"*SALD*" Homework **Part 2, Problems 2.1 – 2.3.**
(This is due in class on Monday, March 18, 2002, with part 1.)

In the following exercises, you will explore several important aspects of convergence for the *EM* algorithm to compute an MLE when the complete data are from the *exponential family* of distributions.

1. The first problem reminds you of some basic properties of the exponential family and asks you to identify two familiar distributions as belonging to this family.

2. The second problem steps you through a result showing that the sequence of *EM* estimates for a (one-dimensional) MLE in the exponential family converges *monotonically* to MLE, either from below or from above the MLE, depending on the starting value for the EM algorithm.

3. The third problem steps you through a result about the *rate of convergence* of the sequence of *EM* estimates for the MLE in the same (one-dimensional) exponential family. That rate is given by the "Missing Information Principle": See Tanner's discussion in section 4.4 for more background on this problem.

Problem 2.1:
Background: Here are some basic facts about the Exponential Family. (See Tanner 4.3, or Casella and Berger's book, where it appears in section 3.3 in the 1^{st} *ed*.)

Defn: A random variable X (or random vector X) has its distribution in the *exponential family* with k-dimensional parameter θ providing that its density function $\boldsymbol{f}$ can be written as:

$$\boldsymbol{f}(x \mid \theta) = \boldsymbol{b}(x)\, exp[\textstyle\sum_{i=1}^{k} \boldsymbol{g_i}(\theta)\, \boldsymbol{t_i}(x)] \,/\, \boldsymbol{a}(\theta)$$

where $\boldsymbol{a}$ (≥ 0) and the $\boldsymbol{t_i}$ are real-valued functions of the data only; where $\boldsymbol{b}$ (≥ 0) and the $\boldsymbol{g_i}$ are real-valued functions of the parameter only.

It is evident from the form of the density for the exponential family that the k-many statistics $\boldsymbol{T} = (\boldsymbol{t_1}(x), \ldots, \boldsymbol{t_k}(x))$ are sufficient for θ.

Defn.: Call $\Gamma = (\boldsymbol{g_1}(\theta), \ldots, \boldsymbol{g_k}(\theta))$, the k-dimensional *natural parameter* of the family,
and $\boldsymbol{T} = (\boldsymbol{t_1}(x), \ldots, \boldsymbol{t_k}(x))$, the k-dimensional *natural sufficient statistic* of the family.
Moreover, the natural sufficient statistic $\boldsymbol{T}$ also has its distribution within the exponential family, using the same natural parameters.

Let $\mathbf{X}_j$ ($j = 1, \ldots, n$) be *iid* sample of size n from an exponential family. Define the k-many statistics $\boldsymbol{T_i} = \Sigma_j\, \boldsymbol{t_i}(\mathrm{x_j})$. It follows that $(\boldsymbol{T_1}, \ldots, \boldsymbol{T_k})$ are jointly sufficient and have a distribution from the exponential family, with the same natural parameters as the $\mathbf{X}_j$.

Exercise 2.1.1: Show that an *iid* sample of size n from the Binomial $Bin(n, \theta)$ distribution ($0 < \theta < 1$) belongs to a exponential family with dimension $k = 1$. What are the natural parameter and sufficient statistic here?

Exercise 2.1.2: Show that an *iid* sample of size n from the Normal $N(\mu, \sigma^2)$ distribution (with $\mu \in \Re$, and $0 < \sigma$) belongs to a exponential family with dimension $k = 2$. What are its natural parameters and sufficient statistics?

Problem 2.2:
Let the observed data $X = x$ come from a statistical model with density $g(x \mid \theta)$. We want to find the *MLE*, $\text{argmax}_\theta \; log\; g(x \mid \theta) = \boldsymbol{L}(\theta)$. We apply the *EM* algorithm with complete data Z, which we assume come from a 1-dimensional exponential family, whose natural parameter is taken for convenience also as θ and whose density, $\boldsymbol{f}(z \mid \theta)$, is described above in problem **2.1**.

First. Show that $\mathbf{E}[\boldsymbol{T}(z) \mid \theta] = \alpha'(\theta)$ and that $\mathbf{E}[\boldsymbol{T}(z) \mid x, \theta] = \alpha'(\theta) + \boldsymbol{L}'(\theta)$.

Hint: Remember that $h(z \mid x, \theta) = f(z \mid \theta) / g(x \mid \theta)$ is the conditional density for the complete data z, given the observed data x.

Thus, $\quad log\; h(z \mid x, \theta) = \boldsymbol{T}(z)\theta + \beta(z) - \alpha(\theta) - \boldsymbol{L}(\theta)$, since

$log\; f(z \mid \theta) = \boldsymbol{T}(z)\theta + \beta(z) - \alpha(\theta)$

where $\alpha(\theta) = log\; \boldsymbol{a}(\theta)$ and likewise $\beta(z) = log\; \mathbf{b}(z)$

Differentiate and take expectations.

Argue that $\mathbf{E}[\partial/\partial\theta \;\; log\; f(z \mid \theta)\,] = \mathbf{E}_x[\partial/\partial\theta \;\; log\; h(z \mid x, \theta)\,] = 0$.

Thus, $\boldsymbol{L}'(\theta) = \mathbf{E}[\boldsymbol{T}(z) \mid x, \theta] - \mathbf{E}[\boldsymbol{T}(z) \mid \theta]$

Side remark: As $\boldsymbol{L}'(\hat{\theta}) = 0$, then $\mathbf{E}[\boldsymbol{T}(z) \mid \hat{\theta}] = \mathbf{E}[\boldsymbol{T}(z) \mid x, \hat{\theta}]$. That is, the MLE $\hat{\theta}$ makes the incomplete and complete data uncorrelated!

Second. Solve for θ_{j+1} which is the $j+1^{st}$ *EM* estimate of the MLE.

Hint: Argue that θ_{j+1} solves $\alpha'(\theta_{j+1}) = \mathbf{E}[\boldsymbol{T}(z) \mid x, \theta_j] = \mathbf{E}[\boldsymbol{T}(z) \mid \theta_{j+1}]$.

Third. Conclude that, because $\delta(\theta) = \mathbf{E}[\boldsymbol{T}(z) \mid x, \theta] - \mathbf{E}[\boldsymbol{T}(z) \mid \theta] > 0$ for $\theta < \hat{\theta}$ and $\delta(\theta) < 0$ for $\theta > \hat{\theta}$, then the sequence of *EM* estimators converges monotonically upwards to $\hat{\theta}$ if started from below $\hat{\theta}$ and converges monotonically downwards to $\hat{\theta}$ if stared from above $\hat{\theta}$.

Problem 2.3:
Last, for determining the *rate of convergence* in the sequence of *EM* estimates of the MLE, $\hat{\theta}$, argue as follows:

Denote by $\mathbf{I}_z(\theta)$ the Fisher Information contained in the complete data with respect to θ, associated with the density $f(z \mid \theta)$. Likewise, denote by $\mathbf{I}_{z|x}(\theta)$ the Fisher information with respect to θ associated with the conditional density $h(z \mid x, \theta)$.

Fourth: Show that $\mathbf{I}_z(\theta) = \alpha''(\theta)$ and that $\mathbf{I}_{z|x}(\theta) = \alpha''(\theta) + \boldsymbol{L}''(\theta)$.

Fifth: Show that as $j \to \infty$, the rate $(\theta_{j+1} - \hat{\theta}) / (\theta_j - \hat{\theta}) = \mathbf{I}_{z|x}(\hat{\theta}) / \mathbf{I}_z(\hat{\theta})$.

Hint: Use these two linear approximations for θ in the neighborhood of $\hat{\theta}$:

$$\mathbf{E}[\boldsymbol{T}(z) \mid x, \theta] = \mathbf{E}[\boldsymbol{T}(z) \mid x, \hat{\theta}\,] + \mathbf{I}_{z|x}(\theta)(\theta - \hat{\theta})$$

$$\mathbf{E}[\boldsymbol{T}(z) \mid \theta] = \mathbf{E}[\boldsymbol{T}(z) \mid \hat{\theta}\,] + \mathbf{I}_z(\theta)(\theta - \hat{\theta}).$$

Explain what this results shows about the rate of convergence in the *EM* estimate of the MLE as a function of how much information is added to X in order to make up the complete data Z.