

# DD2434 Advanced Machine Learning

## Assignment 1AD

Reuben Gezang

November 2025

### D-Level

#### Theory 1.D.1

##### Question 1.1.1

We start by stating the definition of the Kullback-Leibler divergence:

$$KL(q(Z)||p(Z|X)) = \mathbb{E}_{q(Z)}[\log(\frac{q(Z)}{p(Z|X)})] = \int q(Z) \log(\frac{q(Z)}{p(Z|X)})dZ \quad (1)$$

Now we separate the fraction inside the logarithm:

$$KL(q(Z)||p(Z|X)) = \int q(Z) \log(\frac{q(Z)p(X)}{p(X, Z)})dZ = \int q(Z)(\log(q(Z)) - \log(p(Z|X)))dZ \quad (2)$$

Note that  $p(Z|X) = \frac{p(X, Z)}{p(X)}$  and using this we can rewrite the equation as:

$$KL(q(Z)||p(Z|X)) = \int q(Z) \log(q(Z))dZ - \int q(Z) \log(p(X, Z))dZ + \log(p(X)) \int q(Z)dZ \quad (3)$$

Since  $q(Z)$  is a probability distribution, we know that  $\int q(Z)dZ = 1$ . Now we can solve for  $\log(p(X))$ :

$$\log(p(X)) = KL(q(Z)||p(Z|X)) - \int q(Z) \log(q(Z))dZ + \int q(Z) \log(p(X, Z))dZ \quad (4)$$

Combining the logarithm terms, we get:

$$\log(p(X)) = KL(q(Z)||p(Z|X)) + \mathbb{E}_{q(Z)}[\frac{\log(p(X, Z))}{q(Z)}] \quad (5)$$

Now we can identify the Evidence lower bound (ELBO)

$$\mathcal{L}(q) = \mathbb{E}_{q(Z)}[\frac{\log(p(X, Z))}{q(Z)}] \quad (6)$$

and with this we have shown that:

$$\log(p(X)) = \mathcal{L}(q) + KL(q(Z)||p(Z|X)) \quad (7)$$

concluding the proof.

### Question 1.1.2

KOLLA ÖVER IGEN In this question we are to describe (in one sentence) how the choice of variational family  $q(Z)$  affects

- (i) The tightness of the ELBO
- (ii) The accuracy of the posterior approximation

(i) A more expressive variational family can lead to a tighter ELBO as it can better approximate the true posterior, reducing the KL divergence term.

(ii) The choice of variational family directly impacts the accuracy of the posterior approximation, as a limited family may not capture the true posterior's complexity, leading to a less accurate approximation.

### Question 1.D.2

#### 1.1.3

For a mean field assumption and joint distribution

$$q(Z_1, Z_2, Z_3) = q_1(Z_1)q_2(Z_2)q_3(Z_3), \quad p(X, Z)$$

Let  $q_1^*(Z_1)$  be the  $q_1$  that maximizes the ELBO. We want to show that  $q_1^*$  satisfies

$$\log q_1^*(Z_1) = \mathbb{E}_{-Z_1}[\log p(X, Z)]$$

We can start by inspecting the ELBO:

$$\mathcal{L}(q) = \mathcal{L}(q) = \mathbb{E}_{q(Z)}[\log(\frac{p(X, Z)}{q(Z)})] = \mathbb{E}_q[\log p(X, Z)] - \mathbb{E}_q[\log q(Z)]$$

and using the mean field assumption we can rewrite this as:

$$\mathcal{L}(q) = \mathbb{E}_{q(Z)}[\log p(X, Z)] - \mathbb{E}_{q(Z)}[\log(q_1(Z_1)q_2(Z_2)q_3(Z_3))]$$

and by separating the logarithm we get, and taking expectations over the relevant distribution ( $z_i$  is independent of  $z_j$  for  $i \neq j$ ):

$$\mathbb{E}_{q(Z)}[\log(q_1(Z_1)q_2(Z_2)q_3(Z_3))] = \mathbb{E}_{q_1(Z_1)}[\log(q_1(Z_1))] + \mathbb{E}_{q_2(Z_2)}[\log(q_2(Z_2))] + \mathbb{E}_{q_3(Z_3)}[\log(q_3(Z_3))]$$

Since we are maximizing w.r.t  $q_1(Z_1)$  we can ignore the terms that do not depend on it. Thus we can rewrite the ELBO as:

$$\mathcal{L}(q) = \mathbb{E}_{q(Z)}[\log p(X, Z)] - \mathbb{E}_{q_1(Z_1)}[\log(q_1(Z_1))] + C$$

where  $C$  is a constant w.r.t  $q_1(Z_1)$  (and can thus be ignored). Now we can rewrite the expectation over  $q(Z)$  as:

$$\mathbb{E}_{q(Z)}[\log p(X, Z)] = \mathbb{E}_{q_1(Z_1)}[\mathbb{E}_{q_2(Z_2)q_3(Z_3)}[\log p(X, Z)]]$$

meaning that we can rewrite the ELBO as:

$$\mathcal{L}(q) = \mathbb{E}_{q_1(Z_1)}[\mathbb{E}_{q_2(Z_2)q_3(Z_3)}[\log p(X, Z)]] - \mathbb{E}_{q_1(Z_1)}[\log(q_1(Z_1))] + C$$

Using the fact that  $\int_{Z_1} q(Z_1) dZ_1 = 1$  we will now optimize the ELBO w.r.t  $q_1(Z_1)$  and with a lagrange multiplier  $\lambda$ .

$$\frac{\partial}{\partial q_1(Z_1)} \left( \mathbb{E}_{q_1(Z_1)} [\mathbb{E}_{q_2(Z_2)q_3(Z_3)} [\log p(X, Z)]] - \mathbb{E}_{q_1(Z_1)} [\log(q_1(Z_1))] + \lambda \left( \int_{Z_1} q(Z_1) dZ_1 - 1 \right) \right) = 0 \quad (8)$$

giving that

$$\mathbb{E}_{q_2(Z_2)q_3(Z_3)} [\log p(X, Z)] - \log(q_1(Z_1)) - 1 + \lambda = 0 \rightarrow \log(q_1^*(Z_1)) = \mathbb{E}_{q_2(Z_2)q_3(Z_3)} [\log p(X, Z)] + \lambda - 1 \quad (9)$$

where  $\lambda - 1$  is a additive constant that can be ignored when normalizing  $q_1^*(Z_1)$ . Thus we have shown that:

$$\log q_1^*(Z_1) = \mathbb{E}_{-Z_1} [\log p(X, Z)] \quad (10)$$

as required.

## Practice/Implementation - D level (I have chosen 1.D.3)

### 1.2.4

The log likelihood of the data (D) is:

$$\log(P(D|\mu, \tau)) = \frac{N}{2} \log(\tau) - \frac{N}{2} \log(2\pi) - \frac{\tau}{2} \sum_{i=1}^N (x_i - \mu)^2 \quad (11)$$

The log prior for  $\mu$  and  $\tau$  is: (Note that  $\mu|\tau \sim \mathcal{N}(\mu_0, (\lambda_0\tau)^{-1})$  and  $\tau \sim \text{Gamma}(a_0, b_0)$ )

$$\log(P(\mu, \tau)) = \frac{1}{2} \log(\lambda_0\tau) - \frac{1}{2} \log(2\pi) - \frac{\lambda_0\tau}{2} (\mu - \mu_0)^2 + a_0 \log(b_0) - \log(\Gamma(a_0)) + (a_0 - 1) \log(\tau) - b_0\tau \quad (12)$$

The log-variational distribution is (Where  $q(\mu) \sim \mathcal{N}(\mu_N, \lambda_N^{-1})$  and  $q(\tau) \sim \text{Gamma}(a_N, b_N)$ ):

$$\log(q(\mu, \tau)) = \frac{1}{2} \log(\lambda_N) - \frac{1}{2} \log(2\pi) - \frac{\lambda_N}{2} (\mu - \mu_N)^2 + a_N \log(b_N) - \log(\Gamma(a_N)) + (a_N - 1) \log(\tau) - b_N\tau \quad (13)$$

Finally we state the score functions of the variational distributions. Let

$$\omega = (\mu_N, \lambda_N, a_N, b_N)$$

be the variational parameters.

$$\nabla_{\omega} \log(q(\mu, \tau)) = \begin{bmatrix} \lambda_N(\mu - \mu_N) \\ \frac{1}{2\lambda_N} - \frac{1}{2}(\mu - \mu_N)^2 \\ \psi(a_N) - \log(b_N) + \log(\tau) \\ \frac{a_N}{b_N} - \tau \end{bmatrix} \quad (14)$$

where  $\psi$  is the digamma function.

### 1.2.5

In this exercise we implement Algorithm 1 of the BBVI paper [Ranganath et al., 2014]. We reuse the data sampling script from 1.E.3 and prior parameters, meaning that

- $\mu_0 = 1.0$
- $\lambda_0 = 0.1$

- $a_0 = 1.0$
- $b_0 = 2.0$

I have chosen to use the Robbins-monro sequence to be  $\rho_t = \frac{1}{1000+t}$  and the following plots show the ELBO over iterations and the expected value of  $\mu$  and  $\tau$  over iterations for dataset 2 with 100 samples.

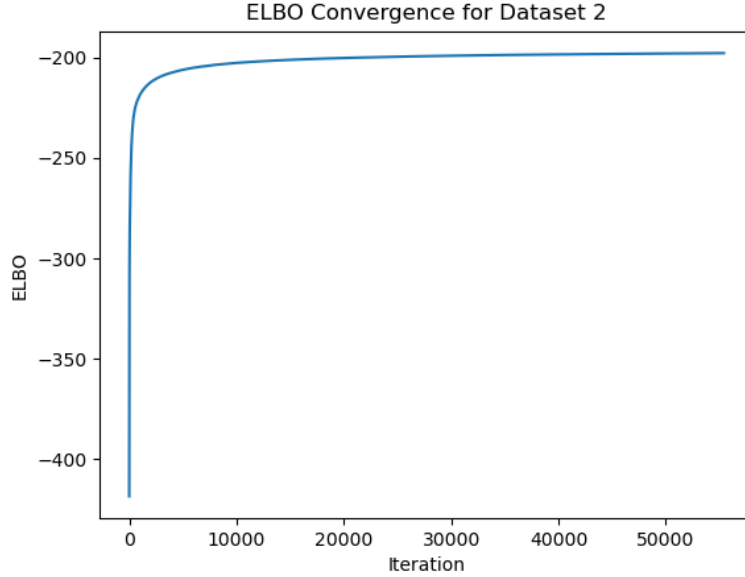


Figure 1: ELBO over iterations for dataset with 100 samples.

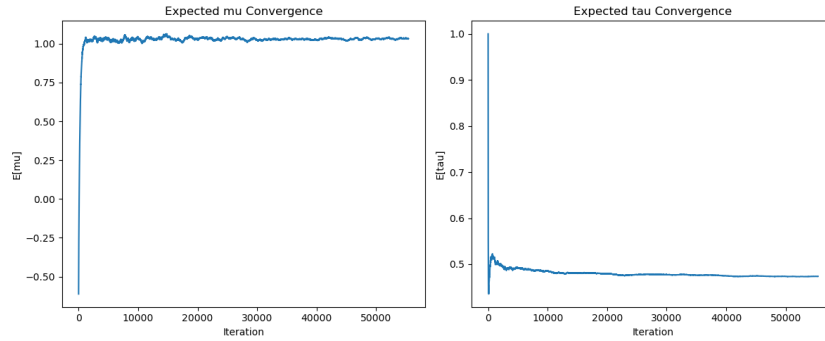


Figure 2: Expected value of  $\mu$  and  $\tau$  over iterations for dataset with 100 samples.

## C-level

### Theory 1.C.1

#### Question 2.1.10

For this exercise we want to show that the IWELBO (Importance-Weighted ELBO) is a valid lower bound on the log marginal likelihood  $\log p(X)$ . First of, we state the IWELBO:

$$\mathcal{L}_K(q) := \mathbb{E}_{Z_1, \dots, Z_K} \left[ \log \left( \frac{1}{K} \sum_{k=1}^K \frac{p(X, Z_k)}{q(Z_k|X)} \right) \right] \quad (15)$$

Now, note that we can rewrite the marginal likelihood as:

$$p(X) = \int p(X, Z) dZ = \int q(Z|X) \frac{p(X, Z)}{q(Z|X)} dZ = \mathbb{E}_{q(Z|X)} \left[ \frac{p(X, Z)}{q(Z|X)} \right]$$

and we can extent this to  $K$  samples:

$$p(X) = \mathbb{E}_{q(Z_1|X), \dots, q(Z_K|X)} \left[ \frac{1}{K} \sum_{k=1}^K \frac{p(X, Z_k)}{q(Z_k|X)} \right]$$

Now, by applying Jensen's inequality that says that for a concave function  $f$  and random variable  $X$ ,  $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$ , (note that log is concave) we get:

$$\log p(X) = \log \left( \mathbb{E}_{q(Z_1|X), \dots, q(Z_K|X)} \left[ \frac{1}{K} \sum_{k=1}^K \frac{p(X, Z_k)}{q(Z_k|X)} \right] \right) \geq \mathbb{E}_{q(Z_1|X), \dots, q(Z_K|X)} \left[ \log \left( \frac{1}{K} \sum_{k=1}^K \frac{p(X, Z_k)}{q(Z_k|X)} \right) \right]$$

Thus, the IWELBO is a valid lower bound on the log marginal likelihood  $\log p(X)$ , as required.

\*2.1.11 Let  $W_K = \frac{p(X, Z_K)}{q(Z_K|X)}$ . The IWELBO can then be rewritten as:

$$\mathcal{L}_K(q) = \mathbb{E}_{Z_1, \dots, Z_K} \left[ \log \left( \frac{1}{K} \sum_{k=1}^K W_k \right) \right]$$

We can also see that the standard ELBO can be rewritten as:

$$\mathcal{L}_1(q) = \mathbb{E}_{Z_1} [\log(W_1)] = \mathbb{E}_{Z_1, \dots, Z_K} \left[ \frac{1}{K} \sum_{k=1}^K \log(W_k) \right]$$

This is because the  $W_k$  are i.i.d. Now we can see that

$$\log \left( \frac{1}{K} \sum_{k=1}^K W_k \right) \geq \frac{1}{K} \sum_{k=1}^K \log(W_k)$$

This follows from the fact that the logarithm is strictly concave and according to Jensens inequality the logarithm of an average is greater than the average of the logaritms. Taking expectations of both sides, we get the final expression:

$$\mathcal{L}_K(q) = \mathbb{E}_{Z_1, \dots, Z_K} \left[ \log \left( \frac{1}{K} \sum_{k=1}^K W_k \right) \right] \geq \mathbb{E}_{Z_1, \dots, Z_K} \left[ \frac{1}{K} \sum_{k=1}^K \log(W_k) \right] = \mathcal{L}_1(q)$$

## Theory 1.C.2

### Question 2.1.12

## A Additional Details

Add derivations, extra figures, or ablation studies here.