Deep Learning Homework 4

eh1885

November 2020

1 Problem 1

1.1 1

We first define ELBO and KL-divergence formulas in expectation $\mathbb{E}$ and integral $\int$ notation:

$$\text{ELBO}(\theta, \phi, x) = \mathbb{E}_{q_\phi(z|x)}[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}] = \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

$$\text{KL}[q_\phi(z|x)||p_\theta(z|x)] = \mathbb{E}_{q_\phi(z|x)}[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}] = \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)}$$

Two important things to note:

1. KL-divergence assumes that if $p_\theta(x|z) = 0$, then $q_\phi(z|x) = 0$; otherwise, the logarithmic term is undefined.
2. Strictly speaking, we should be writing $\int_z$ as $\int_{z:q(z|x)\neq 0}$. But for these integrals, there is no difference. Proving the ELBO in particular, whenever $q(z|x) = 0, q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} = 0$ also.

We can now begin the proof. The proof strategy is to transform $\log p_\theta(x)$ into the ELBO formula; whatever falls out should be mapped into the KL-divergence term.

$$\begin{aligned}
\log p_\theta(x) &= \int_z q_\phi(z|x) \log p_\theta(x) && (1)\\
&= \int_z q_\phi(z|x) \log p_\theta(x) \frac{p_\theta(x,z)}{p_\theta(x,z)} \frac{q_\phi(z|x)}{q_\phi(z|x)} && (2)\\
&= \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} \frac{q_\phi(z|x)}{p_\theta(z|x)} && (3)\\
&= \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} + \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} && (4)\\
&= \text{ELBO}(\theta, \phi, x) + \text{KL}[q_\phi(z|x)||p_\theta(z|x)] && (5)
\end{aligned}$$

(1) follows from the Law of Total Probability; the integral of a probability distribution $q_\phi(z|x)$ with respect to z sums to 1. Thus $\int_z q_\phi(z|x) = 1 \log p_\theta(x) = 1$. Since $\log p_\theta(x)$ is constant with respect to z, it can pass through inside the integral: $\int_z q_\phi(z|x) \log p_\theta(x)$.

(2) is valid because we trivially multiplied top and bottom by those values.

(3) shuffles the formula. Using the definition of conditional probability, $\frac{p_\theta(x)}{p_\theta(x,z)} = \frac{1}{p_\theta(z|x)}$.

(4) splits out the log term into their own formula. We know $\log(bc) = \log b + \log c$, therefore by setting $b = \frac{p_\theta(x,z)}{q_\phi(z|x)}$ and $c = \frac{q_\phi(z|x)}{p_\theta(z|x)}$, we get the result.

This gives us the requested result.

1.2 2

We have the inequality $\log p_\theta \geq \text{ELBO}(\theta, \phi, x)$ because $\text{KL}[q||q] \geq 0$ for any two probability distributions $q(z), p(z)$. We prove this using Jensen's inequality, which states that for any convex function f, measurable function ψ, and nonnegative function ϕ, we have $\int_z \phi_z f(\psi(z))) \geq f(\int_z \phi_z \psi(z))$.

$$\begin{aligned}
\text{KL}[q||p] &= \int_z q \log \frac{q}{p} && (6)\\
&= -\int_z q \log \frac{p}{q} && (7)\\
&\geq -\log \int_z q \frac{p}{q} && (8)\\
&= -\log 1 = 0 && (9)
\end{aligned}$$

(8) is the key application of Jensen's inequality, where we set $f = \log$, $\psi = \frac{p}{q}$ and $\phi = q$.

We can see that in the situation where $q = p$ almost everywhere, the KL-divergence equals 0 because $\log \frac{q}{p} = 0$.

2 2

2.1 1

We just proved $\log p_\theta(x) = \text{ELBO}(\theta, \phi, x) + \text{KL}[q_\phi(z|x)||p_\theta(z|x)]$. A simple rewrite gives us $\text{ELBO}(\theta, \phi, x) = \log p_\theta(x) - \text{KL}[q_\phi(z|x)||p_\theta(z|x)]$. Starting from this equation, we can derive:

$$\text{ELBO}(\theta,\phi,x) = \log p_\theta(x) - \text{KL}[q_\phi(z|x)||p_\theta(z|x)] \tag{10}$$
$$= \int_z q_\phi(z|x) \log p_\theta(x) - \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} \tag{11}$$
$$= \int_z q_\phi(z|x) \log p_\theta(x) - \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} \frac{p_\theta(z)}{p_\theta(z)} \tag{12}$$
$$= \int_z q_\phi(z|x) \log \frac{p_\theta(x)p_\theta(z|x)}{p_\theta(z)} - \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z)} \tag{13}$$
$$= \int_z q_\phi(z|x) \log p_\theta(x|z) - \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z)} \tag{14}$$
$$= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}[q_\phi(z|x)||p_\theta(z)] \tag{15}$$

(11) follows again from the law of total probability. (14) is Bayes theorem.

2.2 2

The $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ term gives the log-likelihood of the decoder p's reconstruction $\hat{x}$ of the true input x corresponding to the sampled latent variable z. Maximizing this value forces the distribution $\hat{x}$ emitted by our decoder p to be similar to the x originally taken in by our encoder q.

The KL-divergence term $\text{KL}[q_\phi(z|x)||p_\theta(z)]$ smoothens the latent space distribution by forcing the encoder q_ϕ to keep its encodings z for different classes x close to the decoder's prior distribution $p_\theta(z)$. This closeness regularization prevents the encoder from keeping the latent space mappings from different classes to be extremely far away from each other, which would prevent smooth interpolation between classes.

2.3 3

Note that the entropy term is given by the equation

$$H[q_\phi(z|x)] = \int_z q_\phi(z|x) \log q_\phi(z|x)$$

We'll start the proof off from (14):

$$\text{ELBO}(\theta,\phi,x) = \int_z q_\phi(z|x) \log p_\theta(x|z) - \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z)} \tag{16}$$
$$= \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{p(z)} - \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z)} \tag{17}$$
$$= \int_z q_\phi(z|x) \log p_\theta(x,z) - \int_z q_\phi(z|x) \log q_\phi(z|x) \tag{18}$$
$$= \int_z q_\phi(z|x) \log p_\theta(x,z) - \mathbb{H}[q_\phi(z|x)] \tag{19}$$

3 3

3.1 1

Note that a Bernoulli$(x_d; \hat{x}_d)$ has probability distribution $\hat{x}_d^{x_d}(1-\hat{x}_d)^{1-x_d}$.

Also recall that the binary cross-entropy loss is given by

$$-\sum_{d=1}^{D} x_d \log \hat{x}_d + (1-x_d)\log(1-\hat{x}_d)$$

Then we have

$$p_\theta(x|\tilde{z}) = -\log(\Pi_{d=1}^{D}\text{Bernoulli}(x_d;\hat{x}_d)) = -\log(\Pi_{d=1}^{D}\hat{x}_d^{x_d}(1-\hat{x}_d)^{1-x_d}) \tag{20}$$

$$= -\sum_{d=1}^{D}\log \hat{x}_d^{x_d}(1-\hat{x}_d)^{1-x_d} = -\sum_{d=1}^{D}\log \hat{x}_d^{x_d} + \log(1-\hat{x}_d)^{1-x_d} \tag{21}$$

$$= -\sum_{d=1}^{D} x_d \log \hat{x}_d + (1-x_d)\log(1-\hat{x}_d) \tag{22}$$

3.2 2

We have $\mathcal{N}(x_d; \hat{x}_d, \sigma^2) = C_d \exp(E_d(x_d - \hat{x}_d))^2$, where C_d and E_d are constants. This can be seen by:

$$\mathcal{N}(x_d;\hat{x}_d;\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp(\frac{-1}{2}(\frac{x_d-\hat{x}_d}{\sigma}^2) \tag{23}$$

$$= C_d \exp(E_d(x_d-\hat{x}_d))^2 \tag{24}$$

where $C_d = \dfrac{1}{\sigma\sqrt{2\pi}}$ and $E_d = \dfrac{-1}{2\sigma^2}$.

Then it is clear that

$$-\log p_\theta(x|\tilde{z}) = -\log \Pi_{d=1}^{D}\mathcal{N}(x_d;\hat{x}_d,\sigma^2) \tag{25}$$

$$= -\log \Pi_{d=1}^{D} C_d \exp(E_d(x_d-\hat{x}_d)^2) \tag{26}$$

$$= -\sum_{d=1}^{D}\log C_d + \log\exp(E_d(x_d-\hat{x}_d)^2) \tag{27}$$

$$= -\sum_{d=1}^{D}\log C_d + E_d(x_d-\hat{x}_d)^2) \tag{28}$$

4 4

4.1 1

VAEs use reparameterization so that it is possible to backpropagate the ELBO gradient from the decoder p to the encoder q. If the decoder randomly sampled from the latent parameters, then it would not be possible to backpropagate gradients through it. Instead, the decoder treats the latent parameters as fixed and forces the randomness into another variable ϵ. ϵ cannot be backpropagated on, but that is fine because the goal is to backpropagate through z into the encoder parameters.

4.2 2

This overlapping is problematic because the decoder p is going to have to emit $x^{(1)}$ for the space around the input $q_\phi(z|x^{(1)})$ and emit $x^{(2)}$ for the space around the input $q_\phi(z|x^{(2)})$. If the spaces $q_\phi(z|x^{(1)})$ and $q_\phi(z|x^{(2)})$ are very close, then with high probability the random sampling process for $q_\phi(z|x^{(1)})$ may select a region of latent space corresponding to x^2, and correspondingly the random sampling process for $q_\phi(z|x^{(2)})$ may select a region of latent space corresponding to x^1. Then with high probability the model would be penalized for generating x^1 when the input is x^2 and vice versa.

4.3 3

One option is to use a self-supervised learning model which takes a random masking of x and y and attempts to predict the masked features. Since both the x and the y are treated equally as features and the goal is to predict missing features given the ones which were not missing, the final model should be able to predict the y values for the data points which are missing.

A second option is to train a VAE on all the data. The latent variables should be able to encode a condensed and more informative representation z of the features x. We can take these latent variables for the subset of data where we have all the labels and train a classification model on top of the latent variables. Then, this classification model should be able to predict the missing labels for the rest of the data.

4.4 4

With a discrete, categorical latent variable, there are a few difficulties. The first is that the reparameterization trick cannot really be used to fix the discrete random variable to be deterministic and force the randomness into a separate noise variable. This makes backpropagation difficult as the categorical latent variable's CDF is piecewise constant, making derivatives 0 almost everywhere and infinite at the breakage points. Thus, the reconstruction loss $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ cannot be backpropagated against.

The REINFORCE algorithm is a method to generate unbiased estimates of the gradient of $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$. It is based off of the log-derivative trick, and only requires samples from $q_\phi(z|x)$ and the gradients of $_\theta$. The log-derivative trick is as follows:

$$\nabla \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z) \nabla \log q_\phi(z|x)]$$

The expectation over $q_\phi(z|x)$ is calculated using random samples from $p_\theta(x|z)$.