Machine Learning 2 — Homework 6

Maurice Frank
11650656
maurice.frank@posteo.de

October 14, 2019

Problem 1.

We have PDF $p(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$ and the approximation of it $q(\boldsymbol{x})$.

a)

Algorithm 1 Rejection Sampler

```
function SAMPLE_REJECTION($q$, $\tilde{p}$, $c$)
    $acc \leftarrow False$
    while $acc == False$ do
        $z_0 \leftarrow \text{sample}(q(z))$
        $u \leftarrow \text{rand}[0, c \cdot q(z_0)]$
        $acc \leftarrow u \leq \tilde{p}(z)$
    end while
    return $z$
end function
```

b)

Yes the generated samples are independent. We sample independently from $q(\boldsymbol{z})$ which makes the accepted samples also independently.

c)

$$w_n = \frac{p(z_n)/q(z_n)}{\sum_m p(z_m)/q(z_m)}$$

d)

$$\begin{aligned}
\alpha(x_{t+1}, x_t) &= \min\left(1, \frac{\tilde{p}(x_{t+1})q(x_t|x_{t+1})}{\tilde{p}(x_t)q(x_{t+1}|x_t)}\right) \\
&= \min\left(1, \frac{\tilde{p}(x_{t+1})q(x_t|x_{t+1})}{\tilde{p}(x_t)q(x_{t+1})}\right) \\
&= \min\left(1, \frac{\tilde{p}(x_{t+1})q(x_{t+1}|x_t)q(x_t)}{\tilde{p}(x_t)q(x_{t+1})^2}\right) \\
&= \min\left(1, \frac{\tilde{p}(x_{t+1})q(x_t)}{\tilde{p}(x_t)q(x_{t+1})}\right) \\
&= \min\left(1, \frac{p(x_{t+1})q(x_t)}{p(x_t)q(x_{t+1})}\right)
\end{aligned}$$

e)

A proposed sample using the proposal distribution is chosing independent of the previous sample: $q(x_{t+1}|x_t) = q(x_t)$. Whether or not the proposal is added to the chain though, is dependent on the previous sample's value. Thus two accepted successive samples in the chain are not independent of each other.

f)

When a proposal is rejected the independence sampler, samples the previous sample again:

$$[x_1, x_1, x_3, x_4, x_4]$$

g)

The rejection sampler compares the volumes between the proposal approximate distribution and the unnormalized true distribution. With higher dimensionality the acceptance ratio will become exponentially small. Therefore the rejection sampler does perform considerably worse with higher dimensionality. Using the importance sampler we will have the problem of the sum in the weights to calculate. With more dimensions this sum will scale exponentially. Further the importance sampler has the problem that the proposal distribution has to have big enough mass over the whole support of the true distribution. As we increase the number of dimensions the mass is distributed over a exponentially growing volume thus increasing the risk of zeroing/near-zeroing an area which will make the weights ununsable.

The independence sampler does not have theses *curse of dimensionality* problems that the rejection and importance sampler have. Through the Markov chain property the sampling for the independence sampler only depends on the previous sampler. The only complexity dependending on the number of dimensions is the sampling from the proposal distribution.

Problem 2.

We have $x \sim \mathcal{N}(x|\mu, \tau^{-1})$, $\mu \sim \mathcal{N}(\mu|\mu_0, s_0)$ and $\tau \sim \text{Gamma}(\tau|a, b)$.

To gibbs sample from the posterior $p(\mu, \tau|x)$ we iteratively sample from the conditionals $p(\mu|\tau, x)$ and $p(\tau|\mu, x)$.

$$\begin{aligned} p(\mu|\tau, x) &= \mathcal{N}\left(\mu|\frac{\tau^{-1}}{s_0 + \tau^{-1}} \cdot \mu_0 + \frac{s_0}{s_0 + \tau^{-1}} \cdot x, \frac{1}{\frac{1}{s_0} + \frac{1}{\tau^{-1}}}\right) \\ p(\tau|\mu, x) &= \text{Gamma}(\tau|a + \frac{1}{2}, b + \frac{1}{2}(x - \mu)2) \end{aligned}$$

(using Bishop 2.141/142 and 2.150/151)

Problem 3.

1.

$$
\begin{aligned}
p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\alpha, \beta) &= p(\boldsymbol{W}|\boldsymbol{Z}, \boldsymbol{\Phi}) \cdot p(\boldsymbol{Z}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Phi}|\beta) \cdot p(\boldsymbol{\Theta}|\alpha) \\
&= p(\boldsymbol{W}|\boldsymbol{\Phi}) \cdot p(\boldsymbol{\Phi}|\beta) \cdot p(\boldsymbol{Z}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}|\alpha) \\
p(\boldsymbol{Z}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}|\alpha) &= \prod_{d}^{D} p(\boldsymbol{z}_d|\boldsymbol{\theta}_d) p(\boldsymbol{\theta}_d|\alpha) \\
&= \prod_{d}^{D} (\prod_{n}^{N_d} \prod_{k}^{K} p(z_{dn} = k|\theta_{dk})^{\delta(z_{dn}=k)}) p(\boldsymbol{\theta}_d|\alpha) \\
&= \prod_{d}^{D} (\prod_{k}^{K} \theta_{dk}^{A_{dk}}) \cdot (\frac{1}{\mathrm{B}(\alpha)} \prod_{k}^{K} \theta_{dk}^{\alpha-1}) \\
&= \prod_{d}^{D} \frac{1}{B(\alpha)} \prod_{k}^{K} \theta_{dk}^{A_{dk}+\alpha-1} \\
p(\boldsymbol{W}|\boldsymbol{\Phi}) \cdot p(\boldsymbol{\Phi}|\beta) &= \prod_{k}^{K} (\prod_{d}^{D} \prod_{n}^{N_d} p(w_{dn}|z_{dn}, \phi_k)) \cdot p(\boldsymbol{\phi}_k|\beta) \\
&= \prod_{k}^{K} (\prod_{d}^{D} \prod_{n}^{N_d} \prod_{w}^{W} p(w_{dn} = w|z_{dn} = k, \phi_{kw})^{\delta(w_{dn}=w)\delta(z_{dn}=k)}) \cdot \frac{1}{B(\beta)} \prod_{w}^{W} \phi_{kw}^{\beta-1} \\
&= \prod_{k}^{K} \prod_{w}^{W} \phi_{kw}^{B_{kw}}, \phi_k \cdot \frac{1}{\mathrm{B}(\beta)} \prod_{w}^{W} \phi_{kw}^{\beta-1} \\
&= \prod_{k}^{K} \frac{1}{\mathrm{B}(\beta)} \prod_{w}^{W} \phi_{kw}^{\mathrm{B}_{kw}+\beta-1} \\
&\implies \\
p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\alpha, \beta) &= \left(\prod_{d}^{D} \frac{1}{\mathrm{B}(\alpha)} \prod_{k}^{K} \theta_{dk}^{A_{dk}+\alpha-1} \right) \left(\prod_{k}^{K} \frac{1}{\mathrm{B}(\beta)} \prod_{w}^{W} \phi_{kw}^{B_{kw}+\beta-1} \right)
\end{aligned}
$$

2.

$$
\begin{aligned}
\int p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\alpha, \beta) d\boldsymbol{\theta}_d &= \int \prod_d^D \frac{1}{\mathrm{B}(\alpha)} \prod_k^K \theta_{dk}^{A_{dk}+\alpha-1} d\boldsymbol{\theta}_d \\
&= \prod_d^D \frac{1}{\mathrm{B}(\alpha)} \int \prod_k^K \theta_{dk}^{A_{dk}+\alpha-1} d\boldsymbol{\theta}_d \\
&= \prod_d^D \frac{1}{\mathrm{B}(\alpha)} \int \frac{\mathrm{B}(\boldsymbol{A}_d + \alpha)}{\mathrm{B}(\boldsymbol{A}_d + \alpha)} \prod_k^K \theta_{dk}^{A_{dk}+\alpha-1} d\boldsymbol{\theta}_d \\
&= \prod_d^D \frac{\mathrm{B}(\boldsymbol{A}_d + \alpha)}{\mathrm{B}(\alpha)} \int \mathrm{Dir}(\boldsymbol{A}_d + \alpha) d\boldsymbol{\theta}_d \\
&= \prod_d^D \frac{\mathrm{B}(\boldsymbol{A}_d + \alpha)}{\mathrm{B}(\alpha)} \\
\int p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\alpha, \beta) d\boldsymbol{\phi}_k &= \int \prod_k^K \frac{1}{\mathrm{B}(\beta)} \prod_w^W \phi_{kw}^{B_{kw}+\beta-1} d\boldsymbol{\phi}_k \\
&= \prod_k^K \frac{1}{\mathrm{B}(\beta)} \int \prod_w^W \phi_{kw}^{B_{kw}+\beta-1} d\boldsymbol{\phi}_k \\
&= \prod_k^K \frac{\mathrm{B}(\boldsymbol{B}_k + \beta)}{\mathrm{B}(\beta)} \int \frac{1}{\mathrm{B}(\boldsymbol{B}_k + \beta)} \prod_w^W \phi_{kw}^{B_{kw}+\beta-1} d\boldsymbol{\phi}_k \\
&= \prod_k^K \frac{\mathrm{B}(\boldsymbol{B}_k + \beta)}{\mathrm{B}(\beta)} \int \mathrm{Dir}(\boldsymbol{B}_k + \beta) d\boldsymbol{\phi}_k \\
&= \prod_k^K \frac{\mathrm{B}(\boldsymbol{B}_k + \beta)}{\mathrm{B}(\beta)}
\end{aligned}
$$

3.

$$
\begin{aligned}
z_{di} &\sim p(z_{di}|z_{d\{N\backslash i\}}, \boldsymbol{W}, \alpha, \beta) \\
p(\boldsymbol{Z}, \boldsymbol{W}, \alpha, \beta) &= \left(\prod_d^D \frac{\mathrm{B}(\boldsymbol{A}_d + \alpha)}{\mathrm{B}(\alpha)} \right) \left(\prod_k^K \frac{\mathrm{B}(\boldsymbol{B}_k + \beta)}{\mathrm{B}(\beta)} \right)
\end{aligned}
$$

$$
\begin{aligned}
p(z_{di}|z_{d\{N\setminus i\}}, \boldsymbol{W}, \alpha, \beta) &= \frac{p(\boldsymbol{Z}_d, \boldsymbol{W}_d, \alpha, \beta)}{p(\boldsymbol{Z}_{d\{N\setminus i\}}, \boldsymbol{W}, \alpha, \beta)} \\
&= \frac{\frac{\mathrm{B}(\boldsymbol{A}_d+\alpha)}{\mathrm{B}(\alpha)}(\prod_k^K \frac{\mathrm{B}(\boldsymbol{B}_k+\beta)}{\mathrm{B}(\beta)})}{\frac{\mathrm{B}(\boldsymbol{A}_{d,K\setminus i}+\alpha)}{\mathrm{B}(\alpha)}(\prod_k^K \frac{\mathrm{B}(\boldsymbol{B}_{k,W\setminus i}+\beta)}{\mathrm{B}(\beta)})} \\
&= \frac{\mathrm{B}(\boldsymbol{A}_{d,\{1,\ldots,K\}} + \alpha)(\prod_k^K \mathrm{B}(\boldsymbol{B}_{k,\{1,\ldots,V\}} + \beta))}{\mathrm{B}(\boldsymbol{A}_{d,\{1,\ldots,K\}\setminus z_{di}} + \alpha)(\prod_k^K \mathrm{B}(\boldsymbol{B}_{k,\{1,\ldots,V\}\setminus D_i} + \beta))}
\end{aligned}
$$

Problem 4.

We have $p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i}$, $\mu_i \in [0,1]$, $x_i \in \{0,1\}$.

a)

$$
\begin{aligned}
\mathbb{E}[x_i] &= \sum_{x_i} x_i \cdot p(x_i|\mu_o i) \\
&= \sum_{x_i} x_i \cdot \mu_i^{x_i}(1-\mu_i)^{1-x_i} \\
&= \mu_i \\
\mathbb{E}[\boldsymbol{x}] &= \boldsymbol{\mu}
\end{aligned}
$$

b)

$$
\mathrm{cov}(\boldsymbol{x}) = \mathrm{diag}(\boldsymbol{\mu}^T(1-\boldsymbol{\mu}))
$$

c)

Now we got an mixture: $p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k)$. And therefore now:

$$
\mathbb{E}[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k
$$

d)

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

e)

A standard maximum likelihood approach would not work here as we have need to compute the logarithm of the sum. We do not have a closed-form solution and the computation of this logarithm is difficult.

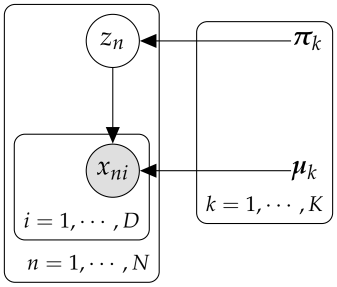
f)

For our variationl approach we have: $p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\mu},\boldsymbol{\pi}) = p(\boldsymbol{z}_n|\boldsymbol{\pi})p(\boldsymbol{x}_n|\boldsymbol{z}_n,\boldsymbol{\mu}) = \prod_{k=1}^{K} \pi_k^{z_{nk}} p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)^{z_{nk}}$.

The data log likelihood becomes:

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

g)



h)

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

i)

$$\begin{aligned}\tilde{\mathcal{B}}(\{q_n(\boldsymbol{z}_n)\}, \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{\boldsymbol{z}_k} q_n(\boldsymbol{z}_n) \Bigg(\Bigg[& \log \pi_{z_n} + \sum_{i=1}^{D} (x_{ni} \log \mu_{z_n i} + (1 - x_{ni}) \log(1 - \mu_{z_n i})) \Bigg] \\ & - \log q_n(\boldsymbol{z}_n) \Bigg) + \alpha \cdot \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \sum_{n=1}^{N} \lambda_n \cdot \left(\sum_{\boldsymbol{z}_N} q_n(\boldsymbol{z}_n) - 1 \right)\end{aligned}$$

j)

$$
\begin{aligned}
0 &= \frac{\partial}{\partial q_n(z_n)} \tilde{\mathcal{B}}(\{q_n(z_n)\}, \boldsymbol{\mu}, \boldsymbol{\pi}) \\
0 &= \left[\log \pi_{z_n} + \sum_{i=1}^{D} \left(x_{ni} \log \mu_{z_n i} + (1 - x_{ni}) \log(1 - \mu_{z_n i})\right)\right] - \log q_n(z_n) - 1 + \lambda_n \\
\log q_n(z_n) &= \left[\log \pi_{z_n} + \sum_{i=1}^{D} \left(x_{ni} \log \mu_{z_n i} + (1 - x_{ni}) \log(1 - \mu_{z_n i})\right)\right] + \lambda_n - 1 \\
q_n(z_n) &= \pi_{z_n} \cdot \prod_{i=1}^{D} \mu_{z_n i}^{x_{ni}} \cdot (1 - \mu_{z_n i})^{(1 - x_{ni})} \cdot \exp\left(\lambda_n - 1\right)
\end{aligned}
$$

In the E-step we want to reduce the KL-divergence between p and q. For that we want to raise the ELBO with the computed $q_n(z_n)$. As such the $q_n(z_n)$ are the best estimation of the likelihood.

k)

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \pi_k} \tilde{\mathcal{B}}(\{q_n(z_n)\}, \boldsymbol{\mu}, \boldsymbol{\pi}) \\
0 &= \frac{\sum_{n=1}^{N} \mathbb{E}[z_{nk}]}{\pi_k} + \alpha \\
0 &= \frac{N_k}{\pi_k} + \alpha \\
-\alpha \sum_{k}^{K} \pi_k &= \sum_{k}^{K} N_k \\
\alpha &= -N \\
&\Longrightarrow \\
\pi_k &= \frac{N_k}{N}
\end{aligned}
$$