Exercises for the tutorials: 2 and 4.

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

Exercise 1. *Mean field variational inference I*

Let $\mathcal{L}_{\mathbf{x}}(q)$ be the evidence lower bound for the marginal $p(\mathbf{x})$ of a joint pdf/pmf $p(\mathbf{x}, \mathbf{y})$,

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right]. \tag{1}$$

Mean field variational inference assumes that the variational distribution $q(\mathbf{y}|\mathbf{x})$ fully factorises, i.e.

$$q(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{d} q_i(y_i|\mathbf{x}), \tag{2}$$

when $\mathbf{y}$ is d-dimensional. An approach to learning the q_i for each dimension is to update one at a time while keeping the others fixed. We here derive the corresponding update equations.

(a) *Show that the evidence lower bound $\mathcal{L}_{\mathbf{x}}(q)$ can be written as*

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q_1(y_1|\mathbf{x})}\mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})}\left[\log p(\mathbf{x}, \mathbf{y})\right] - \sum_{i=1}^{d} \mathbb{E}_{q_i(y_i|\mathbf{x})}\left[\log q_i(y_i|\mathbf{x})\right] \tag{3}$$

where $q(\mathbf{y}_{\setminus 1}|\mathbf{x}) = \prod_{i=2}^{d} q_i(y_i|\mathbf{x})$ is the variational distribution without $q_1(y_1|\mathbf{x})$.

Solution. This follows directly from the definition of the ELBO and the assumed factorisation of $q(\mathbf{y}|\mathbf{x})$. We have

$$\begin{aligned}
\mathcal{L}_{\mathbf{x}}(q) &= \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log q(\mathbf{y}|\mathbf{x}) && \text{(S.1)}\\
&= \mathbb{E}_{\prod_{i=1}^{d} q_i(y_i|\mathbf{x})} \log p(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{\prod_{i=1}^{d} q_i(y_i|\mathbf{x})} \sum_{i=1}^{d} \log q_i(y_i|\mathbf{x}) && \text{(S.2)}\\
&= \mathbb{E}_{\prod_{i=1}^{d} q_i(y_i|\mathbf{x})} \log p(\mathbf{x}, \mathbf{y}) - \sum_{i=1}^{d} \mathbb{E}_{q_i(y_i|\mathbf{x})} \log q_i(y_i|\mathbf{x}) && \text{(S.3)}\\
&= \mathbb{E}_{q_1(y_1|\mathbf{x})}\mathbb{E}_{\prod_{i=2}^{d} q_i(y_i|\mathbf{x})} \log p(\mathbf{x}, \mathbf{y}) - \sum_{i=1}^{d} \mathbb{E}_{q_i(y_i|\mathbf{x})} \log q_i(y_i|\mathbf{x}) && \text{(S.4)}\\
&= \mathbb{E}_{q_1(y_1|\mathbf{x})}\mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})} \left[\log p(\mathbf{x}, \mathbf{y})\right] - \sum_{i=1}^{d} \mathbb{E}_{q_i(y_i|\mathbf{x})} \left[\log q_i(y_i|\mathbf{x})\right] && \text{(S.5)}
\end{aligned}$$

We have here used the linearity of expectation. In case of continuous random variables, for instance, we have

$$\mathbb{E}_{\prod_{i=1}^d q_i(y_i|\mathbf{x})} \sum_{i=1}^d \log q_i(y_i|\mathbf{x}) = \int q_1(y_1|\mathbf{x}) \cdot \ldots \cdot q_d(y_d|\mathbf{x}) \sum_{i=1}^d \log q_i(y_i|\mathbf{x}) dy_1 \ldots dy_d \tag{S.6}$$

$$= \sum_{i=1}^d \int q_1(y_1|\mathbf{x}) \cdot \ldots \cdot q_d(y_d|\mathbf{x}) \log q_i(y_i|\mathbf{x}) dy_1 \ldots dy_d \tag{S.7}$$

$$= \sum_{i=1}^d \int q_i(y_i|\mathbf{x}) \log q_i(y_i|\mathbf{x}) dy_i \underbrace{\int \prod_{j\neq i} q_j(y_j|\mathbf{x}) dy_j}_{=1} \tag{S.8}$$

$$= \sum_{i=1}^d E_{q_i(y_i|\mathbf{x})} \log q_i(y_i|\mathbf{x}) \tag{S.9}$$

For discrete random variables, the integral is replaced with a sum and leads to the same result.

(b) *Assume that we would like to update* $q_1(y_1|\mathbf{x})$ *and that the variational marginals of the other dimensions are kept fixed. Show that*

$$\operatorname*{argmax}_{q_1(y_1|\mathbf{x})} \mathcal{L}_{\mathbf{x}}(q) = \operatorname*{argmin}_{q_1(y_1|\mathbf{x})} KL(q_1(y_1|\mathbf{x})||\bar{p}(y_1|\mathbf{x})) \tag{4}$$

with

$$\log \bar{p}(y_1|\mathbf{x}) = \mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})} [\log p(\mathbf{x}, \mathbf{y})] + const, \tag{5}$$

where const refers to terms not depending on y_1. *That is,*

$$\bar{p}(y_1|\mathbf{x}) = \frac{1}{Z} \exp \left[\mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})} [\log p(\mathbf{x}, \mathbf{y})] \right], \tag{6}$$

where Z *is the normalising constant. Note that variables* $y_2, \ldots, y_d$ *are marginalised out due to the expectation with respect to* $q(\mathbf{y}_{\setminus 1}|\mathbf{x})$.

Solution. Starting from

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q_1(y_1|\mathbf{x})} \mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})} [\log p(\mathbf{x}, \mathbf{y})] - \sum_{i=1}^d \mathbb{E}_{q_i(y_i|\mathbf{x})} [\log q_i(y_i|\mathbf{x})] \tag{S.10}$$

we drop terms that do not depend on q_1. We then obtain

$$J(q_1) = \mathbb{E}_{q_1(y_1|\mathbf{x})} \mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})} [\log p(\mathbf{x}, \mathbf{y})] - \mathbb{E}_{q_1(y_1|\mathbf{x})} [\log q_1(y_1|\mathbf{x})] \tag{S.11}$$

$$= \mathbb{E}_{q_1(y_1|\mathbf{x})} \log \bar{p}(y_1|\mathbf{x}) - \mathbb{E}_{q_1(y_1|\mathbf{x})} [\log q_1(y_1|\mathbf{x})] + \text{const} \tag{S.12}$$

$$= \mathbb{E}_{q_1(y_1|\mathbf{x})} \left[\log \frac{\bar{p}(y_1|\mathbf{x})}{q_1(y_1|\mathbf{x})} \right] \tag{S.13}$$

$$= -\text{KL}(q_1(y_1|\mathbf{x})||\bar{p}(y_1|\mathbf{x})) \tag{S.14}$$

Hence

$$\operatorname*{argmax}_{q_1(y_1|\mathbf{x})} \mathcal{L}_{\mathbf{x}}(q) = \operatorname*{argmin}_{q_1(y_1|\mathbf{x})} \text{KL}(q_1(y_1|\mathbf{x})||\bar{p}(y_1|\mathbf{x})) \tag{S.15}$$

(c) *Conclude that given $q_i(y_i|\mathbf{x})$, $i = 2, \ldots, d$, the optimal $q_1(y_1|\mathbf{x})$ equals $\bar{p}(y_1|\mathbf{x})$.*
This then leads to an iterative updating scheme where we cycle through the different dimensions, each time updating the corresponding marginal variational distribution according to:

$$q_i(y_i|\mathbf{x}) = \bar{p}(y_i|\mathbf{x}), \qquad \bar{p}(y_i|\mathbf{x}) = \frac{1}{Z} \exp\left[\mathbb{E}_{q(\mathbf{y}_{\setminus i}|\mathbf{x})}\left[\log p(\mathbf{x}, \mathbf{y})\right]\right] \tag{7}$$

where $q(\mathbf{y}_{\setminus i}|\mathbf{x}) = \prod_{j \neq i} q(y_j|\mathbf{x})$ is the product of all marginals without marginal $q_i(y_i|\mathbf{x})$.

Solution. This follows immediately from the fact that the KL divergence is minimised when $q_1(y_1|\mathbf{x}) = \bar{p}(y_1|\mathbf{x})$. Side-note: The iterative update rule can be considered to be coordinate ascent optimisation in function space, where each "coordinate" corresponds to a $q_i(y_i|\mathbf{x})$.

Exercise 2. *Mean field variational inference II*

Assume random variables y_1, y_2, x are generated according to the following process

$$y_1 \sim \mathcal{N}(y_1; 0, 1) \qquad y_2 \sim \mathcal{N}(y_2; 0, 1) \tag{8}$$
$$n \sim \mathcal{N}(n; 0, 1) \qquad x = y_1 + y_2 + n \tag{9}$$

where y_1, y_2, n are statistically independent.

(a) *y_1, y_2, x are jointly Gaussian. Determine their mean and their covariance matrix.*

Solution. The expected value of y_1 and y_2 is zero. By linearity of expectation, the expected value of x is

$$\mathbb{E}(x) = \mathbb{E}(y_1) + \mathbb{E}(y_2) + \mathbb{E}(n) = 0 \tag{S.16}$$

The variance of y_1 and y_2 is 1. Since y_1, y_2, n are statistically independent,

$$\mathbb{V}(x) = \mathbb{V}(y_1) + \mathbb{V}(y_2) + \mathbb{V}(n) = 1 + 1 + 1 = 3. \tag{S.17}$$

The covariance between y_1 and x is

$$\begin{aligned} \text{cov}(y_1, x) &= \mathbb{E}((y_1 - \mathbb{E}(y_1))(x - \mathbb{E}(x))) = \mathbb{E}(y_1 x) && \text{(S.18)} \\ &= \mathbb{E}(y_1(y_1 + y_2 + n)) = \mathbb{E}(y_1^2) + \mathbb{E}(y_1 y_2) + \mathbb{E}(y_1 n) && \text{(S.19)} \\ &= 1 + \mathbb{E}(y_1)\mathbb{E}(y_2) + \mathbb{E}(y_1)\mathbb{E}(n) && \text{(S.20)} \\ &= 1 + 0 + 0 && \text{(S.21)} \end{aligned}$$

where we have used that y_1 and x have zero mean and the independence assumptions.

The covariance between y_2 and x is computed in the same way and equals 1 too.

We thus obtain the covariance matrix $\mathbf{\Sigma}$,

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 3 \end{pmatrix} \tag{S.22}$$

(b) The conditional $p(y_1, y_2|x)$ is Gaussian with mean $\mathbf{m}$ and covariance $\mathbf{C}$,

$$\mathbf{m} = \frac{x}{3}\begin{pmatrix}1\\1\end{pmatrix} \qquad \mathbf{C} = \frac{1}{3}\begin{pmatrix}2 & -1\\-1 & 2\end{pmatrix} \tag{10}$$

Since x is the sum of three random variables that have the same distribution, it makes intuitive sense that the mean assigns $1/3$ of the observed value of x to y_1 and y_2. Moreover, y_1 and y_2 are negatively correlated since an increase in y_1 must be compensated with a decrease in y_2.

Let us now approximate the posterior $p(y_1, y_2|x)$ with mean field variational inference. Determine the optimal variational distribution using the method and results from Exercise 1. You may use that

$$p(y_1, y_2, x) = \mathcal{N}((y_1, y_2, x); \mathbf{0}, \mathbf{\Sigma}) \qquad \mathbf{\Sigma} = \begin{pmatrix}1 & 0 & 1\\0 & 1 & 1\\1 & 1 & 3\end{pmatrix} \qquad \mathbf{\Sigma}^{-1} = \begin{pmatrix}2 & 1 & -1\\1 & 2 & -1\\-1 & -1 & 1\end{pmatrix} \tag{11}$$

Solution. The mean field assumption means that the variational distribution is assumed to factorise as

$$q(y_1, y_2|x) = q_1(y_1|x)q_2(y_2|x) \tag{S.23}$$

From Exercise 1, the optimal $q_1(y_1|x)$ and $q_2(y_2|x)$ satisfy

$$q_1(y_1|x) = \bar{p}(y_1|x), \qquad \bar{p}(y_1|x) = \frac{1}{Z}\exp\left[\mathbb{E}_{q_2(y_2|x)}\left[\log p(y_1, y_2, x)\right]\right] \tag{S.24}$$

$$q_2(y_2|x) = \bar{p}(y_2|x), \qquad \bar{p}(y_2|x) = \frac{1}{Z}\exp\left[\mathbb{E}_{q_1(y_1|x)}\left[\log p(y_1, y_2, x)\right]\right] \tag{S.25}$$

Note that these are coupled equations: q_2 features in the equation for q_1 via $\bar{p}(y_1|x)$, and q_1 features in the equation for q_2 via $\bar{p}(y_2|x)$. But we have two equations for two unknowns, which for the Gaussian joint model $p(x, y_1, y_2)$ can be solved in closed form.

Given the provided equation for $p(y_1, y_2, x)$, we have that

$$\log p(y_1, y_2, x) = -\frac{1}{2}\begin{pmatrix}y_1\\y_2\\x\end{pmatrix}^\top \begin{pmatrix}2 & 1 & -1\\1 & 2 & -1\\-1 & -1 & 1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\x\end{pmatrix} + \text{const} \tag{S.26}$$

$$= -\frac{1}{2}\left(2y_1^2 + 2y_2^2 + x^2 + 2y_1y_2 - 2y_1x - 2y_2x\right) + \text{const} \tag{S.27}$$

Let us start with the equation for $\bar{p}(y_1|x)$. It is easier to work in the logarithmic domain, where we obtain:

$$\log \bar{p}(y_1|x) = \mathbb{E}_{q_2(y_2|x)}\left[\log p(y_1, y_2, x)\right] + \text{const} \tag{S.28}$$

$$= -\frac{1}{2}\mathbb{E}_{q_2(y_2|x)}\left[2y_1^2 + 2y_2^2 + x^2 + 2y_1y_2 - 2y_1x - 2y_2x\right] + \text{const} \tag{S.29}$$

$$= -\frac{1}{2}\left(2y_1^2 + 2y_1\mathbb{E}_{q_2(y_2|x)}[y_2] - 2y_1x\right) + \text{const} \tag{S.30}$$

$$= -\frac{1}{2}\left(2y_1^2 + 2y_1m_2 - 2y_1x\right) + \text{const} \tag{S.31}$$

$$= -\frac{1}{2}\left(2y_1^2 - 2y_1(x - m_2)\right) + \text{const} \tag{S.32}$$

where we have absorbed all terms not involving y_1 into the constant. Moreover, we set $\mathbb{E}_{q_2(y_2|x)}[y_2] = m_2$.

Note that an arbitrary Gaussian density $\mathcal{N}(y; m, \sigma^2)$ with mean m and variance σ^2 can be written in the log-domain as

$$\log \mathcal{N}(y; m, \sigma^2) = -\frac{1}{2}\frac{(y-m)^2}{\sigma^2} + \text{const} \tag{S.33}$$
$$= -\frac{1}{2}\left(\frac{y^2}{\sigma^2} - 2y\frac{m}{\sigma^2}\right) + \text{const} \tag{S.34}$$

Comparison with (S.32) shows that $\bar{p}(y_1|x)$, and hence $q_1(y_1|x)$, is Gaussian with variance and mean equal to

$$\sigma_1^2 = \frac{1}{2} \qquad\qquad m_1 = \frac{1}{2}(x - m_2) \tag{S.35}$$

Note that we have not made a Gaussianity assumption on $q_1(y_1|x)$. The optimal $q_1(y_1|x)$ turns out to be Gaussian because the model $p(y_1, y_2, x)$ is Gaussian.

The equation for $\bar{p}(y_2|x)$ gives similarly

$$\log \bar{p}(y_2|x) = \mathbb{E}_{q_1(y_1|x)}[\log p(y_1, y_2, x)] + \text{const} \tag{S.36}$$
$$= -\frac{1}{2}\mathbb{E}_{q_1(y_1|x)}\left[2y_1^2 + 2y_2^2 + x^2 + 2y_1y_2 - 2y_1x - 2y_2x\right] + \text{const} \tag{S.37}$$
$$= -\frac{1}{2}\left(2y_2^2 + 2\mathbb{E}_{q_1(y_1|x)}[y_1]y_2 - 2y_2x\right) + \text{const} \tag{S.38}$$
$$= -\frac{1}{2}\left(2y_2^2 + 2m_1y_2 - 2y_2x\right) + \text{const} \tag{S.39}$$
$$= -\frac{1}{2}\left(2y_2^2 - 2y_2(x - m_1)\right) + \text{const} \tag{S.40}$$

where we have absorbed all terms not involving y_2 into the constant. Moreover, we set $\mathbb{E}_{q_1(y_1|x)}[y_1] = m_1$. With (S.34), this is defines a Gaussian distribution with variance and mean equal to

$$\sigma_2^2 = \frac{1}{2} \qquad\qquad m_2 = \frac{1}{2}(x - m_1) \tag{S.41}$$

Hence the optimal marginal variational distributions $q_1(y_1|x)$ and $q_2(y_2|x)$ are both Gaussian with variance equal to 1/2. Their means satisfy

$$m_1 = \frac{1}{2}(x - m_2) \qquad\qquad m_2 = \frac{1}{2}(x - m_1) \tag{S.42}$$

These are two equations for two unknowns. We can solve them as follows

$$2m_1 = x - m_2 \tag{S.43}$$
$$= x - \frac{1}{2}(x - m_1) \tag{S.44}$$
$$4m_1 = 2x - x + m_1 \tag{S.45}$$
$$3m_1 = x \tag{S.46}$$
$$m_1 = \frac{1}{3}x \tag{S.47}$$

Hence

$$m_2 = \frac{1}{2}x - \frac{1}{6}x = \frac{2}{6}x = \frac{1}{3}x \tag{S.48}$$

In summary, we find

$$q_1(y_1|x) = \mathcal{N}\left(y_1; \frac{x}{3}, \frac{1}{2}\right) \qquad\qquad q_2(y_2|x) = \mathcal{N}\left(y_2; \frac{x}{3}, \frac{1}{2}\right) \tag{S.49}$$

and the optimal variational distribution $q(y_1, y_2|x) = q_1(y_1|x)q_2(y_2|x)$ is Gaussian. We have made the mean field (independence) assumption but not the Gaussianity assumption. Gaussianity of the variational distribution is a consequence of the Gaussianity of the model $p(y_1, y_2, x)$.

Comparison with the true posterior shows that the mean field variational distribution $q(y_1, y_2|x)$ has the same mean but ignores the correlation and underestimates the marginal variances. The true posterior and the mean field approximation are shown in Figure 1.

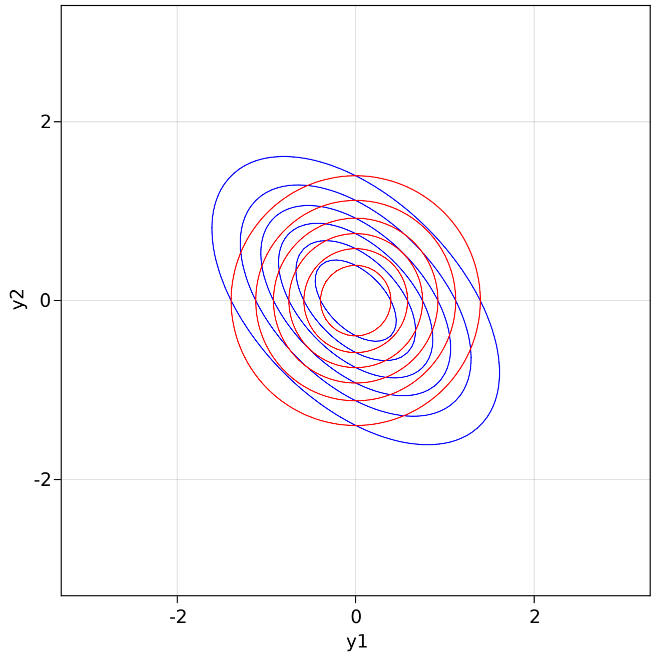


Figure 1: In blue: correlated true posterior. In red: mean field approximation.

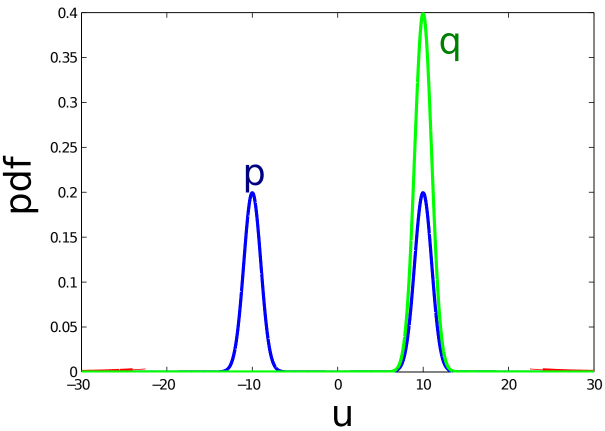
Exercise 3. *Variational posterior approximation I*

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

$$KL(q||p) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x} = \mathbb{E}_q \left[\log \frac{q(\mathbf{x})}{p(\mathbf{x})}\right]. \tag{12}$$

(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,*

$$p(x_1, x_2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{x_1^2}{2\sigma_1^2}\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{x_2^2}{2\sigma_2^2}\right]. \tag{13}$$

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

$$q(x_1, x_2; \lambda^2) = \frac{1}{2\pi\lambda^2} \exp\left[-\frac{x_1^2 + x_2^2}{2\lambda^2}\right] \tag{14}$$

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 4. *Variational posterior approximation II*

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*

$$p(\mathbf{x}) = \mathcal{N}(x_1; \sigma_1^2)\mathcal{N}(x_2; \sigma_2^2) \tag{15}$$

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

$$q(\mathbf{x};\lambda^2) = \mathcal{N}(x_1;\lambda^2)\mathcal{N}(x_2;\lambda^2), \tag{16}$$

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

$$J(\lambda) = KL(q(\mathbf{x};\lambda^2)||p(\mathbf{x})), \tag{17}$$

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

Solution. We write

$$\begin{aligned} \mathrm{KL}(q(\mathbf{x};\lambda^2)||p(\mathbf{x})) &= \mathbb{E}_q\left[\log\frac{q(\mathbf{x};\lambda^2)}{p(\mathbf{x})}\right] && \text{(S.50)}\\ &= \mathbb{E}_q \log q(\mathbf{x};\lambda^2) - \mathbb{E}_q \log p(\mathbf{x}) && \text{(S.51)}\\ &= \mathbb{E}_q \log \mathcal{N}(x_1;\lambda^2) + \mathbb{E}_q \log \mathcal{N}(x_2;\lambda^2)\\ &\quad - \mathbb{E}_q \log \mathcal{N}(x_1;\sigma_1^2) - \mathbb{E}_q \log \mathcal{N}(x_2;\sigma_2^2) && \text{(S.52)} \end{aligned}$$

We further have

$$\begin{aligned} \mathbb{E}_q \log \mathcal{N}(x_i;\lambda^2) &= \mathbb{E}_q \log\left[\frac{1}{\sqrt{2\pi\lambda^2}}\exp\left[-\frac{x_i^2}{2\lambda^2}\right]\right] && \text{(S.53)}\\ &= \log\left[\frac{1}{\sqrt{2\pi\lambda^2}}\right] - \mathbb{E}_q\left[\frac{x_i^2}{2\lambda^2}\right] && \text{(S.54)}\\ &= -\log\lambda - \frac{\lambda^2}{2\lambda^2} + \text{const} && \text{(S.55)}\\ &= -\log\lambda - \frac{1}{2} + \text{const} && \text{(S.56)}\\ &= -\log\lambda + \text{const} && \text{(S.57)} \end{aligned}$$

where we have used that for zero mean x_i, $\mathbb{E}_q[x_i^2] = \mathbb{V}(x_i) = \lambda^2$.
We similarly obtain

$$\begin{aligned} \mathbb{E}_q \log \mathcal{N}(x_i;\sigma_i^2) &= \mathbb{E}_q \log\left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left[-\frac{x_i^2}{2\sigma_i^2}\right]\right] && \text{(S.58)}\\ &= \log\left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\right] - \mathbb{E}_q\left[\frac{x_i^2}{2\sigma_i^2}\right] && \text{(S.59)}\\ &= -\log\sigma_i - \frac{\lambda^2}{2\sigma_i^2} + \text{const} && \text{(S.60)}\\ &= -\frac{\lambda^2}{2\sigma_i^2} + \text{const} && \text{(S.61)} \end{aligned}$$

We thus have

$$\mathrm{KL}(q(\mathbf{x};\lambda^2||p(\mathbf{x})) = -2\log\lambda + \lambda^2\left(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right) + \text{const} \tag{S.62}$$

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

Solution. Taking derivatives of $J(\lambda)$ with respect to λ gives

$$\frac{\partial J(\lambda)}{\partial \lambda} = -\frac{2}{\lambda} + \lambda \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \tag{S.63}$$

Setting it zero yields

$$\frac{1}{\lambda^2} = \frac{1}{2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \tag{S.64}$$

so that

$$\lambda^2 = 2 \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{S.65}$$

or

$$\lambda = \sqrt{2} \sqrt{\frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}} \tag{S.66}$$

This is a minimum because the second derivative of $J(\lambda)$

$$\frac{\partial^2 J(\lambda)}{\partial \lambda^2} = \frac{2}{\lambda^2} + \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \tag{S.67}$$

is positive for all $\lambda > 0$.

The result has an intuitive explanation: the optimal variance λ^2 is the harmonic mean of the variances σ_i^2 of the true posterior. In other words, the optimal precision $1/\lambda^2$ is given by the average of the precisions $1/\sigma_i^2$ of the two dimensions.

If the variances are not equal, e.g. if $\sigma_2^2 > \sigma_1^2$, we see that the optimal variance of the variational distribution strikes a compromise between two types of penalties in the KL-divergence: the penalty of having a bad fit because the variational distribution along dimension two is too narrow; and along dimension one, the penalty for the variational distribution to be nonzero when p is small.