October 22, 1986

BRIEF COMMENTS ON THE EM ALGORITHM

Donald L. Snyder

Electronic Systems and Signals Research Laboratory
Department of Electrical Engineering
Washington University
St. Louis, MO 63130

Abstract

The EM algorithm is being used in several research projects in ESSRL. An overview of this algorithm, references for more careful study of it, and some simple examples using it are given.

I. Introduction

The ''EM'' in ''EM algorithm'' stands for ''expectation-maximization.'' As you might guess from its name, the algorithm proceeds by evaluating an expectation and then performing a maximization. The EM algorithm is but one numerical approach for determining maximum-likelihood estimates of parameters from measured data. It does so recursively, evaluating an expectation then a maximization then an expectation then a maximization then an expectation ... and so on over and over. It is an algorithm with some advantages and some disadvantages compared to alternative algorithms for finding maximum-likelihood estimates numerically.

The EM algorithm is being used in several research projects in ESSRL and its collaborating laboratory, BCL. Some of these projects are as follows.

a. Single-Photon Emission Computed Tomography. The EM algorithm is being used to produce estimates of the spatial distribution of a radioactive tracer that emits a single photon with each radioactive decay event.

b. Positron-Emission Tomography. The EM algorithm is being used to produce estimates of the spatial distribution of a radioactive tracer that results in two photons with each radioactive decay event. It is also used to estimate parameters related to the time dependent behavior of transport and decay of the radiotracer.

c. auditory electrophysiology. The EM algorithm is being used to estimate parameters of the discharge rate of the auditory nerve in response to acoustic stimuli.

d. probability density estimation. The EM algorithm is being used to estimate the probability density of a random variable from noisy measurements of the variable.

e. power-density spectrum estimation. The EM algorithm is being used to estimate the power density of stationary and periodic random processes.

f. direction finding. The EM algorithm is being used to estimate the spatial location of signal sources from data collected with spatially distributed array of sensors.

g. radar-imaging. The EM algorithm is being used to estimate radar images from delay-doppler radar data.

h. magnetic-resonance imaging. The EM algorithm is being used to estimate signals in magnetic resonance imaging.

i. electron-microscopic autoradiography. The EM algorithm is being used to estimate parameters in images obtained in electron-microscopic autoradiography.

The purpose of these notes is to give a brief introduction to the EM algorithm because it is being used in so many ESSRL projects and is not regularly covered in courses.

II. References

Here is a short collection of references that discuss the EM algorithm. It is biased towards publications that have been important to ESSRL projects or reporting on them. Additional references are contained in them.

1. A. P. Dempster, N. M. Laird, and D. B. Rubin, ''Maximum Likelihood from Incomplete Data via the EM Algorithm,'' J. Royal Statistical Society, B., Vol. 39, pp. 1-37, 1977. This is a classic reference in which the algorithm was first defined in a unified manner. It is a ''must read'' for anyone doing research where the EM algorithm is used.

2. C. F. Wu, ''On the Convergence Properties of the EM Algorithm,'' Ann. Statis., Vol. 11, pp. 95-103. This is an important paper in which conditions for the convergence of the EM algorithm are given. It corrects an error in the convergence proof in DLR (reference 1).

3. L. A. Shepp and Y. Vardi, ''Maximum Likelihood Reconstruction for Emission Tomography,'' IEEE Trans. on Medical Imaging, Vol. MI-1, pp. 113-121, October 1982. This paper contains the gives the first use of the EM algorithm in emission tomography.

4. D. L. Snyder and D. G. Politte, ''Image Reconstruction from List-Mode Data in an Emission Tomography System Having Time-of-Flight Measurements,'' IEEE Trans. on Nuclear Science, Vol. NS-20, pp. 1843-1849, June 1983. This is the first paper describing the use of the EM algorithm in the positron-emission tomography project at BCL.

5. D. L. Snyder, ''parameter estimation in dynamic studies.'' This develops an approach for using the EM algorithm to estimate parameters from measurements that can be viewed as a superposition of several separate components.

6. D. L. Snyder and M. I. Miller, ''The Use of Sieves to Stabilize Images Produced with the EM Algorithm for Emission Tomography,'' IEEE Trans. on Nuclear Science, Vol. NS-32, pp. 3864-3872, October 1985. This describes our suggested use of Grenander's sieves with the EM algorithm for reducing a fundamental noise artifact encountered with maximum likelihood estimation.

7. U. Grenander, <u>Abstract Inference</u>, John Wiley, 1981. The ''bible'' on sieves. It's tough but essential if you get serious about sieves.

8. S. Geman and C.-R. Hwang, ''Nonparametric Maximum Likelihood Estimation by the Method of Sieves,'' The Annals of Statistics, Vol. 10, pp. 401-414, 1982. This is a quite readable account about sieves along with some examples.

9. R. A. Tapia and J. R. Thompson, Nonparametric_Probability_Density_Estimation, John Hopkins Univ. Press, 1978. This gives a good account of penalty type sieves and the idea of dimensional instability.

10. M. I. Miller, K. B. Larson, J. E. Saffitz, D. L. Snyder, and L. J. Thomas, Jr., ''Maximum-Likelihood Applied to Electron-Microscopic Autoradiography,'' J. of Electron Microscopy Technique, April 1985. This paper documents the use of the EM algorithm in the EMA project at BCL.

11. M. I. Miller, ''Algorithms for removing Recovery Related Distortion from Auditory-Nerve Discharge Patterns,'' J. Acoust. Soc. America, Vol. 77, pp. 1452-1464, 1985. This paper documents the use of the EM algorithm in the auditory electrophysiology project.

12. M. I. Miller and D. L. Snyder, ''The Application of Maximum-Entropy and Maximum-Likelihood for the Solution of Incomplete and Noisy Data Problems in Estimating Point-Process Intensities, Probability Densities, and Spectral Densities,'' in review for publication in the IEEE Proceedings. This manuscript gives an overview of the use of the EM algorithm in several ESSRL projects, including power density spectrum estimation.

13. M. I. Miller, D. L. Snyder, and T. R. Miller, ''Maximum Likelihood Reconstruction for Single-Photon Emission Computed Tomography,'' IEEE Trans. on Nuclear Science, Vol. NS-32, pp. 769-778, February 1985. This describes our approach being used on the SPECT project.

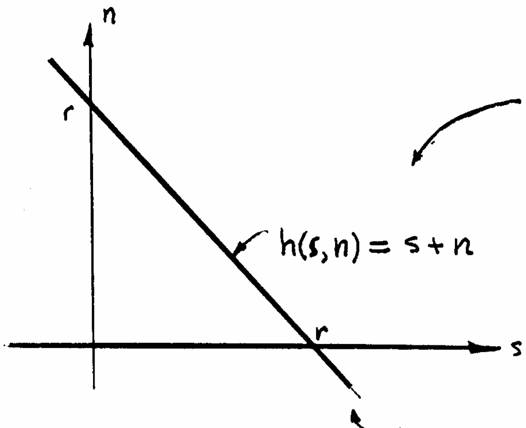
14. R. H. Shumway, ''Some Applications of the EM Algorithm to Analyzing Incomplete Time Series Data.'' I have misplaced the source of this paper, but it is a good one about the estimation of parameters of finite dimensional Gaussian processes. See me if you want to browse through the paper or make a copy.

III. Discussion

It is useful to think of two data spaces for understanding how the EM algorithm works. The first data space is called the incomplete-data space. This is the space where measured data takes its values. For example, if measured data are in the form of points along a line, then the incomplete data space is the line, and the data are in the form of the locations of the points on the line. If the measurements are described as a sample of a Gaussian process, then the incomplete data space is the real line, R^1, and the data is a real number. The second space of importance is called the complete-data space. This is harder to define because it is a hypothetical or contrived space created by the user of the EM algorithm. There are usually several candidate ways to select a complete-data space; all have to result in the same final answer, but some may be considerably easier to use than others. The ''art'' in using the EM algorithm is in selecting this hypothetical, complete-data space to make the problem easy in some sense. The way that has proven most successful in all ESSRL projects is to rely on a good mathematical model for the measured data in the incomplete-data space, a model that describes both how the data are generated physically and how the instrumentation used to measure it operates.

One requirement in defining the complete and incomplete data spaces is that there must be some known function of the complete data that uniquely gives the incomplete data. Here are two examples.

<u>Example A. signal in additive noise</u>. Suppose s is a random variable modeling a signal sample, n is a random variable modeling a noise sample, and r = s + n is measured. The incomplete data space is R^1, the real line. One choice of complete data is the pair (s,n). In this case, the complete data space is R^2, the plane. The function h(s,n) = s + n maps the complete data space into the incomplete data space. The incomplete data, r, is uniquely defined by the complete data, (s,n), and this function, h. On the other hand, the incomplete data, r, and the function, h, place a constraint on the possible values of the incomplete data -- only those (s,n) such that s+n equals the measured incomplete data are possible. The function h defines a many-to-one mapping between the complete and incomplete data spaces. Knowing the complete data specifies the incomplete data uniquely, but knowing the incomplete data only specifies a set of possible complete data. This is illustrated in the Figure 1. Another possible choice for the complete data is the pair (r,s), and the mapping is then defined by h(r,s) = r.



This (n,s)-plane is the complete-data space.

This line is the incomplete-data space.

<u>Example B. pooled points of two point processes</u>. Suppose that there are two homogeneous point processes in time; the first has intensity a and the second intensity b. The points of the two processes are pooled to form a single point process with intensity a + b. The occurrence times of points of the pooled process are observed. Here the incomplete data are the measured occurrence times $\{t_1, t_2, ..., t_N\}$ of the pooled process. For the complete data, we can imagine for one choice that each point has an auxiliary mark indicating whether it came from process a or process b. Then the

complete data would be the marked process $\{(t_1,m_1), (t_2,m_2), \ldots, (t_N,m_N)\}$, where m_i is the mark (either a or b) on the $\underline{i\text{th}}$ point. The mapping between the complete and incomplete spaces is $h[(t_1,m_1), (t_2,m_2), \ldots, (t_N,m_N)] = (t_1, t_2, \ldots, t_N)$.

The next important idea is that of the likelihood function on each data space. We imagine that there are some parameters to be estimated, and we denote the collection of these parameters by the vector $\underline{\lambda}$. Let p_{id}(incomplete data$|\underline{\lambda}$) and p_{cd}(complete data$|\underline{\lambda}$) denote the probability densities of the incomplete and complete data, respectively. The loglikelihood function is then

$$L_{id}(\underline{\lambda}) = \ln[p_{id}(\text{incomplete data}|\underline{\lambda})]$$

on the incomplete-data space and

$$L_{cd}(\underline{\lambda}) = \ln[p_{cd}(\text{complete data}|\underline{\lambda})]$$

on the complete-data space.

The usual parameter-estimation problem is that we are given the incomplete data and asked to estimate the parameters. The method of maximum likelihood developed in EE552 would proceed by taking as the estimate a $\underline{\lambda}$ that maximizes the incomplete-data loglikelihood, $L_{id}(\underline{\lambda})$. In simple problems, the maximizing value of $\underline{\lambda}$ can be written explicitly as a function of the incomplete data, but usually the maximizing value satisfies a difficult-(if not impossible)-to-solve transcendental equation obtained by setting the gradient of the loglikelihood function to zero. This difficulty can sometimes be circumvented indirectly through use of the EM algorithm.

The EM algorithm is a recursive algorithm that proceeds as follows. Suppose that we have arrived somehow at an estimate of the parameters; initially this might be some wild guess based on intuition or some other information. Denote this estimate at stage-k by $\hat{\underline{\lambda}}^{(k)}$. To get the estimate at stage-(k+1), we first perform an E-step by evaluating the following expectation:

<u>E-step</u>: Evaluate the function $Q(\underline{\lambda}|\hat{\underline{\lambda}}^{(k)})$ defined according to

$$Q(\underline{\lambda}|\hat{\underline{\lambda}}^{(k)}) = E[L_{cd}(\underline{\lambda})|\text{incomplete data},\hat{\underline{\lambda}}^{(k)}].$$

Thus, Q is the conditional expectation of the complete-data loglikelihood given the incomplete data and assuming that the parameters governing the data are the stage-k estimates $\hat{\underline{\lambda}}^{(k)}$. We then perform an M-step.

<u>M-step</u>: Find the parameters $\underline{\lambda}$ that maximize $Q(\underline{\lambda}|\hat{\underline{\lambda}}^{(k)})$ as a function of $\underline{\lambda}$.

The maximizers obtained in the M-step are the stage-(k+1) parameter estimates $\hat{\underline{\lambda}}^{(k+1)}$. The E and M steps are then repeated again and again.

The art in identifying the complete data space is to make the E and M steps easy, or at least easier than the brute force approach of maximizing the incomplete data loglikelihood. The interesting fact is that under not very stringent conditions, the iterates $\hat{\underline{\lambda}}^{(k)}$ converge to the maximum likelihood estimates of $\underline{\lambda}$ in terms of the incomplete data. This provides for an indirect way to determine the $\underline{\lambda}$ that makes the gradient of the incomplete data loglikelihood zero.

IV. Examples

In what follows, we give two examples of using the EM algorithm to determine the maximum-likelihood estimates of some parameters. The examples are ''simple'' so that a direct comparison can be made to the maximum-likelihood estimate determined by maximizing the incomplete-data loglikelihood. The examples look a little complicated compared to the methods developed in EE552; the ''simple'' is in the sense that an explicit expression can be found for the maximum-likelihood estimate as a function of the incomplete data by solving for the parameters that zero the gradient of the incomplete-data loglikelihood. In the ESSRL projects where the EM algorithm is being used, the maximum-likelihood estimate cannot be determined directly in this manner, so some recourse to a numerical algorithm, such as the EM algorithm, is necessary.

Example A (continued). Let us again look at Ex. A in which r = s + n. Assume now that s and n are uncorrelated, Gaussian random-variables with zero means and with variances a and b; i.e., s is $N_s(0,a)$ and n is $N_n(o,b)$, where ''x is $N_x(0,c)$'' means that

$$N_x(0,c) = (2\pi c)^{-1/2} \exp[-s^2/2c]$$

is the p.d.f. of x. We suppose that the noise variance, b, is known. The problem is:

Given: measurement r
noise variance b
the model (r = a + b, a uncorrelated with b, etc.)
Estimate: a

solution from the incomplete-data loglikelihood:

Let us determine the maximum-likelihood estimate of a by the direct maximization of the incomplete-data loglikelihood, as developed in EE552. Since the sum of two zero-mean, uncorrelated Gaussian random-variables is a zero-mean Gaussian random variable whose variance is the sum of the variances, it follows that r is N(0,a+b). Hence, the incomplete-data loglikelihood is

$$L_{id}(a) = - (1/2)\ln(a+b) - r^2/2(a+b).$$

Setting the derivative of $L_{id}(a)$ with respect to a equal to zero shows that $L_{id}(a)$ has a maximum at $a = r^2 - b$, and an examination of the graph of

$L_{id}(a)$ as a function of a shows that $L_{id}(a)$ is convex down. Since a must be nonnegative, we conclude that the maximum-likelihood estimate of a is given by

$$\hat{a}_{ML} = \max(0, r^2 - b).$$

<u>solution from the complete-data loglikelihood via the EM algorithm</u>:

Let us now determine the maximimum-likelihood estimate of a by the indirect method of the EM algorithm. Let (s,n) be the complete data. Since the p.d.f. of (s,n) is $N_s(0,a)N_n(0,b)$, the complete-data loglikelihood is given by

$$L_{cd}(a) = -(1/2)\ln(a) - s^2/2a - (1/2)\ln(b) - n^2/2b.$$

The E-step yields

$$Q(a|\hat{a}^{(k)}) = -(1/2)\ln(a) - E[s^2|r,\hat{a}^{(k)}]/2a,$$

where terms that are not a function of a have been dropped because we will eventually maximize Q with respect to a. The M-step yields

$$\hat{a}^{(k+1)} = E[s^2|r,\hat{a}^{(k)}].$$

We next need to do the hard part, namely evaluate $E[s^2|r,\hat{a}^{(k)}]$. This we do by first noting that since

$$s^2 = [s - \hat{s}^{(k)}]^2 + 2s\hat{s}^{(k)} - [\hat{s}^{(k)}]^2,$$

where we define $\hat{s}^{(k)} = E[s|r,\hat{a}^{(k)}]$, we have

$$E[s^2|r,\hat{a}^{(k)}] = E\{[s-\hat{s}^{(k)}]^2|r,\hat{a}^{(k)}] + [\hat{s}^{(k)}]^2.$$

But, from equations 143 and 144 on page 59 of Van Trees Volume I (the textbook for EE552), we have that

$$E\{[s-\hat{s}^{(k)}]^2|r,\hat{a}^{(k)}] = \hat{a}^{(k)}b/[\hat{a}^{(k)} + b]$$

and

$$\hat{s}^{(k)} = \hat{a}^{(k)}r/[\hat{a}^{(k)} + b].$$

Putting all these expressions together, we conclude that

$$\hat{a}^{(k+1)} = \hat{a}^{(k)}b/[\hat{a}^{(k)} + b] + \{\hat{a}^{(k)}r/[\hat{a}^{(k)} + b]\}^2.$$

The iteration would proceed by selecting some starting value at $k = 0$, say $\hat{a}^{(0)} = 1$, and then calculating $\hat{a}^{(1)}$, $\hat{a}^{(2)}$, ... ad infinitum. In practice, one stops at some finite value of k when terms in the iteration cease to change significantly. Let the limit point be denoted by $x = \hat{a}^{(\infty)}$. This limit point satisfies

$$x = xb/(x+b) + [xr/(x+b)]^2.$$

The solutions are $x = 0$ and $x = r^2 - b$. Since $\hat{a}^{(k)}$ is nonnegative at every stage of the iteration, we conclude that

$$x = \max[0, r^2 - b],$$

which is the maximum-likelihood estimate of a in terms of the incomplete data.

Example B (continued). Let us look again at Ex. B in which the incomplete data are $\{t_1, t_2, ..., t_N\}$ and the complete data are $\{(t_1,m_1), (t_2,m_2), ..., (t_N,m_N)\}$, where m_i is the mark (either a or b) on the ith point. We now assume that the two point processes are independent Poisson processes, with mean parameters a and b, respectively, and that the duration of the observation is T seconds. Let N(T), $N_a(T)$, and $N_b(T)$ denote the total number of observed points and the total number of points with marks a and b, respectively; obviously, $N(T) = N_a(T) + N_b(T)$. Suppose that we get to measure the incomplete data, we know b, and we want to estimate a.

solution from the incomplete-data loglikelihood:

Since the point process obtained by pooling the points of two Poisson processes is also a Poisson process, with a mean parameter that is the sum of the mean parameters of the constituent processes, it follows that the incomplete-data loglikelihood is given by

$$L_{id}(a) = - (a+b)T + \ln(a+b)\ N(T).$$

Setting the derivative of $L_{id}(a)$ with respect to a equal to zero shows that $L_{id}(a)$ has a maximum at $a = (1/T)N(T) - b$, and an examination of the graph of $L_{id}(a)$ as a function of a shows that $L_{id}(a)$ is convex down. Since a must be nonnegative, we conclude that the maximum-likelihood estimate of a is given by

$$\hat{a}_{ML} = \max[0,(1/T)N(T) - b].$$

solution from the complete-data loglikelihood via the EM algorithm:

Let us now determine the maximimum-likelihood estimate of a by the indirect method of the EM algorithm. Since the complete data can be separated into two independent Poisson processes by knowing the marks, the complete-data loglikelihood is given by

$$L_{cd}(a) = - aT + \ln(a)\ N_a(T) - bT + \ln(b)\ N_b(T).$$

The E-step yields

$$Q(a|\hat{a}^{(k)}) = - aT + \ln(a)\ E[N_a(T)|N(T),\hat{a}^{(k)}]$$

where terms that are not a function of a have been dropped because we will eventually maximize Q with respect to a. The M-step yields

$$\hat{a}^{(k+1)} = (1/T)\ E[N_a(T)|N(T),\hat{a}^{(k)}].$$

This expectation can be evaluated as

$$E[N_a(T)|N(T),\hat{a}^{(k)}] = \hat{a}^{(k)}N(T)/[\hat{a}^{(k)} + b].$$

Hence,

$$\hat{a}^{(k+1)} = (1/T)\ \hat{a}^{(k)}N(T)/[\hat{a}^{(k)} + b].$$

The iteration would proceed by selecting some starting value at $k = 0$, say $\hat{a}^{(0)} = 1$, and then calculating $\hat{a}^{(1)}$, $\hat{a}^{(2)}$, ... ad infinitum. In practice, one stops at some finite value of k when terms in the iteration cease to change significantly. Let the limit point be denoted by $x = \hat{a}^{(\infty)}$. This limit point satisfies

$$x = (1/T)\ xN(T)/[x + b].$$

The solutions are $x = 0$ and $x = (1/T)N(T) - b$. Since $\hat{a}^{(k)}$ is nonnegative at every stage of the iteration, we conclude that

$$\hat{a}_{ML} = \max[0,(1/T)N(T) - b].$$

<u>solution from the complete-data loglikelihood via the EM algorithm</u>:

Let us now determine the maximimum-likelihood estimate of a by the indirect method of the EM algorithm. Since the complete data can be separated into two independent Poisson processes by knowing the marks, the complete-data loglikelihood is given by

$$L_{cd}(a) = - aT + \ln(a)\ N_a(T) - bT + \ln(b)\ N_b(T).$$

The E-step yields

$$Q(a|\hat{a}^{(k)}) = - aT + \ln(a)\ E[N_a(T)|N(T),\hat{a}^{(k)}]$$

where terms that are not a function of a have been dropped because we will eventually maximize Q with respect to a. The M-step yields

$$\hat{a}^{(k+1)} = (1/T)\ E[N_a(T)|N(T),\hat{a}^{(k)}].$$

This expectation can be evaluated as

$$E[N_a(T)|N(T),\hat{a}^{(k)}] = \hat{a}^{(k)}N(T)/[\hat{a}^{(k)} + b].$$

Hence,

$$\hat{a}^{(k+1)} = (1/T)\ \hat{a}^{(k)}N(T)/[\hat{a}^{(k)} + b].$$

The iteration would proceed by selecting some starting value at k = 0, say $\hat{a}^{(0)} = 1$, and then calculating $\hat{a}^{(1)}$, $\hat{a}^{(2)}$, ... ad infinitum. In practice, one stops at some finite value of k when terms in the iteration cease to change significantly. Let the limit point be denoted by $x = \hat{a}^{(\infty)}$. This limit point satisfies

$$x = (1/T)\ xN(T)/[x + b].$$

The solutions are x = 0 and x = (1/T)N(T) - b. Since $\hat{a}^{(k)}$ is nonnegative at every stage of the iteration, we conclude that

$$x = \max[0,(1/T)N(T) - b],$$

which is the maximum-likelihood estimate of a in terms of the incomplete data.

V. Conclusion

In general, maximum likelihood estimates of parameters are impossible to determine explicitly because they satisfy a nonlinear, transcendental equation obtained by setting the gradient of the incomplete-data loglikelihood to zero. The EM algorithm provides a numerical approach for determining maximum-likelihood estimates when they cannot be found explicitly.