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 minimises the Kullback-Leibler divergence to the true posterior. We here investigate the nature of the approximation if the family of variational distributions does not include the true posterior.

(a) Assume that the true posterior for $\mathbf{x} = (x_1, x_2)$ is given by

$$p(\mathbf{x}) = \mathcal{N}(x_1; 0, \sigma_1^2)\mathcal{N}(x_2; 0, \sigma_2^2) \tag{1}$$

and that our variational distribution $q(\mathbf{x}; 0, \lambda^2)$ is

$$q(\mathbf{x}; \lambda^2) = \mathcal{N}(x_1; 0, \lambda^2)\mathcal{N}(x_2; 0, \lambda^2), \tag{2}$$

where $\lambda > 0$ is the variational parameter. Provide an equation for

$$J(\lambda) = \mathrm{KL}(q(\mathbf{x}; \lambda^2)||p(\mathbf{x})), \tag{3}$$

where you can omit additive terms that do not depend on λ.

(b) Determine the value of λ that minimises $J(\lambda) = \mathrm{KL}(q(\mathbf{x}; \lambda^2)||p(\mathbf{x}))$. Interpret the result and relate it to properties of the Kullback-Leibler divergence.

Exercise 2. *Generalised Variational Inference*

The ELBO can be written as

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

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

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

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^*.

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

©Gutmann, University of Edinburgh, 2026 CC BY 4.0

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

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

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

(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 $\boldsymbol{\theta}$:

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

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

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?

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

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

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

©Gutmann, University of Edinburgh, 2026 CC BY 4.0

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 $p_k(\mathbf{x}_i; \boldsymbol{\theta}_k^l)$.

(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.

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

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

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

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 $\pi_k \geq 0$, $k = 1, \ldots K$. 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 $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ when each data point $\mathbf{x}_i$ has a weight w_i. The weights are non-negative but do not necessarily sum to one.

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

©Gutmann, University of Edinburgh, 2026 CC BY 4.0