Problem Sheet 6

1. **Variation of the Metropolis-Hastings algorithm**

   Let P be a probability distribution and ϱ a function such that $0 \leq \varrho(x) \leq 1$ and $\mathbb{E}_P[1/\varrho(x)] < \infty$. A Markov chain $(x^{(n)})_{n\geq 0}$ is constructed as follows: $x^{(n)}$ is replaced by $x^{(n+1)}$ by generating $y \sim P$ and taking

   $$x^{(n+1)} = \begin{cases} y & \text{with probability } \varrho(x^{(n)}), \\ x^{(n)} & \text{with probability } 1 - \varrho(x^{(n)}). \end{cases}$$

   (a) Show that this variation of the Metropolis-Hastings algorithm converges to the stationary distribution with density

   $$\varrho(x)^{-1}/\mathbb{E}_P[\varrho(x)^{-1}]$$

   with respect to P.

   (b) Apply to the case where P is the $\mathcal{B}e(\alpha+1,1)$ distribution and $\varrho(x) = x$.

2. **Marginal density estimation via importance sampling**

   Let (X,Y) be a pair of random variables with distribution $P_{X,Y}$ and density $f_{X,Y}$ on $\mathbb{R}^2$. Let $(X_1,Y_1),\ldots,(X_N,Y_N)$ be i.i.d. with distribution $P_{X,Y}$ and let w be any density on $\mathbb{R}$.

   (a) Show that

   $$\hat{f}_X(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{f_{X,Y}(x,Y_i)w(X_i)}{f_{X,Y}(X_i,Y_i)}$$

   is a consistent estimator of (i.e. converges in probability to) $f_X(x)$.

   (b) Give an expression for the variance of this estimator.

   (c) In the case where $Y \sim \mathcal{N}(0,1)$ and $X|Y = y \sim \mathcal{N}(y, 1+y^2)$, propose an implementation of the above method to estimate $f_X(x)$.

3. **Variational Inference for Gaussian Mixture Models**

   Suppose we have a Bayesian mixture of unit-variance univariate Gaussian distributions. This mixture consists of 2 components each corresponding to a Gaussian distribution, with means $\boldsymbol{\mu} = \{\mu_1, \mu_2\}$. The mean parameters are drawn independently from a Gaussian prior distribution $\mathcal{N}(0,\sigma^2)$. The prior variance σ^2 is a hyperparameter. Generating an observation x_i from this model is done according to the following generative story:

   1. Choose a cluster assignment c_i for the observation. The cluster assignment is chosen from the distribution Categorical$(\frac{1}{2},\frac{1}{2})$ and indicates which latent cluster x_i comes from. Encode c_i as a one-hot vector where $[1,0]$ indicates that x_i is assigned to cluster 0 and vice versa.

2. Generate x_i from the corresponding Gaussian distribution $\mathcal{N}(c_i^T \boldsymbol{\mu}, 1)$.

The complete hierarchical model is as follows:

$$\begin{aligned}
\mu_k &\sim \mathcal{N}(0, \sigma^2), k \in \{1, 2\} \\
c_i &\sim \text{Categorical}\left(\frac{1}{2}, \frac{1}{2}\right), i \in [1, n] \\
x_i | c_i, \boldsymbol{\mu} &\sim \mathcal{N}(c_i^T \boldsymbol{\mu}, 1), i \in [1, n]
\end{aligned}$$

where n is the number of observations generated from the model.

(a) Let's determine the ELBO (evidence lower-bound) for this model. Recall that the ELBO is given by the following equation:

$$\text{ELBO}(q) = \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z})]$$

To calculate $q(\mathbf{z})$, we will now use the mean-field assumption. Under this assumption, each latent variable is governed by its own latent factor, resulting in the following probability distribution:

$$q(\boldsymbol{\mu}, \mathbf{c}) = \left(\prod_{k=1}^{2} q(\mu_k; m_k, v_k^2)\right)\left(\prod_{i=1}^{n} q(c_i; a_i)\right)$$

Here $q(\mu_k; m_k, v_k^2)$ is the Gaussian distribution for the k-th mixture component with mean and variance m_k and v_k^2. $q(c_i; a_i)$ is the categorical distribution for the i-th observation with assignment probabilities a_i (a_i is a 2-dimensional vector). Given this assumption, write down the ELBO as a function of the variational parameters $\mathbf{m}, \mathbf{v}^2, \mathbf{a}$.

(b) Now that we have the ELBO formulation, let's try to compute coordinate updates for our latent variables. Remember that the optimal variational density of a latent variable z_i is proportional to the exponentiated expected log of the complete conditional given all other latent variables in the model and the observed data. In other words:

$$q_i(z_i) \propto \exp\left(\mathbb{E}_{-j}[\log p(z_j | \mathbf{z}_{-j}, \mathbf{x})]\right)$$

Equivalently, you can also say that the variational density is proportional to the exponentiated expected log of the joint $\mathbb{E}_{-j}[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})]$. This is a valid coordinate update since the expectations on the right side of the equation do not involve z_j due to the mean-field assumption.

i. Show that the variational update for $a_{i1} \propto \exp\left(\mathbb{E}[\mu_1; m_1, v_1^2] x_i - \frac{\mathbb{E}[\mu_1^2; m_1, v_1^2]}{2}\right)$.
Hint: We can write the optimal variational density for cluster assignment variables as $q(c_i; a_{i1}) \propto \exp\left(\log p(c_i) + \mathbb{E}_{\boldsymbol{\mu}}[\log p(x_i | c_i, \boldsymbol{\mu}); \mathbf{m}, \mathbf{v}^2]\right)$. Feel free to drop added constants along the way.

ii. Show that the variational updates for the k-th mixture component are $m_k = \frac{\sum_i a_{ik} x_i}{1/\sigma^2 + \sum_i a_{ik}}$ and $v_k^2 = \frac{1}{1/\sigma^2 + \sum_i a_{ik}}$.
Hint: We can write the optimal variational density for the k-th mixture component as

$$q(\mu_k) \propto \exp\left(\log p(\mu_k) + \sum_i \mathbb{E}_{c_i}[\log p(x_i | c_i, \boldsymbol{\mu}); a_i, \mathbf{m}_{-k}, \mathbf{v}_{-k}^2]\right).$$

Feel free to drop added constants along the way. To obtain the mean and variance updates, you might have to complete the square inside the exponent to bring it into the form of a normal distribution.

4. **Metropolis algorithm and Rejection sampling**

   Consider the density f on $\mathbb{R}^2$ defined by

   $$f(u,v) \propto (\cos u)^2(\sin v)^2 e^{-\frac{u^2+v^2}{20}}.$$

   (a) We want to simulate according to the density f using the Metropolis-Hastings algorithm. Starting from $x = (u, v)$, consider the proposal kernel Q defined such that its density $q(x, \cdot)$ corresponds to that of the $\mathcal{N}(x, \sigma^2 I_2)$ distribution, where I_2 is the identity matrix of size 2 and $\sigma > 0$ is a tuning parameter of the algorithm (in other words, $Q(x, \cdot) = \mathcal{N}(x, \sigma^2 I_2)$). Explicitly state the acceptance probability $\rho(x, x') = \rho((u, v), (u', v'))$.

   (b) Propose a rejection method to simulate according to the distribution with density f starting from a Gaussian instrumental distribution.