Reinforcement Learning - Assignment 0

Ali Abbasi - 98105879

February 17, 2023

Contents

1 Variance of Estimator

1.1

We know that:

$$\begin{aligned} \operatorname{Var}(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ \implies \mathbb{E}\left[X^2\right] &= \operatorname{Var}(X) + \mathbb{E}[X]^2 \end{aligned}$$

Substituting X with $W - \theta$, we have:

$$\begin{aligned} \mathbb{E}_\theta\left[(W-\theta)^2\right] &= \operatorname{Var}_\theta(W-\theta) + \mathbb{E}_\theta[W-\theta]^2 \\ \implies \operatorname{MSE}(W,\theta) &= \operatorname{Var}_\theta(W-\theta) + \mathbb{E}_\theta[W-\theta]^2 \\ &= \operatorname{Var}_\theta(W-\theta) + (\mathbb{E}_\theta[W]-\theta)^2 && (\theta \text{ is a constant in the frequentist view}) \\ &= \operatorname{Var}_\theta(W) + (\mathbb{E}_\theta[W]-\theta)^2 && (\operatorname{Var}(X+a) = \operatorname{Var}(X) \text{ with constant } a) \\ &= \operatorname{Var}_\theta(W) + \operatorname{Bias}_\theta(W)^2 \end{aligned}$$

1.2

We show that its expected value is equal to θ.

$$\begin{aligned} \hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} f(X_i) \\ \implies \mathbb{E}\left[\hat{\theta}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(X_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n} \theta \\ &= \theta \end{aligned}$$

1.3

$$\begin{aligned} \theta_1 &= \mathbb{E}[f_1(X)] \\ &= \int_0^4 \frac{1}{4}\left(1 + (\frac{x}{2})^2\right) dx \\ &= \frac{7}{3} \end{aligned}$$

$$\begin{aligned}\theta_2 &= \mathbb{E}\left[f_2(X)\right] \\ &= \int_0^4 \frac{1}{4}(\frac{x}{4})^{10}dx \\ &= \frac{1}{11}\end{aligned}$$

```python
x1 = np.random.random(100) * 4
f1 = 1 + x1 * x1 / 4
theta1 = 7 / 3

bias1 = np.mean(f1) - theta1
var1 = np.var(f1)
print(f'Bias of first estimator: {bias1}')
print(f'Variance of first estimator: {var1}')
```

✓ 0.0s

```
Bias of first estimator: 0.15819413118115966
Variance of first estimator: 1.4260130224434493
```

```python
x2 = np.random.random(100) * 4
f2 = np.power(x2 / 4, 10)
theta2 = 1 / 11

bias2 = np.mean(f2) - theta2
var2 = np.var(f2)
print(f'Bias of second estimator: {bias2}')
print(f'Variance of second estimator: {var2}')
```

✓ 0.0s

```
Bias of second estimator: -0.03522630285572548
Variance of second estimator: 0.01939871576052484
```

Figure 1: Bias and Variance of two estimators.

The lower value of bias and variance of second estimator is because $\frac{x}{4}$ is between 0 and 1 and the power of 10 makes them much smaller. And function with smaller values will have smaller variance and bias.

1.4

Suppose we want to estimate $\theta = \mathbb{E}_p[f(x)]$. We can easily show that the expected value of the new estimator is equal to θ:

$$\begin{aligned}\mathbb{E}_q\left[\frac{p(x)}{q(x)}f(x)\right] &= \sum_x q(x)\frac{p(x)}{q(x)}f(x)\\ &= \sum_x p(x)f(x)\\ &= \mathbb{E}_p[f(x)]\end{aligned}$$

This trick is called 'change of measure'. The same argument can be applied to continuous random variables by simply using integral instead of sum.

Now we calculate the new estimator's variance:

$$\begin{aligned}\mathrm{Var}_q\left(\frac{p(x)}{q(x)}f(x)\right) &= \mathbb{E}_q\left[\frac{p^2(x)}{q^2(x)}f^2(x)\right] - \mathbb{E}_q\left[\frac{p(x)}{q(x)}f(x)\right]^2\\ &= \mathbb{E}_q\left[\frac{p^2(x)}{q^2(x)}f^2(x)\right] - \theta^2\\ &= \mathbb{E}_p\left[\frac{p(x)}{q(x)}f^2(x)\right] - \theta^2\end{aligned}$$

While the variance of previous estimator was:

$$\mathrm{Var}_p(f(x)) = \mathbb{E}_p\left[f^2(x)\right] - \theta^2$$

1.5

We will find the $q(x)$ that minimizes the variance of the new estimator, by using Lagrange multipliers method, because we have to constrain q to be a probability distribution.

$$\underset{q(x);\sum_x q(x)=1}{\arg\min}\ \mathbb{E}_p\left[\frac{p(x)}{q(x)}f^2(x)\right] - \theta^2 = \underset{q(x);\sum_x q(x)=1}{\arg\min}\ \sum_x \frac{p^2(x)}{q(x)}f^2(x)$$

$$\implies \mathcal{L} = \sum_x \frac{p^2(x)}{q(x)}f^2(x) + \lambda\left(\sum_x q(x) - 1\right)$$

The variables that the minimization is done with respect to, are $q(x)$ for each $x \in \mathcal{X}$ and λ. So we have:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial q(x_0)} &= -\frac{p^2(x_0)}{q^2(x_0)}f^2(x_0) + \lambda = 0\\ \implies q^2(x_0) &= \frac{1}{\lambda}p^2(x)f^2(x)\\ \implies q(x_0) &= \frac{1}{\sqrt{\lambda}}p(x)\,|f(x)|\end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \implies q(x_0) = \frac{p(x_0)\,|f(x_0)|}{\sum_{x'} p(x')\,|f(x')|}$$

$$\implies \forall x \in \mathcal{X}, q(x) = \frac{p(x)\,|f(x)|}{\sum_{x'} p(x')\,|f(x')|}$$

The same argument can be applied to continuous random variables with a bit of change.

1.6

Variance of estimator $I_{X>5}$:

$$\begin{aligned}
&p = \mathbb{E}_P\left[\mathbb{1}\{X > 5\}\right] \implies f(X) = \mathbb{1}\{X > 5\} \\
&p = \int_5^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&\quad \approx 2.9 \times 10^{-7} \mathrm{Var}_P\left(\mathbb{1}\{X > 5\}\right) && = \mathbb{E}_P\left[\mathbb{1}\{X > 5\}^2\right] - p^2 \\
&\quad = \mathbb{E}_P\left[\mathbb{1}\{X > 5\}\right] - p^2 \\
&\quad = p - p^2 \approx 2.9 \times 10^{-7}
\end{aligned}$$

Variance of estimator when data are generated from distribution $Q(x)$:

$$\begin{aligned}
\mathrm{Var}_Q\left(\mathbb{1}\{X > 5\}\right) &= \mathbb{E}_P\left[\frac{P(x)}{Q(x)}\mathbb{1}\{X > 5\}^2\right] - p^2 \\
&= \mathbb{E}_P\left[\frac{P(x)}{Q(x)}\mathbb{1}\{X > 5\}\right] - p^2 \\
&= \int_5^{+\infty} \frac{\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right)^2}{e^{-(x-5)}} dx - p^2 \\
&\approx 2.4 \times 10^{-13} - 8.4 \times 10^{-14} \approx 1.6 \times 10^{-13}
\end{aligned}$$

In this example we can see how a good choice of distribution can reduce the variance of the estimator.

2 Markov Chain

2.1

Assuming initial probability of states to be p, the probability of being in state i after n steps is:

$$\begin{aligned}
P(X_n = i) &= \sum_j P(X_n = i | X_0 = j) P(X_0 = j) \\
&= \sum_j P_{ji}^{(n)} p_j \\
&= (pP^{(n)})_i
\end{aligned}$$

So $P(X_n)$ only depends on $P^{(n)}$ and p.

2.2

Based on the definition of Markov chain, given X_i, X_{i+1} is independent of $X_0, X_1, \ldots, X_{i-1}$. For more simplicity, we define:

$$\begin{aligned} A &\triangleq \{X_{F_1}, X_{F_2}, \ldots, X_{F_n}\} \\ t &\triangleq t_1 + t_2 \\ m &\triangleq \max\{F_i\} \end{aligned}$$

We have:

$$\begin{aligned} P(X_t|A) &= \sum_{x_{t-1},\ldots,x_{m+1}} P(X_t, X_{t-1}, \ldots, X_{m+1}|A) \\ &= \sum_{x_{t-1},\ldots,x_{m+1}} P(X_t|X_{t-1}, \ldots, X_{m+1}, A)P(X_{t-1}|X_{t-2}, \ldots, X_{m+1}, A)\ldots P(X_{m+1}|A) \\ &= \sum_{x_{t-1},\ldots,x_{m+1}} P(X_t|X_{t-1})P(X_{t-1}|X_{t-2})\ldots P(X_{m+1}|X_m) \\ &= \sum_{x_{t-1},\ldots,x_{m+1}} P(X_t, X_{t-1}, \ldots, X_{m+1}|X_m) \\ &= P(X_t|X_m) \end{aligned}$$

2.3

$$\begin{aligned} P_{ij}^{(n+m)} &= P(X_{t+n+m} = j|X_t = i) \\ &= \sum_k P(X_{t+n+m} = j|X_{t+n} = k)P(X_{t+n} = k|X_t = i) \\ &= \sum_k P_{kj}^{(m)} P_{ik}^{(n)} \\ &= (P^{(n)}P^{(m)})_{ij} \\ \implies P^{(n+m)} &= P^{(n)}P^{(m)} \end{aligned}$$

2.4

Its steady state distribution will be $[\pi_1, \pi_2, \ldots, \pi_n]$ (π_is are row vectors). Because:

$$\begin{aligned} [\pi_1, \pi_2, \ldots, \pi_n] \begin{bmatrix} P_1 & 0 & \ldots & 0 \\ 0 & P_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & P_n \end{bmatrix} &= [\pi_1 P_1, \pi_2 P_2, \ldots, \pi_n P_n] \\ &= [\pi_1, \pi_2, \ldots, \pi_n] \end{aligned}$$

If there are several steady state distributions, behavior of the chain will depend on the initial state.

2.5

Assuming the initial distribution of states to be p, we have:

$$\begin{aligned}
\lim_{n\to\infty}(pP^{(n)})_i &= \lim_{n\to\infty}\sum_j p_j P_{ji}^{(n)} \\
&= \sum_j p_j \lim_{n\to\infty} P_{ji}^{(n)} \\
&= \sum_j p_j \pi_l(i) \\
&= \pi_l(i)\sum_j p_j \\
&= \pi_l(i)
\end{aligned}$$

$$\implies \lim_{n\to\infty}(pP^{(n)}) = \pi_l$$

This both shows that the π_l is the steady state distribution of this Markov chain based on the definition in the previous part, and also shows that the initial distribution of states doesn't affect the long term behavior of the chain. Because for any initial distribution p, we have the same distribution of states after a long time.

3 Information Theory

3.1

$$\begin{aligned}
&\forall x: \quad 0 \le P(x) \le 1 \implies \forall x: \quad \log P(x) \le 0 \\
\implies &\forall x: \quad P(x)\log P(x) \le 0 \\
\implies &\sum_x P(x)\log P(x) \le 0 \\
\implies &H(X) \ge 0
\end{aligned}$$

3.2

Intuitively, uniform distribution has the highest entropy, because it has the most uncertainty. We can compute the entropy of a discrete uniform distribution easily as follows:

Suppose U is a uniform discrete distribution that can obtain M distinct value.

$$\begin{aligned} H(U) &= -\sum_x P(x) \log P(x) \\ &= -\sum_x \frac{1}{M} \log \frac{1}{M} \\ &= \frac{1}{M} \log M \sum_x 1 \\ &= \log M \end{aligned}$$

And we can confirm our intuition by showing that for a random variable with M distinct values, the entropy is at most $\log M$:

$$\begin{aligned} H(X) &= \mathbb{E}_P \left[\log \frac{1}{P(x)} \right] \\ \xrightarrow[\text{using Jensen inequality}]{\log \text{ is concave}} &\leq \log \mathbb{E}_P \left[\frac{1}{P(x)} \right] \\ &= \log \sum_x P(x) \frac{1}{P(x)} \\ &= \log \sum_x 1 \\ &= \log M \end{aligned}$$

3.3

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} P(x,y) \log P(x,y) \\ &= -\sum_{x,y} P(x,y) \log P(x) P(y|x) \\ &= -\sum_{x,y} P(x,y) \log P(x) - \sum_{x,y} P(x,y) \log P(y|x) \\ &= \sum_x \log P(x) \sum_y P(x,y) - \sum_{x,y} P(x,y) \log P(y|x) \\ &= \sum_x P(x) \log P(x) - \sum_{x,y} P(x,y) \log P(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$

We can show that $H(X,Y) = H(Y,X) = H(Y) + H(X|Y)$ by using a similar approach.

3.4

$$
\begin{aligned}
-\mathrm{D}_{KL}\left(P\|Q\right) &= -\sum_x P(x)\log\frac{P(x)}{Q(x)} \\
&= \sum_x P(x)\log\frac{Q(x)}{P(x)} \\
&= \mathbb{E}_P\left[\log\frac{Q(x)}{P(x)}\right] \qquad \star \\
&\xrightarrow[\text{using Jensen inequality}]{\log \text{ is concave}} \leq \log\mathbb{E}_P\left[\frac{Q(x)}{P(x)}\right] \\
&= \log\sum_x P(x)\frac{Q(x)}{P(x)} \\
&= \log\sum_x Q(x) = \log 1 = 0 \\
\implies \mathrm{D}_{KL}\left(P\|Q\right) &\geq 0
\end{aligned}
$$

Moreover, we know that Jensen inequality becomes equality, if and only if the argument of the strictly convex (or concave) function is always constant. That means if $\mathrm{D}_{KL} = 0$, in equation ($\star$) we have: $\forall x : \frac{Q(x)}{P(x)} = \text{constant}$. Which implies that $\forall x : P(x) = Q(x)$.

(Of course, showing one side of the above statement, i.e., KL divergence of identical distributions being zero was easy:

$$
\begin{aligned}
\mathrm{D}_{KL}\left(P\|P\right) &= \sum_x P(x)\log\frac{P(x)}{P(x)} \\
&= \sum_x P(x)\log 1 \\
&= 0)
\end{aligned}
$$

3.5

We can show that $\mathrm{D}_{KL}\left(P\|U\right) = -H(P) + \log M$.

$$
\begin{aligned}
\mathrm{D}_{KL}\left(P\|U\right) &= \sum_x P(x)\log\frac{P(x)}{\frac{1}{M}} \\
&= \sum_x P(x)\log P(x) + \sum_X P(x)\log M \\
&= -H(P) + \log M\sum_x P(x) \\
&= -H(P) + \log M
\end{aligned}
$$

Instead of what we did in Section 3.2, we could have used this equivalence to show that for every

distribution P with M possible values, $H(P) \leq \log M$:

$$\begin{aligned} & \mathrm{D}_{KL}\left(P \| U\right) \geq 0 \\ \implies & -H(P) + \log M \geq 0 \\ \implies & H(P) \leq \log M \end{aligned}$$

And by using the argument in last part for $\mathrm{D}_{KL} = 0$, we can show that:

$$H(P) = \log M \text{ iff } \mathrm{D}_{KL}\left(P \| U\right) = 0 \text{ iff } P = U$$

3.6

First we show the following property of Mutual Information and Entropy:

$$\begin{aligned} I(X;Y) &= \mathrm{D}_{KL}\left(P(X,Y) \| P(X)P(Y)\right) \\ &= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\ &= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(y)} - \sum_{x,y} P(x,y) \log P(x) \\ &= \sum_{x,y} P(x,y) \log P(x|y) - \sum_{x,y} P(x,y) \log P(x) \\ &= \sum_{x,y} P(x,y) \log P(x|y) - \sum_{x} \log P(x) \sum_{y} P(x,y) \\ &= \sum_{x,y} P(x,y) \log P(x|y) - \sum_{x} P(x) \log P(x) \\ &= -H(X|Y) + H(X) \end{aligned}$$

Now using the non-negativity of KL divergence, we can show that:

$$\begin{aligned} & I(X;Y) \geq 0 \\ \implies & -H(X|Y) + H(X) \geq 0 \\ \implies & H(X|Y) \leq H(X) \end{aligned}$$

This is known as 'Information never hurts' property of Mutual Information!
(When do we have equality? When $I(X;Y) = 0$, i.e. when $P(X,Y) = P(X)P(Y)$, or in other words, when X and Y are independent.)

3.7

3.7.1

We can prove it using results of Section 3.6:

$$\left.\begin{aligned} & I(X,Y;Z) = H(Z) - H(Z|X,Y) \\ & I(X;Z) = H(Z) - H(Z|X) \\ & H(Z|X,Y) \leq H(Z|X) \end{aligned}\right\} \implies I(X,Y;Z) \geq I(X;Z)$$

We have equality if and only if $H(Z|X,Y) = H(Z|X)$.
That means $I(Z,Y|X) = H(Z|X) - H(Z|X,Y) = 0$, i.e. Z and Y are independent given X.

3.7.2

Using 3.3:

$$\left.\begin{array}{l} H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \\ H(Y|X,Z) \geq 0 \end{array}\right\} \implies H(X,Y|Z) \geq H(X|Z)$$

We have equality if and only if $H(Y|X,Z) = 0$, which means if X and Z are known, then there is no uncertainty about Y, and it is known as well. In other words, Y is totally dependent on X and Z: $Y = f(X,Z)$.

3.8

3.8.1

X	Y	Z	P(X, Y, Z)
1	0	1	$\frac{1}{4}$
0	1	1	$\frac{1}{4}$
0	0	0	$\frac{1}{4}$
1	1	0	$\frac{1}{4}$

Table 1: Joint distribution of X, Y, Z. Probability of other possible values of X, Y, Z is zero.

As you can see, when Z is not given, then X and Y are independent and $I(X;Y) = 0$.
And when $Z = 0$ is given:

$$\begin{aligned} I(X;Y|Z=0) &= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}\frac{1}{2}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}\frac{1}{2}} \\ &= 2 \times \frac{1}{2}\log 2 = 1 \end{aligned}$$

$I(X;Y|Z=1)$ can be shown to be 1 similarly. So:

$$I(X;Y|Z) = P(Z=0)I(X;Y|Z=0) + P(Z=1)I(X;Y|Z=1) = 1$$

3.8.2

X	Y	Z	P(X, Y, Z)
1	1	1	$\frac{1}{2}$
0	0	0	$\frac{1}{2}$

Table 2: Joint distribution of X, Y, Z. Probability of other possible values of X, Y, Z is zero.

In this case, when (for example) $Z = 0$ is given, then X and Y are independent. Because:

$$P(x, y|z = 0) = P(x|z = 0)P(y|z = 0) = \begin{cases} 1 & \text{if } x = y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Similarly when $Z = 1$ is given, then X and Y are independent as well. Therefore, $I(X; Y|Z) = 0$.

And when there is no information about Z, then X and Y are totally dependent and $I(X; Y) = 1$:

$$\begin{aligned} I(X; Y) &= \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{2}\frac{1}{2}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{2}\frac{1}{2}} \\ &= 2 \times \frac{1}{2} \log 2 = 1 \end{aligned}$$

4 Probabilistic Models and Latent Variables

4.1

Suppose $P_{data}(x)$ is empirical distribution of data and $P_\theta(x)$ is our model's distribution. Then:

$$\begin{aligned} \mathrm{D}_{KL}\left(P_{data} \| P_\theta\right) &= \mathbb{E}_{P_{data}}\left[\log \frac{P_{data}(x)}{P_\theta(x)}\right] \\ &= \mathbb{E}_{P_{data}}\left[\log P_{data}(x)\right] - \mathbb{E}_{P_{data}}\left[\log P_\theta(x)\right] \\ &= -H(P_{data}) - \mathbb{E}_{P_{data}}\left[\log P_\theta(x)\right] \end{aligned}$$

So minimizing $\mathrm{D}_{KL}\left(P_{data} \| P_\theta\right)$ is equivalent to maximizing $\mathbb{E}_{P_{data}}\left[\log P_\theta(x)\right]$. And we can show that maximizing $\mathbb{E}_{P_{data}}\left[\log P_\theta(x)\right]$ is equivalent to maximizing likelihood of data:

$$\begin{aligned} \arg\max_\theta \mathbb{E}_{P_{data}}\left[\log P_\theta(x)\right] &= \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} \log P_\theta(x_i) \\ &= \arg\max_\theta \log \prod_{i=1}^{N} P_\theta(x_i) \\ &= \arg\max_\theta \prod_{i=1}^{N} P_\theta(x_i) \\ &= \arg\max_\theta P_\theta(\mathcal{D}) \end{aligned}$$

Where $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ is the data set.

4.2

(We use sum for simplicity, but it can be generalized to integral for continuous variables as well.)

$$\begin{aligned}
\ell(\theta; X) &= \log P_\theta(X) = \log \sum_Z P_\theta(X, Z) \\
&= \log \sum_Z Q(Z) \frac{P_\theta(X, Z)}{Q(Z)} \\
\xrightarrow{\text{Jensen}} &\geq \sum_Z Q(Z) \log \frac{P_\theta(X, Z)}{Q(Z)} \\
&= \sum_Z Q(Z) \log \frac{P_\theta(X, Z)}{Q(Z)} - \sum_Z Q(Z) \log Q(Z) \\
&= \mathbb{E}_{Z\sim Q} \left[\log P_\theta(X, Z)\right] + H(Q) = \mathcal{L}(\theta, Q)
\end{aligned}$$

4.3

We'll show that choosing $Q(Z) = P_\theta(Z|X)$ **maximizes** $\mathcal{L}(\theta, Q)$.

$$\begin{aligned}
\ell(\theta; X) - \mathcal{L}(\theta, Q) &= \log P_\theta(X) - \sum_Z Q(Z) \log \frac{P_\theta(X, Z)}{Q(Z)} \\
&= \sum_Z Q(Z) \log P_\theta(X) - \sum_Z Q(Z) \log \frac{P_\theta(X, Z)}{Q(Z)} \\
&= \sum_Z Q(Z)(\log P_\theta(X) - \log \frac{P_\theta(X, Z)}{Q(Z)}) \\
&= \sum_Z Q(Z)(\log Q(Z) - \log \frac{P_\theta(X, Z)}{P_\theta(X)}) \\
&= \sum_Z Q(Z)(\log Q(Z) - \log P_\theta(Z|X)) \\
&= \sum_Z Q(Z) \log \frac{Q(Z)}{P_\theta(Z|X)} \\
&= \mathrm{D}_{KL}\left(Q \| P_\theta(Z|X)\right)
\end{aligned}$$

So:

$$\ell(\theta; X) = \mathcal{L}(\theta, Q) \iff \mathrm{D}_{KL}\left(Q \| P_\theta(Z|X)\right) = 0 \iff Q(Z) = P_\theta(Z|X)$$

And we know $\ell(\theta; X)$ is an upper bound of $\mathcal{L}(\theta, Q)$, so choosing $Q(Z) = P_\theta(Z|X)$ maximizes $\mathcal{L}(\theta, Q)$.

4.4

According to Bayes' rule, $P_\theta(Z|X) = \frac{P_\theta(X,Z)}{P_\theta(X)} = \frac{P_\theta(X|Z)P_\theta(Z)}{P_\theta(X)}$. But computing $P_\theta(X)$ is intractable may be intractable. Because either $P_\theta(X) = \int_Z P_\theta(X|Z)P_\theta(Z)dZ$ and this integral might be in-

tractable. Or for discrete latent variables, $P_\theta(X) = \sum_Z P_\theta(X|Z)P_\theta(Z)$, and when the latent variable has many possible values, then this sum is intractable.

4.5

Substituting $Q_\phi(Z|X)$ with $P_\theta(Z|X)$ in the above lower bound, we get:

$$\begin{aligned}\sum_Z Q_\phi(Z|X)\log\frac{P_\theta(X,Z)}{Q_\phi(Z|X)} &= \sum_Z Q_\phi(Z|X)\log\frac{P_\theta(X|Z)P_\theta(Z)}{Q_\phi(Z|X)} \\ &= \sum_Z Q_\phi(Z|X)\log P_\theta(X|Z) + \sum_Z Q_\phi(Z|X)\log\frac{P_\theta(Z)}{Q_\phi(Z|X)} \\ &= \mathbb{E}_{Q_\phi(Z|X)}\left[\log P_\theta(X|Z)\right] - \mathrm{D}_{KL}\left(Q_\phi(Z|X)\|P_\theta(Z)\right) = L[\theta,\phi]\end{aligned}$$

4.6

Penalizing the KL divergence between $Q_\phi(Z|X)$ and $P_\theta(Z)$ acts as a regularization. Without it, the model can learn narrow distributions for $Q_\phi(Z|X)$, which actually map every X to a single Z in the latent space. This is overfitting on the training data and the model will not generalize well on the unseen data. Encouraging the model to learn a distribution similar to the prior $P_\theta(Z)$ will help the model avoid this problem.

4.7

For this part, we need to show that $\mathrm{D}_{KL}\left(Q_\phi(Z|X)\|P_\theta(Z)\right) = \mathrm{D}_{KL}\left(Q_\phi(Z)\|P_\theta(Z)\right) + I_{Q_\phi}(X;Z)$. But in my opinion, this equality does not hold. Because LHS of the equality depends on the value of X, but RHS is expectation over all possible values of X and does not depend on a specific value of X.
Instead, I will show that $\mathbb{E}_{Q(x)}\left[\mathrm{D}_{KL}\left(Q_\phi(Z|X)\|P_\theta(Z)\right)\right] = \mathrm{D}_{KL}\left(Q_\phi(Z)\|P_\theta(Z)\right) + I_{Q_\phi}(X;Z)$

$$\begin{aligned}I_{Q_\phi}(X;Z) &= \mathbb{E}_{Q_\phi(X,Z)}\left[\log\frac{Q_\phi(X,Z)}{Q_\phi(X)Q_\phi(Z)}\right] \\ &= \mathbb{E}_{Q_\phi(X,Z)}\left[\log\frac{Q_\phi(Z|X)}{Q_\phi(Z)}\right] \\ &= \mathbb{E}_{Q_\phi(X,Z)}\left[\log\frac{Q_\phi(Z|X)P_\theta(Z)}{Q_\phi(Z)P_\theta(Z)}\right] \\ &= \mathbb{E}_{Q_\phi(X,Z)}\left[\log\frac{Q_\phi(Z|X)}{P_\theta(Z)}\right] + \mathbb{E}_{Q_\phi(X,Z)}\left[\log\frac{P_\theta(Z)}{Q_\phi(Z)}\right] \\ &= \mathbb{E}_{Q_\phi(X)}\left[\mathbb{E}_{Q_\phi(Z|X)}\left[\log\frac{Q_\phi(Z|X)}{P_\theta(Z)}\right]\right] - \mathbb{E}_{Q_\phi(z)}\left[\log\frac{Q_\phi(Z)}{P_\theta(Z)}\right] \\ &= \mathbb{E}_{Q_\phi(X)}\left[\mathrm{D}_{KL}\left(Q_\phi(Z|X)\|P_\theta(Z)\right)\right] - \mathrm{D}_{KL}\left(Q_\phi(Z)\|P_\theta(Z)\right)\end{aligned}$$

$$\implies \mathbb{E}_{Q_\phi(X)}\left[\mathrm{D}_{KL}\left(Q_\phi(Z|X)\|P_\theta(Z)\right)\right] = \mathrm{D}_{KL}\left(Q_\phi(Z)\|P_\theta(Z)\right) + I_{Q_\phi}(X;Z)$$