First, we review 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.

Then, we review 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.

**Background facts for the Exponential Family:**
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 1st ed.)

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

\[
f(x \mid \theta) = b(x) \exp\left[\sum_{i=1}^{k} g_i(\theta) t_i(x)\right] / a(\theta)
\]

where \(a \geq 0\) and the \(t_i\) are real-valued functions of the data only; where \(b \geq 0\) and the \(g_i\) 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 \(T = (t_1(x), \ldots, t_k(x))\) are sufficient for \(\theta\).

*Defn.:* Call \(\Gamma = (g_1(\theta), \ldots, g_k(\theta))\), the \(k\)-dimensional natural parameter of the family, and \(T = (t_1(x), \ldots, t_k(x))\), the \(k\)-dimensional natural sufficient statistic of the family.

Moreover, the natural sufficient statistic \(T\) also has its distribution within the exponential family, using the same natural parameters.

Let \(X_j (j = 1, \ldots, n)\) be iid sample of size \(n\) from an exponential family. Define the \(k\)-many statistics \(T_i = \sum_j t_i(x_j)\). It follows that \((T_1, \ldots, T_k)\) are jointly sufficient and have a distribution from the exponential family, with the same natural parameters as the \(X_j\).

Let the observed data \(X = x\) come from a statistical model with density \(g(x \mid \theta)\). We want to find the MLE, \(\arg\max_{\theta} \log g(x \mid \theta) = L(\theta)\). 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 \(\theta\) and whose density, \(f(z \mid \theta)\), is described above.
**First.** Argue that \( E[T(z) \mid \theta] = \alpha'(\theta) \) and that \( E[T(z) \mid x, \theta] = \alpha'(\theta) + L(\theta) \).

*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, \( \log h(z \mid x, \theta) = T(z)\theta + \beta(z) - \alpha(\theta) - L(\theta) \), since

\[
\log f(z \mid \theta) = T(z)\theta + \beta(z) - \alpha(\theta)
\]

where \( \alpha(\theta) = \log a(\theta) \) and likewise \( \beta(z) = \log b(z) \)

Differentiate and take expectations.

Argue that \( E[\partial/\partial \theta \log f(z \mid \theta)] = E_x[\partial/\partial \theta \log h(z \mid x, \theta)] = 0 \).

Thus, \( L'(\theta) = E[T(z) \mid x, \theta] - E[T(z) \mid \theta] \)

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

**Second.** Solve for \( \theta_{j+1} \) which is the \( j+1 \)th EM estimate of the MLE.

*Hint:* Argue that \( \theta_{j+1} \) solves \( \alpha(\theta_{j+1}) = E[T(z) \mid x, \theta_j] = E[T(z) \mid \theta_{j+1}] \).

**Third.** Conclude that, because \( \delta(\theta) = E[T(z) \mid x, \theta] - E[T(z) \mid \theta] > 0 \) 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 \( \hat{\theta} \).

**Next,** for determining the rate of convergence in the sequence of EM estimates of the MLE, \( \hat{\theta} \), argue as follows:

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

**Fourth:** Show that \( I_z(\theta) = \alpha''(\theta) \) and that \( I_{z|x}(\theta) = \alpha''(\theta) + L''(\theta) \).

**Fifth:** Show that as \( j \to \infty \), the ratio \( (\theta_{j+1} - \hat{\theta}) / (\theta_j - \hat{\theta}) = I_{z|x}(\hat{\theta}) / I_z(\hat{\theta}) \).

*Hint:* Use these two linear approximations for \( \theta \) in the neighborhood of \( \hat{\theta} : \)

\[
E[T(z) \mid x, \theta] = E[T(z) \mid x, \hat{\theta}] + I_{z|x}(\theta - \hat{\theta})
\]

\[
E[T(z) \mid \theta] = E[T(z) \mid \hat{\theta}] + I_z(\theta - \hat{\theta}).
\]

This result shows that the rate of convergence in the EM estimate of the MLE is a function of how much information is added to \( X \) in order to make up the complete data \( Z \), i.e. the more that is added, the larger the ratio (above), and the slower the rate of convergence to the MLE.