Advanced Machine Learning – Deep Generative Models Exercise Sheet 4

Generative Models: Denoising Diffusion

Problem 1: To train a diffusion model, we want to maximize the evidence lower bound of the data

$$\mathcal{L} = \mathop{\mathbb{E}}_{q_{\phi(\boldsymbol{x}_0)}} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0, \boldsymbol{z}_{1:N}) - \log q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_{1:N})\right] \leq \log p(\boldsymbol{x}_0).$$

where $q_{\phi(\boldsymbol{x}_0)}$ and $p_{\boldsymbol{\theta}}$ are the forward and reverse distributions as defined in the lecture. Show that the ELBO is equal to

$$\mathcal{L} = -\mathbb{KL}[q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_N) \mid p(\boldsymbol{z}_N)] - \sum_{n>1} \mathbb{KL}[q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_{n-1} \mid \boldsymbol{z}_n) \mid p_{\boldsymbol{\theta}}(\boldsymbol{z}_{n-1} \mid \boldsymbol{z}_n)] + \mathop{\mathbb{E}}_{q_{\phi(\boldsymbol{x}_0)}} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{z}_1)].$$

Hint: Make use of the Markov property / definition of $p_{\boldsymbol{\theta}}$ and $q_{\phi(\boldsymbol{x}_0)}$.

Problem 2: Given the variational distribution (diffusion process) from the lecture

$$q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_1) = \mathcal{N}(\sqrt{1-\beta_1}\boldsymbol{x}_0, \beta_1 \boldsymbol{I}), \quad 0 < \beta_1 < 1,$$
$$q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_n \mid \boldsymbol{z}_{n-1}) = \mathcal{N}(\sqrt{1-\beta_n}\boldsymbol{z}_{n-1}, \beta_n \boldsymbol{I}), \quad 0 < \beta_n < 1,$$

show that $q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_n)$ has the closed form

$$q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_n) = \mathcal{N}(\sqrt{\bar{\alpha}_n}\boldsymbol{x}_0, (1-\bar{\alpha}_n)\boldsymbol{I}), \quad \text{where } \alpha_n = 1 - \beta_n \text{ and } \bar{\alpha}_n = \prod_i^n \alpha_n.$$

Hint: Construct a sample of $q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_n)$ from a sample $\boldsymbol{z}_{n-1} \sim q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_{n-1})$.

Problem 3: Let $\boldsymbol{x}_n = \boldsymbol{x}_0 + \sigma_n \boldsymbol{\varepsilon}$ be the noising function with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\sigma_n > 0$ that we use in score matching to learn the data distribution. $p(\boldsymbol{x}_n \mid x_0)$ denotes the distribution of the noisy version $\boldsymbol{x}_n$ of $\boldsymbol{x}_0$.

a) Derive the conditional score $\nabla_{\boldsymbol{x}_n} \log p(\boldsymbol{x}_n \mid \boldsymbol{x}_0)$.

b) In the lecture, the model predicts the score $s_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n) \approx \nabla_{\boldsymbol{x}_n} \log p(\boldsymbol{x}_n \mid \boldsymbol{x}_0)$. Now we change (reparameterize) the model so that, instead of predicting the score, it predicts the noise $\boldsymbol{\varepsilon} \approx \boldsymbol{\varepsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n)$ added to $\boldsymbol{x}_0$. Derive how the score function $\nabla_{\boldsymbol{x}_n} \log p(\boldsymbol{x}_n \mid \boldsymbol{x}_0)$ can be expressed as a function $s(\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_n, n))$ of the noise estimate $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_n, n)$.