1. In lectures, we derived the M-step updates for fitting Gaussian mixtures with EM algorithm, for the mixing proportions and for the cluster means, assuming the common covariance \( \sigma^2 I \) is fixed and known.

(a) What happens to the algorithm if we set \( \sigma^2 \) to be very small? How does the resulting algorithm as \( \sigma^2 \to 0 \) relate to K-means?

**Answer:** In the E-step, the posterior probabilities are:

\[
q_{\sigma^2}(z_i = k) \propto \pi_k f(x_i | \mu_k, \sigma^2) \propto \pi_k \exp\left(-\frac{1}{2\sigma^2} \|x_i - \mu_k\|_2^2\right)
\]

As \( \sigma^2 \) approaches zero, the exponentiated term will be dominated by the \( k \) such that \( \mu_k \) is closest to \( x_i \) by Euclidean distance. Thus, if there is a unique such \( k \), as \( \sigma^2 \to 0 \):

\[
q_{\sigma^2}(z_i = k) \to \begin{cases} 1 & \text{for } k = \text{argmin}_{k'} \|x_i - \mu_{k'}\|, \\ 0 & \text{otherwise.} \end{cases}
\]

If there is another \( \mu_{k'} \) at same distance to \( x_i \), \( q(z_i) \) will spread probability mass equally among all such components. This looks exactly like the cluster assignment step of K-means (also note: \( \pi_k \) values have no effect on the cluster assignment in this limit). The M-step is exactly the mean update step, thus K-means can be understood as an EM algorithm for a mixture of Gaussians with infinitesimally small \( \sigma^2 \).

(b) If \( \sigma^2 \) is in fact not known and is a parameter to be inferred as well, derive an M-step update for \( \sigma^2 \).

**Answer:** Differentiating the free energy with respect to \( \nu = \sigma^{-2} \) (you can also differentiate with respect to \( \sigma \) or \( \sigma^2 \), just involves a bit more algebra) and setting to 0 gives:

\[
\nabla_{\nu} F(\theta, q) = \sum_{i=1}^{n} \sum_{k=1}^{K} q(z_i = k) \nabla_{\nu} \left( -\frac{p}{2} \log(2\pi/\nu) - \nu \frac{1}{2} \|x_i - \mu_k\|_2^2 \right) = \sum_{i=1}^{n} \sum_{k=1}^{K} q(z_i = k) \left( \frac{p}{2} \frac{1}{\nu} - \frac{1}{2} \|x_i - \mu_k\|_2^2 \right) = \frac{np}{2\nu} - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} q(z_i = k) \|x_i - \mu_k\|_2^2 = 0 \Rightarrow \sigma^2 = \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} q(z_i = k) \|x_i - \mu_k\|_2^2}{np}.
\]

2. We are given a labelled dataset \( \{(x_i, y_i)\}_{i=1}^{n} \) with \( x_i \in \{0, 1\}^p \) and \( y_i \in \{1, \ldots, K\} \) and the na\"ive Bayes classifier model which assumes that different dimensions/features in vector \( X_i \) are independent given the class label \( Y_i = k \), resulting in the joint probability

\[
p(x_i, y_i; \{\pi_k\}, \{\phi_{kj}\}) = \prod_{k=1}^{K} \left( \pi_k \prod_{j=1}^{p} \left( \phi_{kj} x_{ij}^{(y)} (1 - \phi_{kj})^{1-x_{ij}^{(y)}} \right) \right).
\]
where $\pi_k = \mathbb{P}(Y_i = k)$ are the marginal class probabilities and $\phi_{kj}$ is the probability of feature $j$ being present in the class $k$, i.e., of $x_i^{(j)} = 1$ for an item $x_i$ belonging to class $k$.

(a) Derive the maximum likelihood estimates for $\pi_k$ and $\phi_{kj}$.

**Answer:** Denoting $n_k = \sum_{i=1}^{n} 1(y_i = k)$, direct differentiation of the log-likelihood (with the Lagrange multiplier for $\pi_k$ as before) gives

$$\hat{\pi}_k = \frac{n_k}{n}, \quad \hat{\phi}_{kj} = \frac{\sum_{i=1}^{n} 1(y_i = k)x_i^{(j)}}{n_k}.$$

(b) Assume that we are also given an additional set of unlabelled data items $\{x_i\}_{i=n+1}^{n+m}$. Using the same naïve Bayes model, and by treating missing labels as latent variables, describe an EM algorithm that makes use of this unlabelled dataset and give the E-step update for the variational distribution $q$ and the M-step updates for parameters $\pi_k$ and $\phi_{kj}$. Discuss the difference of these results to those in part (a).

**Answer:**

We model the missing labels as latent random variables $z_i = y_{n+i}$, for $i = 1, \ldots, m$. E-step then sets the variational distribution $q(z)$ to the posterior of the missing labels given data and current parameters:

$$q^{(t+1)}(z_i = k) \propto \pi_k^{(t)} \prod_{j=1}^{p} \left( \phi_{kj}^{(t)} x_{n+i}^{(j)} \right)^{x_i^{(j)}} \left( 1 - \phi_{kj}^{(t)} \right)^{1-x_i^{(j)}}$$

and M-step finds the optimal parameters given $q$. The M-step update is:

$$\hat{\pi}_k^{(t)} = \frac{n_k + \sum_{i=1}^{m} q^{(t)}(z_i = k)}{n+ m}, \quad \hat{\phi}_{kj}^{(t)} = \frac{\sum_{i=1}^{n} 1(y_i = k)x_i^{(j)} + \sum_{i=1}^{m} q^{(t)}(z_i = k)x_{n+i}^{(j)}}{n_k + \sum_{i=1}^{m} q^{(t)}(z_i = k)}.$$

This is an example of a semisupervised problem. In part (a), all labels are observed (fully supervised setting) which is like having $q(z_i = k) = 1\{y_i = k\}$ and there are no latent variables.

3. Verify that in the probabilistic PCA model from the lectures, E-step of the EM algorithm at iteration $t + 1$ can be written as

$$q^{(t+1)}(y_i) = \mathcal{N} \left( y_i; b_i^{(t)}, R^{(t)} \right)$$

where

$$b_i^{(t)} = \left( (L^{(t)})^\top L^{(t)} + (\sigma^2)^{(t)} I \right)^{-1} (L^{(t)})^\top x_i, \quad (1)$$

$$R^{(t)} = (\sigma^2)^{(t)} \left( (L^{(t)})^\top L^{(t)} + (\sigma^2)^{(t)} I \right)^{-1}. \quad (2)$$

**Answer:**
Follows from Gaussian conditioning, i.e., completing the square in the exponent. We omit super-
script \(\cdot\) on parameters \(\theta = (L, \sigma^2)\) for simplicity.

\begin{align*}
p(y_i|x_i, \theta) &\propto p(y_i)p(x_i|y_i, \theta) \\
&\propto \exp\left(-\frac{1}{2}y_i^T y_i\right) \exp\left(-\frac{1}{2\sigma^2}(x_i - Ly_i)^T(x_i - Ly_i)\right) \\
&\propto \exp\left(-\frac{1}{2}y_i^T \left(I + \frac{1}{\sigma^2}L^T L\right)y_i - 2y_i^T \frac{1}{\sigma^2}L^T x_i\right) \\
&\propto N(y_i; b_i, R)
\end{align*}

where

\begin{align*}
R &= \sigma^2 \left(L^T L + \sigma^2 I\right)^{-1}, \\
b_i &= \left(L^T L + \sigma^2 I\right)^{-1}L^T x_i.
\end{align*}

as required.

4. Consider a collaborative filtering model with “implicit feedback” observations \(y_{ij}\) which indicate not the rating but some form of frequency of interaction of user \(j\) with item \(i\) (for example, a user may watch a TV series every week, but that does not necessarily mean that she would rate it higher than a film she has seen only once). We convert the implicit feedback into binary \(b_{ij} = 1\{y_{ij} > 0\}\) and also introduce confidence measures \(c_{ij} = 1 + \alpha y_{ij}\) for \(\alpha > 0\) (note that we do not treat \(y_{ij} = 0\) as missing - we simply have a lower confidence in those observations). For user \(j\), we are then solving the weighted least squares problem:

\[\min_{\psi_j} \sum_{i=1}^{n_1} c_{ij}(b_{ij} - \phi_i^T \psi_j)^2 + \lambda_{\psi}\|\psi_j\|_2^2, \quad j = 1, \ldots, n_2.\] (5)

By expressing the criterion in matrix form, derive a closed form solution of \(\psi_j\).

\textbf{Answer:} The objective can be written as

\[J(\psi) = (b_j - \Phi \psi)^T\text{diag}(c_{.j})(b_j - \Phi \psi) + \lambda_{\psi}\psi^T \psi,\]

Now

\[\frac{\partial J}{\partial \psi} = -2\Phi^T \text{diag}(c_{.j})(b_j - \Phi \psi) + 2\lambda_{\psi}\psi.\]

Setting to zero and solving for \(\psi\) gives

\[\psi = (\Phi^T \text{diag}(c_{.j})\Phi + \lambda_{\psi} I)^{-1}\Phi^T \text{diag}(c_{.j})b_{.j}.\]

5. Consider a collaborative filtering model on binary ratings \(-1\) and \(+1\) with a \textit{probit likelihood}

\[p(y_{ij} = 1|a_i, b_j) = \Phi(a_i^T b_j) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_i^T b_j} \exp\left(-t^2/2\right) dt,\] (6)
where $y_{ij}$ is the rating of item $i$ by user $j$, $a_i \in \mathbb{R}^k$ is the feature vector of item $i$, $b_j$ is the preference vector of user $j$ and $\Phi$ is the standard normal cdf.

Consider an alternative model with additional latent variables $z_{ij}$, given by

$$z_{ij}|a_i, b_j \sim \mathcal{N}(a_i^\top b_j, 1), \quad p(y_{ij} = 1|z_{ij}) = 1\{z_{ij} > 0\}.$$  

(a) Show that these two models are equivalent, i.e. that $p(y_{ij} = 1|a_i, b_j)$ still takes the form in (6).

**Answer:**

We need to marginalise over $z_{ij}$ as follows:

$$p(y_{ij} = +1|a_i, b_j) = \int_{-\infty}^{\infty} p(y_{ij} = 1, z_{ij}|a_i, b_j) \, dz_{ij}$$

$$= \int_{-\infty}^{\infty} p(y_{ij} = 1|z_{ij}) \, p(z_{ij}|a_i, b_j) \, dz_{ij}$$

$$= \int_0^{\infty} \mathcal{N}(z_{ij}; a_i^\top b_j, 1) \, dz_{ij}$$

$$= 1 - \Phi(-a_i^\top b_j)$$

$$= \Phi(a_i^\top b_j),$$

where in the last two lines we used that the normal distribution with mean $\mu$ and variance 1 has cdf $\Phi(x - \mu)$ and the rotational symmetry of $\Phi$.

(b) Derive $p(z_{ij}|a_i, b_j, y_{ij} = \pm 1)$.

**Answer:**

$$p(z_{ij}|a_i, b_j, y_{ij} = 1) \propto p(y_{ij} = 1|z_{ij}) \, p(z_{ij}|a_i, b_j)$$

$$= 1\{z_{ij} > 0\} \mathcal{N}(z_{ij}; a_i^\top b_j, 1)$$

$$= \mathcal{N}^+(z_{ij}; a_i^\top b_j, 1),$$

where $\mathcal{N}^+$ is the normal distribution truncated to the positive part of the real line. Similarly, for $y_{ij} = -1$, we obtain the normal distribution truncated to the negative part of the real line.

(c) Now consider the model that treats feature vectors and preference vectors as model parameters $\theta = \{(a_i)_{i=1}^{n_1}, (b_j)_{j=1}^{n_2}\}$ with latents $Z = \{z_{ij}\}_{e_{ij}=1}$. Describe the resulting EM algorithm.

**Answer:** E-step sets the variational distribution $q(z_{ij}) = \mathcal{N}^+(z_{ij}; a_i^\top b_j, 1)$ to the truncated normal derived in (b). M-step maximises over $A$ and $B$

\[
\mathbb{E}_q \log p(Y, Z|A, B) = \sum_{e_{ij}=1} \mathbb{E}_q \log p(y_{ij}|z_{ij})p(z_{ij}|a_i, b_j)
\]

$$= \text{const}_1 + \frac{1}{2} \sum_{e_{ij}=1} \mathbb{E}_q \left( (z_{ij} - a_i^\top b_j)^2 \right)$$

$$= \text{const}_2 + \frac{1}{2} \sum_{e_{ij}=1} \left( \mathbb{E}_q z_{ij} - a_i^\top b_j \right)^2,$$
where in the last step we added and subtracted \((E_q z_{ij})^2\) (which are independent on \(a_i, b_j\), so can be merged into constant terms). Thus, for M-step we perform alternating least squares on \(E_q Z\). Note that:

\[
E_q z_{ij} = a_i^T b_j + N(a_i^T b_j; 0, 1) / \Phi(a_i^T b_j), \quad y_{ij} = +1, \\
E_q z_{ij} = a_i^T b_j - N(a_i^T b_j; 0, 1) / \Phi(-a_i^T b_j), \quad y_{ij} = -1,
\]

so that the variational means are pushed up for the positive ratings and down for the negative ones.

6. Suppose we have a model \(p(X, z|\theta)\) where \(X\) is the observed dataset and \(z\) are the latent variables. We would like to take a Bayesian approach to learning, treating the parameter \(\theta\) to be a random variable as well, with some prior \(p(\theta)\).

(a) Suppose that \(q(z, \theta)\) is a distribution over both \(z\) and \(\theta\). Explain why the following is a lower bound on \(p(X)\):

\[
F(q) = E_q[\log p(X, z, \theta) - \log q(z, \theta)]
\]

**Answer:** We can write

\[
F(q) = E_q[\log p(z, \theta|X) - \log q(z, \theta)] + \log p(X)
\]

with the first term being the negative of KL divergence, so is non-positive. Thus \(F(q)\) is a lower bound on \(\log p(X)\).

(b) Show that the optimal \(q(z, \theta)\) is simply the posterior \(p(z, \theta|X)\).

**Answer:** The optimal \(q\) is the one that minimises KL divergence to the posterior, i.e. the posterior \(p(z, \theta|X)\) itself.

(c) Typically the posterior is intractable. Consider a factorised distribution \(q(z, \theta) = q_\theta(\theta)q_\theta(z)\). In other words we assume that \(z\) and \(\theta\) are independent. Derive the optimal \(q_\theta\) given a \(q_\theta\), and hence describe an algorithm to optimise \(F(q)\) subject to assumption of independence between \(z\) and \(\theta\).

**Answer:** Making the factorization assumption,

\[
F(q) = \int q_\theta(\theta) q_\theta(z) [\log p(X, z, \theta) - \log q_\theta(z) - \log q_\theta(\theta)] d\theta dz
\]

Differentiate wrt \(q_\theta(z)\), including a Lagrange multiplier to make sure \(q_\theta(z)\) integrates to 1, we get:

\[
\nabla_{q_\theta}(z) F(q) = \int q_\theta(\theta) \log p(X, z, \theta) d\theta - \log q_\theta(z) - 1 + \lambda = 0
\]

\[
q_\theta(z) \propto \exp \left( \int q_\theta(\theta) \log p(X, z, \theta) d\theta \right)
\]

By symmetry, the optimal \(q_\theta\) given \(q_\theta\) is:

\[
q_\theta(\theta) \propto \exp \left( \int q_\theta(z) \log p(X, z, \theta) dz \right)
\]

We can alternate between optimizing \(q_\theta\) given \(q_\theta\) and vice versa to maximise the lower bound. This is similar to the EM algorithm.
7. Verify steps (2) and (3) in the CAVI updates for the Latent Dirichlet Allocation model.

**Answer:** For (2), we note that

$$\log p(\theta_d|z_d) = \text{const} + \sum_{k=1}^{K} \left( \alpha_k + \sum_{n=1}^{N_d} z_{dn}[k] - 1 \right) \log \theta_{dk},$$

so that

$$\exp [ E_{z_d \sim q} \log p(\theta_d|z_d)] \propto \prod_{k=1}^{K} \theta_{dk}^{\alpha_k + \sum_{n=1}^{N_d} \phi_{dn}[k] - 1},$$

which is proportional to the Dirichlet distribution with parameter vector $\gamma_d = \alpha + \sum_{n=1}^{N_d} \phi_{dn}$ and similarly for (3).