DD2434 - Machine Learning, Advanced Course
Assignment 1B

Tristan Perrot
tristanp@kth.se

November 2023

Contents

1 CAVI for Earth quakes

1.1 Question 1.1

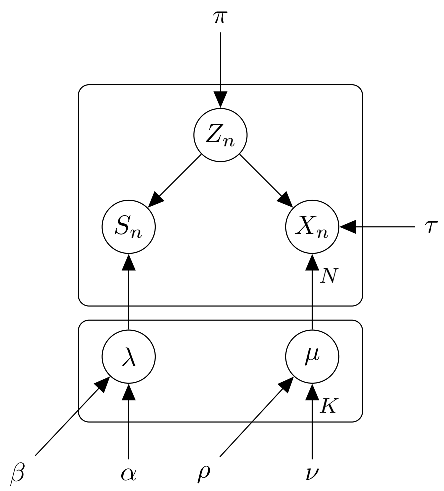

Figure 1: Directed Graphical Model for the Earthquake problem

1.2 Question 1.2

Let us take the Alternative 1 in 2D. Here, we know these distributions:

- $p(Z_n|\pi) = Categorical(\pi)$
- $p(S_n|Z_n = k, \lambda_k) = Poisson(\lambda_k)$
- $p(X_n|Z_n = k, \mu_k, \tau) = Normal(\mu_k, \tau \cdot I)$
- $p(\mu_k|\nu, \rho) = Normal(\nu, \rho \cdot I)$
- $p(\lambda_k|\alpha, \beta) = Gamma(\alpha, \beta)$

Where, ρ and τ define precision and not standard variation. Then we have:

$$\begin{aligned}
\log p(X, S, Z, \lambda, \mu|\pi, \tau, \alpha, \beta, \nu, \rho) &= \log p(X|S, Z, \lambda, \mu, \pi, \tau, \alpha, \beta, \nu, \rho) \\
&\quad + \log p(S, Z, \lambda, \mu|\pi, \alpha, \beta, \nu, \rho) \\
&= \log p(X|Z, \mu, \tau) + \log p(S|Z, \lambda, \mu, \pi, \alpha, \beta, \nu, \rho) \\
&\quad + \log p(Z, \lambda, \mu|\pi\alpha, \beta, \nu, \rho) \\
&= \log p(X|Z, \mu, \tau) + \log p(S|Z, \lambda) + \log p(Z|\pi) \\
&\quad + \log p(\lambda, \mu|\alpha, \beta, \nu, \rho) \\
\log p(X, S, Z, \lambda, \mu|\pi, \tau, \alpha, \beta, \nu, \rho) &= \log p(X|Z, \mu, \tau) + \log p(S|Z, \lambda) + \log p(Z|\pi) \\
&\quad + \log p(\mu|\nu, \rho) + \log p(\lambda|\alpha, \beta)
\end{aligned} \tag{1}$$

Where:

$$\begin{aligned}
\log p(X|Z,\mu,\tau) &= \sum_{n=1}^{N}\sum_{k=1}^{K}\log p(X_n|Z_n=k,\mu_k,\tau) \\
\log p(S|Z,\lambda) &= \sum_{n=1}^{N}\sum_{k=1}^{K}\log p(S_n|Z_n=k,\lambda_k) \\
\log p(Z|\pi) &= \sum_{n=1}^{N}\log p(Z_n|\pi) \\
\log p(\mu|\nu,\rho) &= \sum_{k=1}^{K}\log p(\mu_k|\nu,\rho) \\
\log p(\lambda|\alpha,\beta) &= \sum_{k=1}^{K}\log p(\lambda_k|\alpha,\beta)
\end{aligned} \tag{2}$$

1.3 Question 1.3

Here, the mean field approximation is not an approximation but an equality because Z, μ, λ are independent. Therefore we have:

$$\begin{aligned}
\log q^*(Z_n) &\stackrel{+}{=} \mathbb{E}_{\mu,\lambda}[\log p(X_n, S_n, Z_n, \lambda, \mu|\pi, \tau, \alpha, \beta, \nu, \rho)] \\
&\stackrel{+}{=} \mathbb{E}_{\mu,\lambda}[\log p(X_n|Z_n,\mu,\tau) + \log p(S_n|Z_n,\lambda) + \log p(Z_n|\pi)] \\
&= \mathbb{E}_{\mu}\left[\sum_{k=1}^{K}\mathbb{1}_{\{Z_n=k\}}\left(\log\left(\frac{\tau}{2\pi}\right) - \frac{\tau}{2}\left((x_n-\mu_k)^T(x_n-\mu_k)\right)\right)\right] \\
&\qquad + \mathbb{E}_{\lambda}\left[\sum_{k=1}^{K}\mathbb{1}_{\{Z_n=k\}}\left(\log(\pi_k) - \lambda_k + S_n\log(\lambda_k) - \log(S_n!)\right)\right] \\
&\stackrel{+}{=} \sum_{k=1}^{K}\mathbb{1}_{\{Z_n=k\}}\left(\log\left(\frac{\tau}{2\pi}\right) - \frac{\tau}{2}\mathbb{E}_{\mu}\left[(x_n-\mu_k)^T(x_n-\mu_k)\right]\right. \\
&\qquad \left. + \log(\pi_k) + \mathbb{E}_{\lambda}\left[-\lambda_k + S_n\log(\lambda_k)\right] - \log(S_n!)\right)
\end{aligned} \tag{3}$$

Now, if we take the entire expression that is multiplied by $\mathbb{1}_{\{Z_n=k\}}$ and we call it $u_{n,k}$, we have:

$$q^*(Z_n) \propto \prod_{k=1}^{K} u_{n,k}^{\mathbb{1}_{\{Z_n=k\}}} \tag{4}$$

And if we normalize by taking $r_{n,k} = \frac{u_{n,k}}{\sum_{i=1}^{K} u_{n,i}}$ we get:

$$q^*(Z_n) = \prod_{k=1}^{K} r_{n,k}^{\mathbb{1}_{\{Z_n=k\}}} \tag{5}$$

Wich means that $q^*(Z_n)$ is a categorical distribution with parameters $r_{n,k}$. There for we have the expectation of Z_n easily because $\mathbb{E}[z_{n,k}] = r_{n,k}$ where $z_{n,k} = \mathbb{1}_{\{S_n=k\}}$. Note that $r_{n,k}$ depends of

the expected value of μ_k, μ_k^2, λ_k and $\log\lambda_k$. We will be able to compute these expected values by finding $q^*(\mu_k)$ and $q^*(\lambda_k)$.
Let us compute $q^*(\mu_k)$:

$$\begin{aligned}
\log q^*(\mu_k) &\overset{+}{=} \mathbb{E}_{Z,\lambda}[\log p(X,S,Z=k,\lambda_k,\mu_k|\pi,\tau,\alpha,\beta,\nu,\rho)] \\
&\overset{+}{=} \mathbb{E}_{Z,\lambda}[\log p(X|Z=k,\mu_k,\tau) + \log p(\mu_k|\nu,\rho)] \\
&= \mathbb{E}_{Z,\lambda}\left[\sum_{n=1}^{N} \mathbb{1}_{\{Z_n=k\}}\left(\log\left(\frac{\tau}{2\pi}\right) - \frac{\tau}{2}\left((x_n-\mu_k)^T(x_n-\mu_k)\right)\right)\right] \\
&\qquad + \log\left(\frac{\rho}{2\pi}\right) - \frac{\rho}{2}\left((\mu_k-\nu)^T(\mu_k-\nu)\right) \\
&\overset{+}{=} \sum_{n=1}^{N} r_{n,k}\left(\log\left(\frac{\tau}{2\pi}\right) - \frac{\tau}{2}\left((x_n-\mu_k)^T(x_n-\mu_k)\right)\right) - \frac{\rho}{2}\left((\mu_k-\nu)^T(\mu_k-\nu)\right) \\
&\overset{+}{=} \sum_{n=1}^{N} r_{n,k}\left(-\frac{\tau}{2}\left((x_n-\mu_k)^T(x_n-\mu_k)\right)\right) - \frac{\rho}{2}\left((\mu_k-\nu)^T(\mu_k-\nu)\right) \\
&\overset{+}{=} -\frac{\tau\sum_{n=1}^{N} r_{n,k}}{2}\left(-2\mu_{k,0}x_{n,0} - 2\mu_{k,1}x_{n,1} + \mu_{k,0}^2 + \mu_{k,1}^2\right) \\
&\qquad - \frac{\rho}{2}\left(-2\mu_{k,0}\nu_0 - 2\mu_{k,1}\nu_1 + \mu_{k,0}^2 + \mu_{k,1}^2\right)
\end{aligned} \tag{6}$$

We define $S = \frac{\rho}{\tau\sum_{n=1}^{N} r_{n,k}}$. Then we have:

$$\begin{aligned}
\log q^*(\mu_k) &\overset{+}{=} -\frac{\tau\sum_{n=1}^{N} r_{n,k}}{2}\left[(S+N)\mu_{k,0}^2 + (S+N)\mu_{k,1}^2\right. \\
&\qquad \left. -2\mu_{k,0}(S\nu_0 + \sum_{n=1}^{N} x_{n,0}) - 2\mu_{k,1}(S\nu_1 + \sum_{n=1}^{N} x_{n,1})\right] \\
&\overset{+}{=} -\frac{\tau\sum_{n=1}^{N} r_{n,k}}{2(S+N)}\left[\left(\mu_k - \frac{S\nu + \sum_{n=1}^{N} x_n}{S+N}\right)^T\left(\mu_k - \frac{S\nu + \sum_{n=1}^{N} x_n}{S+N}\right)\right]
\end{aligned} \tag{7}$$

Therefore, we have $q^*(\mu_k) = Normal\left(\mu^*, \rho^* \cdot I\right)$. And we can compute the expected value of μ_k and μ_k^2 easily.

$$\begin{aligned}
\mu^* &= \frac{S\nu + \sum_{n=1}^{N} x_n}{S+N} = \frac{\rho\nu + \tau\sum_{n=1}^{N} r_{n,k}x_n}{\rho + N\tau\sum_{n=1}^{N} r_{n,k}} \\
\rho^* &= \frac{\tau\sum_{n=1}^{N} r_{n,k}}{S+N} = \frac{(\tau\sum_{n=1}^{N} r_{n,k})^2}{\rho + N\tau\sum_{n=1}^{N} r_{n,k}}
\end{aligned} \tag{8}$$

And therefore:

$$\begin{aligned}
\mathbb{E}[\mu_k] &= \mu^* \\
\mathbb{E}[\mu_k^2] &= \frac{1}{\rho^*} + \mu^{*T}\mu^*
\end{aligned} \tag{9}$$

Let us compute $q^*(\lambda_k)$:

$$
\begin{aligned}
\log q^*(\lambda_k) &\stackrel{\pm}{=} \mathbb{E}_{Z,\mu}[\log p(X, S, Z = k, \lambda_k, \mu_k | \pi, \tau, \alpha, \beta, \nu, \rho)] \\
&\stackrel{\pm}{=} \mathbb{E}_{Z,\mu}[\log p(S|Z = k, \lambda_k) + \log p(\lambda_k|\alpha, \beta)] \\
&= \mathbb{E}_Z \left[\sum_{n=1}^{N} \mathbb{1}_{\{Z_n=k\}} \left(-\lambda_k + S_n \log(\lambda_k) - \log(S_n!)\right) \right] \\
&\qquad + \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + (\alpha - 1)\log(\lambda_k) - \beta\lambda_k \\
&\stackrel{\pm}{=} \sum_{n=1}^{N} r_{n,k} \left(-\lambda_k + S_n \log(\lambda_k)\right) + (\alpha - 1)\log(\lambda_k) - \beta\lambda_k \\
&= \left(\alpha + \sum_{n=1}^{N} S_n r_{n,k} - 1\right) \log(\lambda_k) - \left(\beta + \sum_{n=1}^{N} r_{n,k}\right) \lambda_k
\end{aligned}
\tag{10}
$$

Therefore, we have $q^*(\lambda_k) = Gamma\left(\alpha + \sum_{n=1}^{N} S_n r_{n,k}, \beta + \sum_{n=1}^{N} r_{n,k}\right)$. And we can compute the expected value of λ_k and $\log \lambda_k$ easily.

$$
\begin{aligned}
\mathbb{E}[\lambda_k] &= \frac{\alpha + \sum_{n=1}^{N} S_n r_{n,k}}{\beta + \sum_{n=1}^{N} r_{n,k}} \\
\mathbb{E}[\log \lambda_k] &= \psi(\alpha + \sum_{n=1}^{N} S_n r_{n,k}) - \log(\beta + \sum_{n=1}^{N} r_{n,k})
\end{aligned}
\tag{11}
$$

2 VAE image generation

Question 5.1 (in the notebook)

Our objective function is ELBO: $E_{q(z|x)}\big[\log \frac{p(x,z)}{q(z|x)}\big]$
We will show that ELBO can be rewritten as $E_{q(z|x)}\big(\log p(x|z)\big) - D_{KL}\big(q(z|x)||p(z)\big)$ We have:

$$
\begin{aligned}
E_{q(z|x)}\big[\log \frac{p(x,z)}{q(z|x)}\big] &= E_{q(z|x)}\big[\log p(x,z) - \log q(z|x)\big] \\
&= E_{q(z|x)}\big[\log p(x|z) + \log p(z) - \log q(z|x)\big] \\
&= E_{q(z|x)}\big[\log p(x|z)\big] + E_{q(z|x)}\big[\log p(z)\big] - E_{q(z|x)}\big[\log q(z|x)\big] \\
&= E_{q(z|x)}\big[\log p(x|z)\big] - E_{q(z|x)}\big[\log \frac{q(z|x)}{p(z)}\big] \\
&= E_{q(z|x)}\big[\log p(x|z)\big] - D_{KL}\big(q(z|x)||p(z)\big)
\end{aligned}
$$

Question 5.2 (in the notebook)

Consider the second term: $-D_{KL}\big(q(z|x)||p(z)\big)$

Question : *Kullback–Leibler divergence can be computed using the closed-form analytic expression when both the variational and the prior distributions are Gaussian. Write down this KL* *divergence in terms of the parameters of the prior and the variational distributions. Your solution should consider a generic case where the latent space is K-dimensional.*

We have:

$$D_{KL}\left(q(z|x)||p(z)\right) = \int q(z|x)\log\frac{q(z|x)}{p(z)}\,dz$$

And we also have:

$$\begin{aligned} q(z|x) &= \mathcal{N}(z|\mu(x),\sigma(x)) = \prod_{i=1}^{K}\left(\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left(-\frac{(z_i-\mu_i)^2}{2\sigma^2}\right)\right) \\ p(z) &= \mathcal{N}(z|0,I) = \left(\frac{1}{\sqrt{2\pi}}\right)^K \exp\left(-\frac{1}{2}z^Tz\right) = \prod_{i=1}^{K}\left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(-\frac{1}{2}z_i^2\right) \end{aligned} \tag{12}$$

Therefore:

$$\begin{aligned} D_{KL}\left(q(z|x)||p(z)\right) &= \int q(z|x)\log\frac{q(z|x)}{p(z)}\,dz \\ &= \int q(z|x)\log\frac{\prod_{i=1}^{K}\left(2\pi\sigma_i^2\right)^{-\frac{1}{2}}\exp\left(-\frac{(z_i-\mu_i)^2}{2\sigma_i^2}\right)}{\prod_{i=1}^{K}(2\pi)^{-\frac{1}{2}}\exp\left(-\frac{z_i^2}{2}\right)}\,dz \\ &= \int q(z|x)\left(\sum_{i=1}^{K}-\log(\sigma_i)-\frac{(z_i-\mu_i)^2}{2\sigma_i^2}+\frac{z_i^2}{2}\right)dz \\ &= \mathbb{E}_{q(z|x)}\left[\sum_{i=1}^{K}-\log(\sigma_i)-\frac{(z_i-\mu_i)^2}{2\sigma_i^2}+\frac{z_i^2}{2}\right] \\ &= \sum_{i=1}^{K}-\log(\sigma_i)-\frac{\mathbb{E}_{q(z|x)}\left[(z_i-\mu_i)^2\right]}{2\sigma_i^2}+\frac{\mathbb{E}_{q(z|x)}\left[z_i^2\right]}{2} \\ &= \sum_{i=1}^{K}-\log(\sigma_i)-\frac{\sigma_i^2}{2\sigma_i^2}+\frac{\mathbb{E}_{q(z|x)}[z_i^2]}{2} \\ &= \sum_{i=1}^{K}-\log(\sigma_i)-\frac{1}{2}+\frac{\sigma_i^2+\mu_i^2}{2} \\ D_{KL}\left(q(z|x)||p(z)\right) &= \frac{1}{2}\sum_{i=1}^{K}\left(\sigma_i^2+\mu_i^2-\log(\sigma_i^2)-1\right) \end{aligned} \tag{13}$$

The rest of tht implementation could be find in the appendix **??**.

3 Reparameterization and the score function

3.1 Question 3.4

According to the Sticking the Landing paper, we can do the reparameterization by sampling z from a parametric distribution $q_\phi(z)$ and by sampling ϵ from a fixed distribution $p(\epsilon)$ and applying a deterministic transformation $t(\epsilon, \phi) = z$ Therefore, we have for the total derivative of the ELBO term:

$$\begin{aligned}\hat{\nabla}_{\text{TD}}(\epsilon, \phi) &= \nabla_\phi[\log p(z|x) + \log p(x) - \log q_\phi(z|x)] \\ &= \nabla_z[\log p(z|x) - \log q_\phi(z|x)]\nabla_\phi z - \nabla_\phi \log q_\phi(z|x) \\ &= \nabla_z[\log p(z|x) - \log q_\phi(z|x)]\nabla_\phi t(\epsilon, \phi) - \nabla_\phi \log q_\phi(z|x)\end{aligned} \tag{14}$$

Where the left term is the score function.

3.2 Question 3.5

Now, we will show that the expectation of the score function is zero.

$$\begin{aligned}\mathbb{E}_{q_\phi(z|x)}\left[\nabla_\phi \log q_\phi(z|x)\right] &= \int q_\phi(z|x)\nabla_\phi \log q_\phi(z|x)\, dz \\ &= \int \nabla_\phi q_\phi(z|x)\, dz \\ &= \nabla_\phi \int q_\phi(z|x)\, dz \\ &= \nabla_\phi 1 \\ &= 0\end{aligned} \tag{15}$$

3.3 Question 3.6

In the Sticking the Landing paper, the author handle this score function by removing it and it results in an unbiased gradient estimator with variance that approaches zero as the approximate posterior approaches the exact posterior.

3.4 Question 3.7

For particular cases, the score function may actually decrease the variance. The name of the concept that describes how the score function acts in this situation is **control variate**.

4 Reparameterization of common distributions

4.1 Question 4.8 - Exponential distribution

4.2 Question 4.9 - Categorical distribution

A Appendix

A.1 VAE image generation