1. Derivation of the Sliced Score Matching (SSM) loss

Setup:

- Let $\mathbf{x} \in \mathbb{R}^d$ be drawn from the data density $p(\mathbf{x})$.
- We have a model with score-function $S(\mathbf{x};\theta) \approx \nabla_{\mathbf{x}} \log\, p(\mathbf{x})$.
- The original (full) score matching loss is:

$$L_{SM} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\,[\frac{1}{2} \parallel S(\mathbf{x};\theta) \parallel_2^2 + \nabla_{\mathbf{x}} \cdot S(\mathbf{x};\theta)].$$

(Here $\nabla_{\mathbf{x}} \cdot S = \mathrm{tr}\,(\nabla_{\mathbf{x}} S)$.

Motivation for SSM:

- In high dimension d, computing the trace of the Jacobian (or Hessian) of the model score is costly.
- To avoid full-dimensional trace, choose a random direction vector $\mathbf{v} \sim p(\mathbf{v})$(for example isotropic Gaussian or uniform on sphere) and project the score onto $\mathbf{v}$.
- Use the property that for a matrix $A \in \mathbb{R}^{d\times d}$,

$$\mathbb{E}_{\mathbf{v}}[\mathbf{v}^T A\, \mathbf{v}] = \mathrm{tr}\,(A) \quad \text{(if } \mathbf{v} \text{ is appropriate).}$$

Derivation steps:

Start with the original loss written as

$$L_{SM} = \mathbb{E}_{\mathbf{x}}\,[\frac{1}{2} \parallel S(\mathbf{x};\theta) \parallel^2 + \mathrm{tr}\,\,(\nabla_{\mathbf{x}} S(\mathbf{x};\theta))].$$

Replace the $\mathrm{tr}\,(\nabla_{\mathbf{x}} S)$term by using a random vector $\mathbf{v}$:

$$\mathrm{tr}\,\,(\nabla_{\mathbf{x}} S) = \mathbb{E}_{\mathbf{v}}\,[\mathbf{v}^T\,(\nabla_{\mathbf{x}} S)\,\mathbf{v}]$$

(assuming $\mathbf{v}$is zero-mean isotropic).

Also project $S(\mathbf{x};\theta)$to $\mathbf{v}^T S(\mathbf{x};\theta)$. So define the following expectation:

$$L_{SSM} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\ \mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}[\frac{1}{2}(\mathbf{v}^T S(\mathbf{x};\theta))^2 + \mathbf{v}^T\ (\nabla_{\mathbf{x}} S(\mathbf{x};\theta))\ \mathbf{v}].$$

Recognize that

$$\mathbf{v}^T\ (\nabla_{\mathbf{x}} S(\mathbf{x};\theta))\ \mathbf{v} = \mathbf{v}^T \nabla_{\mathbf{x}}\ (\mathbf{v}^T S(\mathbf{x};\theta))$$

because $\mathbf{v}$is constant with respect to $\mathbf{x}$.

So the second term can be rewritten as

$$\mathbf{v}^T \nabla_{\mathbf{x}}\ (\mathbf{v}^T S(\mathbf{x};\theta)).$$

Multiply inside the expectation by 2 (preserving equivalence for optimization up to constant factor) to get the form:

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

That matches form:

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

2. Brief explanation of SDE (Stochastic Differential Equation)

Definition:

- A Stochastic Differential Equation (SDE) describes how a random variable (or vector) $\mathbf{x}_t$evolves over continuous time $t$under both deterministic drift and random diffusion.
- The general (“Itô”) form is

$$d\mathbf{x}_t = f(\mathbf{x}_t, t)\ dt\ +\ g(\mathbf{x}_t, t)\ d\mathbf{W}_t,$$

where:

- $f(\mathbf{x}_t, t)$ is the **drift** term (deterministic rate of change).
- $g(\mathbf{x}_t, t)$ is the **diffusion** coefficient (controls the strength of random perturbation).
- $\mathbf{W}_t$ is a (vector) Wiener process (Brownian motion).

Usage in generative models / score-based modelling:

- In score-based generative modelling, one defines a *forward SDE* that gradually adds noise to the data distribution (so that as t grows the distribution becomes nearly a simple reference, e.g., Gaussian).
- Then one defines or derives a *reverse-time SDE* that uses the model's estimated score function $\nabla_x \log p_t(x)$ to reverse the process and generate clean samples from noise.
- Benefits of using SDEs: continuous time formalism, flexibility in noise scheduling, theoretically rigorous connection to score matching.