An Introduction to Probabilistic Graphical Models

Michael I. Jordan
University of California, Berkeley

June 30, 2003

Chapter 11

The EM algorithm

The expectation-maximization (EM) algorithm provides a general approach to the problem of maximum likelihood parameter estimation in statistical models with latent variables. We have already seen two examples of the EM approach at work in the previous chapter. While these examples are revealing ones, it is important to understand that EM applies much more widely. Indeed, the EM approach goes hand-in-glove with general graphical model machinery, taking advantage of the conditional independence structure of graphical models in a systematic way. As such it occupies a central place in the book.

While in principle one can treat ML parameter estimation as a simple matter of passing a likelihood function to a black-box numerical optimization routine, in practice one would like to take advantage of the structure embodied in the model to break the optimization problem into more manageable pieces. EM provides a systematic way to implement such a divide-and-conquer strategy. As we will see, in this chapter and in later chapters, this approach leads to conceptual clarity and simplicity of algorithmic implementation. It also provides a guide to dealing with models in which issues of computational complexity begin to arise. Indeed, EM will provide a guide to dealing with problems in which the mere calculation of the likelihood or its derivatives appear to be intractable computational challenges.

The main goal of this short chapter is to present a general formulation of the EM algorithm. We show that EM is a rather simple optimization algorithm—it is *coordinate ascent* on an appropriately defined function. Thus, both the E step and the M step can be viewed as maximizations in an abstract space. We show how the *expected complete log likelihood* emerges from this perspective; in particular, we show how the maximization operation that defines the E step can also be viewed as an expectation. We also take the coordinate ascent story a bit further, showing that EM can be viewed as an *alternating minimization algorithm*—a special form of coordinate descent in a Kullback-Leibler divergence.

Finally, we sketch how the EM algorithm applies in the general setting of graphical models. Subsequent chapters will provide many examples of applications to graphical models and will fill in the various details appropriate to these special cases.

11.1 Latent variables and parameter estimation

Recall that latent or hidden variables are generally introduced into a model in order to simplify the model in some way. We may observe a complex pattern of dependency among a set of variables $x = (x_1, \ldots, x_m)$. Rather than modeling this dependency directly, via edges linking these variables, we may find it simpler to account for their dependency via "top-down" dependency on a latent variable z. In the simplest case, we may find it possible to assume that the x_i are conditionally independent given z, and thus restrict our model to edges between the node z and the nodes x_i.

If the latent variables in the model could be observed, then generally the parameter estimation problem would be simplified as well. Indeed, this is one way of characterizing what we mean by the simplification achieved by introducing latent variables into a model. For example, in the case of the mixture of Gaussians model, if we could observe a class label corresponding to each data point, then we would break the data into classes and estimate the mean and covariance matrix separately for each class. The estimation problem would decouple.

But the latent variables are not observed, and this implies that the likelihood function is a marginal probability, obtained by summing or integrating over the latent variables. Marginalization couples the parameters and tends to obscure the underlying structure in the likelihood function.

The EM algorithm essentially allows us to treat latent variable problems using complete data tools, skirting the fact that the likelihood is a marginal probability and exploiting to the fullest the underlying structure induced by the latent variables. EM is an iterative algorithm, consisting of a linked pair of steps. In the *expectation step (E step)*, the values of the unobserved latent variables are essentially "filled in," where the filling-in is achieved by calculating the probability of the latent variables, given the observed variables and the current values of the parameters.[1] In the *maximization step (M step)*, the parameters are adjusted based on the filled-in variables, a problem which is essentially no more complex than it would be if the latent variables had been observed.

11.2 The general setting

Let X denote the observable variables, and let Z denote the latent variables. Often, X and Z decompose into sets of independent, identically-distributed (IID) pairs, in particular X can often be written as $X = (X_1, X_2, \ldots, X_N)$, where the X_i are IID variables and the observed data, $x = (x_1, x_2, \ldots, x_N)$, are the observed values of X. We do not need to make this assumption, however, and indeed we will see many non-IID examples in later chapters. Thus, X represents the totality of observable variables and x is the entire observed dataset. Similarly Z represents the set of all latent variables. The probability model is $p(x, z \mid \theta)$.

If Z could be observed, then the ML estimation problem would amount to maximizing the quantity:

$$l_c(\theta; x, z) \triangleq \log p(x, z \mid \theta), \tag{11.1}$$

[1] We will see that a better way to express this is that in the E step we compute certain expected sufficient statistics, which in the case of multinomial variables reduces to computing the probability of the latent variables. But let us stick with the intuitive and picturesque language of "filling-in" for now.

which is referred to in the context of the EM algorithm as the *complete log likelihood*. If the probability $p(x, z \mid \theta)$ factors in some way, such that separate components of θ occur in separate factors, then the operation of the logarithm has the effect of separating the likelihood into terms that can be maximized independently. As we discussed in Chapter 9, this is what we generally mean by "decoupling" the estimation problem.

Given that Z is not in fact observed, the probability of the data x is a marginal probability, and the log likelihood (referred to in this context as the *incomplete log likelihood*) takes the following form:

$$l(\theta; x) = \log p(x \mid \theta) = \log \sum_z p(x, z \mid \theta), \tag{11.2}$$

where here as in the rest of the chapter we utilize summation to stand for marginalization—the derivation goes through without change if we integrate over continuous z. The logarithm on the right-hand side is separated from $p(x, z \mid \theta)$ by the summation sign, and the problem does not decouple. It is not clear how to exploit the conditional independence structure that may be present in the probability model.

Let us not give up the hope of working with the complete log likelihood. Given that Z is not observed, the complete log likelihood is a random quantity, and cannot be maximized directly. But suppose we average over z to remove the randomness, using an "averaging distribution" $q(z \mid x)$. That is, let us define the *expected complete log likelihood*:

$$\langle l_c(\theta; x, z) \rangle_q \triangleq \sum_z q(z \mid x, \theta) \log p(x, z \mid \theta), \tag{11.3}$$

a quantity that is a deterministic function of θ. Note that the expected complete log likelihood is linear in the complete log likelihood and thus should inherit its favorable computational properties. Moreover, if q is chosen well, then perhaps the expected complete log likelihood will not be too far from the log likelihood and can serve as an effective surrogate for the log likelihood. While we cannot hope that maximizing this surrogate will yield a value of θ that maximizes the likelihood, perhaps it will represent an improvement from an initial value of θ. If so then we can iterate the process and hill-climb. This is the basic idea behind the EM algorithm.

We begin the derivation of the EM algorithm by showing that an averaging distribution $q(z \mid x)$ can be used to provide a lower bound on the log likelihood. Consider the following line of argument:

$$\begin{aligned}
l(\theta; x) &= \log p(x \mid \theta) && (11.4)\\
&= \log \sum_z p(x, z \mid \theta) && (11.5)\\
&= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)} && (11.6)\\
&\geq \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} && (11.7)\\
&\triangleq \mathcal{L}(q, \theta), && (11.8)
\end{aligned}$$

where the last line defines the function $\mathcal{L}(q, \theta)$, a function that we will refer to as an *auxiliary function.*[2] In Eq. 11.7 we have used Jensen's inequality, a simple consequence of the concavity of the logarithm function (see Appendix XXX). What we have shown is that—for an arbitrary distribution $q(z \mid x)$—the auxiliary function $\mathcal{L}(q, \theta)$ is a lower bound for the log likelihood.

The EM algorithm is a coordinate ascent algorithm on the function $\mathcal{L}(q, \theta)$. At the $(t+1)$st iteration, we first maximize $\mathcal{L}(q, \theta^{(t)})$ with respect to q. For this optimizing choice of averaging distribution $q^{(t+1)}$, we then maximize $\mathcal{L}(q^{(t+1)}, \theta)$ with respect to θ, which yields the updated value $\theta^{(t+1)}$. Giving these steps their traditional names, we have:

$$(\textbf{E step}) \qquad q^{(t+1)} = \arg\max_q \ \mathcal{L}(q, \theta^{(t)}) \tag{11.9}$$

$$(\textbf{M step}) \qquad \theta^{(t+1)} = \arg\max_\theta \ \mathcal{L}(q^{(t+1)}, \theta). \tag{11.10}$$

We will soon explain why the first step can be referred to as an "expectation step." We will also explain how a procedure based on maximizing a lower bound on the likelihood $l(\theta; x)$ can maximize the likelihood itself.

The first important point to note is that the M step is equivalently viewed as the maximization of the expected complete log likelihood. To see this, note that the lower bound $\mathcal{L}(q, \theta)$ breaks into two terms:

$$\begin{aligned}
\mathcal{L}(q, \theta) &= \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} && (11.11)\\
&= \sum_z q(z \mid x) \log p(x, z \mid \theta) - \sum_z q(z \mid x) \log q(z \mid x) && (11.12)\\
&= \langle l_c(\theta; x, z) \rangle_q - \sum_z q(z \mid x) \log q(z \mid x), && (11.13)
\end{aligned}$$

and that the second term is independent of θ. Thus, maximizing $\mathcal{L}(q, \theta)$ with respect to θ is equivalent to maximizing $\langle l_c(\theta; x, z) \rangle_q$ with respect to θ.

Let us now consider the E step, the maximization of $\mathcal{L}(q, \theta^{(t)})$ with respect to the averaging distribution q. This maximization problem can be solved once and for all; indeed, we can verify that the choice $q^{(t+1)}(z \mid x) = p(z \mid x, \theta^{(t)})$ yields the maximum. To see this, evaluate $\mathcal{L}(q, \theta^{(t)})$ for this choice of q:

$$\begin{aligned}
\mathcal{L}(p(z \mid x, \theta^{(t)}), \theta^{(t)}) &= \sum_z p(z \mid x, \theta^{(t)}) \log \frac{p(x, z \mid \theta)}{p(z \mid x, \theta^{(t)})} && (11.14)\\
&= \sum_z p(z \mid x, \theta^{(t)}) \log p(x \mid \theta^{(t)}) && (11.15)\\
&= \log p(x \mid \theta^{(t)}) && (11.16)\\
&= l(\theta^{(t)}; x). && (11.17)
\end{aligned}$$

Given that $l(\theta; x)$ is an upper bound for $\mathcal{L}(q, \theta^{(t)})$, this shows that $\mathcal{L}(q, \theta^{(t)})$ is maximized by setting $q(z \mid x)$ equal to $p(z \mid x, \theta^{(t)})$.

[2] Note that $\mathcal{L}(q, \theta)$ is a function of x as well. We omit this dependence, however, to lighten the notation.

There is slightly different way to show this result. We first show that the difference between $l(\theta; x)$ and $\mathcal{L}(q, \theta)$ is a Kullback-Leibler (KL) divergence:

$$\begin{aligned}
l(\theta; x) - \mathcal{L}(q, \theta) &= l(\theta; x) - \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} & (11.18)\\
&= \sum_z q(z \mid x) \log p(x \mid \theta) - \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} & (11.19)\\
&= \sum_z q(z \mid x) \log p(x \mid \theta) - \log \frac{q(z \mid x)}{p(z \mid x, \theta)} & (11.20)\\
&= D(q(z \mid x) \parallel p(z \mid x, \theta)). & (11.21)
\end{aligned}$$

In Appendix XXX we show that the KL divergence is nonnegative (a simple consequence of Jensen's inequality), and that the KL divergence is uniquely minimized by letting $q(z \mid x)$ equal $p(z \mid x, \theta^{(t)})$. Since minimizing the difference between $l(\theta; x)$ and $\mathcal{L}(q, \theta)$ is equivalent to maximizing $\mathcal{L}(q, \theta)$, we again have our result.

The conditional distribution $p(z \mid x, \theta^{(t)})$ is an intuitively appealing choice of averaging distribution. Given the model $p(x, z \mid \theta^{(t)})$, a link betwen the observed data and the latent variables, the conditional $p(z \mid x, \theta^{(t)})$ is our "best guess" as to the values of the latent variables, conditioned on the data x. What the EM algorithm does is to use this "best guess" distribution to calculate an expectation of the complete log likelihood. The M step then maximizes this expected complete log likelihood with respect to the parameters to yield new values $\theta^{(t+1)}$. We then presumably have an improved model, and we can now make a "better guess" $p(z \mid x, \theta^{(t+1)})$, which is used as the averaging distribution in a subsequent EM iteration.

What is the effect of an EM iteration on the log likelihood $l(\theta; x)$? In the M step, we choose the parameters so as to increase a lower bound on the likelihood. Increasing a lower bound on a function does not necessarily increase the function itself, if there is a gap between the function and the bound. In the E step, however, we have closed the gap by an appropriate choice of the q distribution. That is, we have:

$$l(\theta^{(t)}; x) = \mathcal{L}(q^{(t+1)}, \theta^{(t)}), \qquad (11.22)$$

by Eq. 11.17, and thus an M-step increase in $\mathcal{L}(q^{(t+1)}, \theta)$ will also increase $l(\theta; x)$.

In summary, we have shown that the EM algorithm is a hill-climbing algorithm in the log likelihood $l(\theta; x)$. The algorithm achieves this hill-climbing behavior indirectly, by coordinate ascent in the auxiliary function $\mathcal{L}(q, \theta)$. The advantage of working with the latter function is that it involves maximization of the expected complete log likelihood rather than the log likelihood itself, and, as we have seen in examples, this is often a substantial simplication.

11.3 EM and alternating minimization

We can put our results in a slightly more elegant form by working with KL divergences rather than likelihoods.

Recall that in Chapter 8 we noted a simple equivalence between maximization of the likelihood and minimization of the KL divergence between the empirical distribution and the model. Let us return to that equivalence, and bound the KL divergence rather than the log likelihood. We have:

$$\begin{aligned}
D(\tilde{p}(x) \parallel p(x \mid \theta)) &= -\sum_x \tilde{p}(x) \log p(x \mid \theta) + \sum_x \tilde{p}(x) \log \tilde{p}(x) &(11.23)\\
&\leq -\sum_x \tilde{p}(x) \mathcal{L}(q, \theta) + \sum_x \tilde{p}(x) \log \tilde{p}(x) &(11.24)\\
&= -\sum_x \tilde{p}(x) \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} + \sum_x \tilde{p}(x) \log \tilde{p}(x) &(11.25)\\
&= \sum_x \tilde{p}(x) \sum_z q(z \mid x) \log \frac{\tilde{p}(x) q(z \mid x)}{p(x, z \mid \theta)} &(11.26)\\
&= D(\tilde{p}(x) q(z \mid x) \parallel p(x, z \mid \theta)). &(11.27)
\end{aligned}$$

We see that the KL divergence between the empirical distribution and the model—the quantity that we wish to minimize—is upper bounded by a "complete KL divergence," a KL divergence between joint distributions on (x, z).

The term $\sum_x \tilde{p}(x) \log \tilde{p}(x)$ is independent of q and θ and its inclusion in the problem therefore does not change any of our previous results. In particular, minimizing the complete KL divergence with respect to q and θ is equivalent to maximizing the auxiliary function $\mathcal{L}(q, \theta)$ with respect to these variables. We can therefore reformulate the EM algorithm in terms of the KL divergence. Defining $D(q \parallel \theta) \triangleq D(\tilde{p}(x) q(z \mid x) \parallel p(x, z \mid \theta))$ as a convenient shorthand, we have:

$$\begin{aligned}
&(\textbf{E step}) & q^{(t+1)}(z \mid x) &= \arg\min_q D(q \parallel \theta^{(t)}) &(11.28)\\
&(\textbf{M step}) & \theta^{(t+1)} &= \arg\min_\theta D(q^{(t+1)} \parallel \theta) &(11.29)
\end{aligned}$$

We see that EM is a special kind of coordinate descent algorithm—an *alternating minimization* algorithm. We alternate between minimizing over the arguments of a KL divergence.

The alternating minimization perspective and the auxiliary function perspective are essentially the same, and the choice between the two is largely a matter of taste. We will see, however, in Chapter 19, that the alternating minimization view allows us to provide a geometric interpretation of EM as a sequence of projections between manifolds—a perspective reminiscent of our presentation of the LMS algorithm in Chapter 6.

11.4 EM, sufficient statistics and graphical models

[Section not yet written.]

11.5 Historical remarks and bibliography