

# Lecture 30: EM Theory; EM for Gaussian Mixtures

Charles A. Bouman

Purdue University

Wednesday, November 4, 2026

# Outline

- 1 First insight:  $\log p(y|\theta) = Q(\theta; \theta') + H(\theta; \theta')$
- 2 Second insight:  $H(\theta; \theta') \geq H(\theta'; \theta')$ , by Jensen's inequality
- 3 The key bound: the change in log likelihood is at least the change in  $Q$
- 4  $Q$  is a surrogate function for maximizing  $\log p(y|\theta)$
- 5 The EM algorithm: the E-step and the M-step
- 6 Properties of the  $Q$  and  $H$  functions
- 7 EM for Gaussian mixtures: the  $Q$  function, the posterior  $f(m|y_n, \theta')$ , and the update equations

## Core idea

The EM algorithm rests on **two nonobvious mathematical insights**: the log likelihood separates into  $Q + H$ , and  $H$  takes on its minimum value when  $\theta = \theta'$ . Together they make  $Q(\theta; \theta')$  a surrogate function for maximizing  $\log p(y|\theta)$ .

# Today's Big Ideas

- 1 The log likelihood separates into the sum of two functions, called  $Q$  and  $H$ , that are both conditional expectations given the observed data at a previous value of the parameter.
- 2 The function  $H$  takes on its minimum value when the new parameter equals the previous one, which follows from Jensen's inequality.
- 3 These two results together show that the change in the log likelihood is at least the change in  $Q$ , so  $Q$  is a surrogate function for maximizing the log likelihood.
- 4 The EM algorithm computes  $Q$  in the E-step and maximizes it in the M-step, so the log likelihood never decreases from one iteration to the next.
- 5 For a Gaussian mixture, the E-step reduces to computing the posterior probability of each class by Bayes' rule, and the M-step is a weighted form of the usual maximum likelihood estimates of the mean, the variance, and the class probability.

# First insight: the log likelihood separates into two parts

The log likelihood can be expressed as

$$\log p(y|\theta) = Q(\theta; \theta') + H(\theta; \theta') ,$$

where  $\theta'$  is any value of the parameter and the functions  $Q$  and  $H$  are defined as

$$Q(\theta; \theta') \triangleq \mathbb{E}[\log p(y, X|\theta) | Y = y, \theta'] ,$$

$$H(\theta; \theta') \triangleq -\mathbb{E}[\log p(X|y, \theta) | Y = y, \theta'] ,$$

and  $p(y, x|\theta)$  is assumed to be a strictly positive density function.

- Both expectations are conditioned on  $Y = y$  and taken with the parameter set to  $\theta'$ .
- The parameter inside the densities is  $\theta$ ; the parameter of the conditional distribution is  $\theta'$ .

# Two identities from conditional probability

In order to prove that the decomposition holds, we first observe the following two equalities:

$$p(y|\theta) = \frac{p(y, x|\theta)}{p(x|y, \theta)} = \frac{p(y, X|\theta)}{p(X|y, \theta)} .$$

- The first equality is simply a result of conditional probability that must hold for all values of  $x$ .
- Since this equality holds for all  $x$ , it must in particular hold if we substitute in the random variable  $X$ .

# Proof of the decomposition

Using these identities, the following sequence of equalities proves the desired result:

$$\begin{aligned}\log p(y|\theta) &= \mathbb{E}[\log p(y|\theta) | Y = y, \theta'] \\ &= \mathbb{E}\left[\log \left\{ \frac{p(y, X|\theta)}{p(X|y, \theta)} \right\} \middle| Y = y, \theta'\right] \\ &= \mathbb{E}[\log p(y, X|\theta) | Y = y, \theta'] - \mathbb{E}[\log p(X|y, \theta) | Y = y, \theta'] \\ &= Q(\theta; \theta') + H(\theta; \theta') .\end{aligned}$$

- Line 1:  $\log p(y|\theta)$  is not a function of  $X$ , so it equals its own conditional expectation.
- Line 2: substitute the identity from the previous slide.
- Line 3: the log of the ratio is the difference of the logs.

## Second insight: $H$ is minimized at $\theta = \theta'$

The function  $H(\theta; \theta')$  takes on its minimum value when  $\theta = \theta'$ . More precisely, for all  $\theta, \theta' \in \Omega$ , we have that

$$H(\theta; \theta') \geq H(\theta'; \theta') .$$

### Jensen's inequality

A concave function of the expectation is greater than or equal to the expectation of the concave function.

The proof on the next slide applies this inequality with the concave function  $\log(\cdot)$ .

# Proof that $H(\theta; \theta') \geq H(\theta'; \theta')$

To see that this is true, we have the following set of inequalities:

$$\begin{aligned} 0 &= \log \left\{ \int p(x|y, \theta) dx \right\} \\ &= \log \left\{ \int \frac{p(x|y, \theta)}{p(x|y, \theta')} p(x|y, \theta') dx \right\} \\ &\geq \int \log \left\{ \frac{p(x|y, \theta)}{p(x|y, \theta')} \right\} p(x|y, \theta') dx \\ &= \int \log p(x|y, \theta) p(x|y, \theta') dx - \int \log p(x|y, \theta') p(x|y, \theta') dx \\ &= -H(\theta; \theta') + H(\theta'; \theta') , \end{aligned}$$

where the third line depends on the famous **Jensen's inequality**.

# The key bound: likelihood change $\geq Q$ change

The two results now admit the foundation of the EM algorithm.

If we adjust the parameter from an initial value of  $\theta'$  to a new value of  $\theta$ , then the change in the log likelihood is lower bounded by the change in the  $Q$  function. This is formally stated as

$$\log p(y|\theta) - \log p(y|\theta') \geq Q(\theta; \theta') - Q(\theta'; \theta') .$$

- The left side is the change in the log likelihood.
- The right side is the change in the  $Q$  function.

# Proof of the key bound

The proof of this key result is then quite simple:

$$\begin{aligned}\log p(y|\theta) - \log p(y|\theta') &= Q(\theta; \theta') + H(\theta; \theta') - [Q(\theta'; \theta') + H(\theta'; \theta')] \\ &= Q(\theta; \theta') - Q(\theta'; \theta') + H(\theta; \theta') - H(\theta'; \theta') \\ &\geq Q(\theta; \theta') - Q(\theta'; \theta') .\end{aligned}$$

- Line 1 applies the decomposition twice, at  $\theta$  and at  $\theta'$ .
- Line 3 drops  $H(\theta; \theta') - H(\theta'; \theta')$ , which is nonnegative by the second insight.

## $Q$ is a surrogate function

By rearranging the terms of the key bound, we obtain the defining property of a surrogate function, as in equation (8.7) of Chapter 8 (Lecture 19):

$$\log p(y|\theta) \geq Q(\theta; \theta') - Q(\theta'; \theta') + \log p(y|\theta') .$$

So we see that the function  $Q(\theta; \theta')$  is a surrogate function for maximizing  $\log p(y|\theta)$ .

As with any surrogate function, if  $Q$  is optimized, then so is the likelihood, so we have

$$Q(\theta; \theta') > Q(\theta'; \theta') \Rightarrow p(y|\theta) > p(y|\theta') .$$

# The EM algorithm

From this result, the central concept of the EM algorithm becomes clear. Each iteration starts with an initial parameter,  $\theta'$ . Our objective is then to find a new parameter,  $\theta$ , so that  $Q(\theta; \theta') > Q(\theta'; \theta')$ . We then know that this new parameter is guaranteed to produce a larger value of the likelihood.

With this understanding in mind, we can now state the two-step recursion that defines the EM algorithm:

$$\text{E-step: } Q(\theta; \theta^{(k)}) = \mathbb{E} \left[ \log p(y, X | \theta) | Y = y, \theta^{(k)} \right] ;$$

$$\text{M-step: } \theta^{(k+1)} = \arg \max_{\theta \in \Omega} Q(\theta; \theta^{(k)}) .$$

Since each new value of the parameter  $\theta^{(k)}$  is selected to maximize the  $Q$  function, we know that this iteration will produce a monotone increasing sequence of log likelihood values,

$$\log p(y | \theta^{(k+1)}) \geq \log p(y | \theta^{(k)}) .$$

# Properties: decomposition of the likelihood

## Property: Decomposition of likelihood

Let  $X$  and  $Y$  be random vectors with joint density function  $p(x, y|\theta)$  for  $\theta \in \Omega$ . Then we have that

$$\log p(y|\theta) = Q(\theta; \theta') + H(\theta; \theta') ,$$

where  $Q(\theta; \theta')$  and  $H(\theta; \theta')$  are given by the definitions above.

## Property: Properties of $Q$ and $H$ functions

Let  $X$  and  $Y$  be random vectors with joint density function  $p(x, y|\theta)$  for  $\theta \in \Omega$ , and let the functions  $Q$  and  $H$  be given by the definitions above. Then

- $Q(\theta; \theta')$  is a surrogate function for the maximization of  $\log p(y|\theta)$ , and
- $H(\theta'; \theta') = \min_{\theta \in \Omega} H(\theta; \theta')$ .

# EM for Gaussian mixtures

Now that we have the basic tools of the EM algorithm, we can use them to estimate the parameters of the Gaussian mixture distribution.

In order to calculate the  $Q$  function for this problem, we will need a more suitable expression for the joint distribution of  $Y_n$  and  $X_n$ .

Recall that the Gaussian mixture is parameterized by

$$\theta = [\pi_0, \mu_0, \sigma_0^2, \dots, \pi_{M-1}, \mu_{M-1}, \sigma_{M-1}^2] .$$

# Rewriting the complete data log likelihood

It is easily verified that the equalities

$$\begin{aligned}\log p(y_n, x_n | \theta) &= \log \{p(y_n | \mu_{x_n}, \sigma_{x_n}) \pi_{x_n}\} \\ &= \sum_{m=0}^{M-1} \delta(x_n - m) \{ \log p(y_n | \mu_m, \sigma_m) + \log \pi_m \}\end{aligned}$$

hold, where  $p(y_n | \mu_m, \sigma_m)$  denotes a Gaussian density with mean  $\mu_m$  and standard deviation  $\sigma_m$ .

Notice that the delta function in the sum is only one when  $x_n = m$ ; otherwise it is zero.

Since the values of  $Y_n$  and  $X_n$  are assumed independent for different  $n$ , we know that

$$\begin{aligned}\log p(y, x | \theta) &= \sum_{n=1}^N \log p(y_n, x_n | \theta) \\ &= \sum_{n=1}^N \sum_{m=0}^{M-1} \delta(x_n - m) \{ \log p(y_n | \mu_m, \sigma_m) + \log \pi_m \} .\end{aligned}$$

# Calculating the $Q$ function

Using this expression, we can now calculate the  $Q$  function:

$$\begin{aligned} Q(\theta; \theta') &= \mathbb{E}[\log p(y, X|\theta) | Y = y, \theta'] \\ &= \mathbb{E}\left[\sum_{n=1}^N \sum_{m=0}^{M-1} \delta(X_n - m) \{\log p(y_n | \mu_m, \sigma_m) + \log \pi_m\} \middle| Y = y, \theta'\right] \\ &= \sum_{n=1}^N \sum_{m=0}^{M-1} \mathbb{E}[\delta(X_n - m) | Y = y, \theta'] \{\log p(y_n | \mu_m, \sigma_m) + \log \pi_m\} \\ &= \sum_{n=1}^N \sum_{m=0}^{M-1} P\{X_n = m | Y = y, \theta'\} \{\log p(y_n | \mu_m, \sigma_m) + \log \pi_m\} . \end{aligned}$$

- Line 3: the expectation moves inside the two sums.
- Line 4: the expectation of a delta function is the probability that  $X_n = m$ .

# The explicit $Q$ function

Now plugging in the explicit expression for the Gaussian distribution, we get the final expression for the  $Q$  function:

$$Q(\theta; \theta') = \sum_{n=1}^N \sum_{m=0}^{M-1} \left\{ \frac{-1}{2\sigma_m^2} (y_n - \mu_m)^2 - \frac{1}{2} \log(2\pi\sigma_m^2) + \log \pi_m \right\} P\{X_n = m | Y = y, \theta'\}$$

- The terms in braces depend on the new parameter  $\theta$ .
- The posterior probability  $P\{X_n = m | Y = y, \theta'\}$  depends on the old parameter  $\theta'$ .

# The posterior probability by Bayes' rule

The posterior conditional probability,  $P\{X_n = m | Y = y, \theta'\}$ , can be easily computed using Bayes' rule. If we define the new notation that

$$f(m|y_n, \theta') \triangleq P\{X_n = m | Y = y, \theta'\} ,$$

then this function is given by

$$f(m|y_n, \theta') = \frac{\frac{1}{\sqrt{2\pi\sigma'^2_m}} \exp\left\{-\frac{1}{2\sigma'^2_m}(y_n - \mu'_m)^2\right\} \pi'_m}{\sum_{j=0}^{M-1} \frac{1}{\sqrt{2\pi\sigma'^2_j}} \exp\left\{-\frac{1}{2\sigma'^2_j}(y_n - \mu'_j)^2\right\} \pi'_j} .$$

While this expression for  $f(m|y_n, \theta')$  may appear complex, it is quite simple to compute numerically.

# The M-step and the EM update equations

In order to calculate the associated M-step, we must maximize  $Q(\theta; \theta')$  with respect to the parameter  $\theta$ . This maximization procedure is much the same as is required for ML estimation of parameters. So, for example,  $\mu_k$  and  $\sigma_k$  are calculated the same way as is used for the ML estimate of the mean and variance of a Gaussian random variable. This approach results in the final EM update equations for this clustering problem:

$$\begin{aligned}\hat{N}_m^{(k+1)} &= \sum_{n=1}^N f(m|y_n, \theta^{(k)}) , \\ \hat{\pi}_m^{(k+1)} &= \frac{\hat{N}_m^{(k+1)}}{N} , \\ \hat{\mu}_m^{(k+1)} &= \frac{1}{\hat{N}_m^{(k+1)}} \sum_{n=1}^N y_n f(m|y_n, \theta^{(k)}) , \\ [\hat{\sigma}_m^2]^{(k+1)} &= \frac{1}{\hat{N}_m^{(k+1)}} \sum_{n=1}^N \left( y_n - \hat{\mu}_m^{(k+1)} \right)^2 f(m|y_n, \theta^{(k)}) .\end{aligned}$$

The four updates use the posterior probabilities  $f(m|y_n, \theta^{(k)})$  computed in the E-step.

# Summary

- Decomposition:  $\log p(y|\theta) = Q(\theta; \theta') + H(\theta; \theta')$ , with  $Q$  the conditional expectation of  $\log p(y, X|\theta)$  and  $H$  the negative conditional expectation of  $\log p(X|y, \theta)$ .
- $H(\theta; \theta') \geq H(\theta'; \theta')$ , proved with Jensen's inequality.
- Key bound:  $\log p(y|\theta) - \log p(y|\theta') \geq Q(\theta; \theta') - Q(\theta'; \theta')$ , so  $Q$  is a surrogate function for maximizing  $\log p(y|\theta)$ .
- EM: E-step computes  $Q(\theta; \theta^{(k)})$ ; M-step sets  $\theta^{(k+1)} = \arg \max_{\theta \in \Omega} Q(\theta; \theta^{(k)})$ ; the log likelihood is monotone increasing.
- Gaussian mixture:  $Q$  follows from the delta function expansion, the E-step reduces to  $f(m|y_n, \theta')$  by Bayes' rule, and the M-step gives  $\hat{N}_m$ ,  $\hat{\pi}_m$ ,  $\hat{\mu}_m$ , and  $\hat{\sigma}_m^2$ .

Reading: Sections 12.3–12.4.    Next lecture: Sections 12.5–12.6, algorithmic implementation of EM clustering and convergence.