1. Show that the sliced score matching (SSM) loss can also be written as

$$L_{SSM} = \mathbb{E}_{x\sim p(x)}\, \mathbb{E}_{v\sim p(v)}[\|v^T S(x;\theta)\|^2 + 2v^T \nabla_x\, (v^T S(x;\theta))].$$

from week 7-2 notes we have

$$L_{SSM}(\theta) = \mathbb{E}_{x\sim p(x)} \| S(x;\theta)\|^2 + \mathbb{E}_{x\sim p(x)} \mathbb{E}_{v\sim p(v)} [2v^T \nabla_x (v^T S(x;\theta))]$$

We want: $L_{SSM}(\theta) = \mathbb{E}_{x\sim p(x)} \mathbb{E}_{v\sim p(v)} [\| v^T S(x;\theta)\|^2 + 2v^T \nabla_x (v^T S(x;\theta))]$

$\Rightarrow$ we only need to prove that $\mathbb{E}_{v\sim p(v)} \| v^T S(x;\theta)\|^2 = \| S(x;\theta)\|^2$

fixed x, then $S(x;\theta) = (S_1, S_2, \cdots, S_d)$ with $S_1, \cdots, S_d$ are Constant

$\because v \in \mathbb{R}^d \Rightarrow v^T S(x;\theta)$ is a scalar

$$\Rightarrow \| v^T S(x;\theta)\|^2 = (v^T S(x;\theta))^2 = (v^T S(x;\theta))(v^T S(x;\theta))^T$$

$$= v^T (S(x;\theta) S(x;\theta)^T)\, v$$

By Hutchinson's trace estimator, $\mathbb{E}_{v\sim p(v)} \| v^T S(x;\theta)\|^2 = \mathbb{E}[v^T (S(x;\theta) S(x;\theta)^T)\, v]$

$$= tr(S(x;\theta) S(x;\theta)^T)$$

$$= S_1^2 + S_2^2 + \ldots + S_d^2$$

$$= \| S(x;\theta)\|^2$$

$\therefore$ we prove that $\mathbb{E}_{v\sim p(v)} \| v^T S(x;\theta)\|^2 = \| S(x;\theta)\|^2$

then we can derive that

$$L_{SSM}(\theta) = \mathbb{E}_{x\sim p(x)} \mathbb{E}_{v\sim p(v)} [\| v^T S(x;\theta)\|^2 + 2v^T \nabla_x (v^T S(x;\theta))] \quad \#$$

2.Briefly explain SDE

SDE 是在一個標準的微分方程基礎上，增加了一個隨機項，SDE 的通式如下:

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

而這個方程式在說明數據X_t在一段極小的時間dt內的微小變化dX_t，可以由 Drift Term 和 Diffusion Term 組成:

- Drift Term 為$f(x,t)dt$：若方程只有 Drift Term，也就是$dX_t = f(x,t)dt$，那就是一般常見的 ODE，$f(x,t)$代表的則是在不同時間的數據變化速率，相較後面提到的 Diffusion Term，給這個數據比較明確的變化趨勢。

- Diffusion Term 為$g(t)dW_t$：這一項則是建立在 ODE 的基礎上，再額外加入一些隨機噪音，$g(t)$指在時間t時，隨機噪音要放大或縮小多少，dW_t是 Wiener Process(or Brownian Motion) 的變量，也就是再dt時間內要加入的隨機噪音。

而在 Diffusion Model 中，加噪的過程就可以被視為一種 SDE，讓這些數據隨著時間持續且平滑的加入噪音，而解噪的過程(由起始噪音生成的過程)剛好也是一種 SDE，且解噪的 SDE 可以寫成

$$dX_t = [f(x,t) - g(t)^2 \nabla_x \log p_t(x)]dt + g(t)d\overline{W_t}$$

其中$f(x,t)$和$g(t)$是從加噪的過程就知道的，$\overline{W_t}$則是反向的 Wiener Process 同樣是隨機的，所以唯一要求的$\nabla_x \log p_t(x)$，剛好就是之前提到的 score function，所以只要搭配 SDE(連續平滑的特性)與 score function 就能有效訓練出能隨時調整輸出品質或輸出速度的模型，而不像 DDPM，若加噪的過程是 1000 次，生成圖片也需要花 1000 次的離散訓練限制。

3. Unanswered Questions

不同的加噪過程，會訓練出不同的解噪模型，那在不同的實務應用中，該如何選擇出適當的加噪過程?