

# Lecture 32: EM Clustering; Convergence and Majorization

Charles A. Bouman

Purdue University

Monday, November 9, 2026

# Outline

- 1 Pseudocode notation:  $\theta$  and the program variable  $P_{n,m}$
- 2 The E-step as an assignment to  $P_{n,m}$
- 3 The M-step as four assignments:  $N_m, \pi_m, \mu_m, \sigma_m^2$
- 4 The structure of the updates, illustrated
- 5 Soft assignment and partial membership
- 6 Back to the heuristic of the beginning of the chapter
- 7 Convergence: what the surrogate property gives us
- 8 What can still go wrong
- 9 The limit  $L^*$  exists
- 10 A simplified convergence theorem, and its proof

## Core idea

The E-step is a **soft classification** of the data; the M-step is **ML parameter estimation** given those soft classifications.

# Today's Big Ideas

- 1 Written as pseudocode, the EM algorithm is two steps: the E-step assigns each posterior probability, and the M-step assigns the four cluster parameters.
- 2 The assignment is soft, so each data point has partial membership in every cluster, and each cluster is updated from the weighted contributions of all the data.
- 3 EM generalizes the hard classification heuristic from the beginning of the chapter, with the E-step doing soft classification and the M-step doing ML estimation given those soft classifications.
- 4 Because  $Q$  is a surrogate function, the log likelihood never decreases and it reaches a limit, but the iterations can still be trapped in a local maximum.
- 5 Under mild conditions, the EM iterations converge to a parameter value at which the gradient of the log likelihood is zero.

## Section 12.5: the algorithmic view

In order to better understand the EM algorithm, it is useful to take a closer look at its algorithmic implementation.

- We use the algorithmic notation of **pseudocode programming**, where  $\theta$  is the variable that contains the current value of the parameter.
- Let  $P_{n,m}$  be the program variable that contains the current value of the posterior probability,

$$P\{X_n = m | Y = y, \theta\} .$$

- The symbol  $\leftarrow$  indicates that the value is **assigned** to the program variable.

In this notation there are no iteration superscripts: the variables simply hold their current values.

# The E-step

Using these variables, the E-step is computed by evaluating the following expression for all values of  $n$  and  $m$ :

$$\text{E-step: } P_{n,m} \leftarrow \frac{\frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left\{-\frac{1}{2\sigma_m^2}(y_n - \mu_m)^2\right\} \pi_m}{\sum_{j=0}^{M-1} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{1}{2\sigma_j^2}(y_n - \mu_j)^2\right\} \pi_j}.$$

This is Bayes' rule for the label  $X_n$ , evaluated with the current values of  $\pi_m$ ,  $\mu_m$ , and  $\sigma_m^2$ .

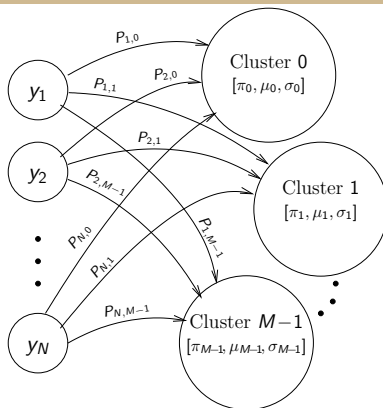
# The M-step

Once  $P_{n,m}$  is computed, we can update the model parameters by evaluating the following expressions for all values of  $m$ .

$$\begin{aligned}\text{M-step: } N_m &\leftarrow \sum_{n=1}^N P_{n,m} , \\ \pi_m &\leftarrow \frac{N_m}{N} , \\ \mu_m &\leftarrow \frac{1}{N_m} \sum_{n=1}^N y_n P_{n,m} , \\ \sigma_m^2 &\leftarrow \frac{1}{N_m} \sum_{n=1}^N (y_n - \mu_m)^2 P_{n,m} .\end{aligned}$$

Later assignments use the values assigned before them.

# The structure of the EM algorithm updates



This figure illustrates the structure of the EM algorithm updates. In the E-step, the matrix  $P_{n,m}$ , which contains the posterior probability that  $X_n = m$ , is computed. The entries  $P_{n,m}$  represent the soft assignment of each sample,  $Y_n$ , to a cluster,  $m$ . Armed with this information, the parameters of each cluster,  $(\pi_m, \mu_m, \sigma_m)$ , may be updated in the M-step.

# Soft assignment and partial membership

In this form, the meaning of the EM algorithm starts to become more clear.

- With each update, we use the current estimate of the parameter  $\theta$  to compute the matrix  $P_{n,m}$ , the probability that  $X_n$  has label  $m$ .
- These posterior probabilities are then used to assign each data point,  $Y_n$ , to the associated  $M$  clusters.
- Since the assignment is **soft**, each data point has **partial membership** in each cluster.

# The iteration

- Once the data points are assigned to the clusters, then the parameters of each cluster,  $[\pi_m, \mu_m, \sigma_m^2]$ , can be updated based on the **weighted contributions** of each sample.
- This results in an updated value for the parameter  $\theta$ .
- However, with this new parameter, we must compute new entries for the matrix  $P_{n,m}$ , and the iterative process repeats.

$$\theta \rightarrow P_{n,m} \rightarrow \theta \rightarrow P_{n,m} \rightarrow \dots$$

## Back to the heuristic of Lecture 29

Recall the hard classification heuristic from the beginning of the chapter:

$$\hat{X}_n = \begin{cases} 1, & Y_n \geq \text{threshold} \\ 0, & Y_n < \text{threshold} \end{cases}.$$

In fact, EM generalizes that heuristic method, except the hard classification heuristic is replaced with the **soft classification** of the posterior probability.

So the E-step can be viewed as a form of soft classification of the data to the available categories, while the M-step can be viewed as ML parameter estimation given these soft classifications.

## Section 12.6: what the surrogate property gives us

From the properties of the  $Q$  and  $H$  functions (Lecture 30), we know that  $Q(\theta; \theta')$  is a surrogate function for maximizing the log likelihood.

So we know that the iterations of the EM algorithm will produce a **monotone nondecreasing sequence of log likelihoods**.

However, this raises the question of whether the EM algorithm actually converges to the ML estimate of  $\theta$ .

# What can still go wrong

It is difficult to give a simple and clear answer to this question because in optimization (as in life) many things can go wrong.

- For example, the iterations can become trapped in a **local maximum** of the likelihood that is not a global maximum.
- Even if the likelihood has no local maximum, strange technical problems can occur, so it is difficult to make a simple statement about the convergence to the ML estimate that is **absolutely true under all conditions**.

## But as a practical matter

But with that said, as a practical matter, the EM algorithm generally does converge to a **local maximum** of the log likelihood in typical estimation problems.

We refer the interested reader to the references of Wu (1983) and Redner and Walker (1984), which provide a more detailed introduction to the theory of convergence for the EM algorithm.

However, in order to give some intuition regarding the convergence, we present some simplified results to illustrate the methods of proof.

# The limit $L^*$ exists

Define the function  $L(\theta) = \log p(y|\theta)$ . Then we know that for each iteration of the EM algorithm,

$$L(\theta^{(k+1)}) \geq L(\theta^{(k)}) .$$

If the ML estimate exists, then for all  $k$ , we also know that  $L(\theta^{(k)}) \leq L(\theta_{ML}) < \infty$ .

Therefore, the sequence  $L(\theta^{(k)})$  must be a monotone increasing sequence that is bounded above, so it must therefore reach a limit, which we denote by

$$L^* = \lim_{k \rightarrow \infty} L(\theta^{(k)}) .$$

# What the limit does not tell us

However, even though the limit exists, proving that the value  $L^*$  is the maximum of the likelihood estimate, or that  $\lim_{k \rightarrow \infty} \theta^{(k)}$  exists, is a much more difficult matter.

The following simplified theorem gives some insight into the convergence of the EM algorithm.

- The basic result is that under normal conditions, the EM algorithm converges to a parameter value for which the gradient of the likelihood is zero.
- Typically, this means that the EM algorithm converges to a local or global maximum of the likelihood function.

# A simplified convergence theorem

## Theorem

Let  $\theta^{(k)} \in \Omega$  be a sequence of parameter values generated by the EM updates. Assume that

- a)  $\Omega$  is an open set,
- b) the functions  $Q$  and  $H$  are continuously differentiable on  $\Omega^2$ , and
- c)  $\lim_{k \rightarrow \infty} \theta^{(k)} = \theta^*$  exists for  $\theta^* \in \Omega$ .

Then

$$\nabla \log p(y|\theta^*) = \mathbf{0} .$$

# Proof

First, we know that by the nature of the M-step, the updated parameter value,  $\theta^{(k+1)}$ , is a global and therefore local maximum of the  $Q$  function. This implies that

$$\begin{aligned} \mathbf{0} &= \nabla Q(\theta; \theta^{(k)}) \Big|_{\theta=\theta^{(k+1)}} \\ &= \lim_{k \rightarrow \infty} \nabla Q(\theta^{(k+1)}; \theta^{(k)}) \\ &= \nabla Q\left(\lim_{k \rightarrow \infty} \theta^{(k+1)}, \lim_{k \rightarrow \infty} \theta^{(k)}\right) \\ &= \nabla Q(\theta^*; \theta^*) \\ &= \nabla \log p(y|\theta^*) , \end{aligned}$$

where the last equality is the result of the gradient property of surrogate functions (Chapter 8, Lecture 19).

# The role of assumption a)

The assumptions of this proof are a bit artificial, but the result serves to illustrate the basic concepts of convergence.

In particular, assumption a) is used to ensure that the solution does not fall on the boundary of a closed set because, if this happened, the gradient would no longer need to be zero at a global maximum.

This is the same interior condition that appears in the gradient property of surrogate functions in Chapter 8.

# Summary

- In pseudocode form,  $\theta$  holds the current parameter and  $P_{n,m}$  holds the current posterior probability  $P\{X_n = m | Y = y, \theta\}$ .
- E-step: assign  $P_{n,m}$  from Bayes' rule using the current  $\pi_m, \mu_m, \sigma_m^2$ .
- M-step: assign  $N_m \leftarrow \sum_n P_{n,m}$ ,  $\pi_m \leftarrow N_m/N$ ,  $\mu_m \leftarrow \frac{1}{N_m} \sum_n y_n P_{n,m}$ , and  $\sigma_m^2 \leftarrow \frac{1}{N_m} \sum_n (y_n - \mu_m)^2 P_{n,m}$ .
- The assignment is soft, so each data point has partial membership in each cluster; the parameters are updated from the weighted contributions, and the process repeats.
- EM generalizes the hard classification heuristic of Lecture 29: E-step = soft classification, M-step = ML estimation given those soft classifications.
- $Q$  is a surrogate function, so the log likelihood is monotone nondecreasing and  $L^* = \lim_k L(\theta^{(k)})$  exists; iterations can still be trapped in a local maximum.
- Under a)  $\Omega$  open, b)  $Q$  and  $H$  continuously differentiable, c)  $\theta^{(k)} \rightarrow \theta^*$ , we get  $\nabla \log p(y|\theta^*) = \mathbf{0}$ .

Reading: Sections 12.5–12.6.    Next lecture: Section 12.7, Simplified EM Update Derivations.