The Expectation-Maximization and Alternating Minimization Algorithms

Shane M. Haas

Research Laboratory of Electronics, and

Laboratory for Information and Decision Systems

Massachusetts Institute of Technology

The EM Algorithm

- The ML Problem
- The EM Solution
- Mixture Example
- Alternating Minimization Algorithms
- Conclusions

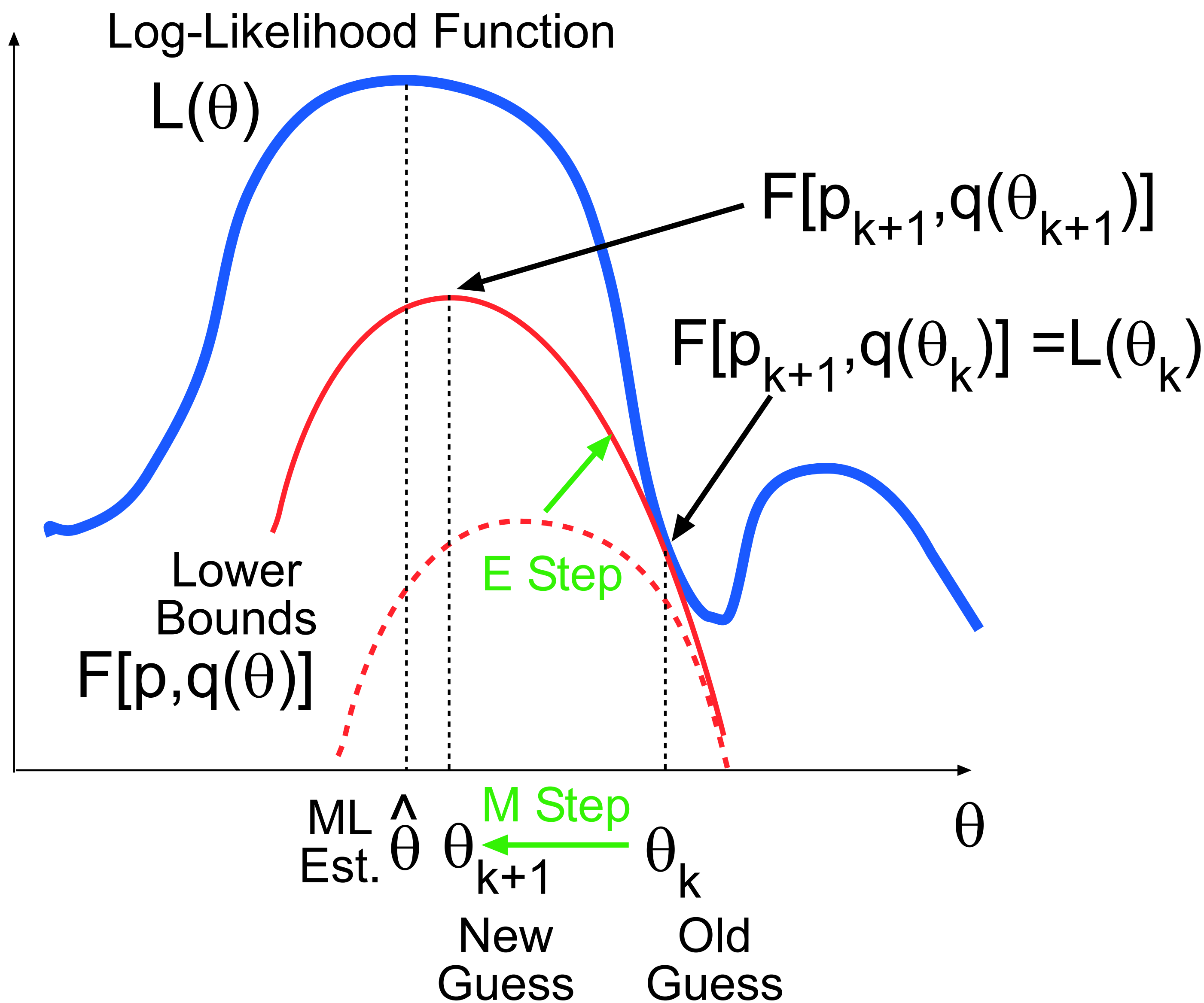
The ML Problem

- The maximum likelihood problem:

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \; \log f\,(\,y \mid \theta\,).$$

- Local solutions:
  - Gradient ascent
  - Newton-Rhapson method
  - Expectation maximization

The EM Algorithm Picture



The EM Algorithm

- Divide and conquer!

$$\textbf{E Step: } p_{k+1} = \underset{p}{\text{argmax}}\, F\,[\,p,\; q(\theta_k)\,],$$

$$\textbf{M Step: } \theta_{k+1} = \underset{\theta}{\text{argmax}}\, F\,[\,p_{k+1},\; q(\theta)\,].$$

Log-Likelihood Lower Bounds

- Introduce unobserved var. x with joint density $q\,(\,x, y \mid \theta\,)$
- Apply Jensen's inequality for arbitrary $p\,(\,x\,)$:

$$
\begin{aligned}
L(\theta) &\equiv \log f\,(\,y \mid \theta\,) \\
&= \log \int q\,(\,x, y \mid \theta\,) dx \\
&= \log \int p\,(\,x\,) \frac{q\,(\,x, y \mid \theta\,)}{p\,(\,x\,)} dx \\
&\geq \int p\,(\,x\,) \log \left[\frac{q\,(\,x, y \mid \theta\,)}{p\,(\,x\,)} \right] dx \\
&\equiv F\,[\,p,\, q(\theta)\,].
\end{aligned}
$$

Relationship to Free Energy

- Energy $\equiv -\log q\,(\,x, y \mid \theta\,)$ for given y
- Free Energy = Average Energy - Entropy
- The lower bound is

$$-F\,[\,p,\; q(\theta)\,] = \underbrace{-E\log q\,(\,x, y \mid \theta\,)}_{\text{Avg. Energy}} - \underbrace{[-E\log p\,(\,x\,)]}_{\text{Entropy}},$$

- **Interpretation:** E step minimizes free energy

Relationship to KL Divergence

- The lower bound is also

$$F\,[\,p,\; q(\theta)\,] = -D[\; p \parallel q(\theta)\;],$$

where

$$D[\; p \parallel q(\theta)\;] \equiv \int p\,(\,x\,) \log \left[\frac{p\,(\,x\,)}{q\,(\,x, y \mid \theta\,)} \right] dx.$$

- As defined, divergence can be negative
- **Interpretation:** E step minimizes KL divergence

The E Step

- The lower bound is also

$$
\begin{aligned}
F[p,\ q(\theta)] &= \int p(x) \log \left[\frac{w(x \mid y, \theta) f(y \mid \theta)}{p(x)} \right] dx \\
&= \log f(y \mid \theta) - D[\, p \,\|\, w(\theta)\,].
\end{aligned}
$$

- Maximize for a fixed θ_k

E Step: $p_{k+1} = w(x \mid y, \theta_k) = \dfrac{h(y \mid x, \theta_k)\pi(x \mid \theta_k)}{\int h(y \mid \chi, \theta_k)\pi(\chi \mid \theta_k) d\chi}$

where

$$h(y \mid x, \theta) = q(x, y \mid \theta)\pi(x \mid \theta)$$

$$\pi(x \mid \theta) = \int q(x, y \mid \theta)\, dy$$

The M Step

- Evaluating the lower bound at p_{k+1}

$$
\begin{aligned}
F\left[\, p_{k+1},\ q(\theta)\,\right] &= \int w\left(\, x \mid y, \theta_k \,\right) \log\left[\frac{q\left(\, x, y \mid \theta \,\right)}{w\left(\, x \mid y, \theta_k \,\right)}\right] dx \\
&= \int w\left(\, x \mid y, \theta_k \,\right) \log q\left(\, x, y \mid \theta \,\right)\, dx \\
&\quad - \int w\left(\, x \mid y, \theta_k \,\right) \log w\left(\, x \mid y, \theta_k \,\right)\, dx.
\end{aligned}
$$

- Second term does not depend on θ:

M Step: $\theta_{k+1} = \underset{\theta}{\text{argmax}} \int w\left(\, x \mid y, \theta_k \,\right) \log q\left(\, x, y \mid \theta \,\right)\, dx.$

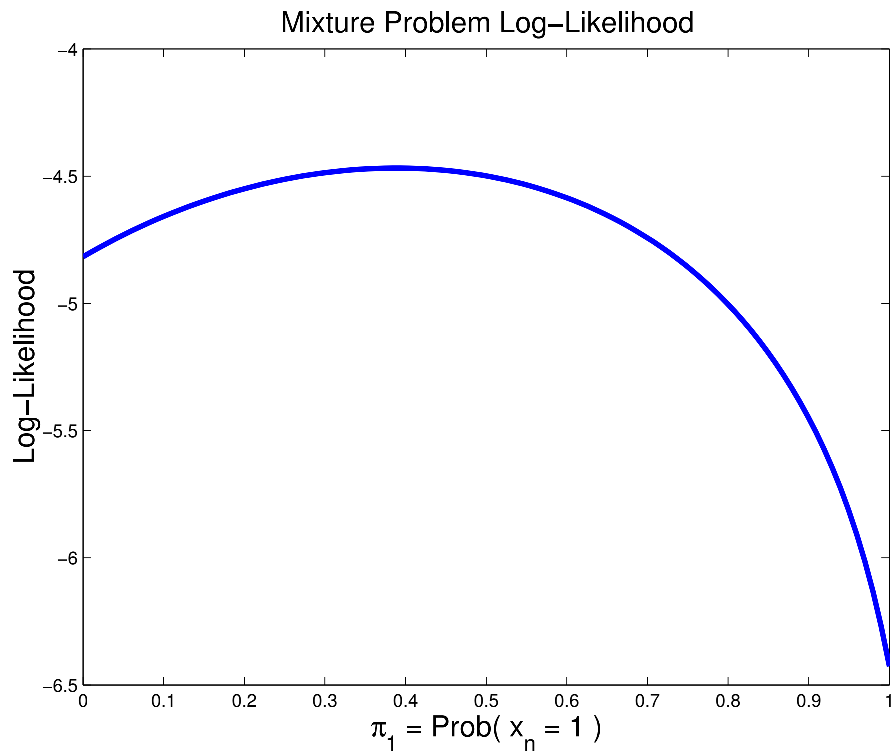
Example: A Mixture Problem

- Observe: $y = \{y_1, \dots, y_N\}$
- Each sample generated independently:
  - Select a group g, $1 \le g \le G$ with unknown prob. π_g
  - Generate y_n according to known $h_g(y_n)$
- Unknown parameters:

$$\theta = \{\pi_1, \dots, \pi_{G-1}\}$$

$$\pi_G = 1 - \sum_{g=1}^{G-1} \pi_g$$

Log-Likelihood



$$L(\theta) = \sum_{n=1}^{N} \log \left\{ \sum_{g=1}^{G} h_g(y_n)\pi_g \right\}$$

Hidden Variable

- Hidden variable: $x_n = g$ if y_n generated from group g
- $x = \{x_1, \ldots, x_N\}$
- Complete log-likelihood:

$$q\,(\,x, y \mid \theta\,) = \prod_{n=1}^{N} h_{x_n}(y_n)\pi_{x_n}$$

The E Step

- Maximizing the lower bound for a fixed $\theta = \theta_k$

$$\textbf{E Step: } p_{k+1} = w\,(\,x \mid y, \theta_k\,) = \prod_{n=1}^{N} \Pr\{\,x_n \mid y_n, \theta_k\},$$

- Baye's rule:

$$\Pr\{\,x_n = g \mid y_n, \theta_k\} = \frac{h_g(y_n)\pi_g^k}{\sum_{\gamma=1}^{G} h_\gamma(y_n)\pi_\gamma^k} \equiv m_n^k(g)$$

- π_g^k is estimate of g-th group probability at iteration k

Preparing for the M Step

- Maximize over θ for exp. w.r.t. $p_{k+1}(x) = w\,(\,x \mid y, \theta_k\,)$:

$$E[\,\log q\,(\,x, y \mid \theta\,)\,] = \sum_{n=1}^{N} (E[\,\log h_{x_n}(y_n)] + E[\,\log \pi_{x_n}])\,,$$

- Second term:

$$\begin{aligned} E[\,\log \pi_{x_n}] &= \sum_{x} w\,(\,x \mid y, \theta_k\,) \log \pi_{x_n} \\ &= \sum_{g=1}^{G} \sum_{x : x_n = g} w\,(\,x \mid y, \theta_k\,) \log \pi_g \\ &= \sum_{g=1}^{G} m_n^k(g) \log \pi_g \end{aligned}$$

The M Step

- Differentiating w.r.t. π_g, $1 \leq g \leq G-1$:

$$\frac{\partial E[\log q(x, y \mid \theta)]}{\partial \pi_g} = \sum_{n=1}^{N} \frac{m_n^k(g)}{\pi_g^{k+1}} - \sum_{n=1}^{N} \frac{m_n^k(G)}{\pi_G^{k+1}} = 0$$

- Solving for π_g:

M Step: $\pi_g^{k+1} = \dfrac{\sum_{n=1}^{N} m_n^k(g)}{\sum_{n=1}^{N} \sum_{g=1}^{G} m_n^k(g)} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} m_n^k(g)$

$=$ (Avg. # of times $x_n = g$ given y and $\theta_k)/N$

Alternating Minimization Algorithms

- Arbitrary sets: $\mathcal{P}$ and $\mathcal{Q}$
- "Distance" measure: $d : \mathcal{P} \times \mathcal{Q} \to \mathcal{R}$
- Alternating minimization seq. $\{P_k\}_{k=0}^{\infty}$ and $\{Q_k\}_{k=0}^{\infty}$:

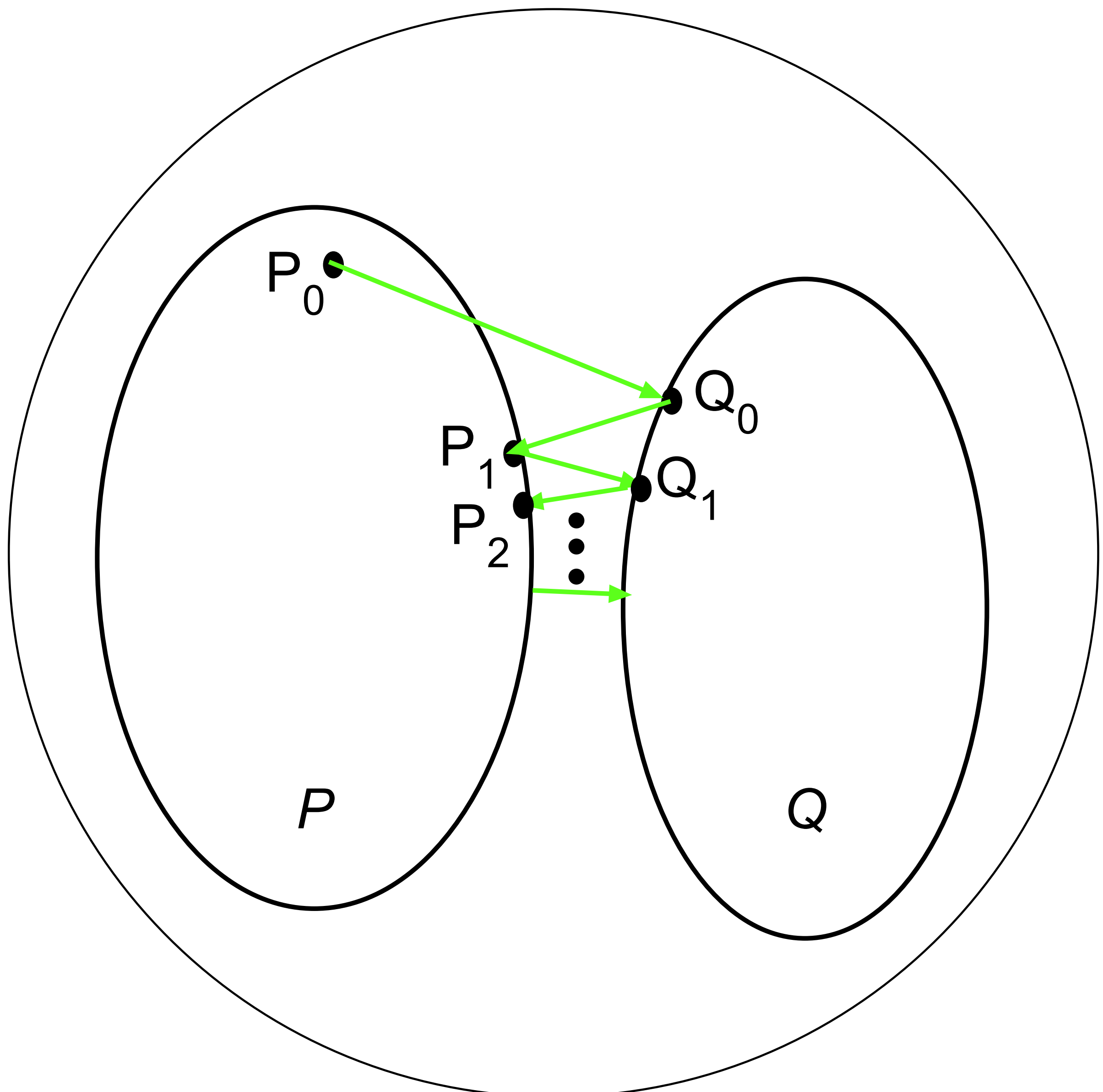
$$P_{k+1} = \underset{P \in \mathcal{P}}{\text{argmin}} \; d(P, Q_k)$$

$$Q_{k+1} = \underset{Q \in \mathcal{Q}}{\text{argmin}} \; d(P_{k+1}, Q)$$

  with starting point P_0 arbitrary

- Notation: $P_0 \to Q_0 \to P_1 \to Q_1 \to \cdots$

Example: Projection on Convex Sets



Convergence: Csiszar and Tusnady

- **[Th. 1]** If $P_0 \to Q_0 \to P_1 \to Q_1 \to \cdots$ such that for every $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$

  $$d(P,Q) + d(P,Q_{k-1}) \geq d(P,Q_k) + d(P_k,Q_k)$$

  then $\lim_{k\to\infty} d(P_k,Q_k) = \inf_{P\in\mathcal{P},Q\in\mathcal{Q}} d(P,Q)$

- **[Th. 3]** If $\mathcal{P}$ and $\mathcal{Q}$ are convex sets of measures and $d(P,Q) = D(P\|Q)$, then divergences from alternating minimization sequences converge.

EM as an Alt. Min. Algorithm

- Define:

$$d(P, Q) = D[\,P \| Q\,]$$

$$\mathcal{P} = \left\{ \int_{-\infty}^{x} p(\chi)\, d\chi \right\}$$

$$\mathcal{Q} = \left\{ \int_{-\infty}^{x} q(\,\chi, y \mid \theta\,)\, d\chi \right\}$$

Note: $\mathcal{Q}$ not necessarily convex

- EM Algorithm:

E Step: $P_{k+1} = \underset{P \in \mathcal{P}}{\text{argmin}}\, D[\,P \| Q(\theta_k)\,],$

M Step: $Q(\theta_{k+1}) = \underset{\theta \in \Theta}{\text{argmin}}\, D[\,P_{k+1} \| Q(\theta)\,].$

Conclusions

- The EM Algorithm
- Mixture Example
- Alternating Minimization Algorithms