Report assignement 1A - Machine Learning Advanced DD2434

Flandre Corentin, Michel Le Dez*

from November 14, 2023 to November 30, 2023

1 1A Assignement

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:

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

*flandre@kth.se, micld@kth.se

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:

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

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:

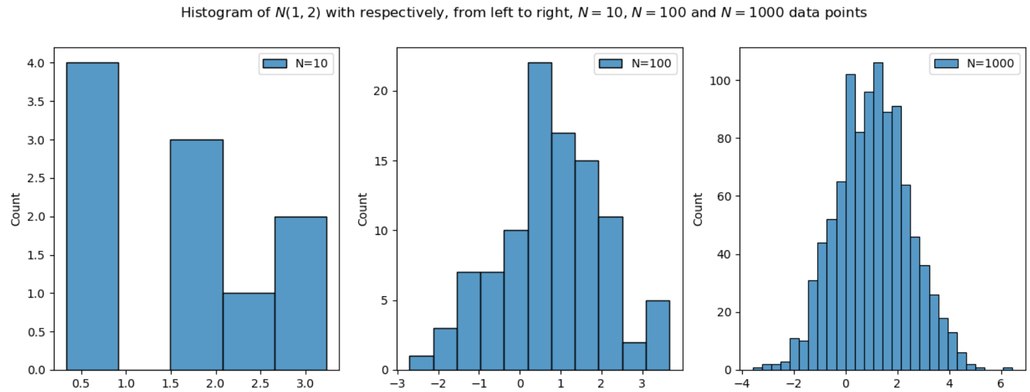

Figure 1: Histogram of $\mathcal{N}(\mu, \frac{1}{\tau})$ with $\mu = 1$ and $\tau = 0.5$ with respectively, from left to right, $N = 10$, $N = 100$, $N = 1000$ data points.

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.

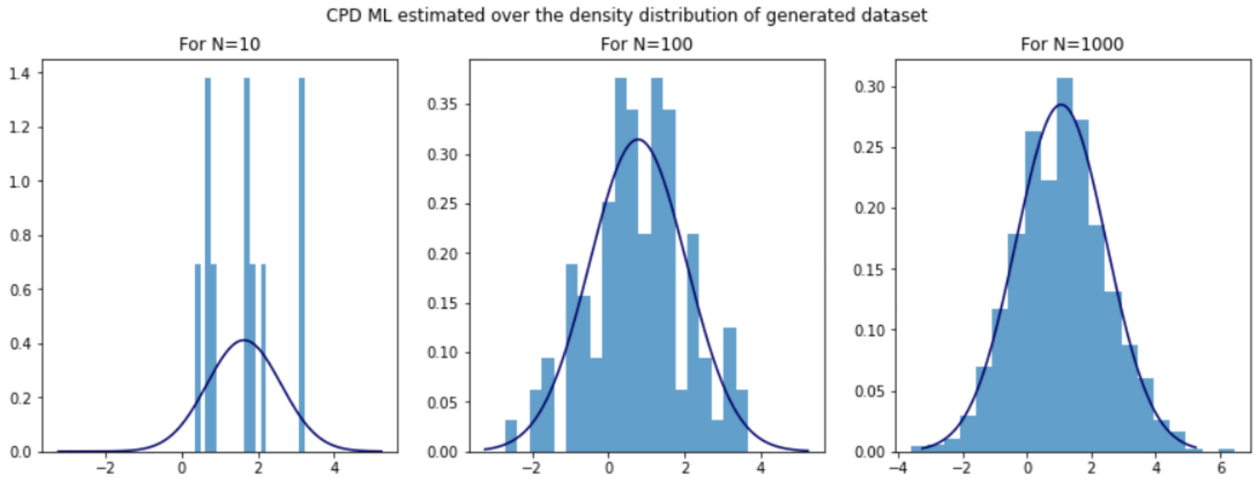

Figure 2: Visualization of Conditional Probability Distribution of posteriors estimates over generated datasets$\mathcal{N}(\mu, \frac{1}{\tau})$ with $\mu = 1$ and $\tau = 0.5$ with respectively, from left to right, $N = 10$, $N = 100$, $N = 1000$ data points.

Question 1.3.14: To compute the posterior, we are going to use the Bayes' theorem, use the equation (1)-(3) and gather in the constant term all the terms that do not depend on μ or τ. Here is the posterior:

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

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

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

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 $\frac{1}{2}\frac{(\sum_{n=1}^{N} x_n + \lambda_0\mu_0)^2}{\lambda_0 + N}$ for completing the square, which yields to:

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

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

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

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 $\mu \sim Normal(\mu_0^*, \lambda_0^*)$, the log of the probability density function is as follows:

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

By identification, we have:

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

1.4 SVI - LDA

Question 1.4.16: Local hidden variables according Hoffman is in a local context (as a subspace of a specific dimensionality, i.e, the document), correspond to latent variables not observerd often named $z_{1:N}$ impacting observations $x_{1:N}$.

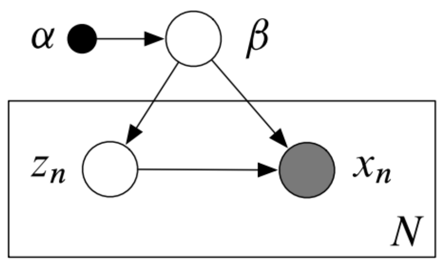

Figure 3: Hoffman's DGMM example to present graphical model with observations $x_{1:N}$, local hidden variables $z_{1:N}$, global hidden variables β and fixed parameters α.

In terms of conditional probability, we can write:

$$p(x_n|\beta, z_{1:N}) = p(x_n|\beta, z_n) \qquad and \qquad p(x_n|\beta, z_{1:N-n}) = p(x_n|\beta)$$

Effectively, all β impact x_n and z_n. On the contrary only z_n impacts x_n among z that's why we can define it local context.

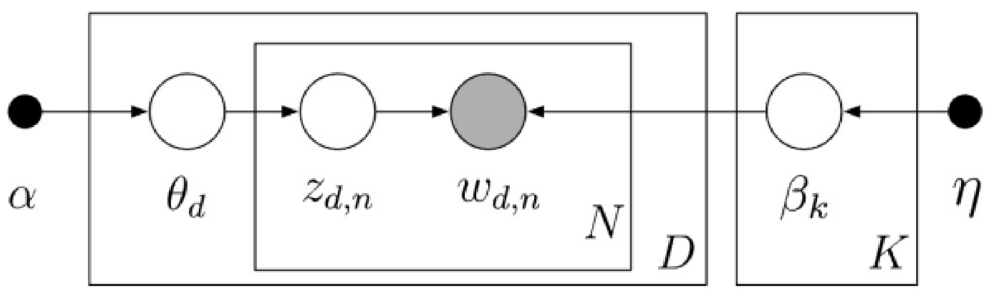

Figure 4: LDA DGM for section 1.4 SVI - LDA

Question 1.4.17: In the LDA model, global hidden variables are $\beta_{1:K}$ (corresponding to topics). Local hidden variables are θ_d and $z_{d,1:N}$ (corresponding to topic proportions and topics assignements). Of course, local observations are words $w_{d,1:N}$.

Question 1.4.18: The final expression of ELBO for the LDA model as a function of variational parameters and natural parameters of the full conditionals is :

$$\sum_{k=1}^{K} E[\log p(\overrightarrow{\beta_k}|\eta)] + \sum_{d=1}^{D} E[\log p(\overrightarrow{\theta_{d|}}\overrightarrow{\alpha})] + \sum_{d=1}^{D}\sum_{n=1}^{N} E[\log p(Z_{d,n}|\overrightarrow{\theta_d})] + \sum_{d=1}^{D}\sum_{n=1}^{N} E[\log p(w_{d,n}|Z_{d,n}, \overrightarrow{B}_{1:K})] + H(q)$$

The source for this derivation is the paper Topic Models from David M.Blei and John D. Lafferty

Question 1.4.19: blablabla todo blabla

1.5 BBVI

Question 1.5.20: The gradient estimate $w.r.t.\ \vartheta$ is

$$\nabla_\vartheta L = E_{q(\theta,\vartheta)}[\nabla_\vartheta \ln q(\theta,\vartheta)(\ln p(X,\theta) - \ln q(\theta,\vartheta))]$$

with

$$\begin{aligned}
\ln q(\theta,\vartheta) &= -\ln(\theta) - \frac{1}{2}(\frac{\ln(\theta)-\vartheta}{\epsilon})^2 + cst \\
\nabla_\vartheta \ln q(\theta,\vartheta) &= \frac{1}{\epsilon^2}(\ln(\theta)-\vartheta) + cst \\
\ln p(X,\theta) = \ln p(X|\theta) * \ln p(\theta) &= \frac{-1}{2}(\frac{X-\theta}{\epsilon})^2 + (\alpha-1)\ln(\theta) - \beta\theta + cst
\end{aligned}$$

So, the final expression can be written as :

$$\nabla_\vartheta L = E_{q(\theta,\vartheta)}[\frac{1}{\epsilon^2}(\ln(\theta)-\vartheta) * (\alpha \ln\theta - \beta\theta - \frac{1}{2}\frac{X-\theta}{\sigma}^2) + \frac{1}{2}\frac{\ln\theta-\vartheta}{\epsilon}^2)]$$

Question 1.5.21: Control variates descibred in the BBVI paper for the Module 5 - Black-Box VI, care used to reduce the variance of Monte Carlo gradient estimates, enhancing the efficiency of stochastic gradient ascent during the optimization of the variational parameters.

1.6 Appendix