# The Expectation-Maximization Algorithm

Gautham Nair

## 1 An approximation to the log likelihood in the presence of latent variables

Suppose we have measurements $x$ and a desire to estimate parameters $\theta$ underlying it. We'll use maximum likelihood estimation. Suppose that $P(x|\theta)$ is however hard to compute and optimize as it is, because we are unable to observe some latent variable $z$. This is a bit of a shame because $P(x,z|\theta)$ is on the other hand easy to compute or optimize.

$$\text{what we have...} \qquad L(\theta) = \log P(x|\theta) = \log \sum_z P(x,z|\theta) \tag{1}$$

so the problem is that the easy thing to optimize is stuck in a sum over all possible values that the latent variable can take. Even for exponential family probability distributions, this sum is computationally intractable. On the other hand it is very easy to add linear combinations of the *logarithms* of exponential family densities.

$$\text{what we'd like...} \qquad L(\theta) = \text{linear combinations of } \log P(x,z|\theta)$$

Of course we can't get this exactly, but we can try to build an approximation to Eqn. 1 of the form

$$L(\theta) \approx f(\theta) = C + \sum_z a_z \log P(x,z|\theta)$$

where we get to pick both C and all the $a_z$. Our concern is optimizing $L(\theta)$. Supposing we have an initial guess $\theta_0$, we can try to pick the coefficients so that our approximation matches both the value *and* the derivative of the likelihood at $\theta_0$.

$$\text{Pick } C \text{ and the } a_z \text{ so that:} \qquad f(\theta_0) = L(\theta_0) \qquad \text{and} \qquad f'(\theta_0) = L'(\theta_0)$$

Comparing the derivative of the log likelihood,

$$L'(\theta_0) = \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \log P(x|\theta) = \frac{1}{P(x|\theta_0)} \sum_z \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} P(x,z|\theta) \tag{2}$$

And in our approximation,

$$f'(\theta_0) = \sum_z \frac{a_z}{P(x,z|\theta_0)} \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} P(x,z|\theta) \tag{3}$$

We see that to match the contributions of every value of $z$ we should pick $a_z = P(z|x,\theta_0) = P(x,z|\theta_0)/P(x|\theta_0)$. In other words, to approximate the log likelihood as a sum of the log probabilities over $z$, we should weigh each value of $z$ by its posterior probability:

$$a_z = P(z|x,\theta_0) \tag{4}$$

Imposing $L(\theta_0) = f(\theta_0)$, we find $C$ as follows,

$$\begin{aligned} C &= \log P(x|\theta_0) - \sum_z P(z|x,\theta_0)\log P(x,z|\theta_0) \\ &= -\sum_z P(z|x,\theta_0)\log P(z|x,\theta_0) \end{aligned}$$

We finally piece together our approximation of the log likelihood near $\theta_0$:

$$L(\theta) \approx f(\theta) = \sum_z P(z|x,\theta_0)\log\frac{P(x,z|\theta)}{P(z|x,\theta_0)} \tag{5}$$

## 2 Expectation-Maximization

We could at this point just use $f(\theta)$ to do a step of gradient descent (ascent in this case) and improve our estimate of $\theta$. Presumably we would have to take only a very small step $\Delta\theta$ because we are so far on unsure footing about how far away from $\theta_0$ our approximation is valid. $f(\theta)$ is much easier to optimize than $L(\theta)$, but if we go so far from $\theta_0$ that $f(\theta) > L(\theta)$, then there is no guarantee that we are making progress towards our actual objective.

Fortunately, we can show that $f(\theta)$ is always $\leq L(\theta)$. First rewrite $f$ as an expected value over the posterior of z:

$$f(\theta) = \left\langle \log\frac{P(x,z|\theta)}{P(z|x,\theta_0)} \right\rangle_{z|x,\theta_0}$$

Meanwhile we somewhat trivially rewrite $L(\theta)$ as:

$$L(\theta) = \log\sum_z P(x,z|\theta) = \log\left\langle \frac{P(x,z|\theta)}{P(z|x,\theta_0)} \right\rangle_{z|x,\theta_0}$$

So our true target is the log of an expectation of something, but our approximation is the expectation of the log of that same thing. Because the logarithm curves downwards, the log of the expectation is always greater than the expectation of the log and therefore $L(\theta) \geq f(\theta)$.

This means we can start at $\theta_0$, calculate $f(\theta)$, maximize it with respect to $\theta$, and we are guaranteed at least an equally big improvement in the log likelihood. Repeating these two steps is the Expectation-Maximization algorithm.