EM in practice (continued)

- Memory problems ---> we can compute means, etc., as running totals so that we do not need to store responsibilities for all points over all clusters.
Analysis of EM

- Maximizing the Q function provided a new parameter estimate which increased the likelihood.
- Showing this typically uses Jensen’s inequality.
  - Bishop (§9.4), instead, uses the fact that the KL divergence between two distributions is non-negative, but showing this also uses Jensen’s.
- Given a bounded likelihood, this means the algorithm converges to a stationary point.
  - Typically a local maximum but examples where it is a saddle point can be constructed.

More generally, if $f$ is convex, then, for $\lambda_i \geq 0$, and $\sum \lambda_i = 1$ we have

$$f\left(\sum \lambda_i x_i\right) \leq \sum \lambda_i f(x_i)$$

(Jensen’s inequality)
Result from calculus (prove via mean value theorem)

If \( f \) is twice differentiable on \([a,b]\) and \( f'' \geq 0 \) on \([a,b]\), then \( f(x) \) is convex on \([a,b]\).

\[
f\left(\sum \lambda_i x_i\right) \leq \sum \lambda_i f(x_i) \quad \text{(Jensen's inequality)}
\]

\[
\log \left(\sum \lambda_i x_i\right) \geq \sum \lambda_i \log(x_i) \quad (-\log(x) \text{ is convex})
\]

Notice that \( f(x) = -\log(x) \) is convex

Proof?

\[
f'(x) = -\frac{1}{x}
\]

\[
f''(x) = \frac{1}{x^2}
\]

In EM, we seek \( \theta \) to maximize \( L(\theta) = \ln P(X|\theta) \)

Suppose at step \( n \) we have \( L(\theta_n) \)

\[
L(\theta) - L(\theta_n) = \ln \left( \sum_{z} \mathcal{P}(X|z, \theta) \mathcal{P}(z|\theta) \right) - \ln \mathcal{P}(X|\theta_n).
\]
\[ \begin{align*}
L(\theta) - L(\theta_n) &= \ln \left( \sum_x \mathcal{P}(X|z, \theta) \mathcal{P}(z|\theta) \right) - \ln \mathcal{P}(X|\theta_n) \\
&= \ln \left( \sum_x \mathcal{P}(X|z, \theta_n) \frac{\mathcal{P}(z|\theta, \theta_n)}{\mathcal{P}(X|\theta_n)} \right) - \ln \mathcal{P}(X|\theta_n) \\
&\geq \sum_x \mathcal{P}(z|X, \theta_n) \ln \frac{\mathcal{P}(X|z, \theta)}{\mathcal{P}(X|\theta_n)} - \ln \mathcal{P}(X|\theta_n) \\
&= \sum_x \mathcal{P}(z|X, \theta_n) \ln \frac{\mathcal{P}(X|z, \theta)}{\mathcal{P}(X|\theta_n)} \\
&\triangleq \Delta(\theta|\theta_n).
\end{align*} \]

Notice that this is our Q function

\[ \ln \mathcal{P}(X|\theta_n) = \sum_x \ln \mathcal{P}(X|\theta_n) \mathcal{P}(z|X, \theta_n) \]

because \( \mathcal{P}(X|\theta_n) \) does not depend on \( z \), and \( \sum_x \mathcal{P}(z|X, \theta_n) = 1 \)

Proof that \( \Delta(\theta_n|\theta_n) = 0 \) and thus \( I(\theta_n|\theta_n) = L(\theta_n) \)

\[ \begin{align*}
l(\theta_n|\theta_n) &= L(\theta_n) + \Delta(\theta_n|\theta_n) \\
&= L(\theta_n) + \sum_x \mathcal{P}(z|X, \theta_n) \ln \frac{\mathcal{P}(X|z, \theta_n) \mathcal{P}(z|\theta_n)}{\mathcal{P}(X|\theta_n)} \\
&= L(\theta_n) + \sum_x \mathcal{P}(z|X, \theta_n) \ln \frac{\mathcal{P}(X|z, \theta_n)}{\mathcal{P}(X|\theta_n)} \\
&= L(\theta_n) + \sum_x \mathcal{P}(z|X, \theta_n) \ln 1 \\
&= L(\theta_n).
\end{align*} \]

Figure 2: Graphical interpretation of a single iteration of the EM algorithm: The function \( l(\theta|\theta_n) \) is bounded above by the likelihood function \( L(\theta) \). The functions are equal at \( \theta = \theta_n \). The EM algorithm chooses \( \theta_{n+1} \) as the value of \( \theta \) for which \( l(\theta|\theta_n) \) is a maximum. Since \( L(\theta) \geq l(\theta|\theta_n) \), increasing \( l(\theta|\theta_n) \) ensures that the value of the likelihood function \( L(\theta) \) is increased at each step.
Proof that maximizing \( l(\theta|\theta_n) \) gives the same \( \theta_{n+1} \) as maximizing \( Q \)

\[
\theta_{n+1} = \arg \max_{\theta} \{ l(\theta|\theta_n) \} = \arg \max_{\theta} \left\{ L(\theta_n) + \sum_z \mathcal{P}(z|X, \theta_n) \ln \frac{\mathcal{P}(X|\theta) \mathcal{P}(z|\theta)}{\mathcal{P}(X|\theta_n) \mathcal{P}(z,X, \theta_n)} \right\}
\]

Now drop terms which are constant w.r.t. \( \theta \)

\[
\begin{align*}
&= \arg \max_{\theta} \left\{ \sum_z \mathcal{P}(z|X, \theta_n) \ln \frac{\mathcal{P}(X|z, \theta) \mathcal{P}(z|\theta)}{\mathcal{P}(X|\theta_n) \mathcal{P}(z,X, \theta_n)} \right\} \quad \text{(def'n conditional probability)} \\
&= \arg \max_{\theta} \left\{ \sum_z \mathcal{P}(z|X, \theta_n) \ln \frac{\mathcal{P}(X,z, \theta) \mathcal{P}(z|\theta)}{\mathcal{P}(z,X, \theta_n)} \right\} \\
&= \arg \max_{\theta} \left\{ \sum_z \mathcal{P}(z|X, \theta_n) \ln \mathcal{P}(X,z|\theta) \right\} \\
&= \arg \max_{\theta} \left\{ \mathcal{Q}(X,Z|\theta, \theta_n) \right\}
\end{align*}
\]

Algebra for “drop terms which are constant w.r.t. \( \theta \)”

\[
\sum_z \mathcal{P}(z|X, \theta_n) \ln \frac{\mathcal{P}(X|z, \theta) \mathcal{P}(z|\theta)}{\mathcal{P}(z|X, \theta_n) \mathcal{P}(z|\theta)} = \sum_z \mathcal{P}(z|X, \theta_n) \ln \mathcal{P}(X|z, \theta) \mathcal{P}(z|\theta) - \sum_z \mathcal{P}(z|X, \theta_n) \ln \mathcal{P}(z|X, \theta_n) \mathcal{P}(z|\theta)
\]

Those ones do not matter for argument.

Summary

Jensen's provides a lower bound for the likelihood, \( L(\theta) \) in terms of a current \( \theta_n \), namely \( l(\theta|\theta_n) \).

The max of \( l(\theta|\theta_n) \) is reached at the same \( \theta_{n+1} \) found by maximizing the Q function

\[
\begin{align*}
L(\theta_{n+1}) &= l(\theta_{n+1}|\theta_n) \\
&\geq l(\theta_n|\theta_n) = L(\theta_n) \\
&= L(\theta_n) \\
&\geq l(\theta_n|\theta_n) = L(\theta_n)
\end{align*}
\]

because we maximized

In other words, \( L(\theta) \) goes up (or stays the same) with the successive \( \theta \) found by EM.