CPSC 440/540 Machine Learning (January-April, 2022)
Variational Autoencoders Assignment

1 "MLE" Derivation

Recall the Kullback-Leibler-divergence,

$$\mathcal{KL}(p \,||\, q) = E_{x\sim p}[\log(p(x)) - \log(q(x))]$$

as well as the definition of the ELBO function,

$$\text{ELBO}(\theta, \phi) = E_{z\sim q_\phi(z \mid x)}[\log(p(x, z)) - \log(q_\phi(z \mid x))]$$

1.1 Evidence Lower Bound

Show that the ELBO function can be written as

$$\text{ELBO}(\theta, \phi) = E_{z\sim q_\phi(z \mid x)}[\log(p(x \mid z))] - \mathcal{KL}(q_\phi(z \mid x) \,||\, p(z))$$

1.2 Log-Evidence

Starting from the KL-divergence between $q_\phi(z \mid x)$ and $p(z \mid x)$, derive the following formula for the log-evidence:

$$\log(p(x)) = \text{ELBO}(\theta, \phi) + \mathcal{KL}(q_\phi(z \mid x) \,||\, p(z \mid x))$$

Hint: use Bayes rule on the $p(z \mid x)$ term, along with the form of the ELBO *function you derived in the previous part*

1.3 Loss Function

Looking at the formula from the previous part, we still have the intractable $p(x)$ term lying around in the KL-divergence term. However, we can safely ignore the $\mathcal{KL}(q_\phi(z \mid x) \,||\, p(z \mid x))$ term. Recall from the lecture that ELBO is supposed to be a lower bound for the log evidence, ELBO $\leq \log(p(x))$. This allows us to try to maximize the evidence by instead maximizing ELBO.

Your task is to prove that ELBO is indeed a lower bound for $\log(p(x))$. Observing the formula for the log-evidence you derived in 1.2, notice that if we can show the KL-divergence term is nonnegative, then we will have proven ELBO $\leq \log(p(x))$. Show that the KL-divergence between any two distributions is always nonnegative, $\mathcal{KL}(p \,||\, q) \geq 0$.

Hint: Start by showing $\mathcal{KL}(p \,||\, q) = E_{x\sim p}\left[-\log\left(\frac{q(x)}{p(x)}\right)\right]$ *and apply Jensen's inequality. (2.12 in the link)*

2 Short Answer

1. Why do VAEs tend to perform better at generating new samples compared to traditional autoencoders?
2. Write a function in Julia or Python which uses the reparametrization trick for sampling z from $q_\phi(z \mid x)$ and submit the code.
3. What is the main advantage of β-VAEs as opposed to VAEs?
4. Suppose we have a perfect optimization algorithm which can find a unique maximum of the ELBO function. If we maximize ELBO in this manner, have we necessarily found a maximum of $\log(p(x))$? Why or why not?

Solutions

"MLE" Derivation

Evidence Lower Bound

Show that the ELBO function can be written as

$$\text{ELBO}(\theta, \phi) = E_{z \sim q_\phi(z \mid x)}[\log(p(x \mid z))] - \mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z))$$

Answer: Applying the chain rule $p(x, z) = p(x \mid z)p(z)$, we obtain

$$\begin{aligned}
\text{ELBO}(\theta, \phi) &= E_{z \sim q_\phi(z \mid x)}[\log(p(x, z)) - \log(q_\phi(z \mid x))] \\
&= E_{z \sim q_\phi(z \mid x)}[\log(p(x \mid z)) + \log(p(z)) - \log(q_\phi(z \mid x))] \\
&= E_{z \sim q_\phi(z \mid x)}[\log(p(x \mid z))] - E_{z \sim q_\phi(z \mid x)}[\log(q_\phi(z \mid x)) - \log(p(z))] \\
&= E_{z \sim q_\phi(z \mid x)}[\log(p(x \mid z))] - \mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z))
\end{aligned}$$

Log-Evidence

Starting from the KL-divergence between $q_\phi(z \mid x)$ and $p(z \mid x)$, derive the following formula for the log-evidence:

$$\log(p(x)) = \text{ELBO}(\theta, \phi) + \mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z \mid x))$$

Hint: use Bayes rule on the $p(z \mid x)$ term, along with the form of the ELBO *function you derived in the previous part*

Answer: Using Bayes rule $p(z \mid x) = \frac{p(x \mid z)p(z)}{p(x)}$, we obtain

$$\begin{aligned}
\mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z \mid x)) &= E_{z \sim q_\phi(z \mid x)}[\log(q_\phi(z \mid x)) - \log(p(z \mid x))] \\
&= E_{z \sim q_\phi(z \mid x)}[\log(q_\phi(z \mid x)) - \log(p(x \mid z)) - \log(p(z)) + \log(p(x))] \\
&= \mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z)) - E_{z \sim q_\phi(z \mid x)}[\log(p(x \mid z)) + \log(p(x))] \\
&= \mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z)) - E_{z \sim q_\phi(z \mid x)}[\log(p(x \mid z))] + \log(p(x))
\end{aligned}$$

where, in the final line, $p(x)$ is taken out of the expectation because it is independent of z. Recalling from our previous work that $\text{ELBO}(\theta, \phi) = E_{z \sim q_\phi(z \mid x)}[\log(p(x \mid z))] - \mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z))$, we have

$$\mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z \mid x)) = -\text{ELBO}(\theta, \phi) + \log(p(x))$$

Rearranging for the log-evidence,

$$\log(p(x)) = \text{ELBO}(\theta, \phi) + \mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z \mid x))$$

Loss Function

Looking at the formula from the previous part, we still have the intractable $p(x)$ term lying around in the KL-divergence term. However, we can safely ignore the $\mathcal{KL}(q_\phi(z \mid x) \mid\mid p(z \mid x))$ term. Recall from the lecture that ELBO is supposed to be a lower bound for the log evidence, ELBO $\leq \log(p(x))$. This allows us to try to maximize the evidence by instead maximizing ELBO.

Your task is to prove that ELBO is indeed a lower bound for $\log(p(x))$. Observing the formula for the log-evidence you derived in 1.2, notice that if we can show the KL-divergence term is nonnegative, then we will have proven ELBO $\leq \log(p(x))$. Show that the KL-divergence between any two distributions is always nonnegative, $\mathcal{KL}(p \mid\mid q) \geq 0$.

Hint: Start by showing $\mathcal{KL}(p \mid\mid q) = E_{x\sim p}\left[-\log\left(\frac{q(x)}{p(x)}\right)\right]$ *and apply Jensen's inequality. (2.12 in the link)*

Answer: Jensen's inequality states that for any convex function φ, we have $E[\varphi(x)] \geq \varphi(E[x])$. Since $-\log$ is convex, we obtain

$$\begin{aligned}
\mathcal{KL}(p \mid\mid q) &= E_{x\sim p}\left[\log(p(x)) - \log(q(x))\right] \\
&= E_{x\sim p}\left[-\log\left(\frac{q(x)}{p(x)}\right)\right] \\
&\geq -\log(E_{x\sim p}\left[\left(\frac{q(x)}{p(x)}\right)\right]) \\
&= -\log\left(\int p(x)\frac{q(x)}{p(x)}dx\right) \\
&= -\log(1) \\
&= 0
\end{aligned}$$

Short Answer

1. Why do VAEs tend to perform better at generating new samples compared to traditional autoencoders?

   Answer: Traditional autoencoders have no regularization; they are simply designed to reproduce data sets as well as possible. This leads to overfitting, in the sense that latent vectors become highly "specialized" - certain latent vectors z reconstruct the data very well, but a randomly chosen latent vector might produce nonsensical data. VAEs on the other hand use latent distributions, and have a regularizing term from the KL-divergence which prevents the distributions from overfitting by straying to far from $\mathcal{N}(0, 1)$.

2. Write a function in Julia or Python which uses the reparametrization trick for sampling z from $q_\phi(z \mid x)$ and submit the code.

```
function renormalize(mu, Sigma)
        d = length(mu)
        eps = randn(d,1)
        ch = cholesky(Sigma)
        A = ch.L
        z = A*eps + mu
        return z
end
```

3. What is the main advantage of β-VAEs as opposed to VAEs?

   Answer: β-VAEs allow for controllability of the regularization strength, and the β hyper-parameter can be tuned to increase disentanglement.

4. Suppose we have a perfect optimization algorithm which can find a unique maximum of the ELBO function. If we maximize ELBO in this manner, have we necessarily found a maximum of $\log(p(x))$? Why or why not?

   Answer: No, not necessarily. Increasing ELBO only guarantees that we are increasing the smallest possible value $\log(p(x))$ can take, since ELBO $\leq \log(p(x))$. However, it is entirely possible that ELBO is maximized for some values θ, ϕ, while $\log(p(x))$ is maximized for some completely different values (θ', ϕ').