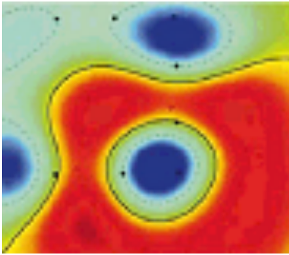
SoSe 2023

Deep Learning 2

Technische Universität Berlin



Solution
Exercise 4

Generative Models - part 2

Global Optimality of the Generator

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

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

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}.

Global Optimality of the Generator - Question a

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

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

Global Optimality of the Generator - Solution a

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 $\frac{a}{a+b}$.

$$
\begin{aligned}
\operatorname*{argmax}_D V(G, D) &= \cdots \\
&= \operatorname*{argmax}_D p_{\text{data}} \log(D) + p_g \log(1 - D)) \\
&= \operatorname*{argmax}_D f(D, p_{data}, p_g)
\end{aligned}
\tag{3}
$$

If we can show this equality we can thereby show, that $D_G^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}$

Global Optimality of the Generator - Solution a

Insert definition of $V(G,D)$:

$$\begin{aligned}
\operatorname*{argmax}_D V(G,D) &= \operatorname*{argmax}_D \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{x}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \\
&= \cdots \\
&= \operatorname*{argmax}_D p_{\text{data}} \log(D) + p_g \log(1-D)) \\
&= \operatorname*{argmax}_D f(D, p_{data}, p_g)
\end{aligned} \tag{4}$$

If we can show this equality we can thereby show, that $D^*_G(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}$

Global Optimality of the Generator - Solution a

Subsitute z with x in right hand term:

$$
\begin{aligned}
\underset{D}{\operatorname{argmax}}\, V(G, D) &= \underset{D}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}\,(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{x}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \\
&= \underset{D}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}\,(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{g}}(\boldsymbol{x})}[\log(1 - D(x))] \\
&= \cdots \\
&= \underset{D}{\operatorname{argmax}}\, p_{\text{data}}\, \log(D) + p_g \log(1 - D)) \\
&= \underset{D}{\operatorname{argmax}}\, f(D, p_{data}, p_g)
\end{aligned}
\tag{5}
$$

If we can show this equality we can thereby show, that $D_G^*(\boldsymbol{x}) = \frac{p_{\text{data}}\,(\boldsymbol{x})}{p_{\text{data}}\,(\boldsymbol{x}) + p_g(\boldsymbol{x})}$

Global Optimality of the Generator - Solution a

Definition of the expected value:

$$
\begin{aligned}
\underset{D}{\operatorname{argmax}}\, V(G,D) &= \underset{D}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{x}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))] \\
&= \underset{D}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{g}}(\boldsymbol{x})}[\log(1-D(x))] \\
&= \underset{D}{\operatorname{argmax}} \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x}))dx + \int_{\boldsymbol{x}} p_{\text{g}}(\boldsymbol{x}) \log(1-D(\boldsymbol{x}))dx \\
&= \cdots \\
&= \underset{D}{\operatorname{argmax}}\, p_{\text{data}} \log(D) + p_g \log(1-D)) \\
&= \underset{D}{\operatorname{argmax}}\, f(D, p_{data}, p_g)
\end{aligned}
\tag{6}
$$

If we can show this equality we can thereby show, that $D_G^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}$

Global Optimality of the Generator - Solution a

Merge Integrals:

$$\begin{aligned}
\underset{D}{\operatorname{argmax}}\, V(G,D) &= \underset{D}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{x}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))] \\
&= \underset{D}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{g}}(\boldsymbol{x})}[\log(1-D(x))] \\
&= \underset{D}{\operatorname{argmax}}\, \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x}))dx + \int_{\boldsymbol{x}} p_{\text{g}}(\boldsymbol{x}) \log(1-D(\boldsymbol{x}))dx \\
&= \underset{D}{\operatorname{argmax}}\, \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x})) + p_g(\boldsymbol{x}) \log(1-D(\boldsymbol{x}))dx \\
&= \cdots \\
&= \underset{D}{\operatorname{argmax}}\, p_{\text{data}} \log(D) + p_g \log(1-D)) \\
&= \underset{D}{\operatorname{argmax}}\, f(D, p_{data}, p_g)
\end{aligned} \tag{7}$$

If we can show this equality we can thereby show, that $D^*_G(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x})+p_g(\boldsymbol{x})}$

Global Optimality of the Generator - Solution a

Maximum of f does not depend on x:

$$
\begin{aligned}
\underset{D}{\operatorname{argmax}}\; V(G,D) &= \underset{D}{\operatorname{argmax}}\; \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{x}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))] \\
&= \underset{D}{\operatorname{argmax}}\; \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{g}}(\boldsymbol{x})}[\log(1-D(x))] \\
&= \underset{D}{\operatorname{argmax}} \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x}))dx + \int_{\boldsymbol{x}} p_{\text{g}}(\boldsymbol{x}) \log(1-D(\boldsymbol{x}))dx \\
&= \underset{D}{\operatorname{argmax}} \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x})) + p_g(\boldsymbol{x}) \log(1-D(\boldsymbol{x}))dx \\
&= \underset{D}{\operatorname{argmax}}\; p_{\text{data}} \log(D) + p_g \log(1-D)) \\
&= \underset{D}{\operatorname{argmax}}\; f(D, p_{data}, p_g)
\end{aligned}
\tag{8}
$$

If we can show this equality we can thereby show, that $D_G^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x})+p_g(\boldsymbol{x})}$

Global Optimality of the Generator - Question b

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

$$C(G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \tag{9}$$

Global Optimality of the Generator - Solution b

Proof Scheme:

$$
\begin{aligned}
C(G) &= \cdots \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right]
\end{aligned} \tag{10}
$$

Global Optimality of the Generator - Solution b

Definition $C(G)$:

$$
\begin{aligned}
C(G) &= \max_D V(G, D) \\
&= \cdots \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \qquad (11)
\end{aligned}
$$

Global Optimality of the Generator - Solution b

Definition $V(G,D)$:

$$\begin{aligned}
C(G) &= \max_D V(G,D) \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log D_G^*(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{x}}}\left[\log\left(1 - D_G^*(G(\boldsymbol{z}))\right)\right] \\
&= \cdots \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right]
\end{aligned} \tag{12}$$

Global Optimality of the Generator - Solution b

Substitute z with x:

$$
\begin{aligned}
C(G) &= \max_D V(G, D) \\
&= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \left[\log D_G^*(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{x}}} \left[\log \left(1 - D_G^*(G(\boldsymbol{z}))\right)\right] \\
&= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \left[\log D_G^*(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x} \sim p_g} \left[\log \left(1 - D_G^*(\boldsymbol{x})\right)\right] \\
&= \cdots \\
&= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x} \sim p_g} \left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right]
\end{aligned}
\tag{13}
$$

Global Optimality of the Generator - Solution b

Insert results from Exercise a:

$$
\begin{aligned}
C(G) &= \max_D V(G, D) \\
&= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \left[\log D^*_G(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{x}}} \left[\log\left(1 - D^*_G(G(\boldsymbol{z}))\right)\right] \\
&= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \left[\log D^*_G(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x} \sim p_g} \left[\log\left(1 - D^*_G(\boldsymbol{x})\right)\right] \\
&= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x} \sim p_g} \left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right]
\end{aligned} \tag{14}
$$

Global Optimality of the Generator - Question c

Show that the global minimum of $C(G)$ is $C^* = -\log(4)$ and that reaching it is equivalent to $p_g = p_{\text{data}}$.

Global Optimality of the Generator - Solution c

Proof idea:

$$
\begin{aligned}
C(G) &= \cdots \\
&= -\log(4) + c \cdot JSD\left(p_{\text{data}} \,\|\, p_g\right)
\end{aligned}
\tag{15}
$$

Since the Jensen-Shannon divergence between two distributions is always non-negative and zero only when they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ where $p_g = p_{\text{data}}$.

Global Optimality of the Generator - Solution c

Results from exercise b

$$
\begin{aligned}
C(G) &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= \cdots \\
&= -\log(4) + c \cdot JSD\left(p_{\text{data}} \| p_g\right)
\end{aligned}
\tag{16}
$$

Since the Jensen-Shannon divergence between two distributions is always non-negative and zero only when they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ where $p_g = p_{\text{data}}$.

Global Optimality of the Generator - Solution c

Add zero and split it into additive inverse terms:

$$
\begin{aligned}
C(G) &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x})+p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x})+p_g(\boldsymbol{x})}\right] \\
&= -\log(4) + \log(4) + \cdots \\
&= \cdots \\
&= -\log(4) + c \cdot JSD\left(p_{\text{data}} \| p_g\right)
\end{aligned}
\tag{17}
$$

Since the Jensen-Shannon divergence between two distributions is always non-negative and zero only when they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ where $p_g = p_{\text{data}}$.

Global Optimality of the Generator - Solution c

Reformulate log(4) with expected value and logarithm rules:

$$
\begin{aligned}
C(G) &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= -\log(4) + \log(4) + \cdots \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log 2] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log 2] + \cdots \\
&= \cdots \\
&= -\log(4) + c \cdot JSD\left(p_{\text{data}} \| p_g\right)
\end{aligned}
\tag{18}
$$

Since the Jensen-Shannon divergence between two distributions is always non-negative and zero only when they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ where $p_g = p_{\text{data}}$.

Global Optimality of the Generator - Solution c

Merge terms with the help of expected value and logarithm rules:

$$
\begin{aligned}
C(G) &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= -\log(4) + \log(4) + \cdots \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log 2] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log 2] + \cdots \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{2p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{2p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= \cdots \\
&= -\log(4) + c \cdot JSD\left(p_{\text{data}} \| p_g\right)
\end{aligned}
\tag{19}
$$

Since the Jensen-Shannon divergence between two distributions is always non-negative and zero only when they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ where $p_g = p_{\text{data}}$.

Global Optimality of the Generator - Solution c

Use known rules about KL divergences:

$$
\begin{aligned}
C(G) &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= -\log(4) + \log(4) + \cdots \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log 2] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log 2] + \cdots \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{2p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{2p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= -\log(4) + KL\left(p_{\text{data}} \,\Big\|\, \frac{p_{\text{data}} + p_g}{2}\right) + KL\left(p_g \,\Big\|\, \frac{p_{\text{data}} + p_g}{2}\right) \\
&= \cdots \\
&= -\log(4) + c \cdot JSD\left(p_{\text{data}} \,\|\, p_g\right)
\end{aligned} \tag{20}
$$

Since the Jensen-Shannon divergence between two distributions is always non-negative and zero only when they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ where $p_g = p_{\text{data}}$.

Global Optimality of the Generator - Solution c

Use definition of Jensen Shannon divergence:

$$
\begin{aligned}
C(G) &= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \right] + \mathbb{E}_{\boldsymbol{x} \sim p_g} \left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \right] \\
&= -\log(4) + \log(4) + \cdots \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} [\log 2] + \mathbb{E}_{\boldsymbol{x} \sim p_g} [\log 2] + \cdots \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \left[\log \frac{2 p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \right] + \mathbb{E}_{\boldsymbol{x} \sim p_g} \left[\log \frac{2 p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \right] \\
&= -\log(4) + KL\left(p_{\text{data}} \left\| \frac{p_{\text{data}} + p_g}{2} \right.\right) + KL\left(p_g \left\| \frac{p_{\text{data}} + p_g}{2} \right.\right) \\
&= -\log(4) + 2 \cdot JSD\left(p_{\text{data}} \| p_g \right) \\
&= \cdots \\
&= -\log(4) + c \cdot JSD\left(p_{\text{data}} \| p_g \right)
\end{aligned}
\tag{21}
$$

Since the Jensen-Shannon divergence between two distributions is always non-negative and zero only when they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ where $p_g = p_{\text{data}}$.

Global Optimality of the Generator - Solution c

Setting the constant $c = 2$ does not change anything about the results:

$$\begin{aligned}
C(G) &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= -\log(4) + \log(4) + \cdots \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log 2] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log 2] + \cdots \\
&= -\log(4) + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{2p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{2p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \\
&= -\log(4) + KL\left(p_{\text{data}} \,\|\, \frac{p_{\text{data}} + p_g}{2}\right) + KL\left(p_g \,\|\, \frac{p_{\text{data}} + p_g}{2}\right) \\
&= -\log(4) + 2 \cdot JSD\left(p_{\text{data}} \,\|\, p_g\right)
\end{aligned} \tag{22}$$

Since the Jensen-Shannon divergence between two distributions is always non-negative and zero only when they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ where $p_g = p_{\text{data}}$.

Reformulating Loss of Diffusion Model - Exercise 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.

Reformulating Loss of Diffusion Model - Solution A

We think about where we want to get:

$$
\begin{aligned}
L_{vlb} &= \mathbb{E}_q\left[-\log \frac{p_\theta\left(\mathbf{x}_{0:T}\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\right] \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right)+\sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&- \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned} \tag{23}
$$

Reformulating Loss of Diffusion Model - Solution A

Replacing the distributions with their definitions given our Markov assumption, we get

$$
\begin{aligned}
L_{vlb} &= \mathbb{E}_q\left[-\log \frac{p_\theta\left(\mathbf{x}_{0:T}\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\right] \\
&= \mathbb{E}_q\left[-\log p\left(\mathbf{x}_T\right) - \log \frac{\prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{\prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}\right] \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&- \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned}
\tag{24}
$$

Reformulating Loss of Diffusion Model - Solution A

We use log rules to transform the expression into a sum of logs, and then we pull out the first term

$$
\begin{aligned}
L_{vlb} &= \mathbb{E}_q \left[-\log \frac{p_\theta\left(\mathbf{x}_{0:T}\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)} \right] \\
&= \mathbb{E}_q \left[-\log p\left(\mathbf{x}_T\right) - \log \frac{\prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{\prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)} \right] \\
&= \mathbb{E}_q \left[-\log p\left(\mathbf{x}_T\right) - \log \frac{p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}{q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right)} - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)} \right] \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&- \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned}
\tag{25}
$$

Reformulating Loss of Diffusion Model - Solution A

Substituting $s_1 = -\log p(\mathbf{x}_T) - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}$ and using Bayes' Theorem and our Markov assumption on the rightmost term, this expression becomes

$$
\begin{aligned}
L_{vlb} &= \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\right] \\
&= \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \log \frac{\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{\prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})}\right] \\
&= \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_1 \mid \mathbf{x}_0)} - \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_t \mid \mathbf{x}_{t-1})}\right] \\
&= \mathbb{E}_q\left[s_1 - \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)}\right] \\
&= \cdots \\
&= D_{\mathrm{KL}}(q(\mathbf{x}_T \mid \mathbf{x}_0) \| p(\mathbf{x}_T)) + \sum_{t=2}^{T} D_{\mathrm{KL}}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)) \\
&\quad - \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)
\end{aligned} \tag{26}
$$

Reformulating Loss of Diffusion Model - Solution A

We then split up the right term using log rules

$$
\begin{aligned}
L_{vlb} &= \cdots \\
&= \mathbb{E}_q \left[-\log p\left(\mathbf{x}_T\right) - \log \frac{p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}{q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right)} - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)} \right] \\
&= \mathbb{E}_q \left[s_1 - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} \cdot \frac{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)} \right] \\
&= \mathbb{E}_q \left[s_1 - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} - \sum_{t=2}^{T} \log \frac{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)} \right] \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&\quad - \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned}
\tag{27}
$$

Reformulating Loss of Diffusion Model - Solution A

Substituting $s_2 = s_1 - \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}$ and splitting up the rightmost term with log rules:

$$
\begin{aligned}
L_{vlb} &= \cdots \\
&= \mathbb{E}_q\left[s_1 - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} \cdot \frac{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)}\right] \\
&= \mathbb{E}_q\left[s_1 - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} - \sum_{t=2}^{T} \log \frac{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)}\right] \\
&= \mathbb{E}_q\left[s_2 - \sum_{t=2}^{T} \log q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right) + \sum_{t=2}^{T} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)\right] \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&\quad - \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned} \tag{28}
$$

Reformulating Loss of Diffusion Model - Solution A

Substitute indices in middle term:

$$
\begin{aligned}
L_{vlb} &= \cdots \\
&= \mathbb{E}_q \left[s_1 - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} - \sum_{t=2}^{T} \log \frac{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)} \right] \\
&= \mathbb{E}_q \left[s_2 - \sum_{t=2}^{T} \log q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right) + \sum_{t=2}^{T} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) \right] \\
&= \mathbb{E}_q \left[s_2 - \sum_{t=1}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \sum_{t=2}^{T} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) \right] \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&- \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned}
\tag{29}
$$

Reformulating Loss of Diffusion Model - Solution A

Draw out edge indices:

$$
\begin{aligned}
L_{vlb} &= \cdots \\
&= \mathbb{E}_q\left[s_2 - \sum_{t=2}^{T} \log q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right) + \sum_{t=2}^{T} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)\right] \\
&= \mathbb{E}_q\left[s_2 - \sum_{t=1}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \sum_{t=2}^{T} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)\right] \\
&= \mathbb{E}_q\left[s_2 - \log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) - \sum_{t=2}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \sum_{t=2}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \log q\left(\mathbf{x}_T \mid \mathbf{x}_0\right.\right. \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&\quad - \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned}
\tag{30}
$$

Reformulating Loss of Diffusion Model - Solution A

Remove terms that cancel each other:

$$
\begin{aligned}
L_{vlb} &= \cdots \\
&= \mathbb{E}_q \left[s_2 - \sum_{t=1}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \sum_{t=2}^{T} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) \right] \\
&= \mathbb{E}_q \left[s_2 - \log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) - \sum_{t=2}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \sum_{t=2}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \log q\left(\mathbf{x}_T \mid \mathbf{x}_0 \right.\right. \\
&= \mathbb{E}_q \left[s_2 - \log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) + \log q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \right] \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&- \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned} \tag{31}
$$

Reformulating Loss of Diffusion Model - Solution A

Resubstitute $s_2 = \log p\left(\mathbf{x}_T\right) - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)} - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1}|\mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0\right)}$:

$$
\begin{aligned}
L_{vlb} &= \cdots \\
&= \mathbb{E}_q \left[s_2 - \log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) - \sum_{t=2}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \sum_{t=2}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \log q\left(\mathbf{x}_T \mid \mathbf{x}_0 \right. \right. \\
&= \mathbb{E}_q \left[s_2 - \log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) + \log q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \right] \\
&= \mathbb{E}_q \left[\log p\left(\mathbf{x}_T\right) - \log \frac{p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}{q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right)} - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} - \log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) + \mathrm{lo} \right. \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&- \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned}
\tag{32}
$$

Reformulating Loss of Diffusion Model - Solution A

Using log rules we rearrange:

$$
\begin{aligned}
L_{vlb} &= \cdots \\
&= \mathbb{E}_q \left[s_2 - \log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) + \log q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \right] \\
&= \mathbb{E}_q \left[\log p\left(\mathbf{x}_T\right) - \log \frac{p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}{q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right)} - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} - \log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) + \text{lo} \right. \\
&= \mathbb{E}_q \left[- \log \frac{p\left(\mathbf{x}_T\right)}{q\left(\mathbf{x}_T \mid \mathbf{x}_0\right)} - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} - \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right) \right] \\
&= \cdots \\
&= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
&- \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned}
\tag{33}
$$

Reformulating Loss of Diffusion Model - Solution A

Using known results about the KL divergence $D_{\mathrm{KL}}\left(p_1(x) \| p_2(x)\right)=\mathbb{E}_{x \sim p_1(x)}\left[\log \frac{p_1(x)}{p_2(x)}\right]$ $=\mathrm{E}_{x \sim p_1(x)}\left[-\log \frac{p_2(x)}{p_1(x)}\right]$ we arrive at the final result:

$$
\begin{aligned}
L_{vlb} & =\cdots \\
& =\mathbb{E}_q\left[s_2-\log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right)+\log q\left(\mathbf{x}_T \mid \mathbf{x}_0\right)\right] \\
& =\mathbb{E}_q\left[\log p\left(\mathbf{x}_T\right)-\log \frac{p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}{q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right)}-\sum_{t=2}^T \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)}-\log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right)+\log \right. \\
& =\mathbb{E}_q\left[-\log \frac{p\left(\mathbf{x}_T\right)}{q\left(\mathbf{x}_T \mid \mathbf{x}_0\right)}-\sum_{t=2}^T \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)}-\log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)\right] \\
& =D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right)+\sum_{t=2}^T D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \\
& -\log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)
\end{aligned} \tag{34}
$$

Appendix: MSE vs KL divergence of gaussians

Let's consider two Gaussian distributions, denoted by P and Q, with means μ_P and μ_Q and variances σ_P^2 and σ_Q^2 respectively. Our goal is to minimize the MSE between these two distributions, which is equivalent to minimizing the squared difference between their means.
The MSE between two Gaussian distributions is given by:

$$\mathrm{MSE} = \frac{1}{2}\left((\mu_P - \mu_Q)^2 + \sigma_P^2 + \sigma_Q^2\right)$$

Now, let's express the KL divergence between P and Q as a function of their means and variances. The KL divergence between two Gaussian distributions is given by:

$$\mathrm{KL}(P\|Q) = \frac{1}{2}\left(\log\left(\frac{\sigma_Q^2}{\sigma_P^2}\right) + \frac{\sigma_P^2}{\sigma_Q^2} + \left(\frac{(\mu_P - \mu_Q)^2}{\sigma_Q^2}\right) - 1\right)$$

Appendix: MSE vs KL divergence of gaussians

To prove that minimizing the MSE is equivalent to minimizing the KL divergence, we need to show that minimizing the MSE is equivalent to setting the gradient of the KL divergence with respect to the means μ_P and μ_Q to zero. Taking the partial derivative of the KL divergence with respect to μ_P and μ_Q, we have:

$$\frac{\partial \mathrm{KL}(P\|Q)}{\partial \mu_P} = \frac{\mu_P - \mu_Q}{\sigma_Q^2}$$

$$\frac{\partial \mathrm{KL}(P\|Q)}{\partial \mu_Q} = -\frac{\mu_P - \mu_Q}{\sigma_Q^2}$$

Appendix: MSE vs KL divergence of gaussians

Setting these partial derivatives to zero, we obtain:

$$\frac{\mu_P - \mu_Q}{\sigma_Q^2} = 0$$

$$\frac{\mu_Q - \mu_P}{\sigma_Q^2} = 0$$

Simplifying these equations, we find that $\mu_P = \mu_Q$. This implies that the means of the two Gaussian distributions are equal, which is achieved by minimizing the squared difference between their means, i.e., minimizing the MSE. Hence, we have shown that minimizing the MSE is equivalent to setting the gradient of the KL divergence with respect to the means to zero. This demonstrates that minimizing the MSE optimizes the KL divergence between two Gaussian distributions.

Appendix: Assistant Function for Exercise 1A

For any $(a, b) \in \mathbb{R}^2\backslash\{0, 0\}, y \in [0, 1]$, consider the function $f(y, a, b) = a \log(y) + b \log(1 - y)$.
To find the maximum of f, we take the derivative with respect to y and set it to zero.
Setting the derivative equal to zero, we get
$\frac{a}{y} - \frac{b}{1-y} = 0.$
Clearing the fractions, we have
$a(1 - y) - b(y) = 0$
Simplifying further, we find
$a - ay - by = 0.$
Rearranging the equation, we obtain
$a - (a + b)y = 0.$
Solving for y, we find
$y = \frac{a}{a+b}.$

Appendix: Assistant Function for Exercise 1A

To confirm that this point is a maximum, we compute the second derivative of f with respect to y.

$$\frac{d^2}{dy^2} f(y, a, b) = -\frac{a}{y^2} - \frac{b}{(1-y)^2}.$$

Evaluating the second derivative at $y = \frac{a}{a+b}$, we have

$$\frac{d^2}{dy^2} f\left(\frac{a}{a+b}, a, b\right) = -\frac{a}{\left(\frac{a}{a+b}\right)^2} - \frac{b}{\left(1 - \frac{a}{a+b}\right)^2} \quad = -\frac{(a+b)^2}{a} - \frac{b(a+b)^2}{a^2}.$$

Since a and b are nonzero, the second derivative is negative. This confirms that $y = \frac{a}{a+b}$ is a maximum for the function $f(y, a, b) = a\log(y) + b\log(1-y)$. Therefore, we have proved that for any $(a, b) \in \mathbb{R}^2 \backslash 0, 0$, the function $f(y, a, b) = a\log(y) + b\log(1-y)$ achieves its maximum at $y = \frac{a}{a+b}$.

Reformulating Loss of Diffusion Model - Solution A

The variational bound is equal to

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

Replacing the distributions with their definitions given our Markov assumption, we get

$$= \mathbb{E}_q \left[-\log p\left(\mathbf{x}_T\right) - \log \frac{\prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{\prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)} \right]$$

We use log rules to transform the expression into a sum of logs, and then we pull out the first term

$$= \mathbb{E}_q \left[-\log p\left(\mathbf{x}_T\right) - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)} - \log \frac{p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}{q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right)} \right]$$

Reformulating Loss of Diffusion Model - Solution A

Using Bayes' Theorem and our Markov assumption, this expression becomes

$$= \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} - \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_1 \mid \mathbf{x}_0)} \right]$$

We then split up the middle term using log rules

$$\begin{aligned} & - \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \\ = & - \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} - \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \end{aligned}$$

Reformulating Loss of Diffusion Model - Solution A

Isolating the second term, we see

$$
\begin{aligned}
& -\sum_{t=2}^{T} \log \frac{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)} \\
&= -\sum_{t=2}^{T} \log q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right) + \sum_{t=2}^{T} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) \\
&= -\sum_{t=1}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \sum_{t=2}^{T} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) \\
&= -\log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) - \sum_{t=2}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \sum_{t=2}^{T-1} \log q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) + \log q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \\
&= -\log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) + \log q\left(\mathbf{x}_T \mid \mathbf{x}_0\right)
\end{aligned}
$$

Plugging this back into our equation for Lvlb, we have

$$
\mathrm{L}_{vlb} = \mathbb{E}_q \left[-\log p\left(\mathbf{x}_T\right) - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} - \log q\left(\mathbf{x}_1 \mid \mathbf{x}_0\right) \right.
$$

$+\log q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) - \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_1 \mid \mathbf{x}_0)}$

Reformulating Loss of Diffusion Model - Solution A

Using log rules, we rearrange

$$= \mathbb{E}_q \left[-\log \frac{p\left(\mathbf{x}_T\right)}{q\left(\mathbf{x}_T \mid \mathbf{x}_0\right)} - \sum_{t=2}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)} - \log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right) \right]$$

Next, we note the following equivalence for the KL divergence for any two distributions:

$$= D_{\mathrm{KL}}\left(p_1(x) \| p_2(x)\right) = \int_{-\infty}^{\infty} p_1(x) \log \frac{p_1(x)}{p_2(x)} \mathrm{dx} = \mathbb{E}_{x \sim p_1(x)} \left[\log \frac{p_1(x)}{p_2(x)} \right]$$

$=\mathrm{E}_{p_1}\left[-\log \frac{p_2(x)}{p_1(x)} \right]$

Finally, applying this equivalence to the previous expression, we arrive at

$$= D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right) + \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)$$

-$\log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)$