Solutions to 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.

Solution: Let the transition kernel of the Markov chain be $K(x, y)$. For $x \neq y$, the probability of transitioning from x to y is governed entirely by proposing y from P and accepting it, which happens with probability $\varrho(x)$. Thus, the transition density is $K(x, y) = P(y)\varrho(x)$, where $P(y)$ is the density of P.

Let $\pi(x) = \frac{1}{Z}\frac{P(x)}{\varrho(x)}$ be the target density, where $Z = \mathbb{E}_P[\varrho(X)^{-1}]$ is the normalization constant. We verify that $\pi(x)$ satisfies the detailed balance condition. For $x \neq y$:

$$\pi(x)K(x, y) = \left(\frac{1}{Z}\frac{P(x)}{\varrho(x)}\right) P(y)\varrho(x) = \frac{1}{Z}P(x)P(y)$$

Since the resulting expression $\frac{1}{Z}P(x)P(y)$ is symmetric with respect to x and y, we have $\pi(x)K(x, y) = \pi(y)K(y, x)$. Detailed balance holds, meaning $\pi(x)$ is the stationary distribution of the chain.

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

Solution: The density of the proposed distribution $P \sim \mathcal{B}e(\alpha + 1, 1)$ is $P(x) = (\alpha + 1)x^\alpha$ for $x \in (0, 1)$.

Given $\varrho(x) = x$, the stationary density $\pi(x)$ with respect to the Lebesgue measure is proportional to:

$$\pi(x) \propto \frac{P(x)}{\varrho(x)} \propto \frac{(\alpha + 1)x^\alpha}{x} \propto x^{\alpha-1}$$

We recognize $x^{\alpha-1}$ as being proportional to the density of a Beta distribution $\mathcal{B}e(\alpha, 1)$. Thus, the algorithm converges to the stationary distribution $\mathcal{B}e(\alpha, 1)$.

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) $f_X(x)$.

Solution: Let $W_i = \frac{f_{X,Y}(x,Y_i)w(X_i)}{f_{X,Y}(X_i,Y_i)}$. Since the pairs (X_i, Y_i) are i.i.d. draws from $P_{X,Y}$, by the Law of Large Numbers, the sample average $\hat{f}_X(x) = \frac{1}{N}\sum_{i=1}^{N} W_i$ converges in probability to its expected value $\mathbb{E}[W_1]$. We compute this expectation by integrating over the joint density $f_{X,Y}(u, y)$:

$$\begin{aligned}
\mathbb{E}[W_1] &= \iint \frac{f_{X,Y}(x, y) w(u)}{f_{X,Y}(u, y)} f_{X,Y}(u, y) du dy \\
&= \iint f_{X,Y}(x, y) w(u) du dy \\
&= \int f_{X,Y}(x, y) dy \int w(u) du
\end{aligned}$$

Since $w(u)$ is a valid density, $\int w(u)du = 1$. Furthermore, the marginal density $f_X(x)$ is precisely $\int f_{X,Y}(x, y)dy$. Therefore, $\mathbb{E}[W_1] = f_X(x)$, proving consistency.

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

Solution: Because the terms W_i are independent, the variance of the estimator is $\frac{1}{N}\operatorname{Var}(W_1)$. The second moment of W_1 is:

$$\begin{aligned}
\mathbb{E}[W_1^2] &= \iint \left(\frac{f_{X,Y}(x, y) w(u)}{f_{X,Y}(u, y)} \right)^2 f_{X,Y}(u, y) du dy \\
&= \iint \frac{f_{X,Y}(x, y)^2 w(u)^2}{f_{X,Y}(u, y)} du dy
\end{aligned}$$

Using $\operatorname{Var}(W_1) = \mathbb{E}[W_1^2] - (\mathbb{E}[W_1])^2$, the variance of $\hat{f}_X(x)$ is:

$$\operatorname{Var}(\hat{f}_X(x)) = \frac{1}{N} \left(\iint \frac{f_{X,Y}(x, y)^2 w(u)^2}{f_{X,Y}(u, y)} du dy - f_X(x)^2 \right)$$

Remark: To minimize the above variance, it suffices to minimize $\mathbb{E}[W_1^2]$, which by Cauchy-Schwarz inequality is lower bounded by

$$\left(\iint \frac{f_{X,Y}(x, y) w(u)}{f_{X,Y}(u, y)} f_{X,Y}(u, y) du dy \right)^2 = f_X(x)^2.$$

We can achieve the lower bound as soon as $\frac{f_{X,Y}(x,y)w(u)}{f_{X,Y}(u,y)}$ is constant as a function of u and y (equality case of Cauchy-Schwarz). This means that

$$w(u) = C\frac{f_{X,Y}(u,y)}{f_{X,Y}(x,y)},$$

for any y. Integrating over u, we find $1 = C\frac{f_Y(y)}{f_{X,Y}(x,y)}$, implying that $C = f_{X|Y}(x \mid y)$ the conditional density of X given $Y = y$. Plugging this into w,

$$w(u) = \frac{f_{X,Y}(u,y)}{f_Y(y)} = f_{X|Y}(u \mid y),$$

for any y. As there is no reason for the conditional density to be constant in the value y, it is impossible to achieve the null variance as in the usual Importance sampling algorithm.

(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 $f_X(x)$.

Solution: 1. Select an arbitrary density $w(u)$ that is easy to evaluate, such as the standard normal density $w(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$.
2. Generate a sample of size N for (X_i, Y_i): First, draw $Y_i \sim \mathcal{N}(0,1)$. Then, given Y_i, draw $X_i \sim \mathcal{N}(Y_i, 1+Y_i^2)$.
3. Note that the joint density is evaluated as $f_{X,Y}(u,y) = f_Y(y)f_{X|Y}(u|y)$, which is:

$$f_{X,Y}(u,y) = \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\right)\left(\frac{1}{\sqrt{2\pi(1+y^2)}}e^{-\frac{(u-y)^2}{2(1+y^2)}}\right)$$

4. For a target point x, compute the estimator:

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

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 $\mathbf{m}, \mathbf{v}^2, \mathbf{a}$.

Solution: We expand the joint log-probability

$$\log p(\mathbf{x}, \boldsymbol{\mu}, \mathbf{c}) = \sum_{k=1}^{2} \log p(\mu_k) + \sum_{i=1}^{n} \log p(c_i) + \sum_{i=1}^{n} \log p(x_i | c_i, \boldsymbol{\mu}).$$

Taking expectations with respect to q:

$$\begin{aligned} \text{ELBO}(q) = &\sum_{k=1}^{2} \mathbb{E}_{q(\mu_k)}\left[\log \mathcal{N}(\mu_k; 0, \sigma^2)\right] + \sum_{i=1}^{n} \mathbb{E}_{q(c_i)}\left[\log\left(\frac{1}{2}\right)\right] \\ &+ \sum_{i=1}^{n} \mathbb{E}_{q(c_i)q(\boldsymbol{\mu})}\left[\log \mathcal{N}(x_i; c_i^T \boldsymbol{\mu}, 1)\right] \\ &- \sum_{k=1}^{2} \mathbb{E}_{q(\mu_k)}\left[\log \mathcal{N}(\mu_k; m_k, v_k^2)\right] - \sum_{i=1}^{n} \mathbb{E}_{q(c_i)}\left[\log q(c_i; a_i)\right] \\ = &\sum_{k=1}^{2} \mathbb{E}_{q(\mu_k)}\left[\frac{\log \mathcal{N}(\mu_k; 0, \sigma^2)}{\log \mathcal{N}(\mu_k; m_k, v_k^2)}\right] - n \log 2 \\ &+ \sum_{i=1}^{n} \mathbb{E}_{q(c_i)q(\boldsymbol{\mu})}\left[\log \mathcal{N}(x_i; c_i^T \boldsymbol{\mu}, 1)\right] - \sum_{i=1}^{n} \mathbb{E}_{q(c_i)}\left[\log q(c_i; a_i)\right]. \end{aligned}$$

(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.

Solution: Using the hint, the unnormalized probability for c_i being assigned to cluster 1 (i.e. $c_{i1} = 1$) is:

$$q(c_{i1} = 1) \propto \exp\left(\log\left(\frac{1}{2}\right) + \mathbb{E}_{\boldsymbol{\mu}}\left[-\frac{1}{2}(x_i - \mu_1)^2\right]\right)$$

Expanding the quadratic term $-\frac{1}{2}(x_i-\mu_1)^2 = -\frac{1}{2}x_i^2+x_i\mu_1-\frac{1}{2}\mu_1^2$. The term $-\frac{1}{2}x_i^2$ and the prior probability $\log(1/2)$ are independent of the cluster assignment and can be absorbed into the proportionality constant. Thus:

$$\begin{aligned} a_{i1} &\propto \exp\left(\mathbb{E}_{\mu_1}[x_i\mu_1 - \frac{1}{2}\mu_1^2]\right) = \exp\left(x_i\mathbb{E}[\mu_1; m_1, v_1^2] - \frac{\mathbb{E}[\mu_1^2; m_1, v_1^2]}{2}\right) \\ &= \exp\left(x_i m_1 - m_1^2/2 - v_1^2/2\right). \end{aligned}$$

Also,

$$a_{i2} \propto \exp\left(x_i m_2 - m_2^2/2 - v_2^2/2\right).$$

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.

Solution: Using the hint, we gather all terms depending on μ_k:

$$\log q(\mu_k) = \log \mathcal{N}(\mu_k; 0, \sigma^2) + \sum_{i=1}^{n} \mathbb{E}_{c_i} \left[-\frac{c_{ik}}{2}(x_i - \mu_k)^2 \right] + \text{const}$$

$$= -\frac{1}{2\sigma^2}\mu_k^2 - \frac{1}{2}\sum_{i=1}^{n} a_{ik}(x_i^2 - 2x_i\mu_k + \mu_k^2) + \text{const}$$

Collecting terms with respect to μ_k^2 and μ_k:

$$= -\frac{1}{2}\mu_k^2 \left(\frac{1}{\sigma^2} + \sum_{i=1}^{n} a_{ik} \right) + \mu_k \left(\sum_{i=1}^{n} a_{ik} x_i \right) + \text{const}$$

We recognize the exponent of a Gaussian distribution $\mathcal{N}(m_k, v_k^2)$ which has the form $-\frac{1}{2v_k^2}\mu_k^2 + \frac{m_k}{v_k^2}\mu_k$. Matching the coefficients yields:

$$\frac{1}{v_k^2} = \frac{1}{\sigma^2} + \sum_{i=1}^{n} a_{ik} \implies v_k^2 = \frac{1}{1/\sigma^2 + \sum_i a_{ik}}$$

and for the mean:

$$\frac{m_k}{v_k^2} = \sum_{i=1}^{n} a_{ik} x_i \implies m_k = v_k^2 \sum_{i=1}^{n} a_{ik} x_i = \frac{\sum_i a_{ik} x_i}{1/\sigma^2 + \sum_i a_{ik}}$$

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 $\rho(x, x') = \rho((u, v), (u', v'))$.

Solution: Because the proposal distribution $Q(x, \cdot) = \mathcal{N}(x, \sigma^2 I_2)$ is symmetric, meaning $q(x, x') = q(x', x)$, the acceptance probability simplifies to the ratio of target densities:

$$\rho(x, x') = \min\left(1, \frac{f(x')q(x', x)}{f(x)q(x, x')}\right) = \min\left(1, \frac{f(x')}{f(x)}\right)$$

Substituting the target density formula for $x = (u, v)$ and $x' = (u', v')$:

$$\rho((u, v), (u', v')) = \min\left(1, \frac{(\cos u')^2 (\sin v')^2 e^{-\frac{(u')^2+(v')^2}{20}}}{(\cos u)^2 (\sin v)^2 e^{-\frac{u^2+v^2}{20}}}\right)$$

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

Solution: We need an instrumental distribution $g(u, v)$ and a constant M such that $f(u, v) \leq Mg(u, v)$ for all (u, v). Notice that $(\cos u)^2(\sin v)^2 \leq 1$. Therefore:

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

We choose $g(u, v)$ to be the density of an isotropic bivariate Gaussian $\mathcal{N}(0, 10I_2)$, which is $g(u, v) = \frac{1}{20\pi} e^{-\frac{u^2+v^2}{20}}$. Assuming f has a normalization constant C, $f(u, v) \leq Ce^{-\frac{u^2+v^2}{20}} = C20\pi g(u, v)$. We select $M = C20\pi$. The ratio for acceptance is then:

$$\frac{f(u, v)}{Mg(u, v)} = (\cos u)^2(\sin v)^2$$

Algorithm:
1. Draw a sample $(U, V) \sim \mathcal{N}(0, 10I_2)$.
2. Draw an independent uniform variable $W \sim \mathcal{U}(0, 1)$.
3. If $W \leq (\cos U)^2(\sin V)^2$, accept (U, V) as a sample from f. Otherwise, reject and return to step 1.