Assignment 8

1. Derivation of the SSM form

Notation: $S(x;\theta) \in \mathbb{R}^d$ is the model score $s_\theta(x)$. Let $v \in \mathbb{R}^d$ be a random projection with distribution $p(v)$. We will choose $v \sim \mathcal{N}(0, I_d)$ (standard Gaussian), which gives the convenient identities $\mathbb{E}_v[v] = 0$ and $\mathbb{E}_v[vv^\top] = I_d$.

Start from the usual (Hyvärinen) score-matching integrand for a single x (up to an additive constant that does not depend on θ):

$$L_{SM} = \frac{1}{2}\|S(x;\theta)\|^2 + \nabla_x \cdot S(x;\theta).$$

Take expectation over $x \sim p(x)$:

$$L_{SM} = \mathbb{E}_{x\sim p(x)}\left[\frac{1}{2}\|S(x;\theta)\|^2 + \nabla_x \cdot S(x;\theta)\right].$$

Now express the two terms by averaging over $v \sim \mathcal{N}(0, I)$.

1. For the quadratic term

$$\mathbb{E}_v[(v^\top S)^2] = S^\top \mathbb{E}_v[vv^\top]S = S^\top I_d S = \|S\|^2.$$

2. For the divergence term, write

$$v^\top \nabla_x\big(v^\top S(x)\big) = \sum_{i,j} v_i v_j \frac{\partial S_j(x)}{\partial x_i}.$$

Taking expectation over v and using $\mathbb{E}_v\big[v_i v_j\big] = \delta_{ij}$ gives

$$\mathbb{E}_v[v^\top \nabla_x (v^\top S)] = \sum_i \frac{\partial S_i(x)}{\partial x_i} = \nabla_x \cdot S(x).$$

Putting these together we get, for each x,

$$\frac{1}{2}\|S(x;\theta)\|^2 + \nabla_x \cdot S(x) = \mathbb{E}_v\left[\frac{1}{2}\big(v^\top S(x)\big)^2 + v^\top \nabla_x\big(v^\top S(x)\big)\right].$$

Taking $\mathbb{E}_{x\sim p(x)}$ on both sides:

$$L_{SM} = \mathbb{E}_x \mathbb{E}_v\left[\frac{1}{2}(v^\top S)^2 + v^\top \nabla_x (v^\top S)\right].$$

Multiply both sides by 2 and define the sliced score matching loss $L_{SSM} \coloneqq 2L_{SM}$. Then

$$L_{SSM} = \mathbb{E}_x \mathbb{E}_v[(v^\top S(x;\theta))^2 + 2v^\top \nabla_x (v^\top S(x;\theta))],$$

which is exactly the expression you asked to show:

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

(Remarks: the factor choices depend on how one defines L_{SSM} relative to the standard SM loss; using $v \sim \mathcal{N}(0, I)$ gives the simple identities above. If a different $p(v)$ is used — e.g., unit sphere — scaling factors like $1/d$ appear.)

---

2. Brief explanation of an SDE

A **stochastic differential equation (SDE)** is like an ordinary differential equation (ODE) but with a random (stochastic) forcing term. In differential form it is usually written

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

Where

- X_t is the (random) state/process,
- $f(X_t, t)$ is the deterministic **drift** term (like the vector field in an ODE),
- $g(X_t, t)$ is the **diffusion** coefficient (controls noise amplitude),
- W_t is a Wiener process (standard Brownian motion); dW_t is the formal stochastic increment.

Key points, briefly:

- **Meaning of the differential:** dW_t is not an ordinary differential; integrals must be interpreted (Ito or Stratonovich). In the Ito interpretation (most common in ML), solutions are defined via stochastic integrals and have nice martingale properties.
- **Solution:** A solution is a stochastic process $(X_t)_{t \geq 0}$ satisfying the integral form

$$X_t = X_0 + \int_0^t f(X_s, s)\, ds + \int_0^t g(X_s, s)\, dW_s.$$

- **Drift vs diffusion:** f moves the mean trajectory; g injects randomness and spreads the distribution.
- **Probability density evolution:** The law $p(x, t)$ of X_t evolves according to a Fokker–Planck (forward Kolmogorov) PDE determined by f and g.
- **Numerical:** SDEs are simulated with schemes such as Euler–Maruyama (Ito analogue of Euler) or higher-order methods.

- **Why used in ML:** SDEs appear in generative modeling (diffusion models / score-based models) to describe how simple noise is transformed into complex data (or vice versa). The score of the density can be related to the drift of a reverse-time SDE, and score matching is used to learn that drift.

Example simple SDE (Ornstein–Uhlenbeck):

$$dX_t = -\lambda X_t dt + \sigma dW_t,$$

which has mean decaying exponentially and stationary Gaussian distribution.