Machine Learning 2 - Homework 6

Andrii Skliar, 11636785

deadline: May 13, 2018

During the process of solving the homework problems, I have collaborated with the following colleagues:

Gabriele Cesa Gabriele Bani Sindy Loẅe

NB: credits for the Latex-format go to Iris Verweij, 2nd year MSc AI Student.

Problem 1: In this question we are interested in generating samples from a probability density $p(x)$ with $x \in \mathbb{R}^d\, d$. We are given an approximation $q(x)$ of $p(x)$. We will denote unnormalized densities as $\widetilde{p}$ and $\widetilde{q}$.

a) Assume that you have a constant c such that $\widetilde{q}(x) = cq(x)$ and $q(x) \geq p(x), \forall x$. Describe with pseudocode the Rejection Sampler algorithm.

Solution: Rejection sampler algorithm looks as following:

Algorithm 1 Rejection sampling

```
Sample $x_i \sim q(x)$
Compute $c \cdot \widetilde{q}(x_i)$
Sample $u_i \sim U[0, c \cdot \widetilde{q}(x_i)]$
Compute $\widetilde{p}(x_i)$
if $u_i > \widetilde{p}(x_i)$ then
    Reject sample
else
    Accept sample
end if
```

b) Are the samples you generate independent from each other?

Solution: Yes, because samples are generated independently and previous sample doesn't have any influence on the next one.

c) An Importance Sampler accepts all samples but weights them using weights w_n . Provide the expression for w_n in terms of $p(x_n)$ and $q(x_n)$.

Solution: $w_n = \frac{p(x_n)}{q(x_n)}$

d) An Independence Sampler uses a proposal distribution of the form $q(x_{t+1}|x_t) = q(x_{t+1})$ (i.e. the proposed new state is independent of the previous state) and subsequently accepts or rejects this proposed state as the next state of the Markov chain. Provide the expression for the Metropolis Hastings accept probability $\alpha(x_{t+1}, x_t)$ in terms of p and q for the Independence Sampler.

Solution: $\alpha(x_{t+1}, x_t) = min\left(1, \frac{\widetilde{p}(x_{t+1})q(x_t)}{\widetilde{p}(x_t)q(x_{t+1})}\right)$

e) Are two subsequent samples from the Independence Sampler independent or dependent in general? Explain your answer.

Solution: Proposal distribution of x_{t+1} is independent of x_t, however, accept probability still depends on the previous state, therefore, two subsequent samples are generally not independent.

f) Imagine we run the Independence sampler for 5 steps and during these 5 steps we propose the states x_1, x_2, x_3, x_4, x_5 (think of these represent as numeric values, e.g. 0.34, 3.5, 2.67, 0.82, 1.60). The MCMC procedure rejects the proposals x_2 and x_5. Which sequence of states will the Independence sampler generate after 5 steps?

Solution: Independence sampler will generate following sequence of states:
x_1, x_1, x_3, x_4, x_4

g) Will any of the three samplers discussed above work in high-dimensional settings (e.g., $d > 20$)? Explain your answer by discussing how this ”curse of dimensionality” will affect each of the three samplers discussed above.

Solution: I would like to discuss three sampling methods separately:

- ”Rejection Sampler” acceptance rate diminishes exponentially with dimensionality, because it is defined by the ratio of volumes under $p(z)$ and $kq(z)$, which becomes smaller as dimensionality increases. Therefore, Rejection Sampler clearly suffers from the ”curse of dimensionality”.

- ”Importance Sampler” also suffers from the ”curse of dimensionality” due to the fact that the success of the importance sampling approach depends on how well the sampling distribution $q(z)$ matches the desired distribution $p(z)$ and finding such distribution gets more and more complicated as the dimensionality grows.

- ”Independence Sampler” suffers from the ”curse of dimensionality” as well due to the same reason mentioned for the ”Importance Sampler”

In general, the reason why all three sampling methods suffer from the curse of dimensionality is that it is getting more and more difficult to find a tight-bounding proposal distribution as dimensionality of space grows.

Problem 2: Consider a simple 3-node graph shown in fig. 1 in which

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

Derive Gibbs sampling for the posterior distribution $p(\mu, \tau|x)$

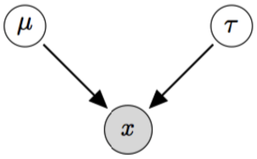


Figure 1: A graph involving an observed Gaussian variable x with prior distributions over its mean μ and precision τ

Solution: To derive Gibbs sampling for the posterior distribution $p(\mu, \tau|x)$, we need to derive equations for $p(\mu|x, \tau)$ and $p(\tau|x, \mu)$ to sample from them.
To do that, we are using the joint distribution, defined as:

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

Using equations (2.140) - (2.142) and (2.149) - (2.151) from Bishop, we can write corresponding equations directly:

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

Problem 3: Consider generative process of LDA.

- Write down the joint probability over the observed data and latent variables.

Solution:

$$\begin{aligned}
p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\alpha, \beta) &= p(\boldsymbol{\theta}|\alpha)p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\phi}|\beta)p(\mathbf{w}|\boldsymbol{\phi}, \mathbf{z}) \\
&= \prod_{k=1}^{K} p(\boldsymbol{\phi}_k|\beta) \cdot \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\boldsymbol{\theta}_d)p(w_{dn}|z_{dn}, \phi_{dn}) \\
&= \prod_{k=1}^{K} Dir(\boldsymbol{\phi}_k|\beta, \dots \beta) \prod_{d=1}^{D} Dir(\boldsymbol{\theta}_d|\alpha, \dots, \alpha) \prod_{n=1}^{N_d} Mult(z_{dn}|\boldsymbol{\theta}_d)Mult(w_{dn}|\phi_{z_{dn}})
\end{aligned}$$

- Integrate out the parameters θ_d's and ϕ_k's from the joint probability. Express this result in terms of the counts N_d, M_k, A_{dk} and B_{kw}.

Solution:

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

Using that Dirichlet distribution is conjugate to the multinomial distribution, $p(\boldsymbol{\theta}|\alpha) \cdot p(\mathbf{z}|\boldsymbol{\theta})$ and $p(\boldsymbol{\phi}|\beta) \cdot p(\mathbf{w}|\boldsymbol{\phi}, \mathbf{z})$ both will result in Dirichlet distribution with an adjusted parameter. **Note**, that $\boldsymbol{\alpha} = (\alpha, \dots \alpha)$ and $\boldsymbol{\beta} = (\beta, \dots \beta)$.

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

Similarly, for the second term we get following:

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

Therefore:

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

- Derive the Gibbs sampling updates for z_{di} with all parameters integrated out.

Solution: We can calculate Gibbs sampling update as following (note, that $(-i)$ means that we exclude i-th token out of calculation):

$$\begin{aligned}
p(z_{di}|\mathbf{z}_d^{(-i)},\mathbf{w}) &= \frac{p(\mathbf{w}_d,\mathbf{z}_d)}{p(\mathbf{w}_d,\mathbf{z}_d^{(-i)})} = \frac{p(\mathbf{z}_d)}{p(\mathbf{z}_d^{(-i)})}\cdot\frac{p(\mathbf{w}_d|\mathbf{z}_d)}{p(\mathbf{w}_d^{(-i)}|\mathbf{z}_d^{(-i)})p(w_{di})} \\
&= \frac{B(\mathbf{A}_d+\boldsymbol{\alpha})\prod_{k=1}^{K}B(\mathbf{B}_k+\boldsymbol{\beta})}{B(\mathbf{A}_d^{(-i)}+\boldsymbol{\alpha})\prod_{k=1}^{K}B(\mathbf{B}_k^{(-i)}+\boldsymbol{\beta})} \\
&= \{\text{Leveraging properties of Gamma functions and assuming,} \\
&\text{that } z_{di} \text{ is in the } w\text{-th position in the vocabulary and denoting topic with } k\text{:}\} \\
&\propto \frac{\Gamma(A_{dk}+\alpha)\Gamma(\sum_{k=1}^{K}A_{dk}^{(-i)}+\alpha)}{\Gamma(A_{dk}^{(-i)}+\alpha)\Gamma(\sum_{k=1}^{K}A_{dk}+\alpha)}\cdot\frac{\Gamma(B_{kw}+\beta)\Gamma(\sum_{w}B_{kw}^{(-i)}+\beta)}{\Gamma(B_{kw}^{(-i)}+\beta)\Gamma(\sum_{w}B_{kw}+\beta)} \\
&\propto (A_{dk}^{(-i)}+\alpha)\frac{B_{kw}^{(-i)}+\beta}{\sum_{w'}B_{kw'}^{(-i)}+\beta}
\end{aligned}$$

Problem 4: Consider a multivariate Bernoulli distribution

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D}\mu_i^{x_i}(1-\mu_i)^{1-x_i}$$

where $\mathbf{x} = (x_1,\dots,x_D)$ and $\boldsymbol{\mu} = (\mu_1,\dots,\mu_D)$, with $\mu_i \in [0,1]$, $x_i \in \{0,1\}$ for $i = 1,\dots,D$.

a) What is the mean of $\mathbf{x}$ under this distribution?

Solution: Using properties of Bernoulli distribution:
$\mathbb{E}[\mathbf{x}] = (\mathbb{E}[x_1], \ldots \mathbb{E}[x_D]) = (\mu_1, \ldots \mu_D) = \boldsymbol{\mu}$

b) What is the covariance matrix of $\mathbf{x}$ under this distribution?

Solution: If we note, that $x_i \perp\!\!\!\perp x_j \forall i \neq j$ and use properties of Bernoulli distribution again, we get following equations:

$$\begin{cases} Cov[x_i, x_j] = 0 & \text{if } i \neq j \\ Cov[x_i, x_i] = Var[x_i] = \mu_i(1-\mu_i) & \text{otherwise} \end{cases} \implies \boldsymbol{\Sigma} = diag(\mu_i(1-\mu_i))_{i=1,\ldots D}$$

Now consider a mixture of K of these multivariate Bernoulli distributions

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, and

$$p(\mathbf{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$$

c) What is the mean of $\mathbf{x}$ under this mixture distribution?

Solution:

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

Suppose we are given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$.

d) Write down the log-likelihood function for this model. Make the expression as explicit as possible, and use brackets to remove any ambiguity regarding what is summed over in the expression.

Solution:

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

e) Why doesnt standard maximum-likelihood work here?

Solution: It doesn't work, because due to the presence of logarithm of summation, there is no closed-form solution.

We will use the Variational EM algorithm to learn the parameters of the model. For each datapoint $\mathbf{x}_n$, introduce a latent variable $z_n = (z_{n1}, \ldots, z_{nK})$ which is a one-of-K coded binary vector that indicates the latent class of that datapoint. In other words: the latent variable $\mathbf{z}_n$ has K components, all of which are 0 except for the kth one that is 1, where k is the latent class for data point $\mathbf{x}_n$. Using these conventions, for data point $\mathbf{x}_n$ and associated latent class $\mathbf{z}_n$, we can write:

$$p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\pi}) = p(\mathbf{z}_n|\boldsymbol{\pi})p(\mathbf{z}_n|\mathbf{z}_n, \boldsymbol{\mu}) = \prod_{k=1}^{K}\pi_k^{z_{nk}}p(\mathbf{x}_n|\boldsymbol{\mu}_k)^{z_{nk}}$$

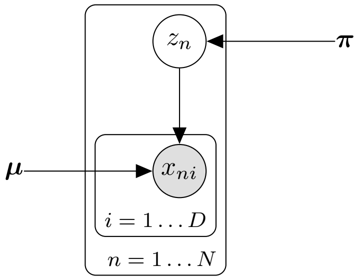
f) Write down the complete-data log-likelihood function for this model. Make the expression as explicit as possible, and use brackets to remove any ambiguity regarding what is summed over in the expression.

Solution:

$$\mathcal{L} = \sum_{n=1}^{N}\sum_{k=1}^{K}\left[z_{nk}\log\pi_k + z_{nk}\sum_{i=1}^{D}(x_{ni}\log\mu_{ki} + (1-x_{ni})\log(1-\mu_{ki}))\right]$$

g) Draw the corresponding graphical model using plate notation. Clearly distinguish observed variables, latent variables, parameters, and make clear which variable subscripts are looped over if you use plates.

Solution:



h) Write down an explicit expression for the VEM objective function $\mathcal{B}(\{q_n(\mathbf{z}_n)\}, \boldsymbol{\mu}, \boldsymbol{\pi})$ for this model.

Solution:

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

i) Include Lagrange multipliers for all constraints in the model and construct the Lagrangian $\widetilde{\mathcal{B}}$ from $\mathcal{B}$. Make the Lagrangian as explicit as possible.

Solution:

$$\widetilde{\mathcal{B}}(\{q_n(\mathbf{z}_n)\}, \boldsymbol{\mu}, \boldsymbol{\pi}) = \mathcal{B}(\{q_n(\mathbf{z}_n)\}, \boldsymbol{\mu}, \boldsymbol{\pi}) + \lambda(\sum_{k=1}^{K} \pi_k - 1) + \sum_{n=1}^{N} \lambda_n (\sum_{\mathbf{z}_n=1}^{K} q_n(\mathbf{z}_n) - 1)$$

j) Work out the details of the E-step, i.e., optimize $\widetilde{\mathcal{B}}$ with respect to q_n for all $n = 1, \ldots, N$. Solve the equation. What is the interpretation of $q_n(\mathbf{z}_n)$?

Solution: To simplify the notation, I will use $\mathbf{m}_k$, which corresponds to the k-th column in the identity matrix $\mathbf{I}_K$. Technically, it means, that we are taking derivative with respect to $q_n(\hat{\mathbf{z}})$, where $\hat{\mathbf{z}}$ is $\mathbf{z}$ with 1 in k-th position.

$$
\begin{aligned}
\frac{\partial \widetilde{\mathcal{B}}}{\partial q_n(\mathbf{m}_k)} &= -\log q_n(\mathbf{m}_k) - 1 + \lambda_n + \log \pi_k + \sum_{i=1}^{D} \left(x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki})\right) = 0 \\
q_n(\mathbf{m}_k) &= \exp\left[\lambda_n - 1 + \log \pi_k + \sum_{i=1}^{D} \left(x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki})\right)\right] \\
&= \pi_k \exp\left[\lambda_n - 1\right] \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}} \\
\sum_{k=1}^{K} q_n(\mathbf{m}_k) &= \sum_{k=1}^{K} \pi_k \exp\left[\lambda_n - 1\right] \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}} = 1 \\
\exp\left[\lambda_n - 1\right] &= \frac{1}{\sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}}} \\
\implies q_n(\mathbf{m}_k) &= \frac{\pi_k \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}}}{\sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1 - x_{ni}}}
\end{aligned}
$$

This represents posterior distribution $p(\mathbf{z}_n | \mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\pi})$.

k) Work out the details of the M-step for $\boldsymbol{\pi}$, i.e., optimize $\widetilde{\mathcal{B}}$ with respect to π_k for all k. Solve the equation.

Solution:

$$
\begin{aligned}
\frac{\partial \widetilde{\mathcal{B}}}{\partial \pi_k} &= \frac{\sum_{n=1}^{N} q_n(\mathbf{m}_k)}{\pi_k} + \lambda = 0 \\
\pi_k &= -\frac{\sum_{n=1}^{N} q_n(\mathbf{m}_k)}{\lambda} \\
\lambda \pi_k &= -\sum_{n=1}^{N} q_n(\mathbf{m}_k) \\
\sum_{k=1}^{K} \lambda \pi_k &= -\sum_{k=1}^{K} \sum_{n=1}^{N} q_n(\mathbf{m}_k) \\
\lambda &= -\sum_{k=1}^{K} \sum_{n=1}^{N} q_n(\mathbf{m}_k) \\
\pi_k &= \frac{\sum_{n=1}^{N} q_n(\mathbf{m}_k)}{\sum_{k=1}^{K} \sum_{n=1}^{N} q_n(\mathbf{m}_k)}
\end{aligned}
$$

Problem 5: Consider a state space z consisting of the integers, with probability

$$\begin{aligned} p(z^{(r+1)} = z^{(r)}) &= 0.5 \\ p(z^{(r+1)} = z^{(r)} + 1) &= 0.25 \\ p(z^{(r+1)} = z^{(r)} - 1) &= 0.25 \end{aligned}$$

where $z^{(r)}$ denotes the state at step r. If the initial state is $z^{(0)} = 0$, prove that

$$\mathbb{E}[(z^{(r)})^2] = \frac{r}{2}$$

Solution: We can use following equation: $z^{(r)} = \sum_{t=1}^{r} x_t$.
Using that

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

This gives us $\mathbb{E}[x_t] = 0$ and $\mathbb{V}[x_t] = \mathbb{E}[x_t^2] = \frac{1}{2}$. Also, it is easy to see, that $\mathbb{E}[z^{(r)}] = 0$. This allows us to write down the following equation:

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