1. concept :

給一份資料集，我們想找出它的pdf $p(x;\theta)=\frac{1}{Z(\theta)}g(x;\theta)$
$X=\{x_1,\dots,x_n\}$

其中θ為參數，$Z(\theta)$為一 normalization term 使得 $\int\frac{1}{Z(\theta)}g(x;\theta)dx=1$

$\Rightarrow$找出最適參數 θ (MLE) 使估計的pdf $p(x;\theta)$ 與 真實pdf $p(x)$ 越相近

$\Rightarrow \theta_{mle}=\arg\max_{\theta}\sum_{i=1}^{n}\log p(x_i;\theta)$
$=\arg\max_{\theta}\sum_{i=1}^{n}\log(\frac{1}{Z(\theta)}g(x_i;\theta))$

但$Z(\theta)$不易求得 $\Rightarrow$ θ_{mle}難以計算出

使用 Score function $s(x)=\nabla_x\log(p(x;\theta))=\nabla_x\log(g(x;\theta))$ $\Rightarrow$ $s(x)$並不依賴$Z(\theta)$ (避免計算$Z(\theta)$)

2. how it is used in scored-based generative models

使用score matching :

目標為讓 $s(x)\approx\nabla_x\log p(x)$

$\Rightarrow\nabla_x\log p(x;\theta)\approx\nabla_x\log p(x)$

因此得出的 loss function $L_{ESM}(\theta)=E_{x\sim p(x)}\|s(x;\theta)-\nabla_x\log p(x)\|^2$ (Explicit score matching)

若知道 $p(x)$為何 $\Rightarrow$用$L_{ESM}(\theta)$作 loss function 訓練 $s(x)$

但大多數情況不知道 $p(x)$，而 $E_{x\sim p(x)}\|s(x;\theta)-\nabla_x\log p(x)\|^2=E_{x\sim p(x)}[\|s(x)\|^2+2\nabla\cdot s(x)]$
$+E_{x\sim p(x)}[\|\nabla_x\log p(x)\|^2]$

$\Rightarrow L_{ISM}(\theta)=E_{x\sim p(x)}[\|s(x;\theta)\|^2+2\nabla_x\cdot s(x;\theta)]$ (minimizing L_{ESM} and L_{ISM} are equivalent)

$\Rightarrow$ 訓練模型用L_{ISM}作 loss function，則無需知道 $p(x)$

Denoising score matching :

- x_0 : original data
- $p_0(x_0)$: data distribution of original data
- x : noisy data (by perturbing the original data)
- $p(x|x_0)$: conditional (noisy) data distribution
- $p_\sigma(x)$: (noisy) data distribution

And note that $p_\sigma(x)=\int_{\mathbb{R}^d}p(x|x_0)p_0(x_0)\,dx_0$

目標將 denoise score function $S_\sigma(x;\theta)=\nabla_x\log p_\sigma(x)$

使用 score matching :

The DSM loss $L_{DSM}(\theta)=E_{x_0\sim p_0(x_0)}E_{x|x_0\sim p(x|x_0)}[\|S_\sigma(x;\theta)-\nabla_x\log p(x|x_0)\|^2]$

(minimizing L_{DSM}, L_{ISM}, L_{ESM} are equivalent)

我們需找出 $\nabla_x\log p(x|x_0)$, by perturbing data by adding Gaussian noise

with isotropic variance

$\Rightarrow x=x_0+\varepsilon_\sigma$, $\varepsilon_\sigma\sim N(0,\sigma^2I)$
$=x_0+\sigma\varepsilon$, $\varepsilon\sim N(0,I)$

$\Rightarrow p(x|x_0)=\frac{1}{(2\pi)^{d/2}\sigma^d}e^{(-\frac{1}{2\sigma^2}\|x-x_0\|^2)}$

$\Rightarrow\nabla_x\log p(x|x_0)=-\frac{1}{\sigma^2}\varepsilon_\sigma$

Then rewrite the DSM loss $L_{DSM}(\theta)=E_{x_0\sim p_0(x_0)}E_{x|x_0\sim p_\sigma(x|x_0)}\frac{1}{\sigma^2}\|\sigma S_\sigma(x_0+\sigma\varepsilon;\theta)+\varepsilon\|^2$

透過此 loss function 訓練 $S_\sigma(x;\theta)$使之逼近 $\nabla_x\log_\sigma(x)$

Unanswered Question