The EM Algorithm for Maximizing a Marginal Likelihood Function

Steffen L. Lauritzen
Aalborg University

May 12, 2000

Abstract

This note gives the main concepts needed to understand the basic elements of the EM algorithm and its elementary convergence properties.

1 Introduction

When collecting data for graphical models, in particular Bayesian networks, data are often incomplete, in the sense that instead of observing the *full data* $(i^1, \ldots, i^n)$, one has only observed partial or *incomplete data* $(i^1_{A_1}, \ldots, i^n_{A_n})$, where A_i are subsets of the set of variables V.

This makes the problem of calculating the maximum likelihood estimate of the unknown probability p harder.

The EM algorithm is an algorithm that enables this computation to be made iteratively in a way which exploits existing algorithms for maximizing the complete data likelihood. The EM algorithm is old and instances were at least known by Thiele (1880), but its full scope and generality was first noted by Sundberg (1972, 1974) and Dempster *et al.* (1977).

The algorithm alternates between two steps: in the **E-step**, the missing data are estimated by computing a conditional expectation; in the **M-step**, the complete data likelihood is maximized, using the estimated data as the full data.

Under rather general conditions, the algorithm can be shown to converge to a local maximum of the likelihood function. The first correct and complete investigation of the convergence properties of the algorithm was given by Wu (1983). Below we shall indicate why convergence is plausible. We are describing the algorithm in the case where all sample spaces are finite, although the algorithm and proofs are not much different in the general case.

2 Information theoretic concepts

The following concepts are fundamental for many things and will help us understand why the algorithm does its job.

2.1 Entropy

The *entropy* of a discrete probability distribution p is the quantity

$$\mathrm{Ent}(P) = -\sum_i P(i) \log P(i).$$

The entropy is always positive and smaller than the entropy of the uniform distribution. To see the latter, we let $P_u(i) = 1/|\mathcal{I}|$ be the uniform distribution and the information inequality then yields

$$\begin{aligned}
\mathrm{Ent}(P) &= -\sum_i P(i) \log P(i) \\
&\leq -\sum_i P(i) \log \frac{1}{|\mathcal{I}|} \\
&= -\sum_i \frac{1}{|\mathcal{I}|} \log \frac{1}{|\mathcal{I}|} \\
&= -\sum_i P_u(i) \log P_u(i) = \mathrm{Ent}(P_u).
\end{aligned}$$

Note that, clearly, $\mathrm{Ent}(P_u) = \log |\mathcal{I}|$. Distributions with high entropy are less concentrated than distributions with low entropy. For a more complete information theoretic analysis of the concept of entropy, see e.g. Cover and Thomas (1991).

2.2 Kullback–Leibler divergence

The *Kullback–Leibler divergence* (or KL divergence for short) between two probability distributions P and Q is determined as

$$KL(P : Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)},$$

see for example Lauritzen (1996), section A.2. The KL divergence is also known as the *relative entropy* of Q with respect to P (Cover and Thomas 1991).

The information inequality says that $KL(P : Q) \geq 0$ and $KL(P : Q) = 0$ if and only if $P = Q$, so the KL divergence defines a type of (assymmetric) distance measure between probability distributions.

3 The EM algorithm

We denote the observed data by x and the unobserved data by y so that the complete data is (x, y). If θ denotes the unknown parameter to be estimated ($\theta = p$ in discrete graphical models), we introduce the following fundamental likelihood functions: The *complete data log-likelihood* is

$$l(\theta) = \log L(\theta \mid x, y) = \log P(x, y \mid \theta).$$

The *marginal log-likelihood* is

$$l_x(\theta) = \log L(\theta \mid x) = \log P(x \mid \theta).$$

We wish to maximize l_x in θ but generally l_x is an unpleasant function. However, we assume that we know how to maximize l. We let θ^* denote an arbitrary but fixed value of the parameter.

The **E-step** of the algorithm now calculates the (conditionally) *expected complete data log-likelihood* $q(\theta \mid \theta^*)$ as

$$q(\theta \mid \theta^*) = \mathbf{E}_{\theta^*}\{l(\theta) \mid x\} = \sum_y P(y \mid x, \theta^*) \log P(x, y \mid \theta).$$

The general applicability of the EM algoithm hinges on the fact than in many cases, this expected complete data log-likelihood has *the same form* as the complete data log-likelihood, just with 'artificial data' y^*, i.e.

$$q(\theta \mid \theta^*) = l(\theta \mid x, y^*),$$

where y^* is suitably chosen. The computational task for the E-step is then to compute y^*.

The **M-step** of the algorithm maximizes $q(\cdot \mid \theta^*)$ as a function of θ for fixed θ^*.

After an E-step and an M-step of the algorithm, the likelihood function has never decreased. In other words, if we let $\theta_0 = \theta^*$ and

$$\theta_{n+1} = \arg\max_\theta q(\theta \mid \theta_n),$$

it holds that

$$l_x(\theta_{n+1}) \geq l_x(\theta_n). \tag{1}$$

To see this we use that $P(x, y \mid \theta) = P(x \mid \theta) P(y \mid x, \theta)$ and get

$$\begin{aligned}
q(\theta \mid \theta^*) &= \sum_y P(y \mid x, \theta^*) \log\{P(x \mid \theta) P(y \mid x, \theta)\} \\
&= \log P(x \mid \theta) + \sum_y P(y \mid x, \theta^*) \log P(y \mid x, \theta) \\
&= l_x(\theta) + \sum_y P(y \mid x, \theta^*) \log P(y \mid x, \theta).
\end{aligned}$$

Adding and subtracting $\sum_y P(y \mid x, \theta^*) \log P(y \mid x, \theta^*)$ to the last line yields further

$$q(\theta \mid \theta^*) = l_x(\theta) - \mathrm{Ent}(P^x_{\theta^*}) - KL(P^x_{\theta^*} : P^x_\theta), \tag{2}$$

where we have let $P^x_{\theta^*}(y) = P(y \mid x, \theta^*)$. From (2) we now easily derive that

$$\begin{aligned} l_x(\theta_{n+1}) &= q(\theta_{n+1} \mid \theta_n) + \mathrm{Ent}(P^x_{\theta_n}) + KL(P^x_{\theta_n} : P^x_{\theta_{n+1}}) \\ &\geq q(\theta_n \mid \theta_n) + \mathrm{Ent}(P^x_{\theta_n}) + 0 \\ &= q(\theta_n \mid \theta_n) + \mathrm{Ent}(P^x_{\theta_n}) + KL(P^x_{\theta_n} : P^x_{\theta_n}) = l_x(\theta_n), \end{aligned}$$

which shows (1). Generally, it is not enough to assure convergence that the likelihood function increases at every step. But indeed, if $q(\theta_1 \mid \theta_2)$ is continuous in (θ_1, θ_2), the EM algorithm will always converge to a stationary point, and if

$$\sup_\theta q(\theta \mid \theta^*) > q(\theta^* \mid \theta^*) \text{ for all } \theta^*$$

then the EM algorithm can be shown to converge to a local maximum of l_x. See Wu (1983) for further details. The conditions needed for convergence to a local maximum is fulfilled in most interesting cases.

If all involved quantities satisfy suitable differentiability conditions, we obtain from (2) by differentiation that

$$\frac{\partial}{\partial \theta_i} l_x(\theta)\bigg|_{\theta=\theta^*} = \frac{\partial}{\partial \theta_i} q(\theta \mid \theta^*)\bigg|_{\theta=\theta^*} - \frac{\partial}{\partial \theta_i} KL(P^x_{\theta^*} : P^x_\theta)\bigg|_{\theta=\theta^*} = \frac{\partial}{\partial \theta_i} q(\theta \mid \theta^*)\bigg|_{\theta=\theta^*}$$

as we must have

$$\frac{\partial}{\partial \theta_i} KL(P^x_{\theta^*} : P^x_\theta)\bigg|_{\theta=\theta^*} = 0$$

because $KL(P^x_{\theta^*} : P^x_\theta)$ is minimized for $\theta = \theta^*$. Thus $l_x(\theta)$ and $q(\theta \mid \theta^*)$ have the same gradient for $\theta = \theta^*$.

References

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* John Wiley and Sons, New York.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press, Oxford, United Kingdom.

Sundberg, R. (1972). *Maximum Likelihood Theory and Applications for Distributions Generated when Observing a Function of an Exponential Family Variable.* PhD thesis, University of Stockholm.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, **1**, 49–58.

Thiele, T. N. (1880). Om Anvendelse af mindste Kvadraters Methode i nogle Tilfælde, hvor en Komplikation af visse Slags uensartede tilfældige Fejlkilder giver Fejlene en 'systematisk' Karakter. *Vidensk. Selsk. Skr. 5. Rk., naturvid. og mat. Afd.*, **12**, 381–408. French version: *Sur la Compensation de quelques Erreurs quasi-systématiques par la Méthode des moindres Carrés.* Reitzel, København, 1880.

Wu, C. F. J. (1983). On the convergence of the EM algorithm. *Annals of Statistics*, **11**, 95–103.