A Variational view of the EM algorithm

Matthew Stephens
University of Chicago
January 9, 2026

See here for a PDF version of this vignette.

Introduction

This vignette describes the EM algorithm, using the very useful view due to Neal and Hinton. [Note on notation: Neal and Hinton use Z to denote observed data, but I use it to denote latent unobserved values. So be sure to keep this in mind if you are trying to relate their paper to my summary here.]

Set-up

Suppose we have data X from a distribution $p(X|\theta)$ where θ denotes parameters whose values are unknown and live in some space Θ. We aim to find the maximum likelihood estimate for θ

$$\hat{\theta} := \arg\max_{\theta \in \Theta} p(X|\theta).$$

Suppose further that we can write $p(X|\theta)$ as an integration (or sum) over latent variables Z. That is

$$p(X|\theta) = \int p(X, Z|\theta) dZ$$

for some joint distribution $p(X, Z|\theta)$.

The pair X, Z is sometimes referred to as the "complete data" (based on the idea that Z is something that we wished we observed, but do not), and $p(X, Z|\theta)$, as a function of θ, is sometimes referred to as "the complete data likelihood" – it is the likelihood we would have if we actually observed Z as well as X.

The EM algorithm provides a convenient way to find $\hat{\theta}$ when the complete data likelihood has certain "friendly" features, which will become apparent below.

Note: throughout this note I treat Z as a continuous random variable, using integrals and density functions. However, the same ideas apply if Z is discrete: you can just replace integrals with sums and replace density functions with probability mass functions.

Note: X and Z willl typically both be multi-dimensional, not just scalars.

The F function

Let Q denote the space of all "distributions on Z". So $q \in Q$ means that q is a distribution on Z, with density $q(Z)$ say.

Define the following function F that maps (Θ, Q) to the real line:

$$F(\theta, q) := E_q \log p(X, Z|\theta) + H(q)$$

where $H(q)$ denotes the entropy of q:

$$H(q) = -E_q \log(q(Z))$$

Note that F can be rewritten as:

$$F(\theta, q) = E_q(\log p(Z|X, \theta)p(X|\theta)/q(Z)) = \log p(X|\theta) - KL[q||p(Z|X, \theta)]$$

where $KL(q, p)$ denotes the Kullback–Leibler divergence from q to p.

Two key results

There are two key results.

1. For any given θ let $\hat{q}_\theta$ denote the value of q that maximizes $F(\theta, q)$. That is,

   $$\hat{q}_\theta := \arg\max_q F(\theta, q).$$

   The first key result is that this optimal q is given by the conditional distribution of the latent variables Z given the data X and θ:

   $$\hat{q}_\theta(z) = p(Z|X, \theta).$$

2. The log-likelihood $l(\theta) := \log p(X|\theta)$ is related to F by:

   $$l(\theta) = \max_q F(\theta, q).$$

Both results follow directly from

$$F(\theta, q) = \log p(X|\theta) - KL[q||p(Z|X, \theta)].$$

and the fact that the minimum of

$$KL(q||p)$$

over q is 0, attained at $q = p$.

The EM algorithm

It follow from Result 2 above that the maximum likelihood estimate $\hat{\theta}$ can be obtained by *jointly maximizing F over q and θ.*

The EM algorithm simply does this by iterating the following two steps:

- Step 1: $q \leftarrow \arg\max_q F(\theta, q)$
- Step 2: $\theta \leftarrow \arg\max_\theta F(\theta, q)$

From result 1, step 1 is achieved by $q = p(Z|X,\theta)$. Step 2 involves *M*aximizing the *E*xpected "complete-data log-likelihood",

$$\theta \leftarrow \arg\max_{\theta} E_q \log p(X,Z|\theta).$$

(Note: the E and M in the EM name come from the "Expectation" and "Maximization" in Step 2.)

Of course the EM algorithm is only useful if both steps are straightforward, which they are in some important problems. The reason for this is that the complete data log-likelihood $\log p(X,Z|\theta)$ is often a simpler function of θ than is the log-likelihod $\log p(X|\theta)$.

Extensions of the EM algorithm

One of the nice things about this view of the EM algorithm is that it naturally suggests extensions of the EM algorithm that may be useful when the regular EM is hard. For example, suppose that in Step 2 it is hard to optimize F over θ, but it is easy to find a new value of θ that *increases* F. Then using this "incremental" step instead of Step 2 may still be enough to produce an algorithm that optimizes F, and therefore optimizes $l(\theta)$.

Another variation on the EM algorithm arises if we replace step 1 with a constrained optimization $q \leftarrow \arg\max_{q \in Q'} F(\theta, q)$ where Q' is a family of distributions chosen to make this optimization tractible. In this case the resulting algorithm would no longer optimize $l(\theta)$ but it can be thought of as an approximation. This approach comes up in many "variational approximation" algorithms, and studying the accuracy of such approximations remains an active area of research.