2. Variational Free Energy

2.1 生成モデル

Today

LLM と DM が 2大巨頭

ChatGPT系　　お絵描き系、

こんなことを考えています

x : 画像データ $(x \in \mathbb{R}^{W \times H \times 3})$

$p(x)$: $\mathbb{R}^{W \times H \times 3}$ 上 prob. dist.

これが手に入れば画像が生成できる

$\rightarrow$ $p(x)$ を $p_\theta(x)$ で近似

How?

$x_1, x_2, \cdots, x_N$: 画像 : N $10^3 \sim 10^6$

$$\operatorname*{argmax}_\theta \prod_i p_\theta(x_i)$$ ← ムリ!

潜在変数を用いる

z の prob. dist.

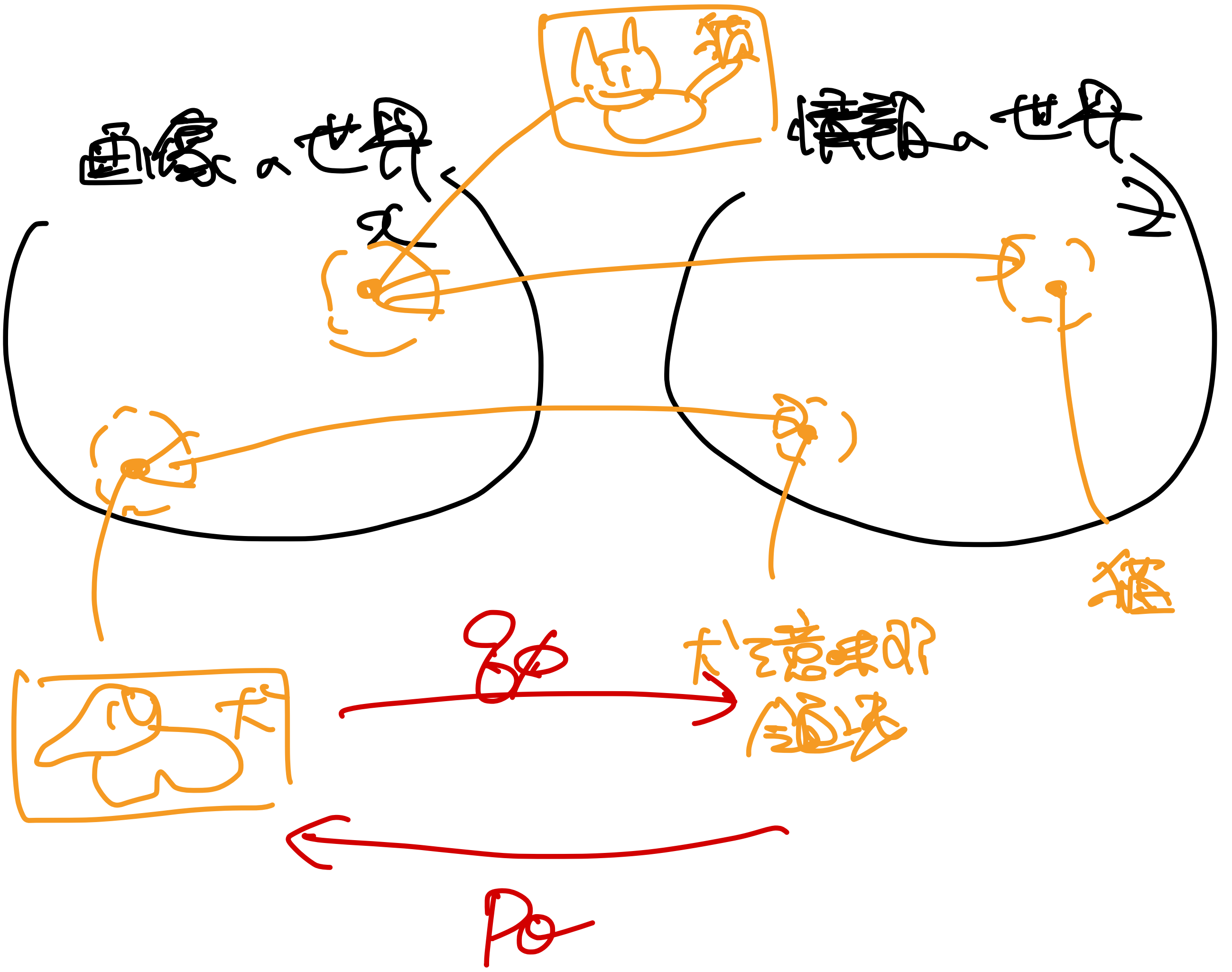
$z \sim p(z)$: latent var.

生成

$x \sim p_\theta(x|z)$: generative model

識別

$z \sim q_\phi(z|x)$: discriminative model



How to learn?

計算機で計算できない量の

$$\cancel{p_\theta}(x) = \int p_\theta(x|z)\,p(z)\,dz$$

$$\underset{\theta}{\mathrm{argmax}} \prod_i \cancel{p_\theta}(x_i) \left(= \prod_i \int p_\theta(x_i|z)\,p(z)\,dz \right)$$

$$\underset{\phi}{\mathrm{argmin}} \sum_i KL\left[q_\phi(z|x_i) \,\|\, \underline{\cancel{p_\theta}(z|x_i)} \right]$$

$$\frac{p_\theta(x_i, z)}{\int p_\theta(x_i, z)\,dz} = \frac{p_\theta(x_i|z)\,p(z)}{\int \text{—}\,dz}$$

impossible!

2.2. Variational Free Energy

__Def__

$$F_i = F_i(\theta, \phi, \{x_i\})$$

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

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

: Variational Free Energy

__Fact__

$\underset{\theta, \phi}{\arg\max}\, F_i$ なるθ,φを求めることができる

Stat. Phys に由来するらしい

by Hinton

Thm

$$F_i = \sum_i E_{q_\phi(z|x_i)}\left[\log \frac{p_\theta(x_i, z)}{q_\phi(z|x_i)}\right] \quad ① \text{ def}$$

$$= \sum_i \Big(\underbrace{\log p_{\hat\theta}(x_i)}_{\Sigma \log \prod_i p_{\hat\theta}(x_i)} - \underbrace{KL[q_\phi(z|x_i) \| p_{\hat\theta}(z|x_i)]}_{\text{最小化KL 即} = -KL} \Big) \quad ②$$

$$= \sum_i \Big(\underbrace{-KL[q_\phi(z|x) \| p(z)]}_{\text{正则化}} + \underbrace{E_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)]}_{q \text{ 与 } p \text{ 的相似性}} \Big) \quad ③$$

$$= \sum_i \Big(E_{q_\phi(z|x_i)}[\log p_\theta(x_i, z)] + \underbrace{E_{q_\phi(z|x_i)}[-\log q_\phi(z|x_i)]}_{Q \text{ 的熵(信息)}} \Big) \quad ④$$

prf

$$p_\theta(x_i, z) = p_\theta(x_i|z)\,p(z) = p_\theta^*(z|x_i)\,p_\theta^*(x_i)$$

正確?

($N=1$ とし $x_i = x$)

$$F_i = E_q\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \quad ①$$

④

$$= E_q[\log p_\theta(x,z)] + E_q[-\log q_\phi(z|x)]$$

$$= E_q\left[\log \frac{p_\theta(x_i|z)\,p(z)}{q_\phi(z|x)}\right]$$

③

$$= -\underbrace{E_q\left[\log \frac{q_\phi(z|x)}{p(z)}\right]}_{=KL[q_\phi(z|x)||p(z)]} + E_q[\log p_\theta(x_i|z)]$$

$$= E_q\left[\log \frac{p_\theta^*(z|x)\,p_\theta^*(x)}{q_\phi(z|x)}\right]$$

②

$$= p_\theta^*(x) - KL[q_\phi(z|x)||p_\theta^*(z|x)]$$

3. Diffusion Models

3.1 Diffusion Models

仮定

▷ q_ϕ と p_θ は多段階に分子

▷ q は人が指定 w/o param.

→ p_θ のみ学習

Diffusion Models

$x_i^{(0)}$: 画像データ

($x_i^{(T)}$: 潜在変数)

単位行列

$q(x_i^{(t)} | x_i^{(t-1)})$: 平均 $\sqrt{\alpha_t}\, x_i^{(t-1)}$
分散 $(1-\alpha_t) I$
(a 多変量正規分布)

$p_\theta(x_i^{(t-1)} | x_i^{(t)}, t)$: 平均 $\mu_\theta(x_i^{(t)}, t)$
分散 $C_t I$
(a 多変量正規分布)

こう設定して

$\hat{\theta} = \underset{\theta}{\arg\max}\, F_t$ を計算(学習)可 ← Possible!!

$$F_t = \sum_i \Big(-\underbrace{KL[q_\phi(z|x_i)\,||\,p(z)]}_{\text{indep. of } \theta} + \underbrace{E_{q_\phi}[\log p_\theta(x_i|z)]}_{\text{こっちを最大化すべし}} \Big)$$

↓めちゃ計算

$F_t = ($大量のKLの和$)$

↓更に計算

$$\underset{\theta}{\arg\min} \sum \| \mu_\theta(x_i^{(t)}, t) - \overbrace{\boxed{}}^{A_t x_i^{(0)} + B_t x_i^{(t)}} \|^2$$

or

$$\underset{\theta}{\arg\min} \sum \| \varepsilon_\theta(\underbrace{\boxed{}}_{\sqrt{\bar{\alpha}_t}\, x_i^{(0)} + \sqrt{1-\bar{\alpha}_t}\, \varepsilon}, t) - \varepsilon \|^2$$

4. SDEとの関係

4.1. SDE

"Def"

$$dx = f(x,t)dt + g(x,t)dw$$

ブラウン運動

: SDE

"Def"

$w(t)$: Brownian motion

"def" $\Longleftrightarrow$ $w(0) = 0$

$t_2 > t_1$ のとき

$w(t_2) - w(t_1)$: 平均0 分散 $t_2 - t_1$ の正規分布

$t_4 > t_3 \geq t_2 > t_1$ のとき

$w(t_4) - w(t_3)$ と $w(t_2) - w(t_1)$ は独立

Rem

SDEは、以下の $\Delta t \to 0$ の極限

$$x(0) = x_0$$

$$x(\Delta t) \fallingdotseq x(0) + f(x(0), 0)\Delta t + g(x(0), 0)\underbrace{\left(w(\Delta t) - w(0)\right)}_{\Delta w}$$

$$x(2\Delta t) \fallingdotseq x(\Delta t) + f(x(\Delta t), \Delta t)\Delta t + g(x(\Delta t), \Delta t)\left(w(2\Delta t) - w(\Delta t)\right)$$

$$\vdots$$

確率変数

4.2 DMとSDE

拡散過程

$$x_i^{(t)} = \sqrt{\alpha_t}\, x_i^{(t-1)} + \sqrt{1-\alpha_t}\, \varepsilon_t$$

↓ 連続時間

$f(t) = a$ なら
$x(t) = x_0 e^{-at}$

$$dx = \underline{-f(t)\,x\,dt} + \underbrace{g(t)\,dw}_{\varepsilon}$$

ノイズ付加。

DMでは、$x_i^{(t)}$ から $x_i^{(t-1)}$ を学習したい

時間反転だ!

Fact

$$dx = -f(t)x\,dt + g(t)\,dw$$

↓ 時間反転

SDEの解の時刻tでのxの分布

$$dx = \left(-f(t)x - \frac{1}{2}g(t)^2 \boxed{\nabla_x \log p_t(x)}\right)dt + g(t)\,d\bar{w}$$

score関数

wの時間反転

DMでは、コレを学習しているのと"等価"である