On the Geometry of EM algorithms

David R. Hunter

Technical report no. 0303
Department of Statistics
Penn State University
University Park, PA 16802-2111
email: dhunter@stat.psu.edu
phone: (814) 863-0979
fax: (814) 863-7114

February 19, 2003

Abstract

An understanding of a simple geometric argument that underlies all EM algorithms gives an appreciation for certain aspects of their behavior. Using several illustrative examples, this paper demonstrates how the geometry of EM algorithms can help explain how their rate of convergence is related to the proportion of missing data and how an EM algorithm can fail in a pathological case. This geometric intuition also helps explain why, contrary to a view expressed by some, there is no reason to exclude EM algorithms from cases in which the likelihood function is zero for certain values of the parameter.

Key Words: EM algorithm, MM algorithm

1 Introduction

The term "EM algorithm" has been around for a long time (Dempster, Laird, and Rubin, 1977), and by now many statisticians probably consider themselves familiar enough with the basic idea of EM to be able to describe it. Nonetheless, for many statisticians an EM algorithm is a bit of a black box, and even many practitioners who use these algorithms in their work might not fully understand what makes them

work. The aim of this paper is to elucidate EM algorithms by showing how they may be viewed as special cases of a class of algorithms called MM algorithms that succeed by a particularly simple geometric intuition. For more on the name MM, which here stands for Minorization-Maximization just as EM stands for Expectation-Maximization, see Hunter and Lange (2000).

An EM algorithm consists of two steps, repeated iteratively: In the E-step, one constructs the conditional expectation of the complete data log-likelihood, which is a function of the parameter; and in the M-step, this function is maximized. Yet there is not general agreement about which algorithms so derived qualify as EM algorithms. Even the choice of terminology is not unanimous: Many sources, including arguably the two most widely known—the Dempster, Laird, and Rubin (1977) paper and the McLachlan and Krishnan (1997) book—use "the EM algorithm" to refer to the entire class of algorithms consisting of an E-step and an M-step. On the other hand, since there are many different examples of algorithms that fall under the EM umbrella, here we adopt the indefinite article instead of the definite article, referring not to "the EM algorithm" but to "an EM algorithm," or, collectively, to "EM algorithms".[1]

In succeeding sections, we first explain how an MM algorithm works, showing that its operation is simple to understand by an elementary argument with a geometric intuition. We then demonstrate that any EM algorithm is also an MM algorithm and exploit this fact to examine the operation of several EM algorithms, using the geometric intuition to help explain concepts such as how the "amount" of missing data affects the rate of convergence and what causes EM algorithms to fail in certain cases. Along the way, we exhibit several examples of EM algorithms that

[1]In a footnote, Dempster, Laird, and Rubin (1977) refer to the comment of a referee, who noted that the use of the word "algorithm" may be criticized since EM is not, strictly speaking, an algorithm. However, EM *is* a recipe for creating algorithms, and thus we consider the set of "EM algorithms" to consist of all algorithms baked according to the EM recipe.

function just as they should despite the fact that the likelihood function vanishes on part of the parameter space.

2 MM algorithms

Suppose one is interested in finding the value of the (scalar or vector) variable θ that maximizes some real-valued function $\ell(\theta)$ for which it is not possible to determine the maximizer explicitly. We describe here an iterative technique, wherein a starting value θ_0 is chosen and then some algorithm is used to give a different value θ_1, the hope being that θ_1 is in some sense a "better" candidate than θ_0 for the maximizer of $\ell(\theta)$. The process is then repeated, with θ_1 giving rise to θ_2, θ_2 to θ_3, and so on until we decide this has gone on long enough and declare, after K iterations, that we believe θ_K to be a maximizer of $\ell(\theta)$.

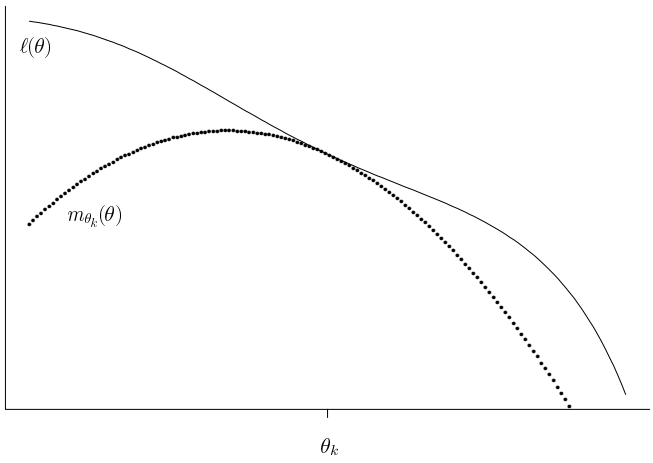

Figure 1: The function $m_{\theta_k}(\theta)$ minorizes the function $\ell(\theta)$ at the point θ_k. Geometrically, it is clear that any θ for which $m_{\theta_k}(\theta) > m_{\theta_k}(\theta_k)$ must satisfy $\ell(\theta) > \ell(\theta_k)$.

Of course, the essence of an iterative algorithm is the method by which θ_{k+1} is determined from a given θ_k. Suppose that it is possible to construct a function

$m_{\theta_k}(\theta)$ as shown in Figure 1. That is, $m_{\theta_k}(\theta)$ is tangent to $\ell(\theta)$ at the point $\theta = \theta_k$ but never exceeds $\ell(\theta)$. A function $m_{\theta_k}(\theta)$ satisfying the condition

$$m_{\theta_k}(\theta) \leq \ell(\theta) \text{ for all } \theta, \text{ with equality when } \theta = \theta_k \tag{1}$$

is said to *minorize* $\ell(\theta)$ at the point θ_k. Suppose further that it is possible to construct the minorizing function in such a way that it is easy to maximize. It is obvious from looking at Figure 1 that any value of θ that increases the value of $m_{\theta_k}(\theta)$ (above $m_{\theta_k}(\theta_k)$) must also increase the value of the objective function $\ell(\theta)$. In other words, the minorization condition (1) implies the following *ascent property*:

$$m_{\theta_k}(\theta) \;>\; m_{\theta_k}(\theta_k) \quad \text{implies} \quad \ell(\theta) \;>\; \ell(\theta_k). \tag{2}$$

The ascent property (2) suggests that the most logical candidate for θ_{k+1} is the maximizer of $m_{\theta_k}(\theta)$. Thus, an MM algorithm proceeds by first selecting a starting value, θ_0, and then alternately constructing a minorizing function $m_{\theta_k}(\theta)$ and maximizing it to give θ_{k+1} for $k = 0, 1, \ldots$. The ascent property (2) implies that $\ell(\theta_0) \leq \ell(\theta_1) \leq \cdots$, which in turn implies (so long as $\ell(\theta)$ is bounded above) that $\lim_k \ell(\theta_k)$ exists and is finite.

Note that it is not necessarily true that $\lim_k \theta_k$ exists, though in practice it usually happens that the iterates θ_k do eventually satisfy some convergence criterion. Even if $\lim_k \theta_k$ exists, it is not guaranteed to be a local maximizer of $\ell(\theta)$, let alone a global maximizer. Fortunately, however, examples in which iterates fail to converge or in which the limit exists but is a saddle point or even a local minimum are quite rare. For the interested reader, several examples of EM algorithms that exhibit these sorts of pathological behaviors are provided by McLachlan and Krishnan (1997). More general versions of MM algorithms have a long history and have been studied under different names, most notably "majorization" or "iterative majorization" algorithms. For surveys of some of this work, see the articles by Heiser (1995) or Lange, Hunter, and Yang (2000).

3 What *is* an EM algorithm?

It may come as a surprise that a class of algorithms as widely studied as EM has become does not have a universally accepted definition. Certainly, the E-step and the M-step are well-defined for a given parametric model for the complete and observed data—however, not all iterative algorithms that alternately apply these two steps are considered universally to be EM algorithms(!) This fact might be due to a passage in the original Dempster, Laird, and Rubin (1977) paper: In discussing the "final level of generality" of an EM algorithm, those authors state, "In particular, we assume that $[f_\theta(X) > 0]$ almost everywhere in [the sample space] for all $[\theta \in \Omega]$." Although this assumption is sufficient for the all-important ascent property of an EM algorithm, it is not necessary, as shown below. Thus, EM algorithms can be usefully extended to the case in which the likelihood function is zero for some $\theta \in \Omega$, as long as the ascent property is preserved. In this article, the term "EM algorithm" is used in this more general sense.

An EM algorithm can arise whenever the data observed in an experiment, say Y, are considered to be some function $Y = t(X)$ of a random vector X, referred to as the complete data. Suppose that the distributions of X and Y come from (usually different) parametric families indexed by a parameter $\theta \in \Omega$; let $f_\theta(x)$ and $g_\theta(y)$ denote the densities of X and Y, respectively (as shown in Lemma 1 in the appendix, a version of $g_\theta(y)$ and its dominating measure may be determined from $f_\theta(x)$ and $t(x)$). We do not dwell here on the measures that dominate these densities, leaving some of these technical considerations to the appendix. We will assume, however, that we are justified in considering the function

$$h_\theta(x \mid y) = \begin{cases} f_\theta(x)/g_\theta(y) & \text{if } g_\theta(y) > 0 \text{ and } y = t(x) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

to be a density for the conditional distribution of X given $Y = y$. Justification for this assumption in most cases may be found in Tjur (1980).

The above setup defines an EM algorithm in the usual way: Denote the observed value of Y by y and denote the starting candidate for the parameter θ by θ_0, chosen so that $g_{\theta_0}(y) > 0$. In particular, we do *not* make the unnecessarily strong assumption that $g_\theta(y) > 0$ for all $\theta \in \Omega$. At the $(k+1)$th iteration of the EM algorithm, $k \geq 0$, the E-step defines the function

$$Q_{\theta_k}(\theta) = \mathrm{E}_{\theta_k}\left[\log f_\theta(X) \mid Y = y\right], \tag{4}$$

in which the random variable X on the right hand side of equation (4) has density $h_{\theta_k}(x \mid y)$. Next, the M-step sets

$$\theta_{k+1} = \arg\max_\theta Q_{\theta_k}(\theta). \tag{5}$$

To reveal EM algorithms as instances of MM algorithms, we first define the function $\ell(\theta) = \log g_\theta(y)$ to be the observed data log-likelihood and then show how to minorize it. Note that $\ell(\theta)$ should be regarded as a function that can take the value $-\infty$ on Ω, since $g_\theta(y)$ might be zero. Allowing the functions $\ell(\theta)$ and $Q_{\theta_k}(\theta)$ to take the value $-\infty$ for some values of θ does not create any technical difficulties since such θ are automatically ignored by the algorithm. In other words, the actual values of $\ell(\theta)$ and $Q_{\theta_k}(\theta)$ whenever

$$Q_{\theta_k}(\theta) < Q_{\theta_k}(\theta_0) \tag{6}$$

are irrelevant, except insofar as inequality (6) is satisfied.

The function $Q_{\theta_k}(\theta)$ is not itself a minorizer of $\ell(\theta)$; however, we can translate $Q_{\theta_k}(\theta)$ by a quantity not depending on θ and create a minorizer, since adding a constant in this way does not change the M-step. Thus, if we set

$$m_{\theta_k}(\theta) = Q_{\theta_k}(\theta) + \ell(\theta_k) - Q_{\theta_k}(\theta_k),$$

then $m_{\theta_k}(\theta)$ minorizes $\ell(\theta)$ at θ_k.

It is obvious that $m_{\theta_k}(\theta_k) = \ell(\theta_k)$, so it remains only to show that $m_{\theta_k}(\theta) \leq \ell(\theta)$ for all $\theta \in \Omega$. For θ satisfying $g_\theta(y) > 0$, this is an immediate consequence of Jensen's inequality using the fact that the conditional density for X given $Y = y$ is given by equation (3). For details of the well-known proof, see, for example, McLachlan and Krishnan (1997). It remains to check, however, that the inequality $m_{\theta_k}(\theta) > \ell(\theta)$ cannot occur if $g_\theta(y) = 0$. This is proven in Lemma 2 in the appendix, which shows that $g_\theta(y) = 0$ implies $Q_{\theta_k}(\theta) = -\infty$.

We conclude this section by observing that when $t(x)$ is the identity function, the conditional distribution of X given $Y = y$ is obviously the degenerate distribution supported on $\{y\}$. Therefore, an EM algorithm in this case merely sets θ_1 equal to the oberved data MLE, regardless of the value of θ_0, and then $\theta_k = \theta_1$ for all $k \geq 1$. As an example, let X be a random sample from uniform $(0, \theta)$, where $\theta \in \Omega = (0, \infty)$. It is well known that the MLE in this case is the maximum observed value, so clearly if $Y = X$, the resulting EM algorithm converges to this value in a single iteration. Yet the likelihood function is zero for any θ less than the MLE. That $\ell(\theta)$ can equal $-\infty$ is no more reason to consider EM algorithms inapplicable to this case than it is to consider the method of maximum likelihood itself inapplicable.

4 Litebulbs and Heavybulbs

This section illustrates the geometric aspects of EM algorithms with three examples that derive from the simple, elegant framework presented by Flury and Zoppè (2000), who imagined experiments involving *litebulbs* and *heavybulbs*. The lifetimes of litebulbs are exponentially distributed with mean θ, whereas the lifetimes of heavybulbs are uniformly distributed on $(0, \theta)$. The parameter θ is unknown and assumed to lie in $\Omega = (0, \infty)$.

As in Flury and Zoppè (2000), suppose that the experiments involve two rooms, here labeled A and B. In each room a group of litebulbs or heavybulbs is switched

on at time 0 and allowed to expire. Let m and n denote the numbers of bulbs in rooms A and B, respectively. The lifetimes $B_1, \ldots, B_n$ of the bulbs in room B are observed exactly. However, in room A, the observations are censored: The observer looks into room A only once, at some fixed time τ, and merely counts how many of the m bulbs are still on. Let Z denote this number.

The goal in each example is to compute the maximum likelihood estimator of θ using an EM algorithm. We make the "obvious" choice for the complete data in these examples: If the lifetimes of the bulbs in room A are $A_1, \ldots, A_m$, then the complete data will be defined to be $X = (A_1, \ldots, A_m, B_1, \ldots, B_n)$. We observe only $Y = (Z, B_1, \ldots, B_n)$, where Z depends on the constant τ and equals $\sum_{i=1}^{m} I\{A_i \geq \tau\}$. Let $y = (z, b_1, \ldots, b_n)$ denote the observed value of the random vector Y.

4.1 Litebulbs only

Consider the first example of Flury and Zoppè (2000), in which both rooms contain only litebulbs. It is not hard to verify that the observed data log-likelihood is

$$\ell(\theta) = -n \ln \theta - \frac{1}{\theta}[n\overline{b} + z\tau] + (m - z) \log \left(1 - e^{-\tau/\theta}\right) \tag{7}$$

and the complete data log-likelihood is

$$\log f_\theta(X) = -(m + n) \log \theta - \frac{1}{\theta}\left[n\overline{B} + m\overline{A}\right]. \tag{8}$$

Conditional on $Y = (z, b_1, \ldots, b_n)$, and denoting the current parameter estimate by θ_k, taking expectations in equation (8) gives (after a nice exercise in conditional probability)

$$Q_{\theta_k}(\theta) = -(m + n) \log \theta - \frac{n\overline{b} + z\tau + m\theta_k + \tau(m - z)/(1 - e^{\tau/\theta_k})}{\theta}. \tag{9}$$

Suppose that we take $n = m = 10$ and $\tau = 2$, and that we observe $z = 5$ and $\overline{b} = 1$. Then the observed data log-likelihood is shown in Figure 2 along with the minorizing curve $m_{\theta_k}(\theta)$ and its translate $Q_{\theta_k}(\theta)$ for the choice $\theta_k = 2.5$.

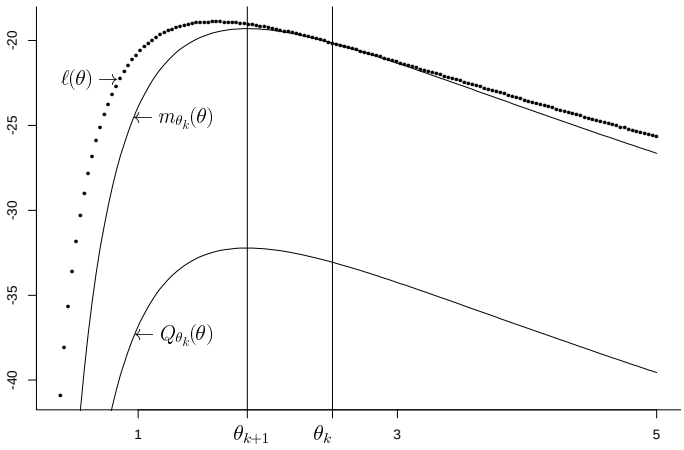

Figure 2: The upper curve is the observed data log-likelihood $\ell(\theta)$ for the case in which all the bulbs in both rooms are litebulbs. Also shown are the E-step function $Q_{\theta_k}(\theta)$ and its translate, the minorizing function $m_{\theta_k}(\theta)$, for the choice $\theta_k = 2.5$. The vertical lines indicate the locations of θ_k and $\theta_{k+1} = 1.842$.

4.2 Heavybulbs only

Consider the same experiment as in the previous subsection, but this time with only heavybulbs in Rooms A and B. This is the second example considered by Flury and Zoppè (2000). In this case, the likelihood functions (both observed data and complete data) will be zero for some positive values of θ, and for this reason Flury and Zoppè argue that no EM algorithm may be defined. Nonetheless, we show here that an EM algorithm with the crucial ascent property *can* be constructed—however, it fails to yield the maximum likelihood estimator. The geometric interpretation of EM will help us to diagnose what goes wrong, and in the next subsection we show that a slight modification of the problem results in an EM algorithm that works perfectly despite the vanishing likelihood functions.

The complete data log-likelihood is

$$\log f_\theta(X) = \begin{cases} -(m+n)\log\theta & \text{if } \theta \geq \max\{A_{\max}, B_{\max}\} \\ -\infty & \text{otherwise.} \end{cases} \tag{10}$$

The case $Z = 0$, where no heavybulbs are still on in Room A at time τ, is not very illuminating (so to speak). Thus, suppose $Z \geq 1$ and define $k = \max\{b_{\max}, \tau\}$. Since clearly θ must be at least as large as k, it is logical to start this EM algorithm at some point $\theta_0 > k$. We will show that the EM algorithm in this example converges to θ_0—in other words, it never goes anywhere once initialized!

Under the assumption that θ_0 is the true value of the parameter, any heavybulb still on in room A at time τ has a uniform (τ, θ_0) distribution, which means that the conditional distribution of $A_{\max}$ given $Z \geq 1$ places positive mass on any open subinterval of (τ, θ_0). Therefore, if $\theta < \theta_0$, there is positive probability that $A_{\max} > \theta$, so clearly if one takes expectations in equation (10) the result is

$$Q_{\theta_0}(\theta) = \begin{cases} -(m+n)\log\theta & \text{if } \theta \geq \theta_0 \\ -\infty & \text{if } 0 < \theta < \theta_0. \end{cases} \tag{11}$$

Since $Q_{\theta_0}(\theta)$ is strictly decreasing on $[\theta_0, \infty)$ and strictly less than $Q_{\theta_0}(\theta_0)$ on $(0, \theta_0)$, setting θ_1 equal to the maximizer of $Q_{\theta_0}(\theta)$ gives $\theta_1 = \theta_0$. By induction, this EM algorithm is forever stuck at the initial value.

Since Z is a binomial random variable with parameters m and $1 - \tau/\max\{\theta, \tau\}$, the observed data log-likelihood is

$$\ell(\theta) = \begin{cases} z\log(1-\tau/\theta) + (m-z)\log(\tau/\theta) & \\ \qquad + \log\binom{m}{z} - n\log\theta & \text{if } \theta \geq b_{\max} \text{ and } \theta > \tau \\ -n\log\theta & \text{if } \theta \leq \tau \text{ and } z = 0 \\ -\infty & \text{otherwise.} \end{cases} \tag{12}$$

This log-likelihood and its minorizer $m_{\theta_0}(\theta)$, obtained by translating the $Q_{\theta_0}(\theta)$ function of equation (11), are shown in Figure 4.2 for the case $m = n = 10$, $z = 5$, $b_{\max} = 1$, $\tau = 1/2$, and $\theta_0 = 2$. The minorizing function $m_{\theta_0}(\theta)$ is nondifferentiable at $\theta = \theta_0$ in this example. If both $\ell(\theta)$ and $m_{\theta_0}(\theta)$ were differentiable at θ_0, on

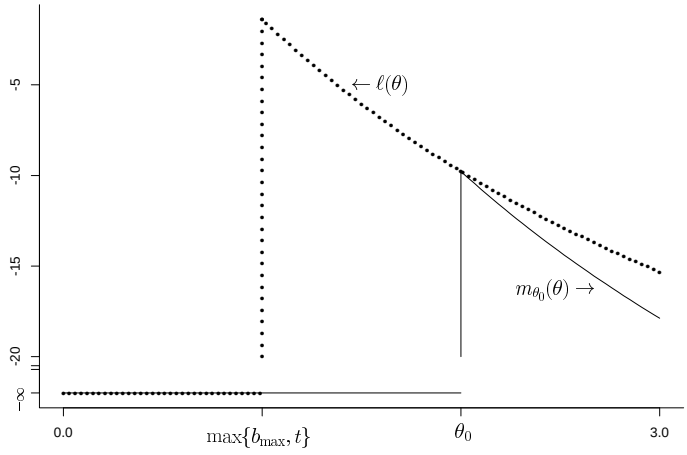

Figure 3: The function $m_{\theta_0}(\theta)$, while guaranteed to minorize $\ell(\theta)$ at θ_0 as shown, is not guaranteed to be differentiable or even continuous at θ_0. Even though the ascent property (2) is vacuously true, the EM algorithm here never leads to an increase in $\ell(\theta)$ because the maximizer of $m_{\theta_0}(\theta)$ is θ_0 itself.

the other hand, these derivatives would coincide by the tangency condition implied by the minorization relationship between the two functions. In that case, it would always be possible to increase the value of $m_{\theta_0}(\theta)$, and hence $\ell(\theta)$, unless the derivative of $\ell(\theta)$ itself were zero at θ_0. The fact that the EM algorithm in this example gets stuck at θ_0, therefore, has to do with the nondifferentiability of $m_{\theta_0}(\theta)$ at θ_0, since $m_{\theta_0}(\theta)$ can in this way have a maximum at θ_0 while $\ell(\theta)$ does not.

4.3 Mixed bulbs

We now extend the ideas of Flury and Zoppè (2000) to create an example (other than the "trivial" example based on $t(x) = x$ given in the last paragraph of section 3) in which the likelihood functions vanish for some $\theta \in \Omega$, yet a correctly defined EM algorithm succeeds in converging to the correct maximum likelihood estimator.

Consider the same experiment as in the previous two subsections, but this time

with litebulbs in Room A and heavybulbs in Room B. That is, the n lifetimes that are actually observed (in Room B) are uniform on $(0, \theta)$ and the m that are censored (in Room A) are exponential with mean θ. Then the observed data log-likelihood and the complete data log-likelihood, respectively, are

$$\ell(\theta) = \begin{cases} -n \log \theta + (m - z) \log(1 - e^{-t/\theta}) - zt/\theta & \text{if } \theta \geq b_{\max} \\ -\infty & \text{if } 0 < \theta < b_{\max} \end{cases} \tag{13}$$

and

$$\log f_\theta(X) = \begin{cases} -(m + n) \log \theta - m\overline{A}/\theta & \text{if } \theta \geq B_{\max} \\ -\infty & \text{if } 0 < \theta < B_{\max}. \end{cases} \tag{14}$$

As in equation (9), taking conditional expectations gives

$$Q_{\theta_k}(\theta) = \begin{cases} -(m + n) \log \theta - c_k/\theta & \text{if } \theta \geq b_{\max} \\ -\infty & \text{if } 0 < \theta < b_{\max}, \end{cases} \tag{15}$$

where

$$c_k = zt + m\theta_k + \frac{t(m - z)}{1 - e^{t/\theta_k}}. \tag{16}$$

Thus, with starting value $\theta_0 > b_{\max}$, the EM algorithm for this example sets

$$\theta_{k+1} = \begin{cases} c_k/(m + n) & \text{if } b_{\max} < c_k/(m + n) \\ b_{\max} & \text{otherwise} \end{cases}$$

for $k \geq 0$. Figure 4 depicts this situation when $\theta_k = 2$, $m = n = 10$, $z = 5$, $t = 3.5$, and $b_{\max} = 1$. Note that if $b_{\max}$ is the MLE, then $Q_k(\theta)$ has a discontinuity (at $b_{\max}$) just as in the example of subsection 4.2, but in this case EM works perfectly. Thus, differentiability of $Q_k(\theta)$ at θ_k is not a necessary condition for a working EM algorithm.

5 Missing data and curvature

It only seems fair, somehow, that an EM algorithm should converge faster in problems with little missing data than in problems with much missing data (indeed,

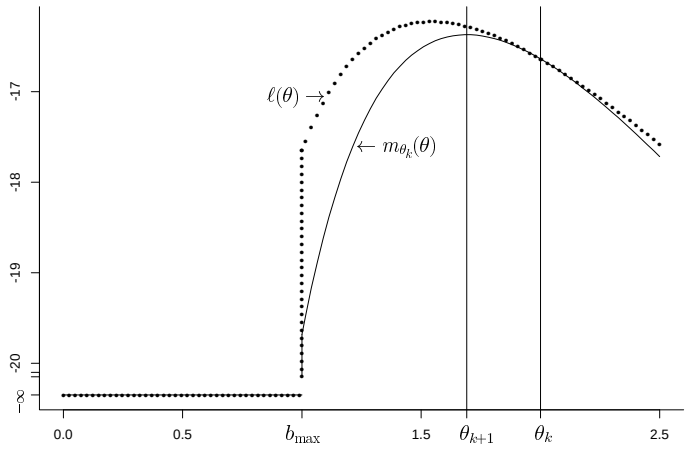

Figure 4: Even though the log-likelihood is undefined for $\theta < b_{\max}$, the EM algorithm performs perfectly in this example. Note that this would still be true if $b_{\max}$ were, say, 1.75, in which case the EM would correctly converge to $b_{\max}$ itself. The vertical lines are drawn at $\theta_k = 2$ and $\theta_{k+1} = 1.691$.

it was argued at the end of Section 3 that when there is no missing data an EM algorithm converges in a single iteration). This section explores how the geometric interpretation of an EM algorithm can aid the understanding of this phenomenon.

It is difficult to quantify the "amount of missing data," but one good way to do it comes from the original Dempster, Laird, and Rubin (1977) paper. The matrix

$$M = \frac{\partial^2 \ell(\theta)}{\partial \theta^2} \left[\frac{\partial^2 m_{\hat{\theta}}(\theta)}{\partial \theta^2} \right]^{-1} \Bigg|_{\theta=\hat{\theta},}$$

where $\hat{\theta}$ denotes the maximum likelihood estimator, makes sense as a measure of the amount of missing data because the second derivates of $\ell(\theta)$ and $m_{\hat{\theta}}(\theta)$, when evaluated at the MLE, give estimates of the (negative) Fisher information about θ contained in the observed data and the complete data, respectively. It is possible to show that $\|\theta_k - \hat{\theta}\|/\|\theta_{k-1} - \hat{\theta}\| \to 1 - \lambda_s$ as $k \to \infty$, where λ_s is the smallest

eigenvalue of M. If there is a great deal of missing data, then $1 - \lambda_s$ will be close to 1 and the convergence of the EM algorithm will be very slow. In the univariate case, M is a scalar and $\lambda_s = M$.

It is easy to understand qualitatively why more missing data leads to slower convergence when we consider that the Hessian matrix of second derivatives may be interpreted as the curvature of a function. If the amount of missing data is large, then $\ell(\theta)$ is much flatter than the minorizer $m_{\theta_k}(\theta)$ and so $m_{\theta_k}(\theta)$ drops away from $\ell(\theta)$ very quickly. But the quicker the minorizer drops away, the smaller the size of the step taken in the M-step when the minorizer is maximized. Smaller steps lead to slower convergence, just as we expect.

Compare two similar examples, one in which the amount of missing data is relatively high and one in which it is relatively low. The two examples are taken from the situation of subsection 4.1, in which each litebulb has exponential lifetime with mean θ. There are n lifetimes fully observed (with sample mean $\overline{b}$) and m lifetimes for which we only know the number z that exceed t. For the purpose of easy visual comparison, both examples have been constructed (by carefully controlling the value of $\overline{b}$) to have the same maximum likelihood estimate, namely $\hat{\theta} = 2$. The parameters and observed data for each of the two examples are summarized below.

Example	m	n	t	z	$\overline{b}$	$\hat{\theta}$
More missing data	18	2	2	7	1.401745	2
Less missing data	2	18	2	1	1.953553	2

Since the values of $\ddot{\ell}(\theta_k)$ and $\ddot{m}_{\theta_k}(\theta_k)$ don't change much as θ_k nears $\hat{\theta}$, we can visualize the effect of different amounts of missing data by plotting $\ell(\theta)$ and $m_{\theta_k}(\theta)$ for θ_k near $\hat{\theta} = 2$. The resulting plots are given in Figure 5. In the "more missing" example, the true value of $\ddot{\ell}(\hat{\theta})/\ddot{m}_{\hat{\theta}}(\hat{\theta})$ equals 0.6103, and the plot shows that the step from θ_k to θ_{k+1} covers about 60% of the distance between θ_k and $\hat{\theta}$ (the exact figure is 60.80%). The corresponding value of $\ddot{\ell}(\hat{\theta})/\ddot{m}_{\hat{\theta}}(\hat{\theta})$ for the "less missing" example is 0.94603, and the exact value of $\|\theta_k - \theta_{k+1}\|/\|\theta_k - \hat{\theta}\|$ is 94.618%.

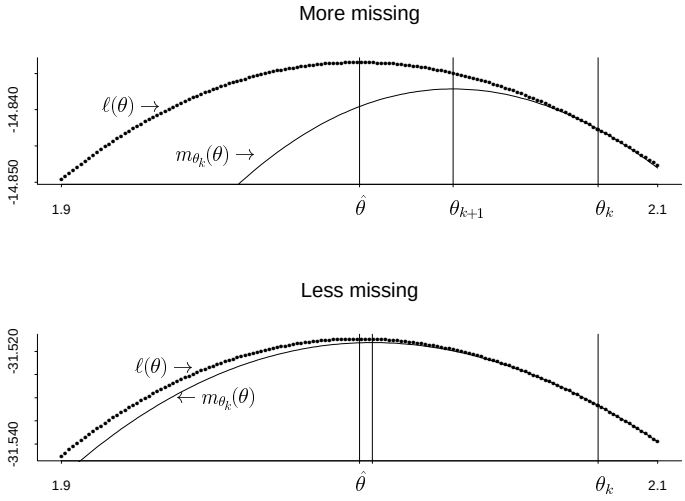

Figure 5: In both examples, the vertical lines mark the values of $\hat{\theta} = 2$, $\theta_k = 2.08$, and θ_{k+1}. In the top graph, $\theta_{k+1} = 2.031$; in the bottom graph, $\theta_{k+1} = 2.004$.

We leave it as an easy exercise for the reader to verify that if $\ell(\theta)$ and $m_{\theta_k}(\theta)$ are quadratic functions and θ a scalar, then the value of $(\theta_k - \theta_{k+1})/(\theta_k - \hat{\theta})$ is exactly $\ddot{\ell}(\hat{\theta})/\ddot{m}_{\theta_k}(\hat{\theta})$.

6 Conclusion

It has been the aim of this paper to exploit the simple geometric intuition underlying MM algorithms, of which EM algorithms are special cases, to help clarify some of the properties of EM algorithms. The graphical depiction of the observed data loglikelihood and its associated minorizing function for several simple examples gives useful insights into how the ascent property of the EM algorithm works and why the ascent property alone does not guarantee convergence to a maximizer of the observed data loglikelihood. Furthermore, the geometry gives an intuitive sense of how the amount of missing data, as represented by the ratio of curvatures of the

two functions, influences the convergence rate of an EM algorithm.

Finally, this paper argues for a broad interpretation of EM algorithms that does not exclude them unnecessarily from problems in which the likelihood function vanishes on some subset of the parameter space. Although this interpretation is at odds with some literature on EM algorithms, nothing prevents this interpretation and nothing is harmed by it.

7 Appendix: Technical notes

Here we prove the two lemmas referred to in section 3.

Lemma 1 *For any random variable X with density $f_\theta(x)$ with respect to measure μ, and any measurable function $t(x)$, there exists a function $g_\theta(y)$ and a measure ν such that $g_\theta(y)$ is a density for $Y = t(X)$ with respect to ν.*

The proof is simple: Define

$$\hat{\nu}(B) \stackrel{\text{def}}{=} \mu[t^{-1}(B)] = \mu[\{x : t(x) \in B\}].$$

Since $P_\theta(Y \in B) = 0$ whenever $\nu(B) = 0$, as may be easily checked, the density $g_\theta(y)$ of the observed data may be taken to be the Radon-Nikodym derivative of the distribution of $Y = t(X)$ with respect to ν. ■

Note that although it may be appealing to take

$$g_\theta(y) = \int_{t^{-1}(y)} f_\theta(x)\, dx$$

as in equation (1.1) of Dempster, Laird, and Rubin (1977), this definition is not always correct, as when X is absolutely continuous and $t^{-1}(y)$ is finite for all y.

Lemma 2 *If $g_\theta(y) = 0$ for some θ, then $Q_{\theta_k}(\theta) = -\infty$.*

To prove this lemma, we demonstrate that the function

$$f_\theta^*(x) = \begin{cases} f_\theta(x) & \text{if } g_\theta[t(x)] > 0 \\ 0 & \text{if } g_\theta[t(x)] = 0 \end{cases} \tag{17}$$

is a density for X. In other words, we may assume without loss of generality that $f_\theta(x) = 0$ whenever $g_\theta[t(x)] = 0$, so the result is obvious.

Let Z_θ denote the set $\{x : g_\theta[t(x)] = 0\}$ and observe that

$$P_\theta[X \in Z_\theta^c] = 1 - P_\theta[Y \in t(Z_\theta)] = 1.$$

Since $f_\theta(x)$ and $f_\theta^*(x)$ agree on Z_θ^c, the proof is complete. ■

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. Ser. B*, **39**, 1–38.

Flury, B. and Zoppè, A. (2000). Exercises in EM, *The American Statistician*, **54**, 207–209.

Heiser, W. J. (1995). Convergent computing by iterative majorization: Theory and applications in multidimensional data analysis, in *Recent Advances in Descriptive Multivariate Analysis*, ed. W. J. Krzanowski. Oxford: Clarendon Press, pp. 157–189.

Hunter, D. R. and Lange, K. (2000). Rejoinder to discussion of "Optimization transfer using surrogate objective functions," *J. Comp. Graph. Stat.*, **9**, 52–59.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions, *J. Comp. Graph. Stat.*, **9**, 1–59.

McLachlan, G. and Krishnan, T. (1997) *The EM Algorithm and Extensions*, New York: Wiley.

Tjur, T. (1980) *Probability Based on Radon Measures*, Chichester: Wiley.