1 Expectation-Maximization (EM) Algorithm

In this problem, you will solve a maximum likelihood estimation problem in two ways. First, you will solve it directly, obtaining a closed form solution. Given a closed form solution, the derivation of an iterative algorithm for this problem is not fundamental. However, in the second part of this problem you will derive an expectation-maximization (EM) algorithm for it. This algorithm may be extended to more complicated scenarios where it would be useful.

Assume that \( \lambda \) is a random variable drawn from a gamma density function

\[
p(\lambda|\theta) = \frac{\theta^M}{\Gamma(M)} \lambda^{M-1} e^{-\lambda \theta}, \lambda \geq 0.
\]  

(1)

Here \( \theta \) is an unknown nonnegative parameter and \( \Gamma(M) \) is the Gamma function. Note that \( \Gamma(M) \) is the normalizing constant,

\[
\Gamma(M) = \int_0^{\infty} x^{M-1} e^{-x} dx.
\]

(2)

The Gamma function satisfies a recurrence formula \( \Gamma(M+1) = M \Gamma(M) \). For \( M \) an integer, \( \Gamma(M) = (M-1)! \). Note that \( M \) is known. This implies that the mean of this gamma density function is \( E[\lambda|\theta] = \frac{M \theta}{\theta} \).

The random variable \( \lambda \) is not directly observable. Given the random variable \( \lambda \), the observations \( x_1, x_2, \ldots, x_N \), are i.i.d. with probability density functions

\[
p_{x_i|\lambda}(X_i|\lambda) = \lambda e^{-\lambda X_i}, X_i \geq 0.
\]

(3)

Note that there is one random variable \( \lambda \) and there are \( N \) random variables \( x_i \) that are i.i.d. given \( \lambda \).

**a.** Find the joint probability density function for \( x_1, x_2, \ldots, x_N \) conditioned on \( \theta \).

**b.** Directly from the probability density function from part a, find the maximum likelihood estimate of \( \theta \). Note that this estimate does not depend on \( \lambda \).

**c.** To start the derivation of the EM algorithm, write down the complete data loglikelihood function, keeping only terms that depend on \( \theta \). Denote this function \( l_{cd}(\lambda|\theta) \).

**d.** Compute the expected value of the complete data loglikelihood function given the observations \( x_1, x_2, \ldots, x_N \), and the previous estimate for \( \theta \) denoted \( \hat{\theta}(k) \). Denote this function by \( Q(\theta|\hat{\theta}(k)) \). Hint: This step requires a little thought. The posterior density function on \( \lambda \) given the measurements is in a familiar form.
e. Maximize the $Q$ function over $\theta$ to obtain $\hat{\theta}^{(k+1)}$. Write down the resulting recursion with $\hat{\theta}^{(k+1)}$ as a function of $\hat{\theta}^{(k)}$.

f. Verify that the maximum likelihood estimate derived in part b is a fixed point of the iterations derived in part e.

### 2 EM Algorithm

Suppose that the random variable $R$ is a sum of two exponentially distributed random variables, $X$ and $Y$,

$$R = X + Y,$$

where $p_X(x) = \alpha \exp(-\alpha x), x \geq 0$, and $p_Y(y) = \beta \exp(-\beta y), y \geq 0$. The value of $\beta$ is known, but $\alpha$ is not known. The goal of the problem is to derive an algorithm to estimate $\alpha$.

a. Write down the loglikelihood function for the data $R$ (this is the incomplete data loglikelihood function). Find a first order necessary condition for $\alpha$ to be a maximum likelihood estimate. Is this equation easy to solve for $\alpha$?

b. Define the complete data to be the pair $(X, R)$, and write down the complete data loglikelihood function.

c. Determine the conditional probability density function on $X$ given $R$, as a function of a nominal value $\tilde{\alpha}$. Denote this probability density function (pdf) $p(x|r, \tilde{\alpha})$.

d. Using the pdf $p(x|r, \tilde{\alpha})$, determine the conditional mean of $X$ given $R$ and $\tilde{\alpha}$.

e. Determine the function $Q(\alpha|\tilde{\alpha})$, the expected value of the complete data loglikelihood function given the incomplete data and $\tilde{\alpha}$.

f. Derive the expectation-maximization algorithm for estimating $\alpha$ given $R$.

### 3 EM Algorithm

In this problem, you will explore some classical issues in density estimation. These are the same types of issues which arise when one tries to estimate a continuous function from discrete data. The difficulty arises because the estimates tend to be concentrated when in fact some prior knowledge leads you to believe that the functions are not “peaked.” To be specific, suppose that you observe $N$ independent identically distributed random variables, and you were supposed to guess what the density was which gave rise to this data. One description of a maximum likelihood solution is

$$\sum_{k=1}^{N} \frac{1}{N} \delta(x - X_k),$$

where the $X_k$ are the observations. If you have some prior knowledge that the original density was in fact smooth, you would immediately reject this solution as infeasible. One approach is to represent your set of possible solutions as a sum of smooth functions (or equivalently as the result of a convolution with a smooth function).

Suppose our set of admissible density functions are

$$p(x) = \frac{1}{N} \sum_{k=1}^{N} f(x - m_k)$$

where $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-x^2/2\sigma^2)$. This set of functions makes sense in that it is the result of a concentrated density being smoothed with a gaussian. The problem of interest is then, given $M$ independent observations of the random variable $x$, to find your best estimate for $N$ and for the $m_k, k = 1, 2, \ldots, N$. 

2
a. Assume that $M = 2$ and that $N = 2$. Thus there are two random variables observed, $X_1$ and $X_2$, and we wish to estimate $m_1$ and $m_2$. Use the method of maximum likelihood to find two equations which the maximum likelihood estimates must satisfy. These equations are difficult to solve. They may be reduced to determining one parameter, however, by noting the symmetry involved in the equations. In particular, if we let $m_1 = X_1 + \Delta$ and $m_2 = X_2 - \Delta$, then the equations to be solved reduce to one (still difficult) equation in $\Delta$. You don’t need to solve these equations. Is $\Delta = 0$ a valid solution? Comment. Is $\Delta = (X_2 - X_1)/2$ a valid solution? Comment. If $X_2 > X_1$, can $\Delta$ ever be negative? Comment.

As a side comment, it is worth noting that one way of finding the solution is to center a Gaussian on each of the observed data points, sum them, then look for the peaks of the sum.

b. In this part of the problem, you will find an EM algorithm for solving for the maximum likelihood solution. A model for the data must be determined which can be put in the usual form of a complete data space and an incomplete data space. In the general problem, one models the observed random variable as having come from one of $N$ equally likely experiments the $k$th one of which has probability density function $f(x - m_k)$. The complete data consists of pairs $(X_i, n_i)$, where $n_i$ specifies which of the $N$ p.d.f.’s $X_i$ came from. The mapping from the complete data to the incomplete data just selects the $X_i$.

b.1. Assume $M = N = 2$. Determine the complete data loglikelihood. This step is crucial as it determines the function we are going to maximize. Define the indicator functions

$$I_k(n) = \begin{cases} 1 & \text{if } k = n \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

(A hint for finding the complete data loglikelihood is that the density on $X_i$ if $n = k$ is $I_k(n) \ln[f(x - m_k)]$; since the $X_i$’s are independent, the complete data loglikelihood is the sum of terms like this.)

b.2. Find the expected value of the complete data loglikelihood given $X_1$, $X_2$, $\hat{m}_1^r$, $\hat{m}_2^r$. This involves finding the expected value of the indicator functions $I_k(n_i)$. Be careful with these. (I use $r$ to indicate the iteration number here.)

b.3. Maximize the result of the last step to get the updates for $\hat{m}_1^{r+1}$ and $\hat{m}_2^{r+1}$.

b.4. Pick a good initial value for your estimates. Justify why this selection is good.

c. This part is almost trivial. Let $M = 2$ and $N = 1$. There is only one parameter to determine here, the mean of the Gaussian density. Find the maximum likelihood estimate for this mean.

d. In this part, you will set up a detection problem to determine how many Gaussians should be included in the sum. Suppose $M = 2$. Under hypothesis $H_1$, there is one Gaussian in the density ($N = 1$). Under hypothesis $H_2$ there are two Gaussians in the density ($N = 2$). Determine the likelihood ratio test for this problem. Assume the prior probabilities on the hypotheses are equal and that the costs of errors are equal. Since there are unwanted parameters in the ratio (namely the means of the Gaussians), and these parameters are nonrandom, substitute for them their appropriate estimates. Usually in hypothesis testing problems the condition that the threshold is exactly equal to the likelihood ratio is not important. Is it important here? For the case of the ratio equalling the threshold, choose $H_1$ (the choice with fewer parameters).

e. Let $X_1 = 1$ and $X_2 = 2$. First, suppose $\sigma^2 = 0.04$. Find the maximum likelihood estimates for $m_1$ and $m_2$ (this is the $N = 2$ case). Now, let $\sigma^2$ get larger. At what point does the hypothesis test in d yield the decision that $N = 1$? What are the maximum likelihood estimates for $m_1$ and $m_2$ at this point? If you cannot determine the point exactly, find a few values nearby.

f. Write and test a Matlab routine to run the EM algorithm derived in part b. Show the performance of the algorithm by running it many times for one choice of the means. Compute the means and covariances of the estimates. Run for different choices of means. Note that it is sufficient to write the code assuming that $\sigma^2 = 1$ and to scale the means by $1/\sigma$. 

3