Generative Models - Assignment 5

Mohammad Mohammadi - 402208592

January 27, 2025

Problem 1

Part A

In Theorem 1 of a paper, it is shown that the KL divergence bound between the distributions p and p_θ^{SDE} decomposes into two main components. One of these components is the cost function $\mathcal{J}_{SM}$, which is associated with the performance of the Score-Based Diffusion Model (SBDM).

Based on Theorem 1, the Kullback–Leibler bound is given by

$$D_{KL}(p \,\|\, p_\theta^{\mathrm{SDE}}) \;\le\; \mathcal{J}_{SM}\big(\theta; g(\cdot)^2\big) \;+\; D_{KL}\big(p_T \,\|\, \pi\big).$$

This bound shows that by decreasing $\mathcal{J}_{SM}$, the KL divergence bound becomes smaller, which implies improved model performance. In other words, the cost function $\mathcal{J}_{SM}$ acts as a principal factor in determining the likelihood bound.

This bound indicates that if the cost function $\mathcal{J}_{SM}$ is reduced, the distribution $p_{\theta_0}^{\mathrm{SDE}}$ becomes closer to the real distribution p. Such closeness translates into an increase in the likelihood of $p_{\theta_0}^{\mathrm{SDE}}$.

From a practical standpoint, the cost function $\mathcal{J}_{SM}$ plays an important role in improving the model's quality. The smaller the value of $\mathcal{J}_{SM}$, the better the performance of the score-based diffusion model. Hence, from a qualitative perspective, reducing $\mathcal{J}_{SM}$ means an improvement in the model's quality.

From a quantitative perspective, this bound clearly shows how reducing $\mathcal{J}_{SM}$ can tighten the KL divergence bound, thereby increasing the likelihood of the model. Therefore, we can say that this bound and the cost function $\mathcal{J}_{SM}$ serve the same ultimate goal, namely increasing the log-likelihood (or equivalently reducing D_{KL}), and thus they are aligned.

Part B

From Part (a), we know that

$$D_{KL}(p \,\|\, p_\theta^{\mathrm{SDE}}) \;\le\; \mathcal{J}_{SM}\big(\theta; g(\cdot)^2\big) \;+\; D_{KL}\big(p_T \,\|\, \pi\big).$$

Also, the KL divergence is defined as

$$D_{KL}(p \,\|\, p_\theta^{\mathrm{SDE}}) \;=\; \mathbb{E}_{p(x)}\Big[\log \frac{p(x)}{p_\theta^{\mathrm{SDE}}(x)}\Big] \;=\; \mathbb{E}_{p(x)}\big[\log p(x)\big] \;-\; \mathbb{E}_{p(x)}\big[\log p_\theta^{\mathrm{SDE}}(x)\big].$$

Hence,

$$\mathbb{E}_{p(x)}\big[\log p(x)\big] \;-\; \mathbb{E}_{p(x)}\big[\log p_\theta^{\mathrm{SDE}}(x)\big] \;\le\; \mathcal{J}_{SM}\big(\theta; g(\cdot)^2\big) \;+\; D_{KL}\big(p_T \,\|\, \pi\big).$$

Rearranging,

$$-\mathbb{E}_{p(x)}\big[\log p_\theta^{\mathrm{SDE}}(x)\big] \;\le\; -\mathbb{E}_{p(x)}\big[\log p(x)\big] \;+\; \mathcal{J}_{SM}\big(\theta; g(\cdot)^2\big) \;+\; D_{KL}\big(p_T \,\|\, \pi\big).$$

Given the stated entropy proposition:

$$H(p) \;=\; \mathcal{H}(p_T(x)) \;-\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[2\,\nabla\cdot f(x,t) \;+\; g(t)^2\,\|\nabla \log p_t(x)\|_2^2\Big]\,dt,$$

and noting that

$$H(p) \;=\; -\,\mathbb{E}_{p(x)}\big[\log p(x)\big],$$

we have

$$-\,\mathbb{E}_{p(x)}\big[\log p(x)\big] \;=\; \mathcal{H}\big(p_T(x)\big) \;-\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[2\,\nabla\cdot f(x,t) \;+\; g(t)^2\,\|\nabla \log p_t(x)\|_2^2\Big]dt.$$

Substituting this entropy expression into the previous inequality, we get

$$\begin{aligned}-\,\mathbb{E}_{p(x)}\big[\log p_\theta^{\mathrm{SDE}}(x)\big] \;\le \mathcal{H}\big(p_T(x)\big) \;-\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[2\,\nabla\cdot f(x,t) \;+\; g(t)^2\,\|\nabla \log p_t(x)\|_2^2\Big]dt \\ +\; \mathcal{J}_{SM}\big(\theta; g(\cdot)^2\big) \;+\; D_{KL}\big(p_T \,\|\, \pi\big).\end{aligned}$$

Since

$$\mathcal{H}\big(p_T(x)\big) \;=\; -\,\mathbb{E}_{p_T(x)}\big[\log p_T(x)\big],$$

we can write

$$\begin{aligned}-\,\mathbb{E}_{p(x)}\big[\log p_\theta^{\mathrm{SDE}}(x)\big] \;\le\; -\,\mathbb{E}_{p_T(x)}\big[\log \pi(x)\big] \;+\; \mathcal{J}_{SM}\big(\theta; g(\cdot)^2\big) - \\ \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[2\,\nabla\cdot f(x,t) \;+\; g(t)^2\,\|\nabla \log p_t(x)\|_2^2\Big]dt.\end{aligned}$$

The cost function $\mathcal{J}_{SM}$ is defined by

$$\begin{aligned}\mathcal{J}_{SM}\big(\theta; g(\cdot)^2\big) \;=\; \mathbb{E}_p\Big[\int_0^T \frac{g(t)^2}{2}\,\big\|s_\theta(x',t) - \nabla \log p_t(x')\big\|_2^2\,dt\Big] \;= \\ \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[g(t)^2\,\big\|s_\theta(x,t) - \nabla \log p_t(x)\big\|_2^2\Big]dt,\end{aligned}$$

where x' is a random variable drawn from $p_t(x)$. Combining all terms, we get

$$\begin{aligned}-\,\mathbb{E}_{p(x)}\big[\log p_\theta^{\mathrm{SDE}}(x)\big] \;\le\; -\;\mathbb{E}_{p_T(x)}\big[\log \pi(x)\big] \;+\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[g(t)^2\,\big\|s_\theta(x,t) - \nabla \log p_t(x)\big\|_2^2\Big]dt \\ -\;\frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[2\,\nabla\cdot f(x,t) \;+\; g(t)^2\,\|\nabla \log p_t(x)\|_2^2\Big]dt \;+\; D_{KL}\big(p_T \,\|\, \pi\big) \\ \le\; -\;\mathbb{E}_{p_T(x)}\big[\log \pi(x)\big] \;+\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[g(t)^2\,\big\|s_\theta(x,t) - \nabla \log p_t(x)\big\|_2^2 - \\ g(t)^2\,\|\nabla \log p_t(x)\|_2^2 \;-\; 2\,\nabla\cdot f(x,t)\Big]dt.\end{aligned}$$

This inequality is exactly the desired one.

Part C

Using the result of Part (b), we have

$$D_{KL}(p \,\|\, p_\theta^{\mathrm{SDE}}) \;-\; \mathbb{E}_{p(x)}\big[\log p(x)\big] \;\le\; -\; \mathbb{E}_{p_T(x)}\big[\log \pi(x)\big] \;+$$
$$\frac{1}{2}\int_0^T \mathbb{E}_{p_t(x'|x)\,p(x)}\Big[g(t)^2 \,\big\|s_\theta(x',t) - \nabla_{x'}\log p_t(x'|x)\big\|_2^2 \;\; -\; g(t)^2\,\|\nabla_{x'}\log p_t(x'|x)\|_2^2 \;-\; 2\,\nabla\cdot f(x',t)\Big]dt.$$

According to the given entropy proposition,

$$-\,\mathbb{E}_{p(x)}\big[\log p(x)\big] \;=\; \mathcal{H}\big(p_T(x)\big) \;-\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[2\,\nabla\cdot f(x,t) \;+\; g(t)^2\,\|\nabla\log p_t(x)\|_2^2\Big]dt.$$

Substituting this into the inequality from Part (a), we obtain

$$\begin{aligned}-\,\mathbb{E}_{p(x)}\big[\log p_\theta^{\mathrm{SDE}}(x)\big] \;\le \mathcal{H}\big(p_T(x)\big) \;-\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[2\,\nabla\cdot f(x,t) \;+\; g(t)^2\,\|\nabla\log p_t(x)\|_2^2\Big]dt \\ +\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x'|x)\,p(x)}\Big[g(t)^2\,\big\|s_\theta(x',t) - \nabla_{x'}\log p_t(x'|x)\big\|_2^2 \\ -\; g(t)^2\,\|\nabla_{x'}\log p_t(x'|x)\|_2^2 \\ -\; 2\,\nabla\cdot f(x',t)\Big]dt \;-\; \mathbb{E}_{p_T(x)}\big[\log\pi(x)\big].\end{aligned}$$

Recalling the definition of the cost function

$$\mathcal{J}_{SM}\big(\theta; g(\cdot)^2\big) \;=\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x)}\Big[g(t)^2\,\big\|s_\theta(x,t) - \nabla\log p_t(x)\big\|_2^2\Big]dt,$$

the inequality can be written as

$$\begin{aligned}-\,\mathbb{E}_{p(x)}\big[\log p_\theta^{\mathrm{SDE}}(x)\big] \;\le\; -\; \mathbb{E}_{p_T(x)}\big[\log\pi(x)\big] \;+\; \mathcal{J}_{SM}\big(\theta; g(\cdot)^2\big) \\ -\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x'|x)}\Big[g(t)^2\,\|\nabla\log p_t(x'|x)\|_2^2 \;+\; 2\,\nabla\cdot f(x',t)\Big]dt.\end{aligned}$$

Combining all these expressions, we obtain

$$-\,\log p_{\theta_0}^{\mathrm{SDE}}(x) \;=\; -\,\mathbb{E}_{\delta_x}\big[\log p_\theta^{\mathrm{SDE}}(x)\big] \;\le\; \mathcal{L}_\theta^{\mathrm{DSM}}(x),$$

where

$$\begin{aligned}\mathcal{L}_\theta^{\mathrm{DSM}}(x) \;=\; -\; \mathbb{E}_{p_T(x'|x)}\big[\log\pi(x')\big] \;+\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x'|x)}\Big[g(t)^2\,\big\|s_\theta(x',t) - \nabla_{x'}\log p_t(x'|x)\big\|_2^2\Big]\,dt \\ -\; \frac{1}{2}\int_0^T \mathbb{E}_{p_t(x'|x)}\Big[g(t)^2\,\|\nabla\log p_t(x'|x)\|_2^2 \;+\; 2\,\nabla\cdot f(x',t)\Big]\,dt.\end{aligned}$$

Thus, the desired inequality is fully established, showing that the cost function $\mathcal{L}_\theta^{\mathrm{DSM}}(x)$ provides an upper bound for $-\log p_{\theta_0}^{\mathrm{SDE}}(x)$.

Problem 2

In the four simulations that were conducted for sampling from a Gaussian Mixture Model (GMM) with 150 components, an initial set of 10^4 samples was generated from the distribution $N(0, I)$, and then the discrete Langevin equation was implemented under various configurations.

The simulations code and their interpretations are available in the file `problem2.ipynb`.

Problem 3

Part A

Diffusion models are recognized as a powerful tool, especially in the context of imputing missing time-series data. The paper CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation introduces a new method for imputing missing data that uses conditional diffusion models. This model is specifically designed to learn the conditional distribution and can utilize the information from available observations.

In this approach, there are two main processes: the **forward process** and the **reverse process**. The forward process gradually converts acceptable data into noisy data, while the reverse process helps reconstruct the original data from its noisy state. The primary objective here is to learn the conditional distribution

$$p(\mathbf{x}_{ta} \mid \mathbf{x}_{co}),$$

where $\mathbf{x}_{ta}$ represents the missing values and $\mathbf{x}_{co}$ represents the observed values.

For training the model, a self-supervised learning approach is employed, which is akin to masked language modeling. In this method, the observed values are split into two parts: one for imputation targets and the other for conditional observations. The model is then trained using these two parts so that it can accurately reconstruct the missing values.

Part B

In Section 4.3 of the paper, several strategies for creating training data are introduced:

- **Target Selection Strategy**: In this method, a subset of the observed values is chosen as imputation targets, and the remainder is used as conditional observations.
- **Random Strategy**: In this strategy, the observed values are chosen randomly to introduce greater variety into the training data.
- **Feature-based Strategy**: This method operates based on specific data characteristics and attempts to identify the features that have the greatest impact on the results.

Comparison of Strategies:

It appears that the **Target Selection Strategy** yields the greatest improvement on the final model, as it offers higher accuracy in capturing existing patterns.

Strategy	Advantages	Disadvantages
Target Selection	Higher imputation accuracy	May reduce diversity of the data
Random	High data diversity	May lead to invalid results
Feature-based	Focuses on important features	Requires deeper data analysis

Table 1: Comparison of strategies for selecting imputation targets

Part C

To evaluate the performance of data imputation models, various metrics are used:

- **Mean Absolute Error (MAE)**: This metric measures the mean of the absolute differences between the actual values and the predicted values, indicating the overall accuracy of the model.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|,$$

  where x_i is the actual data value and $\hat{x}_i$ is the predicted value for point i.

- **Continuous Ranked Probability Score (CRPS)**: This metric is used to assess the quality of the predicted distribution, indicating how accurately predictions account for uncertainty.

$$\text{CRPS}(F, x) = \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}_{\{z \geq x\}}\right)^2 dz,$$

  where $F(z)$ is the predicted distribution function, and $\mathbb{1}_{\{z \geq x\}}$ is the indicator function that is 1 if $z \geq x$ and 0 otherwise.

- **Relative Squared Error (RSE)**: This metric computes the ratio of the error relative to the mean of the actual values, indicating the model's relative performance.

According to the paper's description:

- CRPS is used to evaluate how well the predicted distribution matches the actual values under uncertainty, making it very suitable for probabilistic forecasting tasks.
- MAE focuses on the numerical accuracy in imputing missing data by measuring the mean error.
- **The Target Selection Strategy** has the greatest impact on improving the model's performance because it places more emphasis on accurate data imputation.