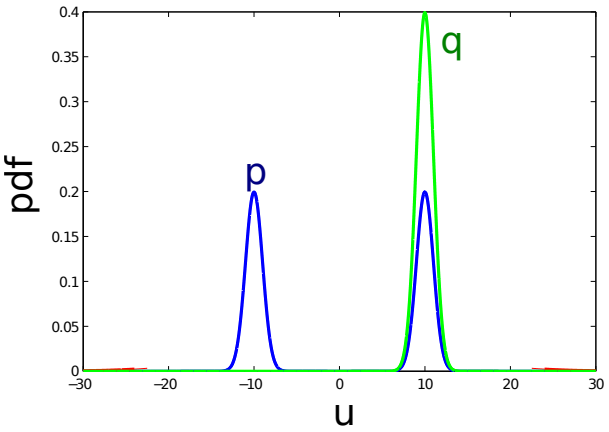
Exercise 1. *Variational posterior approximation*

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 has q in the numerator and p 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, the optimal unimodal q only covers 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 parameterised 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.

©Gutmann, University of Edinburgh, 2026 CC BY 4.0

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

Explanation: 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. *Gaussian mean field variational inference*

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

$$y_1 \sim \mathcal{N}(y_1; 0, 1) \qquad y_2 \sim \mathcal{N}(y_2; 0, 1) \tag{4}$$

$$n \sim \mathcal{N}(n; 0, 1) \qquad x = y_1 + y_2 + n \tag{5}$$

where y_1, y_2, n are statistically independent.

The three questions below can be solved independently of each other.

(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

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

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 $p(y_1, y_2, x) = \mathcal{N}((y_1, y_2, x); \mathbf{0}, \mathbf{\Sigma})$.

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

$$\mathbf{m} = \frac{x}{3}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \mathbf{C} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \tag{6}$$

©Gutmann, University of Edinburgh, 2026 CC BY 4.0

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*

$$q_i(h_i|\mathbf{v}) = \frac{1}{Z} \exp\left[\mathbb{E}_{q(\mathbf{h}_{\setminus i}|\mathbf{v})} \left[\log p(\mathbf{v}, \mathbf{h})\right]\right] \tag{7}$$

   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 $q_i(h_i|\mathbf{v})$. It is typically easier to work with the update rule in the log-domain.

2. *Omitting constants, the log pdf of $p(y_1, y_2, x)$ is*

$$\log p(y_1, y_2, x) = -\frac{1}{2}\left(2y_1^2 + 2y_2^2 + x^2 + 2y_1y_2 - 2y_1x - 2y_2x\right) + const \tag{8}$$

3. *A Gaussian density $\mathcal{N}(y; m, \sigma^2)$ with mean m and variance σ^2 takes, in the log-domain, the following form:*

$$\log \mathcal{N}(y; m, \sigma^2) = -\frac{1}{2}\frac{(y-m)^2}{\sigma^2} + const \tag{9}$$

$$= -\frac{1}{2}\left(\frac{y^2}{\sigma^2} - 2y\frac{m}{\sigma^2}\right) + const \tag{10}$$

Solution. With the first hint, the update rule for the variational distributions $q_1(y_1|x)$ and $q_2(y_2|x)$ in the log-domain is

$$\log q_1(y_1|x) = \left[\mathbb{E}_{q_2(y_2|x)} \left[\log p(y_1, y_2, x)\right]\right] + \text{const} \tag{S.8}$$

$$\log q_2(y_2|x) = \left[\mathbb{E}_{q_1(y_1|x)} \left[\log p(y_1, y_2, x)\right]\right] + \text{const} \tag{S.9}$$

Let us start with the equation for $q_1(y_1|x)$. With (8) from the second hint, we have

$$\log q_1(y_1|x) = \mathbb{E}_{q_2(y_2|x)} \left[\log p(y_1, y_2, x)\right] + \text{const} \tag{S.10}$$

$$= -\frac{1}{2}\mathbb{E}_{q_2(y_2|x)} \left[2y_1^2 + 2y_2^2 + x^2 + 2y_1y_2 - 2y_1x - 2y_2x\right] + \text{const} \tag{S.11}$$

$$= -\frac{1}{2}\left(2y_1^2 + 2y_1\mathbb{E}_{q_2(y_2|x)}[y_2] - 2y_1x\right) + \text{const} \tag{S.12}$$

$$= -\frac{1}{2}\left(2y_1^2 + 2y_1m_2 - 2y_1x\right) + \text{const} \tag{S.13}$$

$$= -\frac{1}{2}\left(2y_1^2 - 2y_1(x - m_2)\right) + \text{const} \tag{S.14}$$

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

We now use the third hint and (10). Comparison with (S.14) shows that $q_1(y_1|x)$ is Gaussian with variance and mean equal to

$$\sigma_1^2 = \frac{1}{2} \qquad\qquad m_1 = \frac{1}{2}(x - m_2) \tag{S.15}$$

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.

©Gutmann, University of Edinburgh, 2026 CC BY 4.0

For $q_2(y_2|x)$, we note that (8) features the same terms for y_2 as for y_1, so that the above calculations carry over, with the roles of indices 1 and 2 reversed. We can thus directly conclude that $q_2(y_2|x)$ is Gaussian with variance and mean equal to

$$\sigma_2^2 = \frac{1}{2} \qquad\qquad m_2 = \frac{1}{2}(x - m_1) \tag{S.16}$$

where $\mathbb{E}_{q_1(y_1|x)}[y_1] = m_1$.

We thus obtain that 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

$$m_1 = \frac{1}{2}(x - m_2) \qquad\qquad m_2 = \frac{1}{2}(x - m_1) \tag{S.17}$$

(c) The update rules from the previous questions shows that 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

$$m_1 = \frac{1}{2}(x - m_2) \qquad\qquad m_2 = \frac{1}{2}(x - m_1) \tag{11}$$

These are two equations for two unknowns that we can solve 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. Substituting the equation for m_2 into m_1 gives

$$2m_1 = x - m_2 \tag{S.18}$$
$$= x - \frac{1}{2}(x - m_1) \tag{S.19}$$
$$4m_1 = 2x - x + m_1 \tag{S.20}$$
$$3m_1 = x \tag{S.21}$$
$$m_1 = \frac{1}{3}x \tag{S.22}$$

Hence

$$m_2 = \frac{1}{2}x - \frac{1}{6}x = \frac{2}{6}x = \frac{1}{3}x \tag{S.23}$$

In summary, we thus have

$$q_1(y_1|x) = \mathcal{N}\left(y_1; \frac{x}{3}, \frac{1}{2}\right) \qquad\qquad q_2(y_2|x) = \mathcal{N}\left(y_2; \frac{x}{3}, \frac{1}{2}\right) \tag{S.24}$$

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.

©Gutmann, University of Edinburgh, 2026 CC BY 4.0

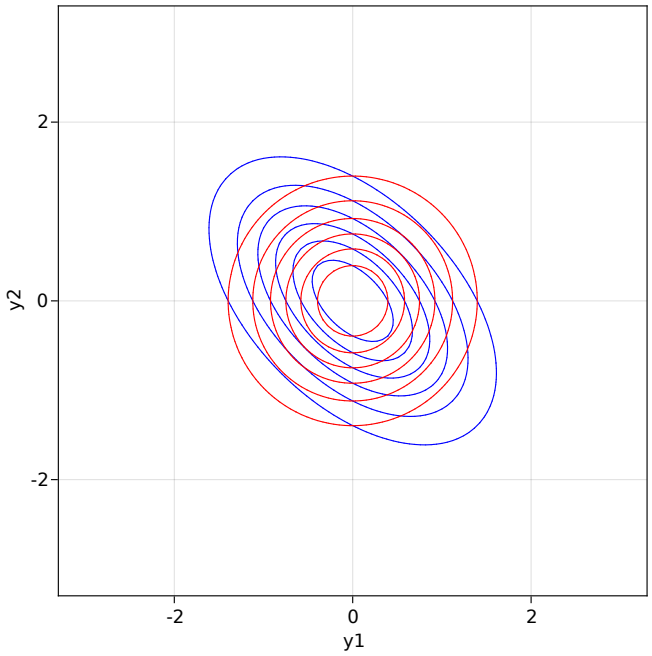


Figure 1: In blue: correlated true posterior. In red: mean field approximation.

©Gutmann, University of Edinburgh, 2026 CC BY 4.0