1. Given a hidden Markov model (HMM, page 610) and observations \( y_1, \ldots, y_{t-1} \), show that the predictive distribution of the observations \( y_t \) at time point \( t \) follows a mixture distribution.

Solution:
Let us first write the joint distribution of all variables:

\[
P(y_1, \ldots, y_t, z_1, \ldots, z_t) = P(z_1)P(y_1 \mid z_1) \prod_{\tau=2}^{t} P(z_\tau \mid z_{\tau-1})P(y_\tau \mid z_\tau). \quad (1)
\]

Then we can manipulate the predictive distribution:

\[
P(y_t \mid y_1, \ldots, y_{t-1}) = \sum_{z_t} P(y_t, z_t \mid y_1, \ldots, y_{t-1})
\]

\[
= \sum_{z_t} P(z_t \mid y_1, \ldots, y_{t-1})P(y_t \mid z_t, y_1, \ldots, y_{t-1})
\]

\[
= \sum_{z_t} P(z_t \mid y_1, \ldots, y_{t-1})P(y_t \mid z_t),
\]

which is clearly a mixture distribution with the posterior distribution of the latent variable \( P(z_t \mid y_1, \ldots, y_{t-1}) \) as the mixture coefficients and \( P(y_t \mid z_t) \) as the component distributions.

2. Show how a second-order Markov chain (page 608) of 3 symbols can be transformed to a hidden Markov model with 9 states and 3 symbols.

Solution:
A second order Markov chain has a model for \( P(y_t \mid y_{t-2}, y_{t-1}) \).

\[
P(y_t \mid y_{t-2}, y_{t-1}) = \begin{array}{cccccccc}
{aa} & {ab} & {ac} & {ba} & {bb} & {bc} & {ca} & {cb} & {cc} \\
y_t = a & . & . & . & . & . & . & . & . \\
y_t = b & . & . & . & . & . & . & . & . \\
\end{array}
\]

By setting the hidden state \( z_t \) to contain both \( y_{t-1} \) and \( y_t \) as a concatenated symbol, we can emulate the second order Markov chain by a hidden Markov model using the following tables:
\[
P(y_t | z_t) \begin{array}{cccccccccc}
\hline
y_t = a & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
y_t = b & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
y_t = c & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\
\hline
\end{array}
\]

\[
P(z_t | z_{t-1}) \begin{array}{cccccccccc}
\hline
z_t = aa & \cdot & 0 & 0 & \cdot & 0 & 0 & \cdot & 0 & 0 \\
z_t = ab & \cdot & 0 & 0 & \cdot & 0 & 0 & \cdot & 0 & 0 \\
z_t = ac & \cdot & 0 & 0 & \cdot & 0 & 0 & \cdot & 0 & 0 \\
z_t = ba & 0 & \cdot & 0 & 0 & \cdot & 0 & 0 & \cdot & 0 \\
z_t = bb & 0 & \cdot & 0 & 0 & \cdot & 0 & 0 & \cdot & 0 \\
z_t = bc & 0 & \cdot & 0 & 0 & \cdot & 0 & 0 & \cdot & 0 \\
z_t = ca & 0 & 0 & \cdot & 0 & 0 & \cdot & 0 & 0 & \cdot \\
z_t = cb & 0 & 0 & \cdot & 0 & 0 & \cdot & 0 & 0 & \cdot \\
z_t = cc & 0 & 0 & \cdot & 0 & 0 & \cdot & 0 & 0 & \cdot \\
\hline
\end{array}
\]

where the values \cdot are copied from the table of the second order Markov chain.

This shows that a hidden Markov model is more general than a second order Markov chain (and similarly of a Markov chain of any order).

3. Let us consider a HMM with a discrete hidden variable \( z \) with 6 states and a Gaussian observation (emission) probability density function. The dimension of the data vectors \( x_1, \ldots, x_T \) is 5 and the covariance function of the Gaussian distribution is diagonal. (a) Quantify the number of parameters in the model, (b) write the joint probability density, (c) and write the \( Q \)-function of the EM-algorithm \( Q(\theta, \theta^{\text{old}}) \) (page 440). Assume that the E-step is done, that is, \( \gamma(z_t) = P(z_t | X, \theta^{\text{old}}) \) and \( \xi(z_{t-1}, z_t) = P(z_{t-1}, z_t | X, \theta^{\text{old}}) \) are given.

Solution:

(a) Parameters \( \theta \) include the starting distribution \( P(z_1) = \pi = P(z_1 | z_0) \) with 6 parameters of which 5 are free, transition matrix \( \bar{A} \) with 36 parameters of which 30 are free, and parameters \( \mu_{ij} \) and \( \sigma^2_{ij} \) for the emission distribution (60 parameters, all of them free). That makes altogether 102 parameters of which 95 are free.

(b) A Gaussian distribution with a diagonal covariance can be repre-
presented as a product of 1-dimensional Gaussians.

\[ p(X, Z \mid \theta) = \prod_{t=1}^{T} P(z_t \mid z_{t-1}, \theta) p(x_t \mid z_t, \theta) \]  
\[ = \prod_{t=1}^{T} a_{z_{t-1}, z_t} \prod_{k=1}^{5} \frac{1}{\sqrt{2\pi\sigma^2_{z_k,k}}} \exp\left[-\frac{(x_{tk} - \mu_{zk,k})^2}{2\sigma^2_{z_k,k}}\right] \]  

\[ Q(\theta, \theta^{\text{old}}) = \sum_{Z} P(Z \mid X, \theta^{\text{old}}) \ln p(X, Z \mid \theta) \]  
\[ = \sum_{Z} P(Z \mid X, \theta^{\text{old}}) [\ln P(Z \mid \theta) + \ln p(X \mid Z, \theta)] \]  
\[ = \sum_{i=1}^{T} \sum_{i=1}^{6} \sum_{j=1}^{6} \xi(z_{t-1,i}, z_{t,j}) \ln a_{ij} \]  
\[ + \sum_{i=1}^{T} \sum_{i=1}^{6} \sum_{k=1}^{5} \gamma(z_{ti}) \ln \left( \frac{1}{\sqrt{2\pi\sigma^2_{ik}}} \exp\left[-\frac{(x_{tk} - \mu_{ik})^2}{2\sigma^2_{ik}}\right]\right) \]  
\[ = Q_z + \sum_{i=1}^{T} \sum_{i=1}^{6} \sum_{k=1}^{5} \gamma(z_{ti}) \left[-\frac{(x_{tk} - \mu_{ik})^2}{2\sigma^2_{ik}} - \frac{1}{2} \ln(2\pi\sigma^2_{ik})\right] \]  
\[ = Q_z + Q_x, \]  

where the division into two parts \( Q_z + Q_x \) will be useful in Problem 4.

4. In the setting of Problem 3, (a) derive the M-step for the Gaussian means \( \mu_{ik} \), where \( i = 1 \ldots 6 \) denotes the state and \( k = 1 \ldots 5 \) denotes the data dimension. (b) Derive the M-step for updating the \( 6 \times 6 \) transition matrix \( A \).

Solution:

(a) As we maximize the Q-function w.r.t. a particular \( \mu_{ik} \), the part \( Q_z \) is constant, and from the sums over \( i \) and \( k \), all the other terms are constant
except the one we are interested in. Therefore we only need:

\[
\frac{\partial}{\partial \mu_{ik}} \sum_{t=1}^{T} \gamma(z_{ti}) \frac{(x_{tk} - \mu_{ik})^2}{2\sigma_{ik}^2} = 0
\]

(13)

\[
\sum_{t=1}^{T} \gamma(z_{ti}) \frac{x_{tk} - \mu_{ik}}{\sigma_{ik}^2} = 0
\]

(14)

\[
\mu_{ik} = \frac{\sum_{t=1}^{T} \gamma(z_{ti}) x_{tk}}{\sum_{t=1}^{T} \gamma(z_{ti})}
\]

(15)

that is, \( \mu \) will be the weighted average of the data points assigned to the cluster (or state) \( i \), the weights being the probabilities \( \gamma \) that this point belongs to this cluster.

(b) Next we should maximize \( Q \) w.r.t. an element of the transition matrix \( a_{ij} \). This time \( Q_x \) is a constant that can be ignored. If we simply try to find the zero of the gradient, we notice that increasing \( a_{ij} \) will always increase \( Q \) so there is no zero of the gradient. We need to take into account the constraint \( \sum_{j=1}^{6} a_{ij} = 1 \forall i \). One way to do this is to introduce Lagrange multipliers \( \lambda_i > 0 \) for each constraint \( i \). We will now maximize

\[
Q_z - \lambda_i \left( \sum_{j=1}^{6} a_{ij} - 1 \right)
\]

(16)

instead. The intuition behind this is to introduce a “counter-force” that balances the ever increasing \( a_{ij} \)s. When the force \( \lambda_i \) is just right, it will set the constraint to be true, and the modified cost function in Eq. (16) will be equal to \( Q_z \) since \( \left( \sum_{j=1}^{6} a_{ij} - 1 \right) = 0 \).

Let us try to maximize (16) by finding the zero of the gradient:

\[
0 = \frac{\partial}{\partial a_{ij}} \left[ \sum_{t=1}^{T} \xi(z_{t-1,i} z_{ij}) \ln a_{ij} - \lambda_i \left( \sum_{j'=1}^{6} a_{ij'} - 1 \right) \right]
\]

(17)

\[
= \frac{\sum_{t=1}^{T} \xi(z_{t-1,i} z_{ij})}{a_{ij}} - \lambda_i
\]

(18)

\[
a_{ij} = \frac{\sum_{t=1}^{T} \xi(z_{t-1,i} z_{ij})}{\lambda_i}
\]

(19)
Thus, $\lambda_i$ turned out to be a normalization constant, whose value we can compute from

$$
\sum_{j=1}^{6} a_{ij} = \sum_{j=1}^{6} \frac{\sum_{t=1}^{T} \xi(z_{t-1,j}, z_{t,j})}{\lambda_i} = 1 \quad (20)
$$

$$
\lambda_i = \sum_{j=1}^{6} \sum_{t=1}^{T} \xi(z_{t-1,i}, z_{t,i}). \quad (21)
$$