These are exercises for self-study and exam preparation. All material is examinable unless otherwise mentioned.

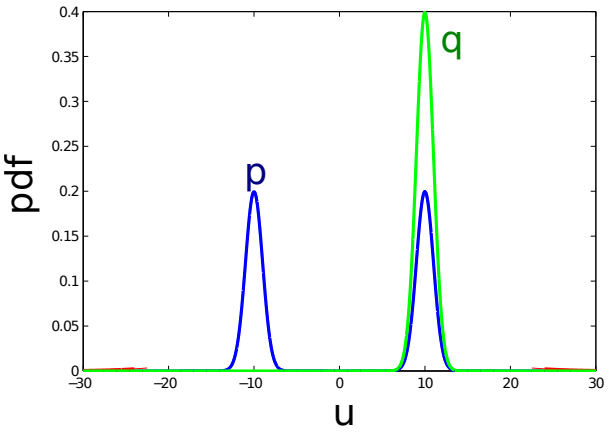
Exercise 1. *Variational posterior approximation*

We have seen that maximising the evidence lower bound (ELBO) with respect to the variational distribution q minimises the Kullback-Leibler divergence to the true posterior p. We here assume that q and p are probability density functions so that the Kullback-Leibler divergence between them is defined as

$$KL(q||p) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x} = \mathbb{E}_q \left[\log \frac{q(\mathbf{x})}{p(\mathbf{x})} \right]. \tag{1}$$

(a) You can here assume that $\mathbf{x}$ is one-dimensional so that p and q are univariate densities. Consider the case where p is a bimodal density but the variational densities q are unimodal. Sketch a figure that shows p and a variational distribution q that has been learned by minimising $KL(q||p)$. Explain qualitatively why the sketched q minimises $KL(q||p)$.

Solution. A possible sketch is shown in the figure below.



Explanation: We can divide the domain of p and q into the areas where p is small (zero) and those where p has significant mass. Since the objective features q in the numerator while p is in the denominator, an optimal q needs to be zero where p is zero. Otherwise, it would incur a large penalty (division by zero). Since we take the expectation with respect to q, however, regions where $p > 0$ do not need to be covered by q; cutting them out does not incur a penalty. Hence, optimal unimodal q only cover one peak of the bimodal p.

(b) Assume that the true posterior $p(\mathbf{x}) = p(x_1, x_2)$ factorises into two Gaussians of mean zero and variances σ_1^2 and σ_2^2,

$$p(x_1, x_2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{x_1^2}{2\sigma_1^2}\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{x_2^2}{2\sigma_2^2}\right]. \tag{2}$$

Assume further that the variational density $q(x_1, x_2; \lambda^2)$ is parametrised as

$$q(x_1, x_2; \lambda^2) = \frac{1}{2\pi\lambda^2} \exp\left[-\frac{x_1^2 + x_2^2}{2\lambda^2}\right] \tag{3}$$

where λ^2 is the variational parameter that is learned by minimising $KL(q||p)$. If σ_2^2 is much larger than σ_1^2, do you expect λ^2 to be closer to σ_2^2 or to σ_1^2? Provide an explanation.

Solution. The learned variational parameter will be closer to σ_1^2 (the smaller of the two σ_i^2).

Explanation: First note that the σ_i^2 are the variances along the two different axes, and that λ^2 is the single variance for both x_1 and x_2. The objective penalises q if it is non-zero where p is zero (see above). The variational parameter λ^2 thus will get adjusted during learning so that the variance of q is close to the smallest of the two σ_i^2.

Exercise 2. *Generalised Variational Inference*

The ELBO can be written as

$$\mathcal{L}(q) = \mathbb{E}_{q(\mathbf{y})}\left[\log p(\mathbf{x}_o|\mathbf{y})\right] - KL(q(\mathbf{y})||p(\mathbf{y})), \tag{4}$$

where $q(\mathbf{y})$ is the variational distribution, $\mathbf{x}_o$ the observed data, and $p(\mathbf{y})$ the prior. The variational distribution $q(\mathbf{y})$ that maximises $\mathcal{L}(q)$ is given by the posterior $p(\mathbf{y}|\mathbf{x}_o)$. The posterior strikes a compromise between explaining $\mathbf{x}_o$, i.e. making the first term large, and staying close to the prior $p(\mathbf{y})$, i.e. making the second term small.

We here consider a generalised version of the ELBO where $\log p(\mathbf{x}_o|\mathbf{y})$ is replaced by some function $r(\mathbf{x}_o, \mathbf{y})$ that "rewards" the variational distribution $q(\mathbf{y})$ for placing probability mass around $\mathbf{y}$ (the values of $r(\mathbf{x}_o, \mathbf{y})$ may be positive or negative). The objective is

$$J(q) = \mathbb{E}_{q(\mathbf{y})}\left[r(\mathbf{x}_o, \mathbf{y})\right] - KL(q(\mathbf{y})||p(\mathbf{y})). \tag{5}$$

Credit: Such objectives were introduced and studied in the paper A general framework for updating belief distributions *by Bissiri, Holmes, and Walker, J. R. Statist. Soc. B (2016).*

(a) What is the distribution q that maximises $J(q)$?

HINT: Write $r(\mathbf{x}_o, \mathbf{y}) = \log\exp(r(\mathbf{x}_o, \mathbf{y}))$ and express $J(q)$ in terms of a KL-divergence between q and some distribution p^*.

Solution. We follow the first hint and write the objective as

$$J(q) = \mathbb{E}_{q(\mathbf{y})} \log\exp(r(\mathbf{x}_o, \mathbf{y})) - \mathrm{KL}(q(\mathbf{y})||p(\mathbf{y})) \tag{S.1}$$

and insert the definition of the KL-divergence

$$J(q) = \mathbb{E}_{q(\mathbf{y})} \log\exp(r(\mathbf{x}_o, \mathbf{y})) - \mathbb{E}_{q(\mathbf{y})} \log\frac{q(\mathbf{y})}{p(\mathbf{y})} \tag{S.2}$$

We then use that $\log(u) = -\log(1/u)$ to obtain

$$J(q) = \mathbb{E}_{q(\mathbf{y})}\left[-\log\frac{1}{\exp(r(\mathbf{x}_o, \mathbf{y}))}\right] - \mathbb{E}_{q(\mathbf{y})} \log\frac{q(\mathbf{y})}{p(\mathbf{y})} \tag{S.3}$$

which allows us to combine the two terms

$$J(q) = -\mathbb{E}_{q(\mathbf{y})} \log\frac{q(\mathbf{y})}{\exp(r(\mathbf{x}_o, \mathbf{y}))p(\mathbf{y})} \tag{S.4}$$

Assuming that $Z(\mathbf{x}_o) = \mathbb{E}_{p(\mathbf{y})}\exp(r(\mathbf{x}_o, \mathbf{y}))$ exists, we then have

$$J(q) = -\mathbb{E}_{q(\mathbf{y})} \log\frac{\frac{1}{Z(\mathbf{x}_o)}q(\mathbf{y})}{\frac{1}{Z(\mathbf{x}_o)}\exp(r(\mathbf{x}_o, \mathbf{y}))p(\mathbf{y})} \tag{S.5}$$

$$= -\mathbb{E}_{q(\mathbf{y})} \log\frac{q(\mathbf{y})}{\frac{1}{Z(\mathbf{x}_o)}\exp(r(\mathbf{x}_o, \mathbf{y}))p(\mathbf{y})} - \mathbb{E}_{q(\mathbf{y})} \log\frac{1}{Z(\mathbf{x}_o)} \tag{S.6}$$

Since $\log \frac{1}{Z(\mathbf{x}_o)}$ does not depend on $\mathbf{y}$, we obtain

$$J(q) = -\mathbb{E}_{q(\mathbf{y})} \log \frac{q(\mathbf{y})}{\frac{1}{Z(\mathbf{x}_o)} \exp(r(\mathbf{x}_o, \mathbf{y}))p(\mathbf{y})} - \log \frac{1}{Z(\mathbf{x}_o)} \tag{S.7}$$

$$= -\mathbb{E}_{q(\mathbf{y})} \log \frac{q(\mathbf{y})}{\frac{1}{Z(\mathbf{x}_o)} \exp(r(\mathbf{x}_o, \mathbf{y}))p(\mathbf{y})} + \text{const} \tag{S.8}$$

$$= -\text{KL}\left(q(\mathbf{y}) || \frac{1}{Z(\mathbf{x}_o)} \exp(r(\mathbf{x}_o, \mathbf{y}))p(\mathbf{y})\right) + \text{const} \tag{S.9}$$

Hence the distribution q^* that maximises $J(q)$ is given by

$$q^*(\mathbf{y}) = \operatorname*{argmin}_q \text{KL}(q(\mathbf{y}) || \frac{1}{Z(\mathbf{x}_o)} \exp(r(\mathbf{x}_o, \mathbf{y}))p(\mathbf{y})).$$

Given the non-negativity properties of the KL-divergence, we thus obtain

$$q^*(\mathbf{y}) = \frac{1}{Z(\mathbf{x}_o)} \exp(r(\mathbf{x}_o, \mathbf{y}))p(\mathbf{y}) \tag{S.10}$$

As a sanity check, let us set $r(\mathbf{x}_o, \mathbf{y}) = \log p(\mathbf{x}_o|\mathbf{y})$: We then obtain obtain $q^*(\mathbf{y}) = \frac{1}{Z} p(\mathbf{x}_o|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\mathbf{x}_o)$.

(b) What constraint does $r(\mathbf{x}_o, \mathbf{y})$ need to satisfy for the optimal $q(\mathbf{y})$ to exist?

Solution. From the above derivation, the expected value $Z(\mathbf{x}_o) = \mathbb{E}_{p(\mathbf{y})} \exp(r(\mathbf{x}_o, \mathbf{y}))$ needs to exist, which places a constraint on the reward function $r(\mathbf{x}_o, \mathbf{y})$.

Exercise 3. *EM algorithm for mixture models (optional, not examinable)*

Mixture models are statistical models of the form

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k) \tag{6}$$

where each $p_k(\mathbf{x}; \boldsymbol{\theta}_k)$ is itself a statistical model parameterised by $\boldsymbol{\theta}_k$ and the $\pi_k \geq 0$ are mixture weights that sum to one. The parameters $\boldsymbol{\theta}$ of the mixture model consist of the parameters $\boldsymbol{\theta}_k$ of each mixture component and the mixture weights π_k, i.e. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \pi_1, \ldots, \pi_K)$. An example is a mixture of Gaussians where each $p_k(\mathbf{x}; \boldsymbol{\theta}_k)$ is a Gaussian with parameters given by the mean vector $\boldsymbol{\mu}_k$ and a covariance matrix $\boldsymbol{\Sigma}_k$.

The mixture model in (6) can be considered to be the marginal distribution of a latent variable model $p(\mathbf{x}, h; \boldsymbol{\theta})$ where h is an unobserved variable that takes on values $1, \ldots, K$ and $p(h = k) = \pi_k$. Defining $p(\mathbf{x}|h = k; \boldsymbol{\theta}) = p_k(\mathbf{x}; \boldsymbol{\theta}_k)$, the latent variable model corresponding to (6) thus is

$$p(\mathbf{x}, h = k; \boldsymbol{\theta}) = p(\mathbf{x}|h = k; \boldsymbol{\theta})p(h = k) = \pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k). \tag{7}$$

In particular note that marginalising out h gives $p(\mathbf{x}; \boldsymbol{\theta})$ in (6).

(a) Verify that the latent variable model in (7) can be written as

$$p(\mathbf{x}, h; \boldsymbol{\theta}) = \prod_{k=1}^{K} [\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)]^{\mathbb{1}(h=k)} \tag{8}$$

where h takes values in $1, \ldots, K$.

Solution. Since $\mathbb{1}(h = k)$ is one if $h = k$ and zero otherwise, we have

$$p(\mathbf{x}, h = j; \boldsymbol{\theta}) = \prod_{k=1}^{K} [\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)]^{\mathbb{1}(j=k)} = \pi_j p_j(\mathbf{x}; \boldsymbol{\theta}_j) \tag{S.11}$$

for any $j \in \{1, \ldots, K\}$, which matches (7).

(b) Since the mixture model in (6) can be seen as the marginal of a latent-variable model, we can use the expectation maximisation (EM) algorithm to estimate the parameters $\boldsymbol{\theta}$.

For a general model $p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})$ where $\mathcal{D}$ are the observed data and $\mathbf{h}$ the corresponding unobserved variables, the EM algorithm iterates between computing the expected complete-data log-likelihood $J^l(\boldsymbol{\theta})$ and maximising it with respect to $\boldsymbol{\theta}$:

$$\textbf{\textit{E-step at iteration l:}} \quad J^l(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}^l)}[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})] \tag{9}$$

$$\textbf{\textit{M-step at iteration l:}} \quad \boldsymbol{\theta}^{l+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} J^l(\boldsymbol{\theta}) \tag{10}$$

Here $\boldsymbol{\theta}^l$ is the value of $\boldsymbol{\theta}$ in the l-th iteration. When solving the optimisation problem, we also need to take into account constraints on the parameters, e.g. that the π_k correspond to a pmf.

Assume that the data $\mathcal{D}$ consists of n iid data points $\mathbf{x}_i$, that each $\mathbf{x}_i$ has associated with it a scalar unobserved variable h_i, and that the tuples $(\mathbf{x}_i, h_i)$ are all iid. What is $J^l(\boldsymbol{\theta})$ under these additional assumptions?

Solution. Since the $(\mathbf{x}_i, h_i)$ are iid, we have that $p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i, h_i; \boldsymbol{\theta})$. Hence

$$J^l\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}^l)}[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})] \tag{S.12}$$

$$= \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}^l)}\left[\sum_{i=1}^{n} \log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})\right] \tag{S.13}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})] \tag{S.14}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{p(h_i|\mathcal{D};\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})] \tag{S.15}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{p(h_i|\mathbf{x}_i;\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})] \tag{S.16}$$

where in the second last step, we have used that each $\log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})]$ only involves one latent variable h_i so that we only need to take the expectation over $p(h_i|\mathcal{D}; \boldsymbol{\theta}^l)$, and in the last step, we have used that $h_i \perp\!\!\!\perp \mathbf{x}_j$, for $j \neq i$.

(c) Show that for the latent variable model in (8), $J^l(\boldsymbol{\theta})$ equals

$$J^l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)], \tag{11}$$

$$w_{ik}^l = \frac{\pi_k^l p_k(\mathbf{x}_i; \boldsymbol{\theta}_k^l)}{\sum_{k=1}^{K} \pi_k^l p_k(\mathbf{x}_i; \boldsymbol{\theta}_k^l)} \tag{12}$$

Note that the w_{ik}^l are defined in terms of the parameters π_k^l and $\boldsymbol{\theta}_k^l$ from iteration l. They are equal to the conditional probabilities $p(h = k|\mathbf{x}_i; \boldsymbol{\theta}^l)$, i.e. the probability that $\mathbf{x}_i$ has been sampled from component $p_k(\mathbf{x}_i; \boldsymbol{\theta}_k^l)$.

Solution. We consider a single term $\mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)}[\log p(\mathbf{x}, h; \boldsymbol{\theta})]$ in (S.16).
Given the form of the model in (8), we have that

$$\log p(\mathbf{x}, h; \boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbb{1}(h = k) \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)] \tag{S.17}$$

and hence

$$\mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)}[\log p(\mathbf{x}, h; \boldsymbol{\theta})] = \mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)} \left[\sum_{k=1}^{K} \mathbb{1}(h = k) \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)] \right] \tag{S.18}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)} \left[\mathbb{1}(h = k) \right] \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)] \tag{S.19}$$

$$= \sum_{k=1}^{K} p(h = k|\mathbf{x}; \boldsymbol{\theta}^l) \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)] \tag{S.20}$$

where we have used that the expectation over an indicator event equals the probability for the event to happen, i.e. $\mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)} \left[\mathbb{1}(h = k) \right] = p(h = k|\mathbf{x}; \boldsymbol{\theta}^l)$.

The probability $p(h = k|\mathbf{x}; \boldsymbol{\theta}^l)$ can be determined via the product (Bayes') rule and Equations (7) and (6)

$$p(h = k|\mathbf{x}; \boldsymbol{\theta}^l) = \frac{p(\mathbf{x}, h = k, \boldsymbol{\theta}^l)}{p(\mathbf{x}; \boldsymbol{\theta}^l)} \tag{S.21}$$

$$= \frac{\pi_k^l p_k(\mathbf{x}; \boldsymbol{\theta}_k^l)}{\sum_{k=1}^{K} \pi_k^l p_k(\mathbf{x}; \boldsymbol{\theta}_k^l)} \tag{S.22}$$

Note that the superscript ^{l} indicates that the π_k^l are the mixture weights and the $\boldsymbol{\theta}_k^l$ the model parameters from iteration l.
The objective $J^l(\boldsymbol{\theta})$ sums over n terms $\mathbb{E}_{p(h|\mathbf{x}_i;\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h; \boldsymbol{\theta})]$. Let us denote $p(h = k|\mathbf{x}_i; \boldsymbol{\theta}^l)$ from (S.22) by w_{ik}^l so that

$$\mathbb{E}_{p(h|\mathbf{x}_i;\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h; \boldsymbol{\theta})] = \sum_{k=1}^{K} w_{ik}^l \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)] \tag{S.23}$$

and

$$J^l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)]. \tag{S.24}$$

The objective $J^l(\boldsymbol{\theta})$ takes the form of a weighted log-likelihood. In more detail, since $\sum_k w_{ik}^l = 1$ for all data points $\mathbf{x}_i$ (and $w_{ik}^l \geq 0$), $\sum_{k=1}^{K} w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)]$ is a convex combination. This means that the different components of the mixture model compete with each other: larger weights for some components mean smaller weights for others. In the extreme case, some components may contribute in a negligible way to the i-th term of the log-likelihood.
The weights w_{ik}^l are sometimes, in particular for mixture of Gaussians, called "soft-assignments" because they specify to which extent a data points $\mathbf{x}_i$ "belongs" to a mixture component p_k. Alternatively, we can interpret the w_{ik}^l to be the "responsibilities" of each mixture component p_k for a datapoint $\mathbf{x}_i$.

In some cases, e.g. for computational reasons, we may determine which of the K weights $w_{i1}^l, \ldots, w_{iK}^l$ is the largest and then set it to one while setting the other weights to zero. This corresponds to "hard-assignments" (and "hard EM") where a data point $\mathbf{x}_i$ is exclusively assigned to a single mixture component p_k.

(d) Assume that the different mixture components $p_k(\mathbf{x}; \boldsymbol{\theta}_k), k = 1, \ldots, K$ do not share any parameters. Show that the updated parameter values $\boldsymbol{\theta}_k^{l+1}$ are given by weighted maximum likelihood estimates.

Solution. We interchange the order of the summations in (11) so that

$$J^l(\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)] \tag{S.25}$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik}^l \log \pi_k + \sum_{k=1}^{K} \underbrace{\sum_{i=1}^{n} w_{ik}^l \log p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)}_{\ell_k^l(\boldsymbol{\theta}_k)} \tag{S.26}$$

When we update the parameters $\boldsymbol{\theta}_k$ of the mixture components, the first term is a constant. The second term is a sum over weighted log-likelihoods $\ell_k^l(\boldsymbol{\theta}_k)$, one for each mixture component. If the mixture components do not share parameters, we thus have

$$\boldsymbol{\theta}_k^{l+1} = \operatorname*{argmax}_{\boldsymbol{\theta}_k} J^l(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}_k} \ell_k^l(\boldsymbol{\theta}_k) \tag{S.27}$$

This means that we can compute $\boldsymbol{\theta}_k^{l+1}$ as if we performed maximum likelihood estimation for the model $p_k(\mathbf{x}; \boldsymbol{\theta}_k)$, expect that the data points $\mathbf{x}_i$ are weighted by the w_{ik}^l.

(e) Show that maximising $J^l(\boldsymbol{\theta})$ with respect to the mixture weights π_k gives the update rule

$$\pi_k^{l+1} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}^l \tag{13}$$

Solution. We start with (11) and drop additive terms that do not depend on the π_k. Since

$$J^l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^l \log \pi_k + \text{terms not depending on the } \pi_k \tag{S.28}$$

we can focus on the objective

$$J_\pi^l(\pi_1, \ldots, \pi_K) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^l \log \pi_k \tag{S.29}$$

$$= \sum_{k=1}^{K} \underbrace{\left(\sum_{i=1}^{n} w_{ik}^l \right)}_{\omega_k^l} \log \pi_k \tag{S.30}$$

$$= \sum_{k=1}^{K} \omega_k^l \log \pi_k. \tag{S.31}$$

Taking into account that the $\pi_k = p(h = k)$ define a pmf, the optimisation problem to solve is

$$\text{maximise} \quad \sum_{k=1}^{K} \omega_k^l \log \pi_k \tag{S.32}$$

$$\text{subject to} \quad \pi_k \geq 0 \tag{S.33}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{S.34}$$

The constrained optimisation problem could be solved via Lagrange multipliers. But we here take another approach and solve the optimisation problem by phrasing it in terms of a KL-divergence minimisation problem.

First, note that the π_k that maximise $J_\pi^l(\pi_1, \ldots, \pi_K)$ will also maximise the re-scaled objective

$$\frac{1}{\sum_{k=1}^K \omega_k^l} J_\pi^l(\pi_1, \ldots, \pi_K) = \frac{1}{\sum_{k=1}^K \omega_k^l} \sum_{k=1}^{K} \omega_k^l \log \pi_k \tag{S.35}$$

$$= \sum_{k=1}^{K} q_k^l \log \pi_k \tag{S.36}$$

where we introduced

$$q_k^l = \frac{\omega_k^l}{\sum_{k=1}^K \omega_k^l}. \tag{S.37}$$

The q_k^l are non-negative and sum to one. Hence, we can consider them to define a pmf.

Second, note that the π_k that maximise $J_\pi^l(\pi_1, \ldots, \pi_K)$ will also maximise

$$\sum_{k=1}^{K} q_k^l \log \pi_k - \sum_{k=1}^{K} q_k^l \log q_k^l = \sum_{k=1}^{K} q_k^l \log \frac{\pi_k}{q_k^l} \tag{S.38}$$

$$= -\sum_{k=1}^{K} q_k^l \log \frac{q_k^l}{\pi_k} \tag{S.39}$$

$$= -\mathrm{KL}(q^l, \pi) \tag{S.40}$$

since adding constants does not change the solution. Hence, the optimal π_k are given by the pmf π that minimises the KL-divergence $\mathrm{KL}(q^l, \pi)$. This means that the optimal π_k are

$$\pi_k = q_k^l = \frac{\omega_k^l}{\sum_{k=1}^K \omega_k^l} = \frac{\sum_{i=1}^n w_{ik}^l}{\sum_{k=1}^K \sum_{i=1}^n w_{ik}^l}. \tag{S.41}$$

The denominator can be simplified by noting that, with (12), $\sum_{k=1}^K w_{ik}^l = 1$ so that

$$\sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik}^l = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^l = n \tag{S.42}$$

The requested update rule thus is

$$\pi_k^{l+1} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}^l \tag{S.43}$$

The update rule does not depend directly on the statistical model $p_k(\mathbf{x};\boldsymbol{\theta}_k)$ that we may choose for the mixture components. Their influence occurs indirectly via the w_{ik}^l.

(f) Summarise the EM-algorithm to learn the parameters $\boldsymbol{\theta}$ of the mixture model in (6) from iid data $\mathbf{x}_1,\dots,\mathbf{x}_n$.

Solution. We collect and summarise the results from the previous questions:

- **E-step at iteration l:** Compute the posterior probabilities (soft assignments)

$$w_{ik}^l = \frac{\pi_k^l p_k(\mathbf{x}_i;\boldsymbol{\theta}_k^l)}{\sum_{k=1}^K \pi_k^l p_k(\mathbf{x}_i;\boldsymbol{\theta}_k^l)} \tag{S.44}$$

for all data points $\mathbf{x}_i$ and and mixture components k. Then formulate the objective function $J^l(\boldsymbol{\theta})$

$$J^l(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i;\boldsymbol{\theta}_k)] \tag{S.45}$$

- **M-step at iteration l:** Compute the new mixture weights

$$\pi_k^{l+1} = \frac{1}{n}\sum_{i=1}^n w_{ik}^l \tag{S.46}$$

To compute the new mixture parameters $\boldsymbol{\theta}_k^{l+1}$, maximise $J^l(\boldsymbol{\theta})$ if some parameters are shared or tied. If the $p_k(\mathbf{x};\boldsymbol{\theta}_k)$ do not share parameters, the new parameters $\boldsymbol{\theta}_k^{l+1}$ are obtained by maximising a weighted log-likelihood for each mixture component separately:

$$\boldsymbol{\theta}_k^{l+1} = \operatorname*{argmax}_{\boldsymbol{\theta}_k} \sum_{i=1}^n w_{ik}^l \log p_k(\mathbf{x}_i;\boldsymbol{\theta}_k) \tag{S.47}$$

for $k = 1,\dots,K$.

Exercise 4. *EM algorithm for mixture of Gaussians (optional, not examinable)*

We here use the results from Exercise 3 to derive the EM update rules for a mixture of Gaussians. This is a mixture model where each mixture component is a Gaussian distribution, i.e.

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^K \pi_k \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k). \tag{14}$$

We consider the case where each $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ can be individually changed (no tying of parameters). The overall parameters of the model are given by the $\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k$ and the mixture weights $\pi_k \geq 0$, $k = 1,\dots K$. As in the case of general mixture models, the mixture weights sum to one.

(a) Determine the maximum likelihood estimates for a multivariate Gaussian $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ for iid data $\mathcal{D} = (\mathbf{x}_1,\dots,\mathbf{x}_n)$ when each data point $\mathbf{x}_i$ has a weight w_i. The weights are non-negative but do not necessarily sum to one.

Solution. The weighted log-likelihood is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} w_i \log \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{S.48}$$

$$= \sum_{i=1}^{n} w_i \log |\det 2\pi\boldsymbol{\Sigma}|^{-1/2} - \frac{1}{2}\sum_{i=1}^{n} w_i (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \tag{S.49}$$

Introducing the normalised weights $W_i = w_i / \sum_{i=1}^n w_i$, we have

$$\frac{1}{\sum_{i=1}^n w_i}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log |\det 2\pi\boldsymbol{\Sigma}|^{-1/2} - \frac{1}{2}\sum_{i=1}^{n} W_i (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \tag{S.50}$$

Let us write out the quadratic term

$$(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_i - 2\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.51}$$

Hence

$$\sum_{i=1}^{n} W_i (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \sum_{i=1}^{n} W_i \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_i - 2\sum_{i=1}^{n} W_i \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \underbrace{\sum_{i=1}^{n} W_i}_{=1} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.52}$$

$$= \operatorname{tr}\left[\left(\sum_{i=1}^{n} W_i \mathbf{x}_i \mathbf{x}_i^\top\right)\boldsymbol{\Sigma}^{-1}\right] - 2\left(\sum_{i=1}^{n} W_i \mathbf{x}_i\right)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.53}$$

$$= \operatorname{tr}\left(\mathbf{R}\boldsymbol{\Sigma}^{-1}\right) - 2\mathbf{b}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.54}$$

where $\mathbf{R} = \sum_{i=1}^n W_i \mathbf{x}_i \mathbf{x}_i^\top$ and $\mathbf{b} = \sum_{i=1}^n W_i \mathbf{x}_i$. Hence

$$\frac{1}{\sum_{i=1}^n w_i}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log |\det 2\pi\boldsymbol{\Sigma}|^{-1/2} - \frac{1}{2}\operatorname{tr}\left(\mathbf{R}\boldsymbol{\Sigma}^{-1}\right) + \mathbf{b}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.55}$$

This has exactly the same form as the unweighted likelihood function, just the sufficient statistics $\mathbf{R}$ and $\mathbf{b}$ are computed using the weights. Hence, the maximum likelihood estimates, when expressed in terms of $\mathbf{R}$ and $\mathbf{b}$ remain the same as in the unweighted case:

$$\hat{\boldsymbol{\mu}} = \mathbf{b} = \sum_{i=1}^{n} W_i \mathbf{x}_i \tag{S.56}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{R} - \mathbf{b}\mathbf{b}^\top = \sum_{i=1}^{n} W_i \mathbf{x}_i \mathbf{x}_i^\top - \mathbf{b}\mathbf{b}^\top \tag{S.57}$$

Moreover, since

$$\sum_{i=1}^{n} W_i (\mathbf{x}_i - \mathbf{b})(\mathbf{x}_i - \mathbf{b})^\top = \sum_{i=1}^{n} W_i \mathbf{x}_i \mathbf{x}_i^\top - \underbrace{\sum_{i=1}^{n} W_i \mathbf{x}_i}_{\mathbf{b}} \mathbf{b}^\top - \mathbf{b}\underbrace{\sum_{i=1}^{n} W_i \mathbf{x}_i^\top}_{\mathbf{b}^\top} + \mathbf{b}\mathbf{b}^\top \tag{S.58}$$

$$= \mathbf{R} - \mathbf{b}\mathbf{b}^\top - \mathbf{b}\mathbf{b}^\top + \mathbf{b}\mathbf{b}^\top \tag{S.59}$$

$$= \mathbf{R} - \mathbf{b}\mathbf{b}^\top \tag{S.60}$$

we find that the weighted maximum likelihood estimates are the weighted average and weighted covariance matrix:

$$\hat{\boldsymbol{\mu}} = \sum_{i=1}^{n} W_i \mathbf{x}_i \qquad \hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} W_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \qquad W_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{S.61}$$

(b) Use the results from Exercise 3 to derive the EM update rules for the parameters of the Gaussian mixture model.

Solution. From the solution to Exercise 3(f) and the derived weighted MLE solutions, we find:

- **E-step at iteration l:** Compute the posterior probabilities (soft assignments)

$$w_{ik}^l = \frac{\pi_k^l \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k^l, \boldsymbol{\Sigma}_k^l)}{\sum_{k=1}^{K} \pi_k^l \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k^l, \boldsymbol{\Sigma}_k^l)} \tag{S.62}$$

  for all data points $\mathbf{x}_i$ and and mixture components k.

- **M-step at iteration l:**
  - Determine the weighted MLEs

$$\boldsymbol{\mu}_k^{l+1} = \sum_{i=1}^{n} W_{ik}^l \mathbf{x}_i \qquad \boldsymbol{\Sigma}_k^{l+1} = \sum_{i=1}^{n} W_{ik}^l (\mathbf{x}_i - \boldsymbol{\mu}_k^{l+1})(\mathbf{x}_i - \boldsymbol{\mu}_k^{l+1})^\top \tag{S.63}$$

    where $W_{ik}^l = w_{ik}^l / (\sum_{i=1}^{n} w_{ik}^l)$.
  - Compute the new mixture weights

$$\pi_k^{l+1} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}^l \tag{S.64}$$