Exercises for the tutorials: 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 $q_1(y_1|\mathbf{x})$.

(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

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

with

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

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

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

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

(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 $q_i(y_i|\mathbf{x})$.

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.

(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

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

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

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

(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 $\mathrm{KL}(q||p)$. Explain qualitatively why the sketched q minimises $\mathrm{KL}(q||p)$.

(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 $\mathrm{KL}(q||p)$. 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.

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

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

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 5. ***Generalised Variational Inference***

The ELBO can be written as

$$\mathcal{L}(q) = \mathbb{E}_{q(\mathbf{y})}\left[\log p(\mathbf{x}_o|\mathbf{y})\right] - \mathrm{KL}(q(\mathbf{y})||p(\mathbf{y})), \tag{18}$$

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

$$J(q) = \mathbb{E}_{q(\mathbf{y})}\left[r(\mathbf{x}_o, \mathbf{y})\right] - \mathrm{KL}(q(\mathbf{y})||p(\mathbf{y})). \tag{19}$$

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?