Machine Learning Advanced Course: Assignment 1A

Michel Le Dez

November 12, 2023

1 Assignment 1A

1.1 Exponential Family

An exponential-family distribution with natural parameters is in the following form:

$$p(x|\theta) = h(x)\exp(\eta(\theta) \cdot T(x) - A(\eta))$$

Question 1.1.1:

We have:

- $\theta = \lambda$
- $\eta(\theta) = \log(\theta) = \log(\lambda)$
- $h(x) = \frac{1}{x!}$
- $T(x) = x$
- $A(\eta) = e^{\eta} = e^{\log \lambda} = \lambda$

Therefore:

$$\begin{aligned} & p(x|\theta) = h(x)\exp(\eta(\theta) \cdot T(x) - A(\eta)) \\ \iff \quad & p(x|\lambda) = \frac{1}{x!}\exp(x\log(\lambda) - \lambda) \\ & \boxed{p(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}} \end{aligned}$$

We recognize the probability mass function of the **Poisson distribution**.

Question 1.1.2:

We have:

- $\theta = [\alpha, \beta]$
- $\eta(\theta) = [\theta_1 - 1, -\theta_2] = [\alpha - 1, -\beta]$
- $h(x) = 1$
- $T(x) = [\log x, x]$
- $A(\eta) = \log\Gamma(\eta_1 + 1) - (\eta_1 + 1)\log(-\eta_2) = \log\Gamma(\alpha) - \alpha\log(\beta) = \log(\frac{\Gamma(\alpha)}{\beta^{\alpha}})$

Therefore:

$$\begin{aligned} & p(x|\theta) = h(x)\exp(\eta(\theta)\cdot T(x) - A(\eta)) \\ \iff \quad & p(x|\alpha,\beta) = \exp((\alpha-1)\log x - \beta x - \log(\frac{\Gamma(\alpha)}{\beta^\alpha})) \\ & \boxed{p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}} \end{aligned}$$

We recognize the probability density function of the **Gamma distribution**.

Question 1.1.3:

We have:

- $\theta = [\mu, \sigma^2]$
- $\eta(\theta) = [\frac{\theta_1}{\theta_2}, -\frac{1}{2\theta_2}] = [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}]$
- $h(x) = \frac{1}{\sqrt{2\pi}}$
- $T(x) = [x, x^2]$
- $A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) = \frac{\mu^2}{2\sigma^2} - \log(\frac{1}{\sigma})$

Therefore:

$$\begin{aligned} & p(x|\theta) = h(x)\exp(\eta(\theta)\cdot T(x) - A(\eta)) \\ \iff \quad & p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}}\exp(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} + \log(\frac{1}{\sigma})) \\ & \boxed{p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)} \end{aligned}$$

We recognize the probability density function of the **Normal distribution**.

Question 1.1.4:

We have:

- $\theta = \lambda$
- $\eta(\theta) = -\theta = -\lambda$
- $h(x) = 2$
- $T(x) = x$
- $A(\eta) = -\log(-\frac{\eta}{2}) = -\log(\frac{\lambda}{2})$

Therefore:

$$\begin{aligned} & p(x|\theta) = h(x)\exp(\eta(\theta)\cdot T(x) - A(\eta)) \\ \iff \quad & p(x|\lambda) = 2\exp(-\lambda x + \log(\frac{\lambda}{2})) \\ & \boxed{p(x|\lambda) = \lambda e^{-\lambda x}} \end{aligned}$$

We recognize the probability density function of the **Exponential distribution**.

Question 1.1.5:

We have:

- $\theta = [\psi_1, \psi_2]$
- $\eta(\theta) = [\theta_1 - 1, \theta_2 - 2] = [\psi_1 - 1, \psi_2 - 2]$
- $h(x) = 1$
- $T(x) = [\log x, \log(1-x)]$
- $A(\eta) = \log\Gamma(\eta_1 + 1) + \log\Gamma(\eta_2 + 1) - \log\Gamma(\eta_1 + \eta_2 + 2) = -\log\frac{\Gamma(\psi_1+\psi_2)}{\Gamma(\psi_1)\Gamma(\psi_2)}$

Therefore:

$$\begin{aligned} & p(x|\theta) = h(x)\exp(\eta(\theta)\cdot T(x) - A(\eta)) \\ \iff \quad & p(x|\psi_1,\psi_2) = \exp((\psi_1 - 1)\log x + (\psi_2 - 1)\log(1-x) + \log\frac{\Gamma(\psi_1+\psi_2)}{\Gamma(\psi_1)\Gamma(\psi_2)}) \\ & \boxed{p(x|\psi_1,\psi_2) = \frac{\Gamma(\psi_1+\psi_2)}{\Gamma(\psi_1)\Gamma(\psi_2)}x^{\psi_1-1}(1-x)^{\psi_2-1}} \end{aligned}$$

We recognize the probability density function of the **Beta distribution**.

1.2 Dependencies in a Directed Graphical Model

Question 1.2.6: Yes.

Question 1.2.7: No.

Question 1.2.8: Yes.

Question 1.2.9: No.

Question 1.2.10: No.

Question 1.2.11: No.

1.3 CAVI

We have the following distribution:

$$p(\tau) = Gam(\tau|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}e^{-b_0\tau} \tag{1}$$

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}) = \frac{\sqrt{\lambda_0\tau}}{\sqrt{2\pi}}\exp(-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2) \tag{2}$$

$$p(D|\mu,\tau) = \prod_{n=1}^{N}\frac{\sqrt{\tau}}{\sqrt{2\pi}}\exp(-\frac{\tau}{2}(x_n-\mu)^2) = (\frac{\tau}{2\pi})^{\frac{N}{2}}\exp(-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2) \tag{3}$$

Question 1.3.12: The function implementation for generating data points, as well as the code for displaying the histogram is in the annex. Here are the the histograms we obtained:

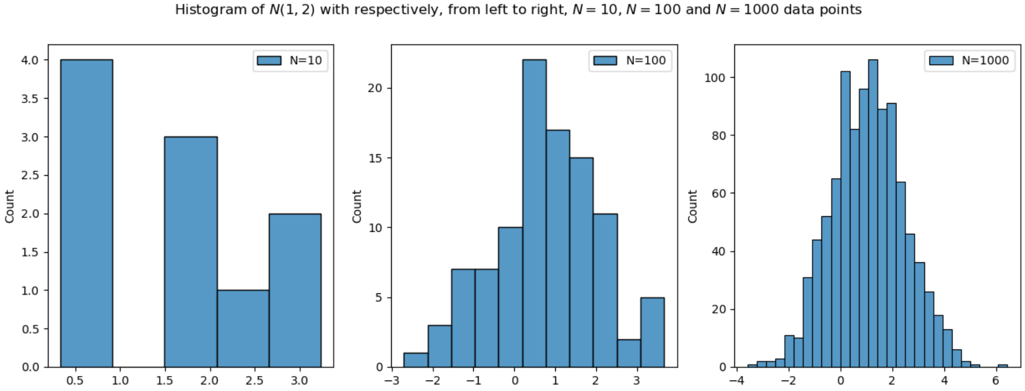

Figure 1: Histogram of $\mathcal{N}(\mu, \frac{1}{\tau})$ with $\mu = 1$ and $\tau = 0.5$ with respectively, from left to right, $N = 10$, $N = 100$, $N = 1000$ data points.

We observe that the more data we have, the closer the histogram is to the normal distribution that generated it.

Question 1.3.13: The likelihood of the data points $D = x_{1:N}$ given the parameter μ, τ is as follows:

$$l(\mu,\tau) := p(D|\mu,\tau) = (\frac{\tau}{2\pi})^{\frac{N}{2}}\exp(-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2)$$

$$\iff \log(l(\mu,\tau)) = \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 + const$$

We are looking for the parameters μ and τ which maximise the likelihood $l(\mu,\tau)$, which is equivalent to maximising the log-likelihood given that the log function is a monotonically increasing one. The constant term above gathers all the terms that do not depend on μ or τ. Deriving the gradient of $l(\mu,\tau)$ and setting it to 0 at (μ_{MLE}, τ_{MLE}) yields the following system of equations:

$$\iff \begin{cases} \tau_{MLE}\sum_{n=1}^{N}x_n - \tau_{MLE}N\mu_{MLE} = 0 \\ \frac{N}{2\tau_{MLE}} - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu_{MLE})^2 = 0 \end{cases}$$

$$\iff \begin{cases} \bar{x} := \mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N}x_n \\ \tau_{MLE} = \frac{1}{\frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})^2} \end{cases}$$

To verify that the point (μ_{MLE}, τ_{MLE}) definitely maximises the likelihood we can compute the hessian of the log-likelihood at this point, which yields to:

$$\begin{bmatrix} -\frac{N^2}{\sum_{n=1}^{N}(x_n-\bar{x})^2} & 0 \\ 0 & -\frac{1}{2N}(\sum_{n=1}^{N}(x_n-\bar{x})^2)^2 \end{bmatrix}$$

We can see the eigenvalues of the hessian are strictly negative, therefore (μ_{MLE}, τ_{MLE}) definitely maximises the likelihood.

Question 1.3.14: To compute the posterior, we are going to use the Bayes' theorem and gather in the constant term all the terms that do not depend on μ or τ. Here is the posterior:

$$\begin{aligned} p(\mu,\tau|D) =& p(D|\mu,\tau)p(\mu|\tau)p(\tau)/p(D) \\ \iff \log p(\mu,\tau|D) =& \log p(D|\mu,\tau) + \log p(\mu|\tau) + \log p(\tau) + \mathit{const} \\ =& \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n^2+\mu^2-2x_n\mu) + \frac{1}{2}\log\tau - \frac{\lambda_0\tau}{2}(\mu^2+\mu_0^2-2\mu\mu_0) \\ &+ (a_0-1)\log\tau - b_0\tau + \mathit{const} \\ =& (a_0+\frac{N}{2}-\frac{1}{2})\log\tau - (b_0+\frac{1}{2}\sum_{n=1}^{N}x_n^2+\frac{\lambda_0\mu_0^2}{2})\tau + (\sum_{n=1}^{N}x_n+\lambda_0\mu_0)\tau\mu \\ &- \frac{\tau}{2}(\lambda_0+N)\mu^2 + \mathit{const} \end{aligned}$$

However, we know that for $\mu,\tau \sim \mathit{NormalGamma}(\mu_0^*, \lambda_0^*, a_0^*, b_0^*)$, the logarithm of the probability density function is as follows:

$$\begin{aligned} \log p(\mu,\tau|\mu_0^*,\lambda_0^*,a_0^*,b_0^*) &= (a_0^*-\frac{1}{2})\log\tau - b_0^*\tau - \frac{\lambda_0^*\mu_0^{2*}}{2} + \lambda_0^*\mu_0^*\tau\mu - \frac{\tau}{2}\lambda_0^*\mu^2 + \mathit{const} \\ &= (a_0^*-\frac{1}{2})\log\tau - b_0^*\tau - \frac{\tau\lambda_0^*}{2}(\mu-\mu_0^*)^2 + \mathit{const} \end{aligned}$$

By identification, we have $a_0^* = a_0 + \frac{N}{2}$, $\lambda_0^* = \lambda_0 + N$ and $\mu_0^* = \frac{\sum_{n=1}^{N}x_n+\lambda_0\mu_0}{\lambda_0+N}$. For b_0^*, let's rewrite the log posterior in the form of the second equality above by adding and subtracting the missing term $\frac{1}{2}\frac{(\sum_{n=1}^{N}x_n+\lambda_0\mu_0)^2}{\lambda_0+N}\tau$ for completing the square, which yields to:

$$\begin{aligned} \log p(\mu,\tau|D) =& (a_0+\frac{N}{2}-\frac{1}{2})\log\tau - (b_0+\frac{1}{2}\sum_{n=1}^{N}x_n^2+\frac{\lambda_0\mu_0^2}{2}-\frac{1}{2}\frac{(\sum_{n=1}^{N}x_n+\lambda_0\mu_0)^2}{\lambda_0+N})\tau \\ &- \frac{\tau(\lambda_0+N)}{2}(\mu-\frac{\sum_{n=1}^{N}x_n+\lambda_0\mu_0}{\lambda_0+N})^2 + \mathit{const} \end{aligned}$$

Here, we can easily identify b_0^* and we summarise the results below:

$$\boxed{\begin{array}{l} \mu,\tau|D \sim \mathit{NormalGamma}(\mu_0^*,\lambda_0^*,a_0^*,b_0^*) \text{ with the following parameters} \\ \qquad \mu_0^* = \frac{\sum_{n=1}^{N}x_n+\lambda_0\mu_0}{\lambda_0+N} \\ \qquad \lambda_0^* = \lambda_0+N \\ \qquad a_0^* = a_0+\frac{N}{2} \\ \qquad b_0^* = b_0+\frac{1}{2}\sum_{n=1}^{N}x_n^2+\frac{\lambda_0\mu_0^2}{2}-\frac{1}{2}\frac{(\sum_{n=1}^{N}x_n+\lambda_0\mu_0)^2}{\lambda_0+N} \end{array}}$$

Question 1.3.15: The mean field approximation for the variational distribution is the following:

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau).$$

The log of the joint distribution can be written as follows:

$$\log p(x, \mu, \tau) = \log p(x|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau),$$

with:

$$\begin{aligned}\log p(x|\mu, \tau) &= \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 + const\\ \log p(\mu|\tau) &= \frac{1}{2}\log\tau - \frac{\lambda_0\tau}{2}(\mu-\mu_0)^2 + const\\ \log p(\tau) &= (a_0-1)\log\tau - b_0\tau + const\end{aligned}$$

where the constant terms include terms that do not depend on μ or τ. Let's derive now the coordinate ascent update for μ by including terms that do not depend on μ in the constant term (i.e $\log p(\tau)$, $\frac{N}{2}\log\tau$, $\frac{1}{2}\log\tau$, $-\frac{1}{2}\mathbf{E}_{q(\tau)}[\tau]\sum_{n=1}^{N}x_n^2$ and $-\frac{1}{2}\mathbf{E}_{q(\tau)}[\tau]\lambda_0\mu_0^2$):

$$\begin{aligned}\log q^*(\mu) &= \mathbf{E}_{q(\tau)}[\log p(x, \mu, \tau)]\\ &= -\mathbf{E}_{q(\tau)}\left[\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 + \frac{\lambda_0\tau}{2}(\mu-\mu_0)^2\right] + const\\ &= -\frac{1}{2}\mathbf{E}_{q(\tau)}[\tau](\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2) + const\\ &= \mathbf{E}_{q(\tau)}[\tau](\sum_{n=1}^{N}x_n + \lambda_0\mu_0)\mu - \frac{1}{2}\mathbf{E}_{q(\tau)}[\tau](\lambda_0+N)\mu^2 + const\end{aligned}$$

However, we know that for $\mu \sim Normal(\tilde{\mu_0}, \tilde{\lambda_0}^{-1})$, the log of the probability density function is as follows:

$$\log p(\mu|\tilde{\mu_0}, \tilde{\lambda_0}^{-1}) = \tilde{\lambda_0}\tilde{\mu_0}\mu - \frac{\tilde{\lambda_0}}{2}\mu^2 + const$$

By identification, we have:

$$\boxed{\begin{aligned}&q^*(\mu) = Normal(\mu|\tilde{\mu_0}, \tilde{\lambda_0}^{-1}) \text{ with the following parameters}\\ &\quad\tilde{\mu_0} = \frac{\sum_{n=1}^{N}x_n + \lambda_0\mu_0}{\lambda_0+N}\\ &\quad\tilde{\lambda_0} = \mathbf{E}_{q(\tau)}[\tau](\lambda_0+N)\\ &\quad\text{with}\\ &\qquad\mathbf{E}_{q(\tau)}[\tau] = \frac{\tilde{a_0}}{\tilde{b_0}}\end{aligned}}$$

Let's derive now the coordinate ascent update for τ by including terms that do not depend on τ in the constant term:

$$\begin{aligned}\log q^*(\tau) &= \mathbf{E}_{q(\mu)}[\log p(x, \mu, \tau)]\\ &= \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}\mathbf{E}_{q(\mu)}[(x_n-\mu)^2] + \frac{1}{2}\log\tau - \frac{\lambda_0\tau}{2}\mathbf{E}_{q(\mu)}[(\mu-\mu_0)^2] + (a_0-1)\log\tau - b_0\tau\\ &= (a_0 + \frac{N+1}{2} - 1)\log\tau - (b_0 + \frac{1}{2}\mathbf{E}_{q(\mu)}[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2])\tau\end{aligned}$$

However, we know that for $\tau \sim Gamma(\tilde{a_0}, \tilde{b_0})$, the log of the probability density function is as follows:

$$\log p(\tau|\tilde{a_0}, \tilde{b_0}) = (\tilde{a_0} - 1)\log\tau - \tilde{b_0}\tau + const$$

By identification, we have:

$q^*(\tau) = Gamma(\tau|\tilde{a_0}, \tilde{b_0})$ with the following parameters

$$\tilde{a_0} = a_0 + \frac{N+1}{2}$$

$$\tilde{b_0} = b_0 + \frac{1}{2}[\sum_{n=1}^{N} x_n^2 + \lambda_0\mu_0^2 - 2(\lambda_0\mu_0 + \sum_{n=1}^{N} x_n)\mathbf{E}_{q(\mu)}[\mu] + (\lambda_0 + \sum_{n=1}^{N} x_n^2)\mathbf{E}_{q(\mu)}[\mu^2]]$$

with

$$\mathbf{E}_{q(\mu)}[\mu] = \tilde{\mu_0}$$

$$\mathbf{E}_{q(\mu)}[\mu^2] = \frac{1}{\tilde{\lambda_0}} + \tilde{\mu_0}^2$$

Now that we have derived the variational distributions, we can compute the ELBO $\mathcal{L}$ to monitor its convergence during the CAVI algorithm:

$$\mathcal{L} = \mathbf{E}_{q(\mu,\tau)}\left[\log\frac{p(x,\mu,\tau)}{q(\mu,\tau)}\right] = \mathbf{E}_{q(\mu)q(\tau)}\left[\log\frac{p(x|\mu,\tau)p(\mu|\tau)p(\tau)}{q(\mu)q(\tau)}\right]$$

which finally gives:

$$\mathcal{L} = \mathbf{E}_{q(\mu)q(\tau)}[\log p(x|\mu,\tau)] + \mathbf{E}_{q(\mu)q(\tau)}[\log p(\mu|\tau)] + \mathbf{E}_{q(\tau)}[\log p(\tau)] - \mathbf{E}_{q(\mu)}[\log q(\mu)] - \mathbf{E}_{q(\tau)}[\log q(\tau)]$$

with:

$$\mathbf{E}_{q(\mu)q(\tau)}[\log p(x|\mu,\tau)] = -\frac{N}{2}\log(2\pi) + \frac{N}{2}\boxed{\mathbf{E}_{q(\tau)}}[\log \boxed{q(\tau)}] - \frac{1}{2}\boxed{\mathbf{E}_{q(\tau)}[\tau]}\sum_{n=1}^{N}(x_n^2 + \mathbf{E}_{q(\mu)}[\mu^2] - 2x_n\mathbf{E}_{q(\mu)}[\mu])$$

+

$$\mathbf{E}_{q(\mu)q(\tau)}[\log p(\mu|\tau)] = \frac{1}{2}\log(\frac{\lambda_0}{2\pi}) + \frac{1}{2}\mathbf{E}_{q(\tau)}[\log q(\tau)] - \frac{\lambda_0}{2}\mathbf{E}_{q(\tau)}[\tau](\mathbf{E}_{q(\mu)}[\mu^2] + \mu_0^2 - 2\mu_0\mathbf{E}_{q(\mu)}[\mu])$$

+

$$\mathbf{E}_{q(\tau)}[\log p(\tau)] = (a_0 - 1)\mathbf{E}_{q(\tau)}[\log q(\tau)] + a_0\log b_0 - b_0\mathbf{E}_{q(\tau)}[\tau] - \log\Gamma(a_0)$$

−

$$\mathbf{E}_{q(\mu)}[\log q(\mu)] = -H(\mu) = -\frac{1}{2}\log(\frac{2\pi}{\tilde{\lambda_0}}) - \frac{1}{2}$$

−

$$\mathbf{E}_{q(\tau)}[\log q(\tau)] = -(\tilde{a_0} - \log\tilde{b_0} + \log\Gamma(\tilde{a_0}) + (1 - \tilde{a_0})\psi(\tilde{a_0}))$$

and:

$$\mathbf{E}_{q(\tau)}[\log q(\tau)] = \psi(\tilde{a_0}) - \log\tilde{b_0}$$

$$\mathbf{E}_{q(\tau)}[\tau] = \frac{\tilde{a_0}}{\tilde{b_0}}$$

$$\mathbf{E}_{q(\mu)}[\mu] = \tilde{\mu_0}$$

$$\mathbf{E}_{q(\mu)}[\mu^2] = \frac{1}{\tilde{\lambda_0}} + \tilde{\mu_0}^2$$

We implemented the CAVI algorithm using the variational distributions found above, and we monitored the convergence of the ELBO. Finally, we compared the variational distribution $q(\mu,\tau)$ obtained at the end of the CAVI algorithm with the true posterior that we computed at **Question 1.3.14**. Here are the results we obtained for the 3 data sets generated at **Question 1.13.12**. We also display the ML estimate computed at **Question 1.13.13**.

First data set:

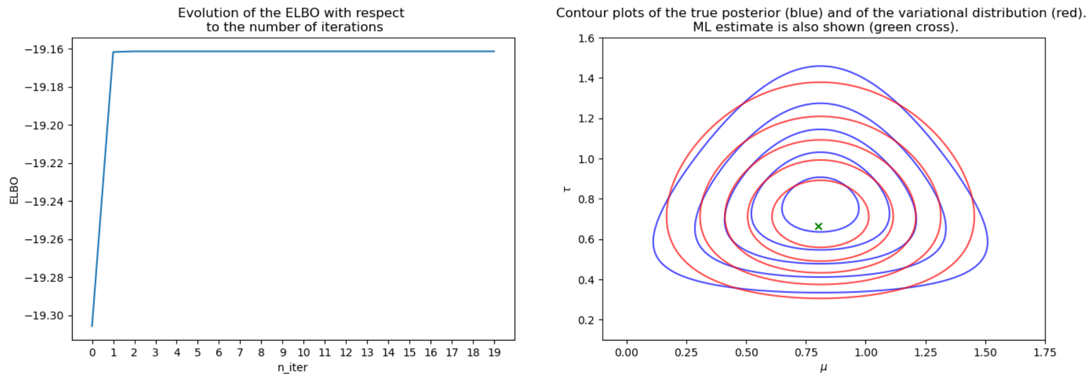

Figure 2: Left: Evolution of the ELBO with respect to the number of iterations. Right: Comparison of the true posterior (blue) with the variational distribution (red) obtained with the CAVI algorithm. The ML estimate (green) is also displayed.

Second data set:

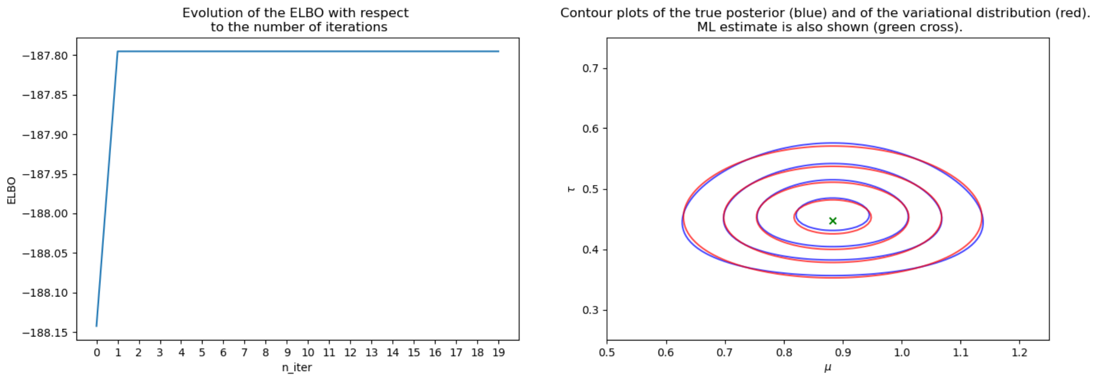

Figure 3: Left: Evolution of the ELBO with respect to the number of iterations. Right: Comparison of the true posterior (blue) with the variational distribution (red) obtained with the CAVI algorithm. The ML estimate (green) is also displayed.

Third data set:

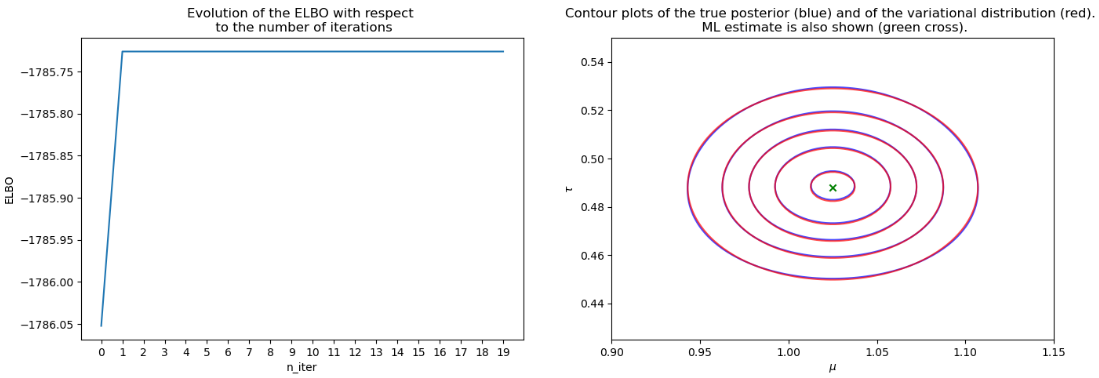

Figure 4: Left: Evolution of the ELBO with respect to the number of iterations. Right: Comparison of the true posterior (blue) with the variational distribution (red) obtained with the CAVI algorithm. The ML estimate (green) is also displayed.

Discussion of the results:

Concerning the ELBO, we notice that its convergence is very fast, a small number of iterations is enough to converge. Concerning the comparison of the variational distribution obtained at the end of the CAVI algorithm with the true posterior distribution, we notice that the more data points we have in our data set, the more accurate will be the variational distribution as the right plot of Figure 5 shows where they overlap almost perfectly. It is the same observation concerning the ML estimate, the more data points we have in our data set, the more accurate it will be as we can see with the right plot of Figure 5 where the ML estimate is effectively at the center of the true posterior.

1.4 SVI - LDA

Question 1.4.16: According to the Hoffman paper, local hidden variables $z_{n,j}$ are defined by the fact that their complete conditional probability distribution are determined by β the global hidden variables, α the fixed parameters and the other local variables $x_n,\ z_{n,-j}$ in the n$^{\text{th}}$ context. Mathematically, the definition is the following:

$$p(z_{n,j}|x_n, x_{-n}, z_{-n}, z_{n,-j}, \beta, \alpha) = p(z_{n,j}|x_n, z_{n,-j}, \beta, \alpha)$$

Question 1.4.17: In the LDA model of the Hoffman paper, the global hidden variables are the $\beta_k, k = 1 \cdots K$ and the local hidden variables are the $\theta_d, d = 1 \cdots D$ and $z_{d,n}, d = 1 \cdots D, n = 1 \cdots N$.

Question 1.4.18: To compute the ELBO, we used this source [1].

$$\boxed{\begin{aligned}
ELBO = &\sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{w=1}^{W}\phi_{dnk}(\psi(\lambda_{kw}) - \psi(\sum_{w'}\lambda_{kw'}))w_{dnw} + \sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{dnk}(\psi(\gamma_{dk}) - \psi(\sum_{k'}\gamma_{dk'})) \\
&- D\log B(\alpha) + \sum_{k=1}^{K}\sum_{d=1}^{D}(\alpha_k - 1)(\psi(\gamma_{dk}) - \psi(\sum_{k'}\gamma_{dk'})) \\
&- K\log B(\eta) + \sum_{k=1}^{K}\sum_{w=1}^{W}(\eta_w - 1)(\psi(\lambda_{kw}) - \psi(\sum_{w'}\lambda_{kw'})) \\
&+ \sum_{d=1}^{D}\left[\log B(\gamma_d) + (\sum_{k=1}^{K}\gamma_{dk} - K)\psi(\sum_{k=1}^{K}\gamma_{dk}) - \sum_{k=1}^{K}(\gamma_{dk} - 1)\psi(\gamma_{dk})\right] \\
&+ \sum_{k=1}^{K}\left[\log B(\lambda_k) + (\sum_{w=1}^{W}\lambda_{kw} - W)\psi(\sum_{w=1}^{W}\lambda_{kw}) - \sum_{w=1}^{W}(\lambda_{kw} - 1)\psi(\lambda_{kw})\right] \\
&- \sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{dnk}\log\phi_{dnk}
\end{aligned}}$$

Question 1.4.19:

1.5 BBVI

Question 1.5.20: First, let's derive the gradient of the ELBO with respect to ν:

$$
\begin{aligned}
\nabla_\nu \mathcal{L} &= \nabla_\nu \int q(\theta|\nu,\epsilon^2)\left[\log p(x,\theta) - \log q(\theta|\nu,\epsilon^2)\right] d\theta \\
&= \int \nabla_\nu q(\theta|\nu,\epsilon^2)\left[\log p(x,\theta) - \log q(\theta|\nu,\epsilon^2)\right] d\theta - \int q(\theta|\nu,\epsilon^2)\nabla_\nu \log q(\theta|\nu,\epsilon^2) d\theta \\
&= \int q(\theta|\nu,\epsilon^2)\nabla_\nu \log q(\theta|\nu,\epsilon^2)\left[\log p(x,\theta) - \log q(\theta|\nu,\epsilon^2)\right] d\theta
\end{aligned}
$$

where we used $\nabla_\nu \log q(\theta|\nu,\epsilon^2) = \frac{\nabla_\nu q(\theta|\nu,\epsilon^2)}{q(\theta|\nu,\epsilon^2)}$ and $\int q(\theta|\nu,\epsilon^2)\nabla_\nu \log q(\theta|\nu,\epsilon^2)d\theta = \int \nabla_\nu q(\theta|\nu,\epsilon^2)d\theta$ which equals 0 by commuting the gradient and the integral and noticing that the integral equals 1. Therefore, an estimate of the gradient of the ELBO using one sample $z \sim q(\theta|\nu,\epsilon^2)$ is obtained as follows:

$$\nabla_\nu \mathcal{L} \approx \nabla_\nu \log q(z|\nu,\epsilon^2)\left[\log p(x,z) - \log q(z|\nu,\epsilon^2)\right]$$

with

$$
\begin{aligned}
\nabla_\nu \log q(z|\nu,\epsilon^2) &= \frac{1}{\epsilon^2}(\log z - \nu) \\
\log q(z|\nu,\epsilon^2) &= -\log(z\epsilon\sqrt{2\pi}) - \frac{1}{2\epsilon^2}(\log z - \nu)^2 \\
\log p(x,z) &= -\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}(x-z)^2 + \alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log z - \beta z
\end{aligned}
$$

Question 1.5.21: Control variates is a variance reduction method used to compute a noisy estimate of the components of the gradient of the ELBO (obtained by the Rao-Blackwellization method) with low variance. Mathematically, if we want to estimate the expected value of f, it introduces a family $\hat{f}(z) = f(z) - a(h(z) - \mathbf{E}[h(z)])$ such that $\mathbf{E}[\hat{f}] = \mathbf{E}[f]$ but $Var[\hat{f}] < Var[f]$.

References

1. Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022. `https://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf`.