left[\frac{p_\phi(x,z)}{q_\theta(z|x)}\right]$$

// log $\geq \underbrace{\mathbb{E}_{z\sim q_\theta}\left[\log \frac{p_\phi(x,z)}{q_\theta(z|x)}\right]}_{\text{ELBO}}$

$$p_\phi(x,z) = p_\phi(x|z)\cdot p(z)$$

$$= \mathbb{E}_{z\sim q_\theta}\left[\log \frac{\boxed{p_\phi(x|z)}\cdot \boxed{p(z)}}{\boxed{q_\theta(z|x)}}\right]$$

$$= \underbrace{\boxed{\mathbb{E}_{z\sim q_\theta}[\log p_\phi(x|z)]}}_{\text{Reconstruction term}} - \underbrace{\boxed{D_{KL}(q_\theta(z|x) \,\|\, p(z))}}_{\text{Latent Regularization}}$$

(4) Normalizing flow:

$$\log p_\phi(x) = \log p(z) + \log\left|\det \frac{\partial f_\phi^{-1}(x)}{\partial x}\right|$$

(4) GAN:

$$\begin{cases} \text{generator}: G_\phi \\ \text{discriminator}: D_\delta \end{cases} \qquad z \rightarrow G_\phi(z) \rightarrow x'$$

4. Summary:

- Explicit models:

  $p_\phi(x) \sim$ via tractable or approximately tractable

  ARs, NFs, VAEs, DMs

  define $p_\phi(x)$ exactly or via a bound.

- Implicit models:

  specify a distribution via sampling:

  $$x = G_\phi(z), \quad z \sim p_{prior}$$

  $p_\phi(x)$: may not defined at all.

- NFs. ARs: objective MLE, tractable
- VAEs, DMs: objective ELBO, Bound / Approximate
- GANs: objective intractable, not directly modeled

Next topic: variational perspective

(VAEs → DDPMs)

① VAE

② DDPM

③ ODE / SDE