Machine Learning Group
Faculty IV – Electrical Engineering and Computer Science
Technische Universität Berlin

Exercise Sheet 4

Exercise 1: Global Optimality of the GAN objective (10 + 10 + 20 P)

In this exercise, we want to show that the global optimal solution for the minimax game

$$\min_G \max_D V(D,G) = \min_G \max_D \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{x}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{1}$$

for training Generative Adversarial Networks is that the data distribution gained from sampling from p_g is equal to the real data distribution p_{data}.

(a) Therefore, we first consider the optimal discriminator D for any given generator G. Show that for fixed G, the optimal discriminator D is

$$D^*_G(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \tag{2}$$

Hint: For any $(a,b) \in \mathbb{R}^2 \backslash \{0,0\}$, $y \in [0,1]$, the function $f(y,a,b) = a\log(y) + b\log(1-y)$ achieves its maximum at $\frac{a}{a+b}$.

$$\begin{aligned}
\operatorname*{argmax}_D V(G,D) &= \operatorname*{argmax}_D \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{x}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \\
&= \operatorname*{argmax}_D \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{g}}(\boldsymbol{x})}[\log(1 - D(x))] \\
&= \operatorname*{argmax}_D \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x}))dx + \int_{\boldsymbol{x}} p_{\text{g}}(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))dx \\
&= \operatorname*{argmax}_D \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x})) + p_g(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))dx \\
&= \operatorname*{argmax}_D p_{\text{data}} \log(D) + p_g \log(1 - D)) \\
&= \operatorname*{argmax}_D f(D, p_{data}, p_g)
\end{aligned}$$

(b) Show that the maximum $C(G) = \max_D V(G,D)$ of the training criterion can be reformulated to:

$$C(G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \tag{3}$$

$$\begin{aligned}
C(G) &= \max_D V(G,D) \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D^*_G(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{x}}}[\log(1 - D^*_G(G(\boldsymbol{z})))] \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D^*_G(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log(1 - D^*_G(\boldsymbol{x}))] \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right]
\end{aligned}$$

(c) Show that the global minimum of $C(G)$ is $C^* = -\log(4)$ and that reaching it is equivalent to $p_g = p_{\text{data}}$.
Hint: Use the fact, that the the Jensen Shannon Divergence $JSD(P\|Q) = \frac{1}{2}(KL(P\|M) + KL(Q\|M))$ is always positive.

$$\begin{aligned}
C(G) &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{2p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{2p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= -\log(4) + KL\left(p_{\text{data}} \| \frac{p_{\text{data}} + p_g}{2}\right) + KL\left(p_g \| \frac{p_{\text{data}} + p_g}{2}\right) \\
&= -\log(4) + 2 \cdot JSD(p_{\text{data}} \| p_g)
\end{aligned}$$

Exercise 2: Reformulating the loss function of diffusion models (20 P)

(a) Show that

$$L_{vlb} = \mathbb{E}_q \left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} \right]$$

can be reformulated to:

$$L_{vlb} = L_0 + L_1 + \ldots + L_{T-1} + L_T$$

where

$$\begin{aligned} L_0 &= -\log p_\theta(x_0 \mid x_1) \\ L_{t-1} &= D_{KL}(q(x_{t-1} \mid x_t, x_0) \| p_\theta(x_{t-1} \mid x_t)) \\ L_T &= D_{KL}(q(x_T \mid x_0) \| p(x_T)) \end{aligned}$$

with the help of the Markov assumption in Diffusion models.
Substituting $s_1 = -\log p(\mathbf{x}_T) - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}$.

$$\begin{aligned}
L_{vlb} &= \mathbb{E}_q \left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} \right] \\
&= \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \log \frac{\prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{\prod_{t=1}^T q(\mathbf{x}_t \mid \mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_1 \mid \mathbf{x}_0)} - \sum_{t=2}^T \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_t \mid \mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_q \left[s_1 - \sum_{t=2}^T \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \right] \\
&= \mathbb{E}_q \left[s_1 - \sum_{t=2}^T \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} - \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \right] \\
&= \mathbb{E}_q \left[s_2 - \sum_{t=2}^T \log q(\mathbf{x}_{t-1} \mid \mathbf{x}_0) + \sum_{t=2}^T \log q(\mathbf{x}_t \mid \mathbf{x}_0) \right] \\
&= \mathbb{E}_q \left[s_2 - \sum_{t=1}^{T-1} \log q(\mathbf{x}_t \mid \mathbf{x}_0) + \sum_{t=2}^T \log q(\mathbf{x}_t \mid \mathbf{x}_0) \right] \\
&= \mathbb{E}_q \left[s_2 - \log q(\mathbf{x}_1 \mid \mathbf{x}_0) - \sum_{t=2}^{T-1} \log q(\mathbf{x}_t \mid \mathbf{x}_0) + \sum_{t=2}^{T-1} \log q(\mathbf{x}_t \mid \mathbf{x}_0) + \log q(\mathbf{x}_T \mid \mathbf{x}_0) \right] \\
&= \mathbb{E}_q \left[s_2 - \log q(\mathbf{x}_1 \mid \mathbf{x}_0) + \log q(\mathbf{x}_T \mid \mathbf{x}_0) \right] \\
&= \mathbb{E}_q \left[\log p(\mathbf{x}_T) - \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_1 \mid \mathbf{x}_0)} - \sum_{t=2}^T \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} - \log q(\mathbf{x}_1 \mid \mathbf{x}_0) + \log q(\mathbf{x}_T \mid \mathbf{x}_0) \right] \\
&= \mathbb{E}_q \left[-\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T \mid \mathbf{x}_0)} - \sum_{t=2}^T \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} - \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1) \right] \\
&= D_{\mathrm{KL}}(q(\mathbf{x}_T \mid \mathbf{x}_0) \| p(\mathbf{x}_T)) + \sum_{t=2}^T D_{\mathrm{KL}}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)) \\
&\quad - \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)
\end{aligned}$$

Exercise 3: Programming (40 P)

Download the programming files on ISIS and follow the instructions.
Generate MNIST with diffusion model in PyTorch.