Theory and Use of the EM Algorithm

Full text available at: http://dx.doi.org/10.1561/2000000034

Theory and Use of the EM Algorithm

Maya R. Gupta

Department of Electrical Engineering
University of Washington
Seattle, WA 98195
USA

gupta@ee.washington.edu

Yihua Chen

Department of Electrical Engineering
University of Washington
Seattle, WA 98195
USA

yhchen@ee.washington.edu

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is M. R. Gupta and Y. Chen, Theory and Use of the EM Algorithm, Foundations and Trends® in Signal Processing, vol 4, no 3, pp 223–296, 2010

ISBN: 978-1-60198-430-2

Foundations and Trends® in
Signal Processing
Volume 4 Issue 3, 2010

Editorial Scope

Foundations and Trends® in Signal Processing will publish survey and tutorial articles on the foundations, algorithms, methods, and applications of signal processing including the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital and multirate signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
- classification and detection
- estimation and regression
- tree-structured methods

Information for Librarians

Foundations and Trends® in Signal Processing, 2010, Volume 4, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in
Signal Processing
Vol. 4, No. 3 (2010) 223–296
© 2011 M. R. Gupta and Y. Chen
DOI: 10.1561/2000000034

Theory and Use of the EM Algorithm

Maya R. Gupta[1] **and Yihua Chen**[2]

[1] *Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA, gupta@ee.washington.edu*
[2] *Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA, yhchen@ee.washington.edu*

Abstract

This introduction to the expectation–maximization (EM) algorithm provides an intuitive and mathematically rigorous understanding of EM. Two of the most popular applications of EM are described in detail: estimating Gaussian mixture models (GMMs), and estimating hidden Markov models (HMMs). EM solutions are also derived for learning an optimal mixture of fixed models, for estimating the parameters of a compound Dirichlet distribution, and for dis-entangling superimposed signals. Practical issues that arise in the use of EM are discussed, as well as variants of the algorithm that help deal with these challenges.

Contents

1 The Expectation-Maximization Method 1

1.1 The EM Algorithm 3
1.2 Contrasting EM with a Simple Variant 6
1.3 Using a Prior with EM (MAP EM) 7
1.4 Specifying the Complete Data 7
1.5 A Toy Example 10

2 Analysis of EM 15

2.1 Convergence 15
2.2 Maximization–Maximization 19

3 Learning Mixtures 23

3.1 Learning an Optimal Mixture of Fixed Models 23
3.2 Learning a GMM 26
3.3 Estimating a Constrained GMM 34

4 More EM Examples 41

4.1 Learning a Hidden Markov Model 41
4.2 Estimating Multiple Transmitter Locations 51
4.3 Estimating a Compound Dirichlet Distribution 54

5 EM Variants **63**

5.1 EM May Not Find the Global Optimum 63
5.2 EM May Not Simplify the Computation 64
5.3 Speed 66
5.4 When Maximizing the Likelihood Is Not the Goal 66

6 Conclusions and Some Historical Notes **69**

Acknowledgments **71**

References **73**

1

The Expectation-Maximization Method

Expectation–maximization (EM) is an iterative method that attempts to find the maximum likelihood estimator of a parameter θ of a parametric probability distribution. Let us begin with an example. Consider the temperature outside your window for each of the 24 hours of a day, represented by $x \in \mathbb{R}^{24}$, and say that this temperature depends on the season $\theta \in \{\text{summer, fall, winter, spring}\}$, and that you know the seasonal temperature distribution $p(x \,|\, \theta)$. But what if you could only measure the average temperature $y = \bar{x}$ for some day, and you would like to estimate what season θ it is (for example, is spring here yet?). In particular, you might seek the maximum likelihood estimate of θ, that is, the value $\hat{\theta}$ that maximizes $p(y \,|\, \theta)$. If this is not a trivial maximum likelihood problem, you might call upon EM. EM iteratively alternates between making guesses about the complete data x, and finding the θ that maximizes $p(x \,|\, \theta)$ over θ. In this way, EM *tries to find* the maximum likelihood estimate of θ given y. We will see in later sections that EM does not actually promise to find the θ that maximizes $p(y \,|\, \theta)$, but there are some theoretical guarantees, and it often does a good job in practice, though it may need a little help in the form of multiple random starts.

This exposition is designed to be useful to both the EM novice and the experienced EM user looking to better understand the method and its use. To this end, we err on the side of providing too many explicit details rather than too few.

First, we go over the steps of EM, breaking down the usual two-step description into a five-step description. Table 1.1 summarizes the key notation. We recommend reading this document linearly up through Section 1.4, after which sections can generally be read out-of-order. Section 1 ends with a detailed version of a historical toy example for EM. In Section 2 we show that EM never gets worse as it iterates in terms of the likelihood of the estimate it produces, and we explain the *maximization–maximization* interpretation of EM. We also explain the general advantages and disadvantages of EM compared to other options for maximizing the likelihood, like the Newton–Raphson method. The

Table 1.1. Notation summary.

$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of positive real numbers
$\mathbb{N}$	Set of natural numbers
$y \in \mathbb{R}^d$	Given measurement or observation
$Y \in \mathbb{R}^d$	Random measurement; y is a realization of Y
$x \in \mathbb{R}^{d_1}$	Complete data you wish you had
$X \in \mathbb{R}^{d_1}$	Random complete data; x is a realization of X
$z \in \mathbb{R}^{d_2}$	Missing data; in some problems $x = (y,z)$
$Z \in \mathbb{R}^{d_2}$	Random missing data; z is a realization of Z
$\theta \in \Omega$	Parameter(s) to estimate, Ω is the parameter space
$\theta^{(m)} \in \Omega$	mth estimate of θ
$p(y \mid \theta)$	Density of y given θ; also written as $p(Y = y \mid \theta)$
$\mathcal{X}$	Support of X (closure of the set of x where $p(x \mid \theta) > 0$)
$\mathcal{X}(y)$	Support of X conditioned on y (closure of the set of x where $p(x \mid y, \theta) > 0$)
$\triangleq$	"Is defined to be"
$L(\theta)$	Likelihood of θ given y, that is, $p(y \mid \theta)$
$\ell(\theta)$	Log-likelihood of θ given y, that is, $\log p(y \mid \theta)$
$E_{X \mid y,\theta}[X]$	Expectation of X conditioned on y and θ, that is, $\int_{\mathcal{X}(y)} x p(x \mid y, \theta) dx$
$1_{\{\cdot\}}$	Indicator function: equals 1 if the expression $\{\cdot\}$ is true, and 0 otherwise
$\mathbf{1}$	Vector of ones
$D_{\mathrm{KL}}(P \,\Vert\, Q)$	Kullback–Leibler divergence (a.k.a. relative entropy) between distributions P and Q

advantages of EM are made clearer in Sections 3 and 4, in which we derive a number of popular applications of EM and use these applications to illustrate practical issues that can arise with EM. Section 3 covers learning the optimal combination of fixed models to explain the observed data, and fitting a Gaussian mixture model (GMM) to the data. Section 4 covers learning hidden Markov models (HMMs), separating superimposed signals, and estimating the parameter for the compound Dirichlet distribution. In Section 5, we categorize and discuss some of the variants of EM and related methods, and we conclude this manuscript in Section 6 with some historical notes.

1.1 The EM Algorithm

To use EM, you must be given some observed data y, a parametric density $p(y\,|\,\theta)$, a description of some complete data x that you wish you had, and the parametric density $p(x\,|\,\theta)$.[1] In Sections 3 and 4 we will explain how to define the complete data x for some standard EM applications.

We assume that the complete data can be modeled as a continuous[2] random vector X with density $p(x\,|\,\theta)$,[3] where $\theta \in \Omega$ for some set Ω. You do not observe X directly; instead, you observe a realization y of the random vector Y that depends[4] on X. For example, X might be a random vector and Y the mean of its components, or if X is a complex number then Y might be only its magnitude, or Y might be the first component of the vector X.

[1] A different standard choice of notation for a parametric density would be $p(y;\theta)$, but we prefer $p(y\,|\,\theta)$ because this notation is clearer when one wants to find the maximum *a posteriori* estimate rather than the maximum likelihood estimate—we will talk more about the maximum *a posteriori* estimate of θ in Section 1.3.

[2] The treatment of discrete random vectors is a straightforward special case of the continuous treatment: one only needs to replace the probability density function with probability mass function and integral with summation.

[3] We assume that the support of X, denoted by $\mathcal{X}$, which is the closure of the set $\{x \mid p(x\,|\,\theta) > 0\}$, does not depend on θ. An example where the support does depend on θ is if X is uniformly distributed on the interval $[0,\theta]$. If the support does depend on θ, then the monotonicity of the EM algorithm might not hold. See Section 2.1 for details.

[4] A rigorous description of this dependency is deferred to Section 1.4.

Given that you only have y, the goal here is to find the maximum likelihood estimate (MLE) of θ:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta \in \Omega} \; p(y \,|\, \theta). \tag{1.1}$$

It is often easier to calculate the θ that maximizes the *log-likelihood* of y:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Omega} \log p(y \,|\, \theta). \tag{1.2}$$

Because log is a monotonically increasing function, the solution to (1.1) will be the same as the solution to (1.2). However, for some problems it is difficult to solve either (1.1) or (1.2). Then we can try EM: we make a guess about the complete data X and solve for the θ that maximizes the (expected) log-likelihood of X. And once we have an estimate for θ, we can make a better guess about the complete data X, and iterate.

EM is usually described as two steps (the E-step and the M-step), but let us first break it down into five steps:

Step 1: Let $m = 0$ and make an initial estimate $\theta^{(m)}$ for θ.

Step 2: Given the observed data y and pretending for the moment that your current guess $\theta^{(m)}$ is correct, formulate the conditional probability distribution $p(x \,|\, y, \theta^{(m)})$ for the complete data x.

Step 3: Using the conditional probability distribution $p(x \,|\, y, \theta^{(m)})$ calculated in Step 2, form the *conditional expected log-likelihood*, which is called the Q-function[5]:

$$\begin{aligned} Q(\theta \,|\, \theta^{(m)}) &= \int_{\mathcal{X}(y)} \log p(x \,|\, \theta) p(x \,|\, y, \theta^{(m)}) dx \\ &= E_{X|y,\theta^{(m)}}[\log p(X \,|\, \theta)], \end{aligned} \tag{1.3}$$

[5] Note this Q-function has nothing to do with the sum of the tail of a Gaussian, which is also called the Q-function. People call (1.3) the Q-function because the original paper [11] used a Q to notate it. We like to say that the Q stands for *quixotic* because it is a bit crazy and hopeful and beautiful to think you can find the maximum likelihood estimate of θ in this way that iterates round-and-round like a windmill, and if Don Quixote had been a statistician, it is just the sort of thing he might have done.

where the integral is over the set $\mathcal{X}(y)$, which is the closure of the set $\{x \mid p(x \,|\, y,\theta) > 0\}$, and we assume that $\mathcal{X}(y)$ does not depend on θ.

Note that θ is a free variable in (1.3), so the Q-function is a function of θ, but also depends on your current guess $\theta^{(m)}$ implicitly through the $p(x \,|\, y,\theta^{(m)})$ calculated in Step 2.

Step 4: Find the θ that maximizes the Q-function (1.3); the result is your new estimate $\theta^{(m+1)}$.

Step 5: Let $m := m + 1$ and go back to Step 2. (The EM algorithm does not specify a stopping criterion; standard criteria are to iterate until the estimate stops changing: $\|\theta^{(m+1)} - \theta^{(m)}\| < \epsilon$ for some $\epsilon > 0$, or to iterate until the log-likelihood $\ell(\theta) = \log p(y \,|\, \theta)$ stops changing: $|\ell(\theta^{(m+1)}) - \ell(\theta^{(m)})| < \epsilon$ for some $\epsilon > 0$.)

The EM estimate is *only guaranteed to never get worse* (see Section 2.1 for details). Usually, it will find a peak in the likelihood $p(y \,|\, \theta)$, but if the likelihood function $p(y \,|\, \theta)$ has multiple peaks, EM will not necessarily find the global maximum of the likelihood. In practice, it is common to start EM from multiple random initial guesses, and choose the one with the largest likelihood as the final guess for θ.

The traditional description of the EM algorithm consists of only two steps. The above Steps 2 and 3 combined are called the *E-step* for *expectation*, and Step 4 is called the *M-step* for *maximization*:

E-step: Given the estimate from the previous iteration $\theta^{(m)}$, compute the conditional expectation $Q(\theta \,|\, \theta^{(m)})$ given in (1.3).

M-step: The $(m + 1)$th guess of θ is:

$$\theta^{(m+1)} = \arg\max_{\theta\in\Omega} \; Q(\theta \,|\, \theta^{(m)}). \tag{1.4}$$

Since the E-step is just to compute the Q-function which is used in the M-step, EM can be summarized as just iteratively solving the M-step given by (1.4). When applying EM to a particular problem, this is usually the best way to think about EM because then one does not waste time computing parts of the Q-function that do not depend on θ.

1.2 Contrasting EM with a Simple Variant

As a comparison that may help illuminate EM, we next consider a simple variant of EM. In Step 2 above, one computes the conditional distribution $p(x\,|\,y,\theta^{(m)})$ over all possible values of x, and this entire conditional distribution is taken into account in the M-step. A simple variant is to instead use only the mth maximum likelihood estimate $x^{(m)}$ of the complete data x:

$$\begin{aligned}
&\text{E-like-step:} && x^{(m)} = \arg\max_{x\in\mathcal{X}(y)} p(x\,|\,y,\theta^{(m)}),\\
&\text{M-like-step:} && \theta^{(m+1)} = \arg\max_{\theta\in\Omega} p(x^{(m)}\,|\,\theta).
\end{aligned}$$

We call this variant the *point-estimate variant of EM*; it has also been called *classification EM*. More on this variant can be found in [7, 9].

Perhaps the most famous example of this variant is *k-means clustering*[6] [21, 35]. In k-means clustering, we have n observed data points $y = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^{\mathrm{T}}$, where each $y_i \in \mathbb{R}^d$, and it is believed that the data points belong to k clusters. Let the complete data be the observed data points and the missing information that specifies which of the k clusters each observed data point belongs to. The goal is to estimate the k cluster centers θ. First, one makes an initial guess $\hat{\theta}^0$ of the k cluster centers. Then in the E-like step, one assigns each of the n points to the closest cluster based on the estimated cluster centers $\theta^{(m)}$. Then in the M-like step, one takes all the points assigned to each cluster, and computes the mean of those points to form a new estimate of the cluster's centroid. Underlying k-means is a model that the clusters are defined by Gaussian distributions with unknown means (the θ to be estimated) and identity covariance matrices.

EM clustering differs from k-means clustering in that at each iteration you do not choose a single $x^{(m)}$, that is, one does not force each observed point y_i to belong to only one cluster. Instead, each observed point y_i is probabilistically assigned to the k clusters by estimating $p(x\,|\,y,\theta^{(m)})$. We treat EM clustering in more depth in Section 3.2.

[6] The k-means clustering algorithm dates to 1967 [35] and is a special case of *vector quantization*, which was first proposed as Lloyd's algorithm in 1957 [32]. See [17] for details.

1.3 Using a Prior with EM (MAP EM)

The EM algorithm can fail due to singularities of the log-likelihood function — for example, for learning a GMM with 10 components, it may decide that the most likely solution is for one of the Gaussians to only have one data point assigned to it, with the bad result that the Gaussian is estimated as having zero covariance (see Section 3.2.5 for details).

A straightforward solution to such degeneracies is to take into account or impose some prior information on the solution for θ. One approach would be to restrict the set of possible θ. Such a restriction is equivalent to putting a uniform prior probability over the restricted set. More generally, one can impose any prior $p(\theta)$, and then modify EM to maximize the posterior rather than the likelihood:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta \in \Omega} \log p(\theta \,|\, y) = \arg\max_{\theta \in \Omega} (\log p(y \,|\, \theta) + \log p(\theta)).$$

The EM algorithm is easily extended to maximum *a posteriori* (MAP) estimation by modifying the M-step:

E-step: Given the estimate from the previous iteration $\theta^{(m)}$, compute as a function of $\theta \in \Omega$ the conditional expectation

$$Q(\theta \,|\, \theta^{(m)}) = E_{X|y,\theta^{(m)}}[\log p(X \,|\, \theta)].$$

M-step: Maximize $Q(\theta \,|\, \theta^{(m)}) + \log p(\theta)$ over $\theta \in \Omega$ to find

$$\theta^{(m+1)} = \arg\max_{\theta \in \Omega} (Q(\theta \,|\, \theta^{(m)}) + \log p(\theta)).$$

An example of MAP EM is given in Section 3.3.

1.4 Specifying the Complete Data

Practically, the complete data should be defined so that given x it is relatively easy to maximize $p(x \,|\, \theta)$ with respect to θ. Theoretically, the complete data X must satisfy the Markov relationship $\theta \to X \to Y$ with respect to the parameter θ and the observed data Y, that is, it must be that

$$p(y \,|\, x, \theta) = p(y \,|\, x).$$

A special case is when Y is a function of X, that is, $Y = T(X)$; in this case, $X \to Y$ is a deterministic function, and thus the Markov relationship always holds.

1.4.1 EM for Missing Data Problems

For many applications of EM, including GMM and HMM, the complete data X is the observed data Y plus some missing (sometimes called *latent* or *hidden*) data Z, such that $X = (Y, Z)$. This is a special case of $Y = T(X)$, where the function T simply removes Z from X to produce Y. In general when using EM with missing data, one can write the Q-function as an integral over the domain of Z, denoted by $\mathcal{Z}$, rather than over the domain of X, because the only random part of the complete data X is the missing data Z. Then, for missing data problems where $x = (y, z)$,

$$\begin{aligned} Q(\theta \,|\, \theta^{(m)}) &= \int_{\mathcal{X}} \log p(x \,|\, \theta) p(x \,|\, y, \theta^{(m)}) dx \\ &= \int_{\mathcal{X}} \log p(y, z \,|\, \theta) p(y, z \,|\, y, \theta^{(m)}) dx \\ &= \int_{\mathcal{Z}} \log p(y, z \,|\, \theta) p(z \,|\, y, \theta^{(m)}) dz \\ &= E_{Z|y,\theta^{(m)}}[\log p(y, Z \,|\, \theta)]. \end{aligned} \tag{1.5}$$

1.4.2 EM for Independently, Identically Distributed Samples

For many common applications such as learning a GMM or HMM, the complete data X is a set of n independent and identically distributed (i.i.d.) random vectors, $X = \begin{bmatrix} X_1 & X_2 & \ldots & X_n \end{bmatrix}^{\mathrm{T}}$ and the ith observed sample y_i is only a function of x_i. Then the following proposition is useful for decomposing the Q-function into a sum:

Proposition 1.1. Suppose $p(x \,|\, \theta) = \prod_{i=1}^{n} p(x_i \,|\, \theta)$ for all $x \in \mathcal{X}^n$ and all $\theta \in \Omega$, and the Markov relationship $\theta \to X_i \to Y_i$ holds for all $i = 1, \ldots, n$, that is,

$$p(y_i \,|\, x, y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n, \theta) = p(y_i \,|\, x_i), \tag{1.6}$$

then

$$Q(\theta \,|\, \theta^{(m)}) = \sum_{i=1}^{n} Q_i(\theta \,|\, \theta^{(m)}),$$

where

$$Q_i(\theta \,|\, \theta^{(m)}) = E_{X_i|y_i,\theta^{(m)}}[\log p(X_i \,|\, \theta)], \quad i = 1,\ldots,n.$$

Proof. First, we show that given θ, the elements of the set $\{(X_i, Y_i)\}$, $i = 1,\ldots,n$, are mutually independent, that is,

$$p(x,y \,|\, \theta) = \prod_{i=1}^{n} p(x_i, y_i \,|\, \theta). \tag{1.7}$$

This mutual independence holds because

$$\begin{aligned}
p(x,y \,|\, \theta) &= p(y_1 \,|\, y_2,\ldots,y_n,x,\theta)\cdots p(y_n \,|\, x,\theta)p(x \,|\, \theta) \\
&\quad \text{(by the chain rule)} \\
&= p(y_1 \,|\, x_1,\theta)\cdots p(y_n \,|\, x_n,\theta)p(x \,|\, \theta) \\
&\quad \text{(by (1.6), but keep } \theta \text{ in the condition)} \\
&= p(y_1 \,|\, x_1,\theta)\cdots p(y_n \,|\, x_n,\theta)\prod_{i=1}^{n} p(x_i \,|\, \theta) \\
&\quad \text{(by the independence assumption on } X\text{)} \\
&= \prod_{i=1}^{n} p(y_i \,|\, x_i,\theta)p(x_i \,|\, \theta) \\
&= \prod_{i=1}^{n} p(x_i, y_i \,|\, \theta).
\end{aligned}$$

Then we show that for all $i = 1,\ldots,n$, we have

$$p(x_i \,|\, y,\theta) = p(x_i \,|\, y_i,\theta). \tag{1.8}$$

This is because

$$\begin{aligned}
p(x_i \,|\, y,\theta) &= \frac{p(x_i, y \,|\, \theta)}{p(y \,|\, \theta)} \\
&\quad \text{(by Bayes' rule)} \\
&= \frac{\int_{\mathcal{X}^{n-1}} p(x,y \,|\, \theta) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n}{\int_{\mathcal{X}^n} p(x,y \,|\, \theta) dx}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\int_{\mathcal{X}^{n-1}} \prod_{j=1}^{n} p(x_j, y_j \,|\, \theta) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n}{\int_{\mathcal{X}^n} \prod_{j=1}^{n} p(x_j, y_j \,|\, \theta) dx_1 \ldots dx_n} \\
&\quad \text{(by (1.7))} \\
&= \frac{p(x_i, y_i \,|\, \theta) \prod_{j=1, j\neq i}^{n} \int_{\mathcal{X}} p(x_j, y_j \,|\, \theta) dx_j}{\prod_{j=1}^{n} \int_{\mathcal{X}} p(x_j, y_j \,|\, \theta) dx_j} \\
&= \frac{p(x_i, y_i \,|\, \theta) \prod_{j=1, j\neq i}^{n} p(y_j \,|\, \theta)}{\prod_{j=1}^{n} p(y_j \,|\, \theta)} \\
&= \frac{p(x_i, y_i \,|\, \theta)}{p(y_i \,|\, \theta)} \\
&= p(x_i \,|\, y_i, \theta).
\end{aligned}
$$

Then,

$$
\begin{aligned}
Q(\theta \,|\, \theta^{(m)}) &= E_{X|y,\theta^{(m)}} [\log p(X \,|\, \theta)] \\
&= E_{X|y,\theta^{(m)}} \left[\log \prod_{i=1}^{n} p(X_i \,|\, \theta) \right] \\
&\quad \text{(by the independence assumption on } X\text{)} \\
&= E_{X|y,\theta^{(m)}} \left[\sum_{i=1}^{n} \log p(X_i \,|\, \theta) \right] \\
&= \sum_{i=1}^{n} E_{X_i|y,\theta^{(m)}} [\log p(X_i \,|\, \theta)] \\
&= \sum_{i=1}^{n} E_{X_i|y_i,\theta^{(m)}} [\log p(X_i \,|\, \theta)],
\end{aligned}
$$

where the last line holds because of (1.8). □

1.5 A Toy Example

We next present a fully worked-out version of a "toy example" of EM that was used in the seminal EM paper [11]. Here, we give more details, and we have changed it to literally be a toy example.

Imagine you ask n kids to choose a toy out of four choices. Let $Y = \begin{bmatrix} Y_1 & \ldots & Y_4 \end{bmatrix}^T$ denote the histogram of their n choices, where Y_i is the number of the kids that chose toy i, for $i = 1, \ldots, 4$. We can model this random histogram Y as being distributed according to a multinomial distribution. The multinomial has two parameters: the *number of kids asked*, denoted by $n \in \mathbb{N}$, and the *probability that a kid will choose each of the four toys*, denoted by $p \in [0,1]^4$, where $p_1 + p_2 + p_3 + p_4 = 1$. Then the probability of seeing some particular histogram y is:

$$P(y \,|\, p) = \frac{n!}{y_1! y_2! y_3! y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4}. \tag{1.9}$$

Next, say that we have reason to believe that the unknown probability p of choosing each of the toys is parameterized by some hidden value $\theta \in (0,1)$ such that

$$p_\theta = \begin{bmatrix} \frac{1}{2} + \frac{1}{4}\theta & \frac{1}{4}(1-\theta) & \frac{1}{4}(1-\theta) & \frac{1}{4}\theta \end{bmatrix}^{\mathrm{T}}, \quad \theta \in (0,1). \tag{1.10}$$

The estimation problem is to guess the θ that maximizes the probability of the observed histogram y of toy choices.

Combining (1.9) and (1.10), we can write the probability of seeing the histogram $y = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \end{bmatrix}^{\mathrm{T}}$ as

$$P(y \,|\, \theta) = \frac{n!}{y_1! y_2! y_3! y_4!} \left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2} \left(\frac{1-\theta}{4}\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4}.$$

For this simple example, one could directly maximize the log-likelihood $\log P(y \,|\, \theta)$, but here we will instead illustrate how to use the EM algorithm to find the maximum likelihood estimate of θ.

To use EM, we need to specify what the complete data X is. We will choose the complete data to enable us to specify the probability mass function (pmf) in terms of only θ and $1 - \theta$. To that end, we define the complete data to be $X = \begin{bmatrix} X_1 & \dots & X_5 \end{bmatrix}^{\mathrm{T}}$, where X has a multinomial distribution with number of trials n and the probability of each event is:

$$q_\theta = \begin{bmatrix} \frac{1}{2} & \frac{1}{4}\theta & \frac{1}{4}(1-\theta) & \frac{1}{4}(1-\theta) & \frac{1}{4}\theta \end{bmatrix}^{\mathrm{T}}, \quad \theta \in (0,1).$$

By defining X this way, we can then write the observed data Y as:

$$Y = T(X) = \begin{bmatrix} X_1 + X_2 & X_3 & X_4 & X_5 \end{bmatrix}^{\mathrm{T}}.$$

The likelihood of a realization x of the complete data is

$$P(x\,|\,\theta) = \frac{n!}{\prod_{i=1}^{5} x_i!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}. \tag{1.11}$$

For EM, we need to maximize the Q-function:

$$\theta^{(m+1)} = \arg\max_{\theta\in(0,1)} Q(\theta\,|\,\theta^{(m)}) = \arg\max_{\theta\in(0,1)} E_{X|y,\theta^{(m)}}[\log p(X\,|\,\theta)].$$

To solve the above equation, we actually only need the terms of $\log p(x\,|\,\theta)$ that depend on θ, because the other terms are irrelevant as far as maximizing over θ is concerned. Take the log of (1.11) and ignore those terms that do not depend on θ, then

$$\begin{aligned}\theta^{(m+1)} &= \arg\max_{\theta\in(0,1)} E_{X|y,\theta^{(m)}}[(X_2+X_5)\log\theta + (X_3+X_4)\log(1-\theta)]\\ &= \arg\max_{\theta\in(0,1)} (E_{X|y,\theta^{(m)}}[X_2] + E_{X|y,\theta^{(m)}}[X_5])\log\theta\\ &\quad + (E_{X|y,\theta^{(m)}}[X_3] + E_{X|y,\theta^{(m)}}[X_4])\log(1-\theta).\end{aligned}$$

To solve the above maximization problem, we need the expectation of the complete data X conditioned on the already known incomplete data y, which only leaves the uncertainty about X_1 and X_2. Since we know that $X_1 + X_2 = y_1$, we can use the indicator function $1_{\{\cdot\}}$ to write that given y_1, the pair (X_1, X_2) is binomially distributed with X_1 "successes" in y_1 events:

$$\begin{aligned}&P(x\,|\,y,\theta^{(m)})\\ &= \frac{y_1!}{x_1!x_2!}\left(\frac{\frac{1}{2}}{\frac{1}{2}+\frac{\theta^{(m)}}{4}}\right)^{x_1}\left(\frac{\frac{\theta^{(m)}}{4}}{\frac{1}{2}+\frac{\theta^{(m)}}{4}}\right)^{x_2} 1_{\{x_1+x_2=y_1\}}\prod_{i=3}^{5}1_{\{x_i=y_{i-1}\}}\\ &= \frac{y_1!}{x_1!x_2!}\left(\frac{2}{2+\theta^{(m)}}\right)^{x_1}\left(\frac{\theta^{(m)}}{2+\theta^{(m)}}\right)^{x_2} 1_{\{x_1+x_2=y_1\}}\prod_{i=3}^{5}1_{\{x_i=y_{i-1}\}}.\end{aligned}$$

Then the conditional expectation of X given y and $\theta^{(m)}$ is

$$E_{X|y,\theta^{(m)}}[X] = \begin{bmatrix}\frac{2}{2+\theta^{(m)}}y_1 & \frac{\theta^{(m)}}{2+\theta^{(m)}}y_1 & y_2 & y_3 & y_4\end{bmatrix}^{\mathrm{T}},$$

and the M-step becomes

$$\theta^{(m+1)} = \arg\max_{\theta\in(0,1)} \left(\left(\frac{\theta^{(m)}}{2+\theta^{(m)}} y_1 + y_4 \right) \log\theta + (y_2 + y_3) \log(1-\theta) \right)$$

$$= \frac{\frac{\theta^{(m)}}{2+\theta^{(m)}} y_1 + y_4}{\frac{\theta^{(m)}}{2+\theta^{(m)}} y_1 + y_2 + y_3 + y_4}.$$

Given an initial estimate $\theta^{(0)} = 0.5$, the above algorithm reaches $\hat{\theta}_{\mathrm{MLE}}$ to MATLAB's numerical precision on the 18th iteration.

References

[1] M. M. Ali, C. Khompatraporn, and Z. B. Zabinsky, "A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems," *Journal of Global Optimization*, vol. 31, no. 4, pp. 635–672, April 2005.

[2] T. L. Bailey and C. Elkan, "Unsupervised learning of multiple motifs in biopolymers using expectation maximization," *Machine Learning*, vol. 21, pp. 51–80, 1995.

[3] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2664–2669, July 2005.

[4] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, February 1970.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge, UK: Cambridge University Press, 2004.

[6] R. A. Boyles, "On the convergence of the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 45, no. 1, pp. 47–50, 1983.

[7] P. Bryant and J. A. Williamson, "Asymptotic behavior of classification maximum likelihood estimates," *Biometrika*, vol. 65, no. 2, pp. 273–281, 1978.

[8] G. Celeux and J. Diebolt, "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem," *Computational Statistics Quaterly*, vol. 2, pp. 73–82, 1985.

[9] G. Celeux and G. Govaert, “A classification EM algorithm for clustering and two stochastic versions,” *Computational Statistics Data Analysis*, vol. 14, pp. 315–332, 1992.

[10] Y. Chen and J. Krumm, “Probabilistic modeling of traffic lanes from GPS traces,” in *Proceedings of 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[12] M. Feder and E. Weinstein, “Parameter estimation of superimposed signals using the EM algorithm,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 4, pp. 477–489, April 1988.

[13] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*. New York, NY: John Wiley & Sons, 2nd Edition, 1999.

[14] B. A. Frigyik, A. Kapila, and M. R. Gupta, “Introduction to the Dirichlet distribution and related processes,” Department of Electrical Engineering, University of Washigntον, UWEETR-2010-0006, 2010.

[15] B. A. Frigyik, S. Srivastava, and M. R. Gupta, “Functional Bregman divergence and Bayesian estimation of distributions,” *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5130–5139, November 2008.

[16] M. Gales and S. Young, “The application of hidden Markov models in speech recognition,” *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[17] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1991.

[18] A. Goldsmith, *Wireless Communications*. Cambridge, UK: Cambridge University Press, 2005.

[19] M. I. Gurelli and L. Onural, “On a parameter estimation method for Gibbs-Markov random fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 4, pp. 424–430, April 1994.

[20] H. O. Hartley, “Maximum likelihood estimation from incomplete data,” *Biometrics*, vol. 14, no. 2, pp. 174–194, June 1958.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer, 2nd Edition, 2009.

[22] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,” *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, February 2005.

[23] M. Hazen and M. R. Gupta, “A multiresolutional estimated gradient architecture for global optimization,” in *Proceedings of the IEEE Congress on Evolutionary Computation*, pp. 3013–3020, 2006.

[24] M. Hazen and M. R. Gupta, “Gradient estimation in global optimization algorithms,” in *Proceedings of the IEEE Congress on Evolutionary Computation*, pp. 1841–1848, 2009.

[25] T. Hebert and R. Leahy, “A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors,” *IEEE Transactions on Medical Imaging*, vol. 8, no. 2, pp. 194–202, June 1989.

[26] T. J. Hebert and K. Lu, "Expectation–maximization algorithms, null spaces, and MAP image restoration," *IEEE Transactions on Image Processing*, vol. 4, no. 8, pp. 1084–1095, August 1995.

[27] M. Jamshidian and R. I. Jennrich, "Acceleration of the EM algorithm by using quasi-Newton methods," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 59, no. 3, pp. 569–587, 1997.

[28] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of Global Optimization*, vol. 21, no. 4, pp. 345–383, December 2001.

[29] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Transactions on Medical Imaging*, vol. 9, no. 4, pp. 439–446, December 1990.

[30] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 57, no. 2, pp. 425–437, 1995.

[31] J. Li and R. M. Gray, *Image Segmentation and Compression Using Hidden Markov Models.* New York, NY: Springer, 2000.

[32] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, First published in 1957 as a Bell Labs technical note, 1982.

[33] L. B. Lucy, "An iterative technique for the rectification of observed distributions," *Astronomical Journal*, vol. 79, no. 6, pp. 745–754, June 1974.

[34] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms.* Cambridge, UK: Cambridge University Press, 2003.

[35] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[36] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions.* New York, NY: John Wiley & Sons, 2nd Edition, 2008.

[37] G. J. McLachlan and D. Peel, *Finite Mixture Models.* New York, NY: John Wiley & Sons, 2000.

[38] R. Mendes, J. Kennedy, and J. Neves, "The fully informed particle swarm: simpler, maybe better," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 204–210, June 2004.

[39] X.-L. Meng and D. B. Rubin, "On the global and componentwise rates of convergence of the EM algorithm," *Linear Algebra and its Applications*, vol. 199, pp. 413–425, March 1994.

[40] X.-L. Meng and D. A. van Dyk, "The EM algorithm — an old folk-song sung to a fast new tune," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 59, no. 3, pp. 511–567, 1997.

[41] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, (M. I. Jordan, ed.), MIT Press, November 1998.

[42] J. K. Nelson and M. R. Gupta, "An EM technique for multiple transmitter localization," in *Proceedings of the 41st Annual Conference on Information Sciences and Systems*, pp. 610–615, 2007.

[43] J. K. Nelson, M. R. Gupta, J. Almodovar, and W. H. Mortensen, "A quasi EM method for estimating multiple transmitter locations," *IEEE Signal Processing Letters*, vol. 16, no. 5, pp. 354–357, May 2009.

[44] S. Newcomb, "A generalized theory of the combination of observations so as to obtain the best result," *American Journal of Mathematics*, vol. 8, no. 4, pp. 343–366, August 1886.

[45] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, NY: Springer, 2nd Edition, 2006.

[46] P. M. Pardalos and H. E. Romeijn, eds., *Handbook of Global Optimization*. Vol. 2, Norwell, MA: Kluwer, 2002.

[47] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*. November 2008. `http://matrixcookbook.com/`.

[48] W. Qian and D. M. Titterington, "Stochastic relaxations and EM algorithms for Markov random fields," *Journal of Statistical Computation and Simulation*, vol. 40, pp. 55–69, 1992.

[49] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[50] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, April 1984.

[51] W. H. Richardson, "Bayesian-based iterative method of image restoration," *Journal of Optical Society of America*, vol. 62, no. 1, pp. 55–59, 1972.

[52] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York, NY: Springer, 2nd Edition, 2004.

[53] A. Roche, "EM algorithm and variants: An informal tutorial," Unpublished (available online at `ftp://ftp.cea.fr/pub/dsv/madic/publis/Roche_em.pdf`), 2003.

[54] G. Ronning, "Maximum Likelihood estimation of Dirichlet distributions," *Journal of Statistical Computation and Simulation*, vol. 32, no. 4, pp. 215–221, 1989.

[55] H. Stark and Y. Yang, *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics*. New York, NY: John Wiley & Sons, 1998.

[56] C. A. Sugar and G. M. James, "Finding the number of clusters in a dataset: An information-theoretic approach," *Journal of the American Statistical Association*, vol. 98, no. 463, pp. 750–763, September 2003.

[57] M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 528–540, June 1987.

[58] D. A. van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, March 2001.

[59] J. Wang, A. Dogandzic, and A. Nehorai, "Maximum likelihood estimation of compound-Gaussian clutter and target parameters," *IEEE Transactions on Signal Processing*, vol. 54, no. 10, pp. 3884–3898, October 2006.

[60] G. C. G. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 699–704, September 1990.

[61] L. R. Welch, “Hidden Markov Models and the Baum-Welch Algorithm,” *IEEE Information Theory Society Newsletter*, vol. 53, no. 4, pp. 1–13, December 2003.

[62] C. F. J. Wu, “On the convergence properties of the EM algorithm,” *The Annals of Statistics*, vol. 11, no. 1, pp. 95–103, March 1983.

[63] L. Xu and M. I. Jordan, “On convergence properties of the EM algorithm for Gaussian mixtures,” *Neural Computation*, vol. 8, no. 1, pp. 129–151, January 1996.

[64] R. W. Yeung, *A First Course in Information Theory*. New York, NY: Springer, 2002.

[65] J. Zhang, “The mean field theory in EM procedures for Markov random fields,” *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2570–2583, October 1992.