COMS 4771 Spring 2015

# Expectation-Maximization

Daniel Hsu *Warning: notes are incomplete and unchecked for correctness*

## 1 Expectation-Maximization

Expectation-Maximization (E-M) is a general method for parameter estimation in statistical models where some of the data are unobserved (i.e., "hidden"). E-M iteratively improves the log-likelihood of a parameter given the observed data until a stationary point of the log-likelihood function is reached (in many cases, a local maximum).

### 1.1 Definitions

An E-M algorithm for a particular statistical model can be derived using a general recipe. Let $\mathcal{X} \times \mathcal{Y}$ be a sample space[1], and let the model $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be a family of distributions over $\mathcal{X} \times \mathcal{Y}$ indexed by the parameter space $\Theta$. We interpret $\mathcal{X}$ as the sample space for the observed data, and $\mathcal{Y}$ as the sample space for the hidden data.[2]

**Example 1** (Mixture of $K$ Poisson distributions)**.** The sample spaces are $\mathcal{X} = \mathbb{Z}_+ := \{0, 1, 2, \dots\} =$ and $\mathcal{Y} = [K] := \{1, 2, \dots, K\}$. The parameter space is $\Theta = \Delta^{K-1} \times \mathbb{R}^K_{++}$, where $\Delta^{K-1} := \{\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K) \in \mathbb{R}^K_+ : \sum_{j=1}^K \pi_j = 1\}$, $\mathbb{R}_+ := \{t \in \mathbb{R} : t \geq 0\}$, and $\mathbb{R}_{++} := \{t \in \mathbb{R} : t > 0\}$. Each distribution $P_{\boldsymbol{\theta}}$ in $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is as follows. If $(X, Y) \sim P_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} = (\boldsymbol{\pi}, \lambda_1, \lambda_2, \dots, \lambda_K)$, then

$$\begin{aligned} Y &\sim \text{Categorical}(\boldsymbol{\pi}), \\ X \,|\, Y = j &\sim \text{Pois}(\lambda_j), \quad j = 1, 2, \dots, K. \end{aligned}$$

In other words, $\Pr_{\boldsymbol{\theta}}(Y = j) = \pi_j$, and $\Pr_{\boldsymbol{\theta}}(X = k \,|\, Y = j) = \lambda_j^k e^{-\lambda_j}/k!$. The joint probability of $(x, y) \in \mathcal{X} \times \mathcal{Y}$ under $P_{\boldsymbol{\theta}}$ is

$$P_{\boldsymbol{\theta}}(x, y) = \prod_{j=1}^K \left( \pi_j \cdot \frac{\lambda_j^x e^{-\lambda_j}}{x!} \right)^{\mathbb{1}\{y=j\}} .$$

[1] For simplicity, we assume $\mathcal{X}$ and $\mathcal{Y}$ are discrete spaces so that summations suffice for marginalization, expectation, etc. For more general sample spaces, replace sums with integrals, and make any necessary measureability assumptions.

[2] If the observed data are an i.i.d. sample, then $\mathcal{X}$ will be a product space. In fact, for many models, the observed data and hidden data come as i.i.d. pairs, and so both $\mathcal{X}$ and $\mathcal{Y}$ are product spaces, and each $P_{\boldsymbol{\theta}}$ is a product distribution. See Example 1.

For a sample size $n \in \mathbb{N}$, we naturally extend the model for $n$ i.i.d. copies $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ of $(X, Y)$, so that $\mathcal{X} = \mathbb{Z}_+^n$, $\mathcal{Y} = [K]^n$. For $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathcal{Y}$,

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} \prod_{j=1}^{K} \left( \pi_j \cdot \frac{\lambda_j^{x_i} e^{-\lambda_j}}{x_i!} \right)^{\mathbb{1}\{y_i = j\}} .$$

□

(Extensions of a probability model to an i.i.d. sample—as in Example 1—are so commonplace that the formalities done there are usually omitted. It is usually straightforward to reason about the implied independence properties of an i.i.d. sample.)

Let $x$ be the observed data. For any $\boldsymbol{\theta} \in \Theta$, the marginal probability of $x$ under $P_{\boldsymbol{\theta}}$ is

$$P_{\boldsymbol{\theta}}(x) = \sum_{y \in \mathcal{Y}} P_{\boldsymbol{\theta}}(x, y).$$

Throughout, we shall assume that $P_{\boldsymbol{\theta}}(x) > 0$. The log-likelihood of the parameter $\boldsymbol{\theta} \in \Theta$ given the observed data is

$$\mathcal{L}(\boldsymbol{\theta}) := \ln P_{\boldsymbol{\theta}}(x) = \ln \left( \sum_{y \in \mathcal{Y}} P_{\boldsymbol{\theta}}(x, y) \right).$$

## 1.2 Deriving an E-M algorithm

The idea behind E-M is to construct a function that (i) is a lower-bound on the log-likelihood function and (ii) is tight at the current parameter; and then to update the parameter to be the maximizer of this lower-bound function. See Figure 1 for an illustration of this process. Iteratively constructing such a lower-bound function and updating to its maximizer guarantees a sequence of parameters whose log-likelihood values are non-decreasing. This iterative process begins at some initial parameter that is furnished by the user; the limit point of the process depends entirely on this initial parameter.

Suppose we have some current parameter $\boldsymbol{\theta}^{(t)} \in \Theta$ in hand, which we wish to update using E-M to $\boldsymbol{\theta}^{(t+1)}$. The lower-bound function is constructed using the conditional distribution (under $P_{\boldsymbol{\theta}^{(t)}}$) over $\mathcal{Y}$ given the observed data $x$. These conditional probabilities,

$$q^{(t)}(y) := \Pr_{\boldsymbol{\theta}^{(t)}}(Y = y \,|\, X = x), \quad y \in \mathcal{Y},$$

can be computed using just the observed data $x$ and the current parameter $\boldsymbol{\theta}^{(t)}$. Note that

$$q^{(t)}(y) = \frac{P_{\boldsymbol{\theta}^{(t)}}(x,y)}{P_{\boldsymbol{\theta}^{(t)}}(x)} = \frac{P_{\boldsymbol{\theta}^{(t)}}(x,y)}{\sum_{y' \in \mathcal{Y}} P_{\boldsymbol{\theta}^{(t)}}(x,y')}, \quad y \in \mathcal{Y}.$$

The lower-bound function, based on $q^{(t)}$, is given by

$$\mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}) := \sum_{y \in \mathcal{Y}} q^{(t)}(y) \cdot \ln\left(\frac{P_{\boldsymbol{\theta}}(x,y)}{q^{(t)}(y)}\right), \quad \boldsymbol{\theta} \in \Theta.$$

This is indeed a lower-bound on the log-likelihood, since

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \ln\left(\sum_{y \in \mathcal{Y}} P_{\boldsymbol{\theta}}(x,y)\right) \\
&= \ln\left(\sum_{y \in \mathcal{Y}} q^{(t)}(y) \cdot \frac{P_{\boldsymbol{\theta}}(x,y)}{q^{(t)}(y)}\right) \\
&\geq \sum_{y \in \mathcal{Y}} q^{(t)}(y) \cdot \ln\left(\frac{P_{\boldsymbol{\theta}}(x,y)}{q^{(t)}(y)}\right) \qquad (1) \\
&= \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}),
\end{aligned}$$

where the inequality in (1) follows from the concavity of the logarithm function and Jensen's inequality. Moreover, the lower-bound function is tight at $\boldsymbol{\theta}^{(t)}$, since

$$\begin{aligned}
\mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}^{(t)}) &= \sum_{y \in \mathcal{Y}} q^{(t)}(y) \cdot \ln\left(\frac{P_{\boldsymbol{\theta}^{(t)}}(x,y)}{q^{(t)}(y)}\right) \\
&= \sum_{y \in \mathcal{Y}} q^{(t)}(y) \cdot \ln\left(\frac{P_{\boldsymbol{\theta}^{(t)}}(x,y)}{P_{\boldsymbol{\theta}^{(t)}}(x,y)/P_{\boldsymbol{\theta}^{(t)}}(x)}\right) \\
&= \sum_{y \in \mathcal{Y}} q^{(t)}(y) \cdot \ln P_{\boldsymbol{\theta}^{(t)}}(x) \\
&= \mathcal{L}(\boldsymbol{\theta}^{(t)}).
\end{aligned}$$

Therefore, updating $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)} := \arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta})$ ensures that

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \geq \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}^{(t+1)}) \geq \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t)}).$$

The computation of the $q^{(t)}(y)$ is called the "E step", and the maximization of $\mathcal{L}_{\mathrm{LB}}^{(t)}$ is called the "M step".

**Example 2** (E-M for mixture of $K$ Poisson distributions)**.** We consider the mixture of $K$ Poisson distributions on an i.i.d. sample of size $n$ (Example 1). Let $\boldsymbol{x} \in \mathbb{Z}_+^n$ be the observed data. Given some current parameter $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\pi}^{(t)}, \lambda_1^{(t)}, \lambda_2^{(t)}, \dots, \lambda_K^{(t)})$, the conditional probability of $\boldsymbol{y} \in [K]^n$ given $\boldsymbol{x}$ is

$$\begin{aligned} q^{(t)}(\boldsymbol{y}) &= \Pr_{\boldsymbol{\theta}^{(t)}}(Y_1 = y_1, \dots, Y_n = y_n \mid X_1 = x_1, \dots, X_n = x_n) \\ &= \prod_{i=1}^n \Pr_{\boldsymbol{\theta}^{(t)}}(Y_i = y_i \mid X_i = x_i) \\ &= \prod_{i=1}^n q_i^{(t)}(y_i), \end{aligned}$$

where

$$q_i^{(t)}(j) := \frac{\pi_j^{(t)} \cdot \dfrac{(\lambda_j^{(t)})^{x_i} e^{-\lambda_j^{(t)}}}{x_i!}}{\displaystyle\sum_{j'=1}^K \pi_{j'}^{(t)} \cdot \dfrac{(\lambda_{j'}^{(t)})^{x_i} e^{-\lambda_{j'}^{(t)}}}{x_i!}}, \quad j \in [K].$$

Computing these $q_i^{(t)}(j)$ for all $i \in [n]$ and $j \in [K]$ is the “E step”.

To simplify notation, *define* a new random vector $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_n)$ taking values in $[K]^n$ with $\Pr(\boldsymbol{Z} = \boldsymbol{z}) = q^{(t)}(\boldsymbol{z})$. Observe that $\mathbb{E}[\mathbb{1}\{Z_i = j\}] = q_i^{(t)}(j)$. The lower-bound function $\mathcal{L}_{\mathrm{LB}}^{(t)}$ is as follows: for any $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned} \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}) &= \mathbb{E}\left[\ln\left(\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Z})}{q^{(t)}(\boldsymbol{Z})}\right)\right] \\ &= \mathbb{E}\left[\ln\left(\prod_{i=1}^n \prod_{j=1}^K \left(\pi_j \cdot \frac{\lambda_j^{x_i} e^{-\lambda_j}}{x_i!}\right)^{\mathbb{1}\{Z_i = j\}}\right)\right] + \text{const} \\ &= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^K \mathbb{1}\{Z_i = j\}(\ln \pi_j + x_i \ln \lambda_j - \lambda_j)\right] + \text{const} \\ &= \sum_{i=1}^n \sum_{j=1}^K q_i^{(t)}(j)(\ln \pi_j + x_i \ln \lambda_j - \lambda_j) + \text{const}. \end{aligned}$$

This is easily maximized with respect to $\boldsymbol{\theta}$: the maximizer is given by

$$\pi_j^{(t+1)} \propto \sum_{i=1}^n q_i^{(t)}(j), \qquad \lambda_j^{(t+1)} = \frac{\sum_{i=1}^n q_i^{(t)}(j) x_i}{\sum_{i=1}^n q_i^{(t)}(j)}, \quad j = 1, 2, \dots, K.$$

These parameter updates constitute the “M step”. □

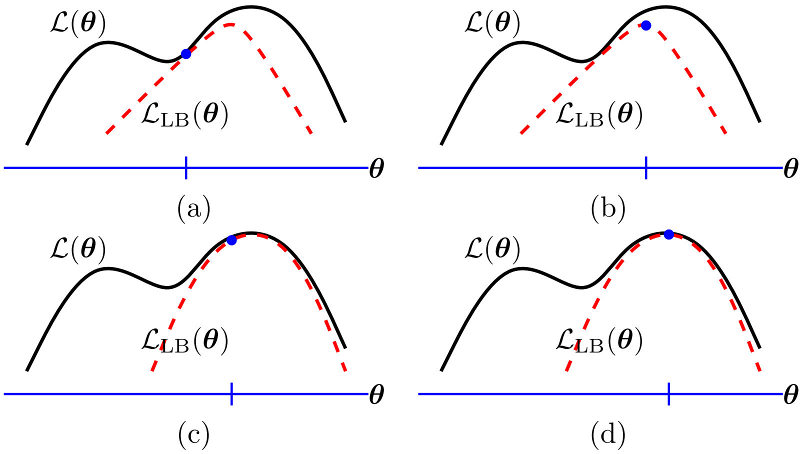
$\mathcal{L}(\boldsymbol{\theta})$
$\mathcal{L}_{\mathrm{LB}}(\boldsymbol{\theta})$
$\boldsymbol{\theta}$
(a)
$\mathcal{L}(\boldsymbol{\theta})$
$\mathcal{L}_{\mathrm{LB}}(\boldsymbol{\theta})$
$\boldsymbol{\theta}$
(b)
$\mathcal{L}(\boldsymbol{\theta})$
$\mathcal{L}_{\mathrm{LB}}(\boldsymbol{\theta})$
$\boldsymbol{\theta}$
(c)
$\mathcal{L}(\boldsymbol{\theta})$
$\mathcal{L}_{\mathrm{LB}}(\boldsymbol{\theta})$
$\boldsymbol{\theta}$
(d)

Figure 1: Cartoon illustrating two rounds of the E-M algorithm. The horizontal axis represents the parameter space $\Theta$. (a) Construction of the lower-bound function $\mathcal{L}_{\mathrm{LB}}$ that is tight at the current parameter (located at the blue tick mark; the blue dot shows the log-likelihood value), but otherwise lower-bounds the log-likelihood function $\mathcal{L}$. (b) The current parameter is updated to be the maximizer of the lower-bound function $\mathcal{L}_{\mathrm{LB}}$. (c) A new lower-bound function is constructed, again tight at the current parameter but otherwise lower-bounds the log-likelihood. (d) The current parameter is updated again to maximize the lower-bound function. At this point, it appears to be very close to a local maximizer of the log-likelihood function.

### 1.3 Further analysis

For any $\boldsymbol{\theta} \in \Theta$, we write, for any $x \in \mathcal{X}$ with $P_{\boldsymbol{\theta}}(x) > 0$,

$$P_{\boldsymbol{\theta}}(y|x) := \frac{P_{\boldsymbol{\theta}}(x,y)}{P_{\boldsymbol{\theta}}(x)} = \Pr_{\boldsymbol{\theta}}(Y = y \mid X = x), \quad y \in \mathcal{Y}.$$

We can write the gap between the log-likelihood $\mathcal{L}$ and the lower-bound function $\mathcal{L}_{\mathrm{LB}}^{(t)}$ as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta}) &= \ln P_{\boldsymbol{\theta}}(x) - \sum_{y \in \mathcal{Y}} q^{(t)}(y) \cdot \ln\left(\frac{P_{\boldsymbol{\theta}}(x,y)}{q^{(t)}(y)}\right) \\
&= \ln P_{\boldsymbol{\theta}}(x) - \sum_{y \in \mathcal{Y}} q^{(t)}(y) \cdot \ln\left(\frac{P_{\boldsymbol{\theta}}(y|x) P_{\boldsymbol{\theta}}(x)}{q^{(t)}(y)}\right) \\
&= \sum_{y \in \mathcal{Y}} q^{(t)}(y) \cdot \ln\left(\frac{q^{(t)}(y)}{P_{\boldsymbol{\theta}}(y|x)}\right) \\
&= \mathrm{RE}\Big(q^{(t)} \| P_{\boldsymbol{\theta}}(\cdot|x)\Big) \geq 0
\end{aligned}$$

with equality if and only if $q^{(t)} = P_{\boldsymbol{\theta}}(\cdot|x)$; equality holds (i.e., $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{\mathrm{LB}}^{(t)}(\boldsymbol{\theta})$) when $\theta = \theta^{(t)}$ since $q^{(t)} = P_{\boldsymbol{\theta}^{(t)}}(\cdot|x)$.