Math 152 - Statistical Theory - not Homework 3

write your name here

not due

As an example, consider a simple coin-flipping experiment in which we are given a pair of coins A and B of unknown biases, θ_A and θ_B, respectively (that is, on any given flip, coin A will land on heads with probability θ_A and tails with probability $1 - \theta_A$ and similarly for coin B). Our goal is to estimate $\theta = (\theta_A, \theta_B)$ by repeating the following procedure five times: randomly choose one of the two coins (with equal probability), and perform ten independent coin tosses with the selected coin. Thus, the entire procedure involves a total of 50 coin tosses.

During our experiment, suppose that we keep track of two vectors $x = (x_1, x_2, ..., x_5)$ and $z = (z_1, z_2, ..., z_5)$, where $x_i \in \{0, 1, ..., 10\}$ is the number of heads observed during the ith set of tosses, and $z_i = 1$ if coin A and 0 is coin B identifies the coin used during the ith set of tosses. Parameter estimation in this setting is known as the complete data case in that the values of all relevant random variables in our model (that is, the result of each coin flip and the type of coin used for each flip) are known.

1. Find the MLE of θ_A and θ_B in the complete data setting described above.
2. Now consider a more challenging variant of the parameter estimation problem in which we are given the recorded head counts x but not the identities z of the coins used for each set of tosses.

2a. Write down the complete data log likelihood with the missing information ($P(x, z; \theta)$). [Note that this problem is very similar to the normal mixture model, except that we don't have a mixing parameter, π. Instead, the example says that the coins are randomly chosen, so each coin happens with probability 0.5.] Convince yourself that the complete data log likelihood is linear in z_j.

2b. Find $E_{q_{Z,\theta^{i-1}}}[P_{X,Z}(x, z; \theta)]$

2c. Maximize $Q(\theta|\theta^{i-1}) = E_{q_{Z,\theta^{i-1}}}[P_{X,Z}(x, z; \theta)]$ with respect to θ_A and then also with respect to θ_B.

2d. Consider the following data: ($x_1 = 5$, $x_2 = 9$, $x_3 = 8$, x_4, $x_5 = 7$). Program the EM algorithm to estimate θ_A and θ_B.

2e. Choose any two (different) values for θ_A and θ_B. Simulate 25 Binomial random variables twice: both times with $n = 100$ but once with θ_A and once with θ_B. Estimate the two values of θ_A and θ_B.

Problem from: Do & Batzoglou, (2008) "What is the expectation maximization algorithm?" *Nature Biotechnology* 26, 897-899.