Exercise 1. *Gaussian mean field variational inference*

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

$$y_1 \sim \mathcal{N}(y_1; 0, 1) \qquad y_2 \sim \mathcal{N}(y_2; 0, 1) \tag{1}$$

$$n \sim \mathcal{N}(n; 0, 1) \qquad x = y_1 + y_2 + n \tag{2}$$

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, and hence their joint distribution.

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

$$\mathbb{E}(x) = \mathbb{E}(y_1) + \mathbb{E}(y_2) + \mathbb{E}(n) = 0 \tag{S.1}$$

The mean of the joint is thus zero. The variance of y_1 and y_2 is 1. Since y_1, y_2, n are statistically independent,

$$\mathbb{V}(x) = \mathbb{V}(y_1) + \mathbb{V}(y_2) + \mathbb{V}(n) = 1 + 1 + 1 = 3. \tag{S.2}$$

The covariance between y_1 and x is

$$\begin{aligned}
\text{cov}(y_1, x) &= \mathbb{E}((y_1 - \mathbb{E}(y_1))(x - \mathbb{E}(x))) = \mathbb{E}(y_1 x) && \text{(S.3)}\\
&= \mathbb{E}(y_1(y_1 + y_2 + n)) = \mathbb{E}(y_1^2) + \mathbb{E}(y_1 y_2) + \mathbb{E}(y_1 n) && \text{(S.4)}\\
&= 1 + \mathbb{E}(y_1)\mathbb{E}(y_2) + \mathbb{E}(y_1)\mathbb{E}(n) && \text{(S.5)}\\
&= 1 + 0 + 0 && \text{(S.6)}
\end{aligned}$$

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 $\mathbf{\Sigma}$,

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 3 \end{pmatrix} \tag{S.7}$$

The distribution thus is $p(y_1, y_2, x) = \mathcal{N}\left((y_1, y_2, x); \mathbf{0}, \mathbf{\Sigma}\right)$.

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

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

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

We now approximate the posterior $p(y_1, y_2|x)$ with mean field variational inference. Denoting the variational distribution by $q(\mathbf{y}|x) = q(y_1|x)q(y_2|x)$, derive the update rules for the marginals $q(y_i|x)$.

Hints:

1. *For a model $p(\mathbf{v}, \mathbf{h})$ on observed variables $\mathbf{v}$ and unobserved variables $\mathbf{h}$, in mean-field variational inference, each q_i is iteratively updated as*

$$q_i(h_i|\mathbf{v}) = \frac{1}{Z} \exp\left[\mathbb{E}_{q(\mathbf{h}_{\setminus i}|\mathbf{v})}\left[\log p(\mathbf{v}, \mathbf{h})\right]\right] \tag{4}$$

   where $q(\mathbf{h}_{\setminus i}|\mathbf{v}) = \prod_{j \neq i} q_j(h_j|\mathbf{v})$ is the product of all marginals without marginal $q_i(h_i|\mathbf{v})$.
2. *You may use that*

$$p(y_1, y_2, x) = \mathcal{N}\left((y_1, y_2, x); \mathbf{0}, \mathbf{\Sigma}\right) \qquad \mathbf{\Sigma} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 3 \end{pmatrix} \qquad \mathbf{\Sigma}^{-1} = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & -1 \\ -1 & -1 & 1 \end{pmatrix} \tag{5}$$

Solution. With the hint, the variational distributions $q_1(y_1|x)$ and $q_2(y_2|x)$ are iteratively updated as

$$q_1(y_1|x) = \frac{1}{Z} \exp\left[\mathbb{E}_{q_2(y_2|x)}\left[\log p(y_1, y_2, x)\right]\right] \tag{S.8}$$

$$q_2(y_2|x) = \frac{1}{Z} \exp\left[\mathbb{E}_{q_1(y_1|x)}\left[\log p(y_1, y_2, x)\right]\right] \tag{S.9}$$

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

$$\log p(y_1, y_2, x) = -\frac{1}{2} \begin{pmatrix} y_1 \\ y_2 \\ x \end{pmatrix}^\top \begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & -1 \\ -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ x \end{pmatrix} + \text{const} \tag{S.10}$$

$$= -\frac{1}{2}\left(2y_1^2 + 2y_2^2 + x^2 + 2y_1y_2 - 2y_1x - 2y_2x\right) + \text{const} \tag{S.11}$$

Let us start with the equation for $q_1(y_1|x)$. It is easier to work in the logarithmic domain, where we obtain:

$$\log q_1(y_1|x) = \mathbb{E}_{q_2(y_2|x)}\left[\log p(y_1, y_2, x)\right] + \text{const} \tag{S.12}$$

$$= -\frac{1}{2}\mathbb{E}_{q_2(y_2|x)}\left[2y_1^2 + 2y_2^2 + x^2 + 2y_1y_2 - 2y_1x - 2y_2x\right] + \text{const} \tag{S.13}$$

$$= -\frac{1}{2}\left(2y_1^2 + 2y_1\mathbb{E}_{q_2(y_2|x)}[y_2] - 2y_1x\right) + \text{const} \tag{S.14}$$

$$= -\frac{1}{2}\left(2y_1^2 + 2y_1m_2 - 2y_1x\right) + \text{const} \tag{S.15}$$

$$= -\frac{1}{2}\left(2y_1^2 - 2y_1(x - m_2)\right) + \text{const} \tag{S.16}$$

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

$$\log \mathcal{N}(y; m, \sigma^2) = -\frac{1}{2}\frac{(y - m)^2}{\sigma^2} + \text{const} \tag{S.17}$$

$$= -\frac{1}{2}\left(\frac{y^2}{\sigma^2} - 2y\frac{m}{\sigma^2}\right) + \text{const} \tag{S.18}$$

Comparison with (S.16) shows that $q_1(y_1|x)$ is Gaussian with variance and mean equal to

$$\sigma_1^2 = \frac{1}{2} \qquad\qquad m_1 = \frac{1}{2}(x - m_2) \tag{S.19}$$

Note that we have not made a Gaussianity assumption on $q_1(y_1|x)$. The variational distribution $q_1(y_1|x)$ turns out to be Gaussian because the model $p(y_1, y_2, x)$ is Gaussian.
The equation for $q_2(y_2|x)$ gives similarly

$$\log q_2(y_2|x) = \mathbb{E}_{q_1(y_1|x)}\left[\log p(y_1, y_2, x)\right] + \text{const} \tag{S.20}$$
$$= -\frac{1}{2}\mathbb{E}_{q_1(y_1|x)}\left[2y_1^2 + 2y_2^2 + x^2 + 2y_1y_2 - 2y_1x - 2y_2x\right] + \text{const} \tag{S.21}$$
$$= -\frac{1}{2}\left(2y_2^2 + 2\mathbb{E}_{q_1(y_1|x)}[y_1]y_2 - 2y_2x\right) + \text{const} \tag{S.22}$$
$$= -\frac{1}{2}\left(2y_2^2 + 2m_1y_2 - 2y_2x\right) + \text{const} \tag{S.23}$$
$$= -\frac{1}{2}\left(2y_2^2 - 2y_2(x - m_1)\right) + \text{const} \tag{S.24}$$

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.18), this is defines a Gaussian distribution with variance and mean equal to

$$\sigma_2^2 = \frac{1}{2} \qquad\qquad m_2 = \frac{1}{2}(x - m_1) \tag{S.25}$$

Hence the marginal variational distributions $q_1(y_1|x)$ and $q_2(y_2|x)$ are both Gaussian with variance equal to $1/2$ and their means are iteratively updated as

$$m_1 = \frac{1}{2}(x - m_2) \qquad\qquad m_2 = \frac{1}{2}(x - m_1) \tag{S.26}$$

(c) In this example, the update rules result in two equations for two unknowns that can be solved in closed form. Derive the closed-form expression for the optimal mean-field approximation and compare it with the true conditional $p(y_1, y_2|x)$.

Solution. We found that the marginal variational distributions $q_1(y_1|x)$ and $q_2(y_2|x)$ are Gaussian with variance equal to $1/2$ and their means being iteratively updated as

$$m_1 = \frac{1}{2}(x - m_2) \qquad\qquad m_2 = \frac{1}{2}(x - m_1) \tag{S.27}$$

We note that the update rule involves two unknowns, m_1 and m_2 and two equations. We can thus solve them, which gives

$$2m_1 = x - m_2 \tag{S.28}$$
$$= x - \frac{1}{2}(x - m_1) \tag{S.29}$$
$$4m_1 = 2x - x + m_1 \tag{S.30}$$
$$3m_1 = x \tag{S.31}$$
$$m_1 = \frac{1}{3}x \tag{S.32}$$

Hence

$$m_2 = \frac{1}{2}x - \frac{1}{6}x = \frac{2}{6}x = \frac{1}{3}x \tag{S.33}$$

In summary, we thus have

$$q_1(y_1|x) = \mathcal{N}\left(y_1; \frac{x}{3}, \frac{1}{2}\right) \qquad\qquad q_2(y_2|x) = \mathcal{N}\left(y_2; \frac{x}{3}, \frac{1}{2}\right) \tag{S.34}$$

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.

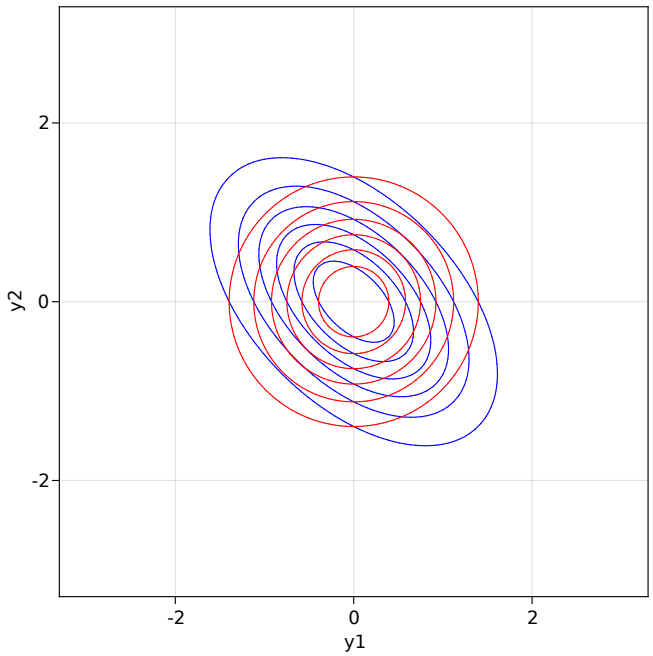

Figure 1: In blue: correlated true posterior. In red: mean field approximation.

Exercise 2. ***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*

$$p(\mathbf{x}) = \mathcal{N}(x_1; 0, \sigma_1^2)\mathcal{N}(x_2; 0, \sigma_2^2) \tag{6}$$

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

$$q(\mathbf{x}; \lambda^2) = \mathcal{N}(x_1; 0, \lambda^2)\mathcal{N}(x_2; 0, \lambda^2), \tag{7}$$

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

$$J(\lambda) = KL(q(\mathbf{x}; \lambda^2)||p(\mathbf{x})), \tag{8}$$

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

Solution. We write

$$\mathrm{KL}(q(\mathbf{x};\lambda^2)||p(\mathbf{x})) = \mathbb{E}_q\left[\log\frac{q(\mathbf{x};\lambda^2)}{p(\mathbf{x})}\right] \tag{S.35}$$

$$= \mathbb{E}_q \log q(\mathbf{x};\lambda^2) - \mathbb{E}_q \log p(\mathbf{x}) \tag{S.36}$$

$$= \mathbb{E}_q \log \mathcal{N}(x_1;0,\lambda^2) + \mathbb{E}_q \log \mathcal{N}(x_2;0,\lambda^2) - \mathbb{E}_q \log \mathcal{N}(x_1;0,\sigma_1^2) - \mathbb{E}_q \log \mathcal{N}(x_2;0,\sigma_2^2) \tag{S.37}$$

We further have

$$\mathbb{E}_q \log \mathcal{N}(x_i;0,\lambda^2) = \mathbb{E}_q \log\left[\frac{1}{\sqrt{2\pi\lambda^2}}\exp\left[-\frac{x_i^2}{2\lambda^2}\right]\right] \tag{S.38}$$

$$= \log\left[\frac{1}{\sqrt{2\pi\lambda^2}}\right] - \mathbb{E}_q\left[\frac{x_i^2}{2\lambda^2}\right] \tag{S.39}$$

$$= -\log\lambda - \frac{\lambda^2}{2\lambda^2} + \text{const} \tag{S.40}$$

$$= -\log\lambda - \frac{1}{2} + \text{const} \tag{S.41}$$

$$= -\log\lambda + \text{const} \tag{S.42}$$

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

$$\mathbb{E}_q \log \mathcal{N}(x_i;0,\sigma_i^2) = \mathbb{E}_q \log\left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left[-\frac{x_i^2}{2\sigma_i^2}\right]\right] \tag{S.43}$$

$$= \log\left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\right] - \mathbb{E}_q\left[\frac{x_i^2}{2\sigma_i^2}\right] \tag{S.44}$$

$$= -\log\sigma_i - \frac{\lambda^2}{2\sigma_i^2} + \text{const} \tag{S.45}$$

$$= -\frac{\lambda^2}{2\sigma_i^2} + \text{const} \tag{S.46}$$

We thus have

$$\mathrm{KL}(q(\mathbf{x};\lambda^2||p(\mathbf{x})) = -2\log\lambda + \lambda^2\left(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right) + \text{const} \tag{S.47}$$

(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

$$\frac{\partial J(\lambda)}{\partial\lambda} = -\frac{2}{\lambda} + \lambda\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) \tag{S.48}$$

Setting it zero yields

$$\frac{1}{\lambda^2} = \frac{1}{2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) \tag{S.49}$$

so that

$$\lambda^2 = 2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{S.50}$$

or

$$\lambda = \sqrt{2}\sqrt{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}} \tag{S.51}$$

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

$$\frac{\partial^2 J(\lambda)}{\partial \lambda^2} = \frac{2}{\lambda^2} + \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) \tag{S.52}$$

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.