Machine Learning 2 - Homework 6

Sindy Löwe (11594969)

May 13, 2018

Collaborators: Pascal Esser, Andrii Skliar, Gabriele Cesa, Davide Belli

Problem 1

a) The pseudocode for the Rejection Sampler is given in algorithm 1.

Algorithm 1 Rejection Sampling

Assume that we can sample from q
Assume that $q(x) \neq 0$ where $p(x) \neq 0$

Sample $x_i \sim q$
Compute $\tilde{q}(x_i) \cdot c$
Sample $u_i \sim U[0, c \cdot \tilde{q}(x_i)]$
Compute $\tilde{p}(x_i)$
if $u_i > \tilde{p}(x_i)$ **then**
    Reject the sample
else
    Keep the sample x_i

b) Yes, the generated samples are independent from each other. This is due to the fact that they only depend on uniform samples, which are drawn independently.

c) w_n can be expressed as:

$$w_n = \frac{\tilde{p}(x_n)}{\tilde{q}(x_n)} = \frac{Z_p \cdot p(x_n)}{Z_q \cdot q(x_n)}$$

Where Z_p and Z_q are the normalization constants for p and q respectively.

d) The acceptance probability for the Independence Sampler can be expressed as:

$$
\begin{aligned}
\alpha(x_{t+1}, x_t) &= \min\left(1, \frac{\tilde{p}(x_{t+1})}{\tilde{p}(x_t)} \frac{q(x_t|x_{t+1})}{q(x_{t+1}|x_t)}\right) \\
&= \min\left(1, \frac{\tilde{p}(x_{t+1})}{\tilde{p}(x_t)} \frac{\frac{q(x_{t+1}|x_t)\cdot q(x_t)}{q(x_{t+1})}}{q(x_{t+1})}\right) \\
&= \min\left(1, \frac{\tilde{p}(x_{t+1})}{\tilde{p}(x_t)} \frac{\frac{q(x_{t+1})\cdot q(x_t)}{q(x_{t+1})}}{q(x_{t+1})}\right) \\
&= \min\left(1, \frac{\tilde{p}(x_{t+1})}{\tilde{p}(x_t)} \frac{q(x_t)}{q(x_{t+1})}\right)
\end{aligned}
$$

e) Two subsequent samples are dependent, as the acceptance probability is calculated using both the current and the proposed states (x_t and x_{t+1})

f) The sequence of states generated by the Independence sampler will be: x_1, x_1, x_3, x_4, x_4

g) Rejection Sampling and Importance Sampling both do not work well in high-dimensional settings, since they depend on the fact that $\tilde{q}$ needs to approximate $\tilde{p}$ as closely as possible in order to provide a good performance. Now, the higher-dimensional the setting is, the bigger the volume between the surfaces becomes and the harder it is to find a $\tilde{q}$ that is a close enough approximation. The Independence Sampler works better in a higher-dimensional setting, due to the combination of a Markov Chain with the Monte Carlo approach.

Problem 2

In order to use Gibbs sampling for the posterior distribution $p(\mu, \tau|x)$, we need to derive the distributions for $p(\mu|\tau, x)$ and $p(\tau|\mu, x)$ in order to sample from them. First, we use the joint distribution:

$$
\begin{aligned}
p(\mu, \tau, x) &= p(\mu) \cdot p(\tau) \cdot p(x|\mu, \tau) \\
&= \mathcal{N}(\mu|\mu_0, s_0) \cdot \text{Gamma}(\tau|a, b) \cdot \mathcal{N}(x|\mu, \tau^{-1})
\end{aligned}
$$

For $p(\mu|\tau, x)$, we get:

$$
\begin{aligned}
p(\mu|\tau, x) &= \frac{p(\mu, \tau, x)}{p(\tau, x)} \\
&= \frac{p(\mu, \tau, x)}{\int p(\mu, \tau, x) d\mu} \\
&= \frac{p(\mu) \cdot p(\tau) \cdot p(x|\mu, \tau)}{p(\tau) \cdot \int p(\mu) \cdot p(x|\mu, \tau) d\mu} \\
&= \frac{p(\mu) \cdot p(x|\mu, \tau)}{\int p(\mu) \cdot p(x|\mu, \tau) d\mu} \\
&\propto \mathcal{N}(\mu|\mu_0, s_0) \cdot \mathcal{N}(x|\mu, \tau^{-1}) \\
&= (2\pi s_0)^{-\frac{1}{2}} \exp\left(-\frac{1}{2s_0}(\mu - \mu_0)^2\right) \cdot (2\pi\tau^{-1})^{-\frac{1}{2}} \exp\left(-\frac{1}{2\tau^{-1}}(x - \mu)^2\right) \\
&\propto -\frac{1}{2s_0}(\mu - \mu_0)^2 - \frac{1}{2\tau^{-1}}(x - \mu)^2 \\
&= -\frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{2s_0} - \frac{\tau(x^2 - 2x\mu + \mu^2)}{2} \\
&= -\mu^2\left(\frac{1}{2s_0} + \frac{\tau}{2}\right) + \mu\left(\tau x + \frac{\mu_0}{s_0}\right) + \text{const} \\
&= -\left(\frac{1}{2s_0} + \frac{\tau}{2}\right)\left(\mu - \frac{\tau x + \frac{\mu_0}{s_0}}{\frac{1}{s_0} + \tau}\right)^2 + \text{const} \\
&\propto \mathcal{N}\left(\mu \mid \frac{\tau x + \frac{\mu_0}{s_0}}{\frac{1}{s_0} + \tau}, \left(\frac{1}{s_0} + \tau\right)^{-1}\right)
\end{aligned}
$$

For $p(\tau|\mu, x)$, we get:

$$
\begin{aligned}
p(\tau|\mu, x) &= \frac{p(\mu, \tau, x)}{p(\mu, x)} \\
&= \frac{p(\mu) \cdot p(\tau) \cdot p(x|\mu, \tau)}{p(\mu) \cdot \int p(\tau) \cdot p(x|\mu, \tau) d\tau} \\
&\propto \mathcal{N}(x|\mu, \tau^{-1}) \cdot \text{Gamma}(\tau|a, b) \\
&= (2\pi\tau^{-1})^{-\frac{1}{2}} \exp\left(-\frac{1}{2\tau^{-1}}(x - \mu)^2\right) \cdot \frac{b^a}{\Gamma(a)}\tau^{a-1} \exp(-b\tau) \\
&= \frac{b^a}{\Gamma(a)\sqrt{2\pi}}\tau^{a-\frac{1}{2}} \exp\left(-\tau\left(\frac{1}{2}(x - \mu)^2 + b\right)\right) \\
&\propto \text{Gamma}\left(\tau \mid a + \frac{1}{2}, \frac{1}{2}(x - \mu)^2 + b\right)
\end{aligned}
$$

Problem 3

1. The joint probability over the observed data and latent variables is:

$$
p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \boldsymbol{\phi} \mid \alpha, \beta) = p(\boldsymbol{\theta}|\alpha) \cdot p(\mathbf{z}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\phi}|\beta) \cdot p(\mathbf{w}|\boldsymbol{\phi}, \mathbf{z})
$$

2. Integrating out the parameters gives us:

$$\begin{aligned}
p(\mathbf{z}, \mathbf{w} \mid \alpha, \beta) &= \iint p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \boldsymbol{\phi} \mid \alpha, \beta) d\boldsymbol{\theta} d\boldsymbol{\phi} \\
&= \iint p(\boldsymbol{\theta}|\alpha) \cdot p(\mathbf{z}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\phi}|\beta) \cdot p(\mathbf{w}|\boldsymbol{\phi}, \mathbf{z}) d\boldsymbol{\theta} d\boldsymbol{\phi} \\
&= \int p(\boldsymbol{\theta}|\alpha) \cdot p(\mathbf{z}|\boldsymbol{\theta}) d\boldsymbol{\theta} \cdot \int p(\boldsymbol{\phi}|\beta) \cdot p(\mathbf{w}|\boldsymbol{\phi}, \mathbf{z}) d\boldsymbol{\phi}
\end{aligned}$$

We can see that both terms constitute the integral over the product of a multinomial with a Dirichlet distribution. Evaluating the first term, we get:

$$\begin{aligned}
\int p(\boldsymbol{\theta}|\alpha) \cdot p(\mathbf{z}|\boldsymbol{\theta}) d\boldsymbol{\theta} &= \prod_{d=1}^{D} \int p(\boldsymbol{\theta}_d|\alpha) \cdot \prod_{n=1}^{N_d} p(z_{dn}|\boldsymbol{\theta}_d) d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \int \frac{1}{B(\alpha)} \prod_{k=1}^{K} \boldsymbol{\theta}_{dk}^{\alpha-1} \cdot \prod_{k=1}^{K} \theta_{dk}^{A_{dk}} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \frac{1}{B(\alpha)} \int \prod_{k=1}^{K} \boldsymbol{\theta}_{dk}^{\alpha-1+A_{dk}} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \frac{B(\mathbf{A}_d + \alpha)}{B(\alpha)} \int \frac{1}{B(\mathbf{A}_d + \alpha)} \prod_{k=1}^{K} \boldsymbol{\theta}_{dk}^{A_{dk}+\alpha-1} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \frac{B(\mathbf{A}_d + \alpha)}{B(\alpha)} \int \mathrm{Dir}(\boldsymbol{\theta}_d \mid \mathbf{A}_d + \alpha) d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \frac{B(\mathbf{A}_d + \alpha)}{B(\alpha)}
\end{aligned}$$

For the second term, we get:

$$\begin{aligned}
\int p(\boldsymbol{\phi}|\beta) \cdot p(\mathbf{w}|\boldsymbol{\phi}, \mathbf{z}) d\boldsymbol{\phi} &= \int \prod_{k=1}^{K} (p(\boldsymbol{\phi}_k|\beta)) \cdot \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(\mathbf{w}_{dn}|\boldsymbol{\phi}_k, \mathbf{z}_{dn}) d\boldsymbol{\phi}_k \\
&= \prod_{k=1}^{K} \frac{1}{B(\beta)} \int \prod_{w} \phi_{kw}^{\beta-1} \cdot \prod_{w} \phi_{kw}^{B_{kw}} d\boldsymbol{\phi}_k \\
&= \prod_{k=1}^{K} \frac{1}{B(\beta)} \int \prod_{w} \phi_{kw}^{\beta-1+B_{kw}} d\boldsymbol{\phi}_k \\
&= \prod_{k=1}^{K} \frac{B(\mathbf{B}_k + \beta)}{B(\beta)}
\end{aligned}$$

Therefore:

$$p(\mathbf{z}, \mathbf{w} \mid \alpha, \beta) = \prod_{d=1}^{D} \frac{B(\mathbf{A}_d + \alpha)}{B(\alpha)} \cdot \prod_{k=1}^{K} \frac{B(\mathbf{B}_k + \beta)}{B(\beta)}$$

3. We get the Gibbs sampling update by calculating (denoting all elements but i with $-i$ and leaving out α, β for clarity):

$$
\begin{aligned}
p(z_{di}|\mathbf{z}_{d-i}, \mathbf{w}) &= \frac{p(\mathbf{z}_d, \mathbf{w}_d)}{p(\mathbf{z}_{d-i}, \mathbf{w}_d)} \\
&= \frac{\frac{B(\mathbf{A}_d+\alpha)}{B(\alpha)} \cdot \prod_{k=1}^{K} \frac{B(\mathbf{B}_k+\beta)}{B(\beta)}}{\frac{B(\mathbf{A}_d(-i)+\alpha)}{B(\alpha)} \cdot \prod_{k=1}^{K} \frac{B(\mathbf{B}_k(-i)+\beta)}{B(\beta)}} \\
&= \frac{B(\mathbf{A}_d + \alpha) \cdot \prod_{k=1}^{K} B(\mathbf{B}_k + \beta)}{B(\mathbf{A}_d(-i) + \alpha) \cdot \prod_{k=1}^{K} B(\mathbf{B}_k(-i) + \beta)}
\end{aligned}
$$

Problem 4

a) The mean of $\mathbf{x}$ under this distribution is:

$$
\underset{p(\mathbf{x}|\boldsymbol{\mu})}{\mathbb{E}}[\mathbf{x}] = \left[\underset{p(x_1|\mu_1)}{\mathbb{E}}[x_1], ..., \underset{p(x_D|\mu_D)}{\mathbb{E}}[x_D] \right]^T = \boldsymbol{\mu}
$$

b) For $i \neq j : x_i \perp\!\!\!\perp x_j$, therefore, we get $\mathbf{\Sigma}_{ij} = 0$. For $\mathbf{\Sigma}_{ii}$, we get $\mathbf{\Sigma}_{ii} = \mathrm{var}(x_i) = \mu_i(1 - \mu_i)$

c) The mean of $\mathbf{x}$ under this mixture distribution is:

$$
\begin{aligned}
\underset{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\pi})}{\mathbb{E}}[\mathbf{x}] &= \sum_{k=1}^{K} \pi_k \underset{p(\mathbf{x}|\boldsymbol{\mu}_k)}{\mathbb{E}}[\mathbf{x}] \\
&= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k
\end{aligned}
$$

d) The log-likelihood for this model is:

$$
\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k) \right)
$$

e) Since there is a summation inside the logarithm, there is no closed form solution for the standard maximum-likelihood approach.

f) The complete-data log-likelihood function for this model is:

$$
\begin{aligned}
\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) &= \sum_{n=1}^{N} \left(\sum_{k=1}^{K} z_{nk} (\ln(\pi_k) + \ln p(\mathbf{x}_n|\boldsymbol{\mu}_k)) \right) \\
&= \sum_{n=1}^{N} \left(\sum_{k=1}^{K} z_{nk} \left(\ln(\pi_k) + \sum_{i=1}^{D} [x_{ni} \ln(\mu_{ki}) + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right) \right)
\end{aligned}
$$

g) We can draw the corresponding graphical model using plate notation as in fig. 1.

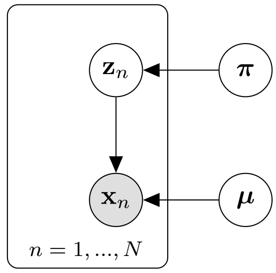

Figure 1: Graphical model

h) The VEM objective function is:

$$\begin{aligned}
\mathcal{B}(\{q_n(\mathbf{z}_n)\}, \boldsymbol{\mu}, \boldsymbol{\pi}) &= \sum_{n=1}^{N} H(q_n) + \sum_{n=1}^{N} \mathop{\mathbb{E}}_{q_n}[\ln p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\mu}, \boldsymbol{\pi})] \\
&= \sum_{n=1}^{N} \sum_{\mathbf{z}_n} q_n(\mathbf{z}_n) \log p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\mu}, \boldsymbol{\pi}) - \sum_{n=1}^{N} \sum_{\mathbf{z}_n} q_n(\mathbf{z}_n) \log q_n(\mathbf{z}_n) \\
&= \sum_{n=1}^{N} \sum_{\mathbf{z}_n} q_n(\mathbf{z}_n) (\sum_{k=1}^{K} z_{nk} \left(\ln(\pi_k) + \sum_{i=1}^{D} [x_{ni} \ln(\mu_{ki}) + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right) \\
&- \log q_n(\mathbf{z}_n))
\end{aligned}$$

i) The VEM objective function including Lagrangian multipliers is:

$$\begin{aligned}
\tilde{\mathcal{B}}(\{q_n(\mathbf{z}_n)\}, \boldsymbol{\mu}, \boldsymbol{\pi}) &= \sum_{n=1}^{N} \sum_{\mathbf{z}_n} q_n(\mathbf{z}_n) (\sum_{k=1}^{K} z_{nk} \left(\ln(\pi_k) + \sum_{i=1}^{D} [x_{ni} \ln(\mu_{ki}) + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right) \\
&- \log q_n(\mathbf{z}_n)) + \alpha(\sum_{k=1}^{K} \pi_k - 1) + \sum_{n=1}^{N} \lambda_n (\sum_{\mathbf{z}_n} q_n(\mathbf{z}_n) - 1)
\end{aligned}$$

j) For the E-Step, we optimize $\tilde{\mathcal{B}}$ with respect to q_n as follows:

$$\begin{aligned}
\frac{\partial \tilde{\mathcal{B}}}{\partial q_n} &= \sum_{k=1}^{K} z_{nk} \left(\ln(\pi_k) + \sum_{i=1}^{D} \left[x_{ni} \ln(\mu_{ki}) + (1 - x_{ni}) \ln(1 - \mu_{ki}) \right] \right) \\
&- 1 - \log q_n(\mathbf{z}_n) + \lambda_n = 0
\end{aligned}$$

$$\begin{aligned}
\lambda_n - 1 &= -\sum_{k=1}^{K} z_{nk} \left(\ln(\pi_k) + \sum_{i=1}^{D} \left[x_{ni} \ln(\mu_{ki}) + (1 - x_{ni}) \ln(1 - \mu_{ki}) \right] \right) \\
&+ \log q_n(\mathbf{z}_n) \\
\exp(\lambda_n - 1) &= \left(\prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \prod_{i=1}^{D} \left[\mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}} \right] \right)^{-1} \cdot q_n(\mathbf{z}_n) \\
\exp(\lambda_n - 1) &= \left(\sum_{\mathbf{z}_n} \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \prod_{i=1}^{D} \left[\mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}} \right] \right)^{-1}
\end{aligned}$$

$$\begin{aligned}
q_n(\mathbf{z}_n) &= \exp(\sum_{k=1}^{K} z_{nk} \left(\ln(\pi_k) + \sum_{i=1}^{D} \left[x_{ni} \ln(\mu_{ki}) + (1 - x_{ni}) \ln(1 - \mu_{ki}) \right] \right) \\
&- 1 + \lambda_n) \\
&= \prod_{k=1}^{K} \pi_k^{z_{nk}} \left(\prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}} \right)^{z_{nk}} \exp(\lambda_n - 1) \\
&= \frac{\prod_{k=1}^{K} \pi_k^{z_{nk}} \left(\prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}} \right)^{z_{nk}}}{\sum_{\mathbf{z}_n} \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \prod_{i=1}^{D} \left[\mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}} \right]} \\
&= \frac{\prod_{k=1}^{K} \pi_k^{z_{nk}} p(\mathbf{x}_n | \boldsymbol{\mu}_k)^{z_{nk}}}{\sum_{k=1}^{K} \pi_k \cdot p(\mathbf{x}_n | \boldsymbol{\mu}_k)} \\
&= \frac{p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\mu}, \boldsymbol{\pi})}{p(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\pi})} \\
&= p(\mathbf{z}_n | \mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\pi})
\end{aligned}$$

This means, that in the E-Step, we update q_n by setting it to the posterior probability of z_n.

k) For the M-Step, we optimize $\tilde{\mathcal{B}}$ with respect to π_k as follows:

$$\begin{aligned}
\frac{\partial \tilde{\mathcal{B}}}{\partial \pi_k} &= \frac{\sum_{n=1}^{N} \sum_{\mathbf{z}_n} q_n(\mathbf{z}_n) z_{nk}}{\pi_k} + \alpha = 0 \\
\sum_{k=1}^{K} \alpha \pi_k &= -\sum_{k=1}^{K} N_k \\
\alpha &= -N \\
\pi_k &= \frac{N_k}{N}
\end{aligned}$$

Since $\sum_{\mathbf{z}_n} q_n(\mathbf{z}_n) = 1$ and $\sum_{n=1}^{N} z_{nk} = N_k$

Problem 5

As a first step, we can write $z^{(r)} = \sum_{t=1}^{r} x_t$, where:

$$\begin{aligned} p(x_t = 1) &= 0.25 \\ p(x_t = 0) &= 0.5 \\ p(x_t = -1) &= 0.25 \end{aligned}$$

This gives us $\mathbb{E}[x_t] = 0$ and $\text{var}(x_t) = \mathbb{E}[x_t^2] = \frac{1}{2}$.

Additionally, we can see that $\mathbb{E}[(z^{(r)})] = 0$. This gives:

$$\frac{r}{2} = \sum_{t=1}^{r} \text{var}(x_t) = \text{var}\left(\sum_{t=1}^{r} x_t\right) = \text{var}(z^{(r)}) = \mathbb{E}[(z^{(r)})^2] - \mathbb{E}[(z^{(r)})]^2 = \mathbb{E}[(z^{(r)})^2]$$