

## 12. Expectation Maximization

Expectation Maximization (EM) is a very powerful method in statistics for dealing with a broad family of problems that are immune to attacks with Maximum Likelihood. The characteristic of these problems is that they all have hidden variables, or at least they can be thought as having. Let us consider these problems to see why they are hard.

Let us find the parameters  $\theta$  given data  $x$  for a problem whose probability density can only be expressed with the use of some parameters  $z$ . We cannot avoid this parameter  $z$  and we are not interested in it. Then the likelihood is

$$L(\theta | x) = p(x | \theta) = \prod_i p(x_i | \theta) = \prod_i \int p(x_i, z | \theta) dz$$

which is very difficult to handle unless the integral has a closed form solution. Taking the logarithm of the likelihood does not help since

$$\ln L(\theta | x) = \sum_i \ln \int p(x_i, z | \theta) dz \quad (12.1)$$

and there is no general way to simplify the logarithm of an integral. Time to roll out the heavy equipment.

### 12.1. Jensen Inequality

This is a very useful inequality that holds for all concave functions. A concave function  $f$  is one for which

$$f(\lambda_1 x_1 + \lambda_2 x_2) \geq \lambda_1 f(x_1) + \lambda_2 f(x_2) \quad (12.2)$$

where  $x_1$  and  $x_2$  are positive real numbers,  $\lambda_1$  and  $\lambda_2$  are also positive real numbers and  $\lambda_1 + \lambda_2 = 1$ . We can easily show that logarithm is such a function. We know that for the log function Eq. (12.2) holds for  $\lambda_1 = \frac{1}{2}$  and  $\lambda_2 = \frac{1}{2}$ ,

$$\ln\left(\frac{1}{2} x_1 + \frac{1}{2} x_2\right) \geq \frac{1}{2} \ln x_1 + \frac{1}{2} \ln x_2$$

since this is essentially the good old theorem that the arithmetic average is greater than the geometric. It is easy to show it also for  $\lambda_1 = \frac{1}{4}$  and  $\lambda_2 = \frac{3}{4}$ , and every pair of  $\lambda$ s by setting up some clever induction.

The *discrete* version of the Jensen inequality says that for every concave function  $f$ , positive real numbers  $x_i$ , for  $i = 1 \cdots N$  and positive real numbers  $\lambda_i$  for  $i = 1 \cdots N$  such that

$$\sum_i^N \lambda_i = 1$$

the following holds

$$f\left(\sum_i^N \lambda_i x_i\right) \geq \sum_i^N \lambda_i f(x_i)$$

and  $N$  can also be infinity. We already know it for  $N = 2$  and we can easily show it for  $N = 3$  etc with induction. With a bit of clever mathematical juggling we can show the continuous version of it

$$f\left(\int p(\omega)x(\omega)d\omega\right) \geq \int p(\omega)f(x(\omega))d\omega$$

where  $x(\omega)$  is a real positive function and  $p(\omega)$  is a real positive function such that

$$\int p(\omega)d\omega = 1$$

The function  $f$  in our discussion is the log function and in such a case the inequality becomes equality  $x_i = x_1$  for all  $i = 1 \cdots N$  in the discrete case and  $p(\omega) = \text{const}$  in the continuous case. We can already see that such a relation is our only hope to manhandle Eq. (12.1) into something manageable.

Before we go on to the derivation of EM it is worth noticing that both the discrete version and the continuous can be written as

$$f(E_\omega\{x(\omega)\}) \geq E_\omega\{f(x(\omega))\}$$

the form that statisticians prefer.

## 12.2. Derivation of EM

We start from Eq. (12.1) and multiply and divide by  $p(z_i | x_i, \theta^t)$  so that the expression looks like the expected value of  $z$  given a guess  $\theta^t$  for the parameter  $\theta$

$$\ln L(\theta | x) = \sum_i \ln \int \left( \frac{p(x_i, z | \theta)}{p(z | x_i, \theta^t)} \right) p(z | x_i, \theta^t) dz$$

and since

$$\int p(z | x_i, \theta^t) dz = 1$$

we apply the Jensen inequality and have

$$\begin{aligned} \ln L(\theta | x) &\geq \sum_i \int \ln \left( \frac{p(x_i, z | \theta)}{p(z | x_i, \theta^t)} \right) p(z | x_i, \theta^t) dz = \\ &\sum_i \left( \int \ln(p(x_i, z | \theta)) p(z | x_i, \theta^t) dz \right) - \sum_i \left( \int \ln(p(z_i | x_i, \theta^t)) p(z | x_i, \theta^t) dz \right). \end{aligned}$$

We rename the last two terms  $Q(\theta, \theta^t)$  and  $Q'(\theta^t)$  and write

$$\ln L(\theta | x) \geq Q(\theta, \theta^t) - Q'(\theta^t).$$

We now notice that this inequality becomes equality when  $\theta = \theta^t$  since

$$\begin{aligned}
Q(\theta^t, \theta^t) - Q'(\theta^t) &= \\
\sum_i \int \ln \left( \frac{p(x_i, z | \theta^t)}{p(z | x_i, \theta^t)} \right) p(z | x_i, \theta^t) dz &= \\
\sum_i \int \ln \left( \frac{p(x_i | \theta^t) p(z | x_i, \theta^t)}{p(z | x_i, \theta^t)} \right) p(z | x_i, \theta^t) dz &= \\
\sum_i \int \ln \left( p(x_i | \theta^t) \right) p(z | x_i, \theta^t) dz &= \\
\sum_i \ln p(x_i | \theta^t) \int p(z | x_i, \theta^t) dz &= \\
\sum_i \ln p(x_i | \theta^t) &= \\
\ln L(\theta^t | x). &
\end{aligned}$$

Let  $\theta^{t+1}$  be the value of  $\theta$  maximizing  $Q(\theta, \theta^t)$ . Then

$$\ln L(\theta^t | x) = Q(\theta^t, \theta^t) - Q'(\theta^t) \leq Q(\theta^{t+1}, \theta^t) - Q'(\theta^t) \leq \ln L(\theta^{t+1} | x)$$

which shows that by successively maximizing  $Q(\theta, \theta^t)$  we are guaranteed that the likelihood will keep growing, what mathematicians call monotonicity. But monotonicity alone gives us only a nice warm feeling that the method will converge and is by no means proof it *will* converge.

### 12.3. Convergence

To prove convergence we have to prove at least one more thing, which is that

$$\left[ \nabla_{\theta} Q(\theta, \theta^t) \right]_{\theta=\theta^t} = 0$$

if and only if

$$\left[ \nabla_{\theta} \ln L(\theta | x) \right]_{\theta=\theta^t} = 0$$

which is not very hard. The difference

$$D(\theta) = \ln L(\theta | x) - (Q(\theta, \theta^t) - Q'(\theta^t))$$

is always non negative and is equal to zero for  $\theta = \theta^t$ . Since its minimum is achieved for  $\theta = \theta^t$  the derivative of this difference

$$\left[ \frac{\partial D(\theta)}{\partial \theta} \right]_{\theta=\theta^t} = 0$$

so we know

$$\left[ \frac{\partial Q(\theta, \theta^t)}{\partial \theta} \right]_{\theta=\theta^t} = \left[ \frac{\partial \ln L(\theta|x)}{\partial \theta} \right]_{\theta=\theta^t}$$

since of course

$$\left[ \frac{\partial Q'(\theta^t)}{\partial \theta} \right]_{\theta=\theta^t} = 0.$$

So if for some  $\theta^t$  we cannot maximize  $Q(\theta, \theta^t)$ , we know that we have reached the maximum for  $\ln L(\theta|x)$  as well.

This says EM will converge, but does not say how fast. To find out how fast we do up to second order Taylor series expansion for both  $L(\theta|x)$  and  $Q(\theta, \theta^t)$  around  $\theta^t$ . Assuming all gradient operators are with respect to  $\theta$

$$\begin{aligned} \ln L(\theta) &= \ln L(\theta^t) + \left[ \nabla^T \ln L(\theta) \right]_{\theta=\theta^t} (\theta - \theta^t) + \frac{1}{2} (\theta - \theta^t)^T \left[ \nabla \nabla^T \ln L(\theta) \right]_{\theta=\theta^t} (\theta - \theta^t) \\ Q(\theta, \theta^t) &= Q(\theta^t, \theta^t) + \left[ \nabla^T Q(\theta, \theta^t) \right]_{\theta=\theta^t} (\theta - \theta^t) + \frac{1}{2} (\theta - \theta^t)^T \left[ \nabla \nabla^T Q(\theta, \theta^t) \right]_{\theta=\theta^t} (\theta - \theta^t) \end{aligned}$$

and since we hate long series of weird symbols we define

$$\begin{aligned} V &= \left[ \nabla \ln L(\theta) \right]_{\theta=\theta^t} = \left[ \nabla Q(\theta, \theta^t) \right]_{\theta=\theta^t} \\ G_L &= \left[ \nabla \nabla^T \ln L(\theta) \right]_{\theta=\theta^t} \\ G_Q &= \left[ \nabla \nabla^T Q(\theta, \theta^t) \right]_{\theta=\theta^t} \end{aligned}$$

and the Taylor series become

$$\begin{aligned} \ln L(\theta) &= \ln L(\theta^t) + V^T (\theta - \theta^t) + \frac{1}{2} (\theta - \theta^t)^T G_L (\theta - \theta^t) \\ Q(\theta, \theta^t) &= Q(\theta^t, \theta^t) + V^T (\theta - \theta^t) + \frac{1}{2} (\theta - \theta^t)^T G_Q (\theta - \theta^t) \end{aligned}$$

from which we can easily infer that  $\theta_0$  which maximizes the likelihood, assuming we are sufficiently close to the maximum for the second order Taylor series to be enough, is

$$\begin{aligned} V &= G_L (\theta_0 - \theta^t) \\ \theta_0 &= \theta^t - G_L^{-1} V \end{aligned}$$

and similarly for  $\theta^{t+1}$

$$\theta^{t+1} = \theta^t - G_Q^{-1} V.$$

So the distance between the maximum of the likelihood and the maximum of the  $Q$  is

$$\theta_0 - \theta^{t+1} = \left( G_L^{-1} - G_Q^{-1} \right) V = \left( G_L^{-1} - G_Q^{-1} \right) G_L (\theta_0 - \theta^t)$$

so the distance from the maximum likelihood decreases every iteration by matrix

$$M = \left( G_L^{-1} - G_Q^{-1} \right) G_L = \mathbf{1} - G_Q^{-1} G_L$$

assuming of course that the radius of the matrix is less than unity, which it is. To see this we notice that  $G_Q^{-1} G_L$  has a radius less than unity otherwise the  $Q$  curve would not fit under the  $L$  curve, and  $G_Q^{-1} G_L$  is also a positive definite matrix as a product of two negative definite matrices. So  $M$  is both positive and has radius less than unity. In other words all eigenvalues of  $M$  are between 0 and 1, and therein lies the rub. In a problem with many dimensions the chances of having an eigen value between .99 and 1.0 are high, in which case the radius is at least .99, in other words the distance from the destination decreases by less than 1 percent in every iteration. A rather slow rate. This type of convergence is called linear, for although it looks exponential, it is rather slow and thus the term linear makes sense.

On the other hand, since we maximize an easy to maximize curve that lies wholly under the likelihood curve, we have little trouble moving toward the minimum, especially since early on we are most likely not aligned with the eigenvector of the accursed .99 eigenvalue. And so we can approach the minimum, very often global, very fast, but once we get close we slow down.

Newton Raphson is just the opposite. If we are far from the destination we essentially move randomly inverting at every step a matrix that takes us nowhere. But once we get close, the convergence is rapid. Within a couple of iterations we are done.