Probabilistic Inference & Learning

Exercise Sheet #11

EM & Variational Inference

1. **EM for MAP:** Suppose we wish to use the EM algorithm to maximize the *posterior* distribution over parameters $p(\theta \mid x)$ for a model containing latent variables z, where x is the observed data set (rather than the *likelihood* $p(x \mid \theta)$, as done in the lectures). Show that the E step remains the same as in the maximum likelihood case, whereas in the M step the quantity to be maximized in θ is given (up to constants) by $\mathcal{L}(q,\theta) + \log p(\theta)$, where the ELBO (negative variational free energy) $\mathcal{L}(q,\theta)$ is defined as in the lecture, and $q(z) = p(z \mid x, \theta)$. *20 points*

2. **Softer k-means:** Consider a special case of a Gaussian mixture model in which the covariance matrices Σ_k of the components are all constrained to have a common value Σ. Derive the EM equations for maximizing the likelihood function under such a model. *20 points*

3. **Gibbs' inequality:** Jensen's inequality states that, for a convex function $f : \mathbb{R} \to \mathbb{R}$ and a probability density $p(x)$,
$$\mathbb{E}_p(f(x)) \geq f(\mathbb{E}_p(x)). \tag{1}$$
Use Jensen's inequality to show that the Kullback-Leibler divergence $D_{\mathrm{KL}}(p\|q)$ between two arbitrary[1] probability distributions p, q is nonnegative:
$$D_{\mathrm{KL}}(p\|q) := \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \geq 0 \tag{2}$$
This result is known as *Gibbs' inequality* and was quoted frequently in the lecture. Hint: Consider the convex function $f(u) = \log(1/u)$ of the variable $u(x) = p(x)/q(x) > 0$. *20 points*

4. **Free energy for Gaussians:** In the lecture, a sketch was shown to argue that an approximation q to a distribution p found by $q$minimizing $D_{\mathrm{KL}}(p\|q)$ tends to be "too wide", while an approximation found by minimizing $D_{\mathrm{KL}}(q\|p)$ tends to be "too narrow". The following argument supports this insight using the simple case of Gaussian distributions:

   (a) Show that the Kullback-Leibler divergence between two scalar, centered Gaussian distributions is given by *10 points*
   $$D_{\mathrm{KL}}(\mathcal{N}(x; 0, \sigma_q^2)\|\mathcal{N}(x; 0, \sigma_p^2)) = \frac{1}{2}\left(\log\left(\frac{\sigma_p^2}{\sigma_q^2}\right) - 1 + \frac{\sigma_q^2}{\sigma_p^2}\right). \tag{3}$$

   (b) Consider the two-dimensional Gaussian distribution p and a spherical approximation q, each given by (with $\sigma_1 \neq \sigma_2$)
   $$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right), \qquad p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_q^2 & 0 \\ 0 & \sigma_q^2 \end{bmatrix}\right), \tag{4}$$
   Find the values of σ_q that minimize $D_{\mathrm{KL}}(p\|q)$ and $D_{\mathrm{KL}}(q\|p)$, respectively. *30 points*

[1] You may assume that $p(x)/q(x) > 0$ exists for all x.