Expectation Maximization (EM) Algorithm

Motivating Example:

- Have two coins: Coin 1 and Coin 2
- Each has it's own probability of seeing "H" on any one flip. Let

$$\begin{aligned} p_1 &= P(\text{ H on Coin 1 }) \\ p_2 &= P(\text{ H on Coin 2 }) \end{aligned}$$

- Select a coin at random and flip that one coin m times.
- Repeat this process n times.
- Now have data

$$\begin{array}{cccccc} X_{11} & X_{12} & \cdots & X_{1m} & & Y_1 \\ X_{21} & X_{22} & \cdots & X_{2m} & & Y_2 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} & & Y_n \end{array}$$

Here, the X_{ij} are Bernoulli random variables taking values in $\{0, 1\}$ where

$$X_{ij} = \begin{cases} 1 & , \text{ if the jth flip for the ith coin chosen is H} \\ 0 & , \text{ if the jth flip for the ith coin chosen is T} \end{cases}$$

and the Y_i live in $\{1, 2\}$ and indicate which coin was used on the nth trial.

Note that all the X's are independent and, in particular

$$X_{i1}, X_{i2}, \ldots, X_{im} | Y_i = j \overset{iid}{\sim} Bernoulli(p_j)$$

We can write out the joint pdf of all $nm + n$ random variables and formally come up with MLEs for p_1 and p_2. Call these MLEs $\widehat{p}_1$ and $\widehat{p}_2$. They will turn out as expected:

$$\begin{aligned} \widehat{p}_1 &= \frac{\text{total \# of times Coin 1 came up H}}{\text{total \# times Coin 1 was flipped}} \\ \widehat{p}_2 &= \frac{\text{total \# of times Coin 2 came up H}}{\text{total \# times Coin 2 was flipped}} \end{aligned}$$

- Now suppose that the Y_i are not observed but we still want MLEs for p_1 and p_2. The data set now consists of only the X's and is "incomplete".
- The goal of the EM Algorithm is to find MLEs for p_1 and p_2 in this case.

Notation for the EM Algorithm:

- Let X be observed data, generated by some distribution depending on some parameters. Here, X represents something high-dimensional. (In the coin example it is an $n \times m$ matrix.) These data may or may not be iid. (In the coin example it is a matrix with iid observations in each row.) X will be called an "incomplete data set".

- Let Y be some "hidden" or "unobserved data" depending on some parameters. Here, Y can have some general dimension. (In the coin example, Y is a vector.)

- Let $Z = (X, Y)$ represent the "complete" data set. We say that it is a "completion" of the data given by X.

- Assume that the distribution of Z (likely a big fat joint distribution) depends on some (likely high-dimensional) parameter θ and that we can write the pdf for Z as

  $$f(z;\theta) = f(x,y;\theta) = f(y|x;\theta)f(x;\theta).$$

  It will be convenient to think of the parameter θ as "given" and to write this instead as

  $$f(z|\theta) = f(x,y|\theta) = f(y|x,\theta)f(x|\theta).$$

  (Note: Here, the f's are different pdfs identified by their arguments. For example $f(x) = f_X(x)$ and $f(y) = f_Y(y)$. We will use subscripts only if it becomes necessary.)

- We usually use $L(\theta)$ to denote a likelihood function and it always depends on some random variables which are not shown by this notation. Because there are many groups of random variables here, we will be more explicit and write $L(\theta|Z)$ or $L(\theta|X)$ to denote the **complete likelihood** and **incomplete likelihood** functions, respectively.

- The complete likelihood function is

  $$L(\theta|Z) = L(\theta|X,Y) = f(X,Y|\theta).$$

- The incomplete likelihood function is

  $$L(\theta|X) = f(X|\theta).$$

The Algorithm

The EM Algorithm is a numerical iterative for finding an MLE of θ. The rough idea is to start with an initial guess for θ and to use this and the observed data X to "complete" the data set by using X and the guessed θ to postulate a value for Y, at which point we can then find an MLE for θ in the usual way. The actual idea though is slightly more sophisticated. We will use an initial guess for θ and postulate an entire distribution for Y, ultimately averaging out the unknown Y. Specifically, we will look at the expected complete likelihood (or log-likelihood when it is more convenient) $\mathsf{E}[L(\theta|X,Y)]$ where the expectation is taken over the conditional distribution for the random vector Y given X and our guess for θ.

We proceed as follows.

$\boxed{1}$ Let $k = 0$. Give an initial estimate for θ. Call it $\widehat{\theta}^{(k)}$.

$\boxed{2}$ Given observed data X and assuming that $\widehat{\theta}^{(k)}$ is correct for the parameter θ, find the conditional density $f(y|X,\widehat{\theta}^{(k)})$ for the completion variables.

$\boxed{3}$ Calculate the conditional expected log-likelihood or "Q-function":

$$Q(\theta|\widehat{\theta}^{(k)}) = \mathsf{E}[\ln f(X,Y|\theta)|X,\widehat{\theta}^{(k)}].$$

Here, the expectation is with respect to the conditional distribution of Y given X and $\widehat{\theta}^{(k)}$ and thus can be written as

$$Q(\theta|\widehat{\theta}^{(k)}) = \int \ln(f(X,y|\theta)) \cdot f(y|X,\widehat{\theta}^{(k)})\, dy.$$

(The integral is high-dimensional and is taken over the space where Y lives.)

$\boxed{4}$ Find the θ that maximizes $Q(\theta|\widehat{\theta}^{(k)})$. Call this $\widehat{\theta}^{(k+1)}$.

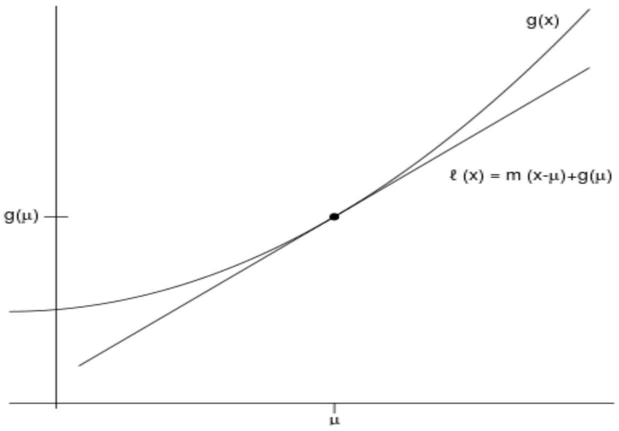
Let $k = k+1$ and return to Step $\boxed{2}$.

The EM Algorithm is iterated until the estimate for θ stops changing. Usually, a tolerance ε is set and the algorithm is iterated until

$$||\widehat{\theta}^{(k+1)} - \widehat{\theta}^{(k)}|| < \varepsilon.$$

We will show that this stopping rule makes sense in the sense that once that distance is less than ε it will remain less than ε.

Figure 1: Visualization for Jensen's Inequality for a Convex Function



Jensen's Inequality

The EM algorithm is derived from **Jensen's inequality**, so we review it here.

Let X be a random variable with mean $\mu = \mathsf{E}[X]$, and let g be a convex function. Then

$$g(\mathsf{E}[X]) \leq \mathsf{E}[g(X)].$$

To prove Jensen's inequality, visualize the convex function g and a tangent line at the point $(\mu, g(\mu))$, as despicted in Figure 1.

Note that, by convexity of g, the line is always below g:

$$\ell(x) \leq g(x) \qquad \forall x.$$

So,

$$m(x-\mu) + g(\mu) \leq g(x) \qquad \forall x,$$

and therefore, we can plug in the random variable X to get

$$m(X-\mu) + g(\mu) \leq g(X).$$

Taking the expected value of both sides leaves us with

$$g(\mu) \leq \mathsf{E}[g(X)]$$

which is the desired result

$$g(\mathsf{E}[X]) \leq \mathsf{E}[g(X)].$$

Note that, if g is concave, then the negative of g is convex. Applying Jensen's inequality to $-g$ and then multiplying through by -1 gives

$$g(\mathsf{E}[X]) \geq \mathsf{E}[g(X)].$$

So, we now know, for example, that

$$\ln(\mathsf{E}[X]) \geq \mathsf{E}[\ln(X)].$$

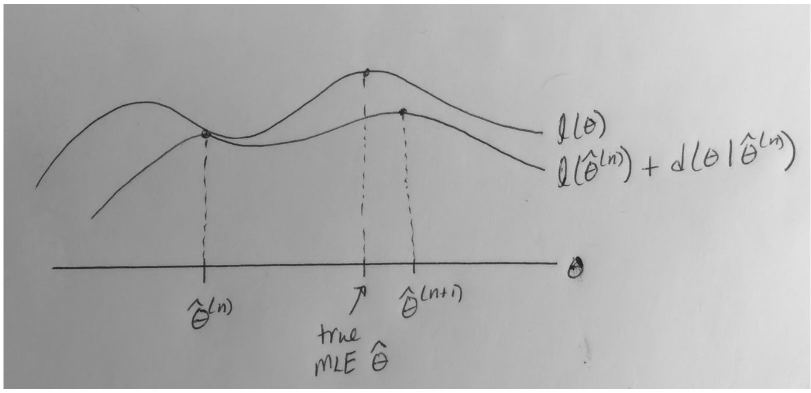
Derivation of the EM Algorithm

- Imagine we have some data X with joint pdf $f(X|\theta)$.
- Let $\ell(\theta) = \ln f(X|\theta)$ denote the log-likelihood.
- Suppose we are trying to guess at θ and improve our guesses through some sort of iteration. Let $\widehat{\theta}^{(n)}$ be our nth iteration guess.
- We would like to find a new value for θ that satisfies
$$\ell(\theta) \geq \ell(\widehat{\theta}^{(n)}).$$
- We will introduce some hidden variables Y, either because we are actually working with a model that has hidden (unobserved) variables or "artificially" because they make the maximization more tractable.
- So, we may write
$$\begin{aligned}\ell(\theta) - \ell(\widehat{\theta}^{(n)}) &= \ln f(X|\theta) - \ln f(X|\widehat{\theta}^{(n)}) \\ &= \ln\left(\int f(X|y,\theta)f(y|\theta)\,dy\right) - \ln f(X|\widehat{\theta}^{(n)}) \\ &= \ln\left(\int \frac{f(X|y,\theta)f(y|\theta)}{f(y|X,\widehat{\theta}^{(n)})} f(y|X,\widehat{\theta}^{(n)})\,dy\right) - \ln f(X|\widehat{\theta}^{(n)})\end{aligned}$$
- Note that the thing in parentheses is an expectation with respect to the distribution of $Y|X,\widehat{\theta}^{(n)}$. So, applying Jensen's inequality, we have
$$\begin{aligned}\ell(\theta) - \ell(\widehat{\theta}^{(n)}) &\geq \int \ln\left(\frac{f(X|y,\theta)f(y|\theta)}{f(y|X,\widehat{\theta}^{(n)})}\right) f(y|X,\widehat{\theta}^{(n)})\,dy - \ln f(X|\widehat{\theta}^{(n)}) \\ &= \int \ln\left(\frac{f(X|y,\theta)f(y|\theta)}{f(y|X,\widehat{\theta}^{(n)})f(X|\widehat{\theta}^{(n)})}\right) f(y|X,\widehat{\theta}^{(n)})\,dy\end{aligned}$$

  Call that entire integral $d(\theta|\widehat{\theta}^{(n)})$.

  We then have
$$\ell(\theta) \geq \ell(\widehat{\theta}^{(n)}) + d(\theta|\widehat{\theta}^{(n)})$$

Figure 2: A Horrible Temporary Image



- Note that

$$\begin{aligned} d(\widehat{\theta^{(n)}}|\widehat{\theta}^{(n)}) &= \int \ln\left(\frac{f(X|y,\widehat{\theta}^{(n)})f(y|\widehat{\theta}^{(n)})}{f(y|X,\widehat{\theta}^{(n)})f(X|\widehat{\theta}^{(n)})}\right) f(y|X,\widehat{\theta}^{(n)})\,dy \\ &= \int \ln\left(\frac{f(X,y|\widehat{\theta}^{(n)})}{f(X,y|\widehat{\theta}^{(n)})}\right) f(y|X,\widehat{\theta}^{(n)})\,dy \\ &= \int \ln(1)\ f(y|X,\widehat{\theta}^{(n)})\,dy = 0 \end{aligned}$$

- So, we have that $\ell(\widehat{\theta}^{(n)}) + d(\theta|\theta^{(n)})$ is bounded above by $\ell(\theta)$ and that it is equal to this upper bound when $\theta = \widehat{\theta}^{(n)}$. This is visualized in Figure 2.

  If we maximize $\ell(\widehat{\theta}^{(n)}) + d(\theta|\theta^{(n)})$ with respect to θ (equivalently, maximize $d(\theta|\theta^{(n)})$ with respect to θ), we may improve towards maximizing $\ell(\theta)$. We know, at least, that we will not get worse.

- Maximizing

$$d(\theta|\widehat{\theta}^{(n)}) = \int \ln\left(\frac{f(X|y,\theta)f(y|\theta)}{f(y|X,\widehat{\theta}^{(n)})f(X|\widehat{\theta}^{(n)})}\right) f(y|X,\widehat{\theta}^{(n)})\,dy$$

  with respect to θ is equivalent to maximizing

$$\int \ln\left(f(X|y,\theta)f(y|\theta)\right)\ f(y|X,\widehat{\theta}^{(n)})\,dy$$

  with respect to θ.

  Note that this is

$$\int \ln f(X,y|\theta)\, f(y|X,\widehat{\theta}^{(n)})\,dy$$

  which is

$$\mathsf{E}_Y[\ln f(X,Y|\theta)|X\widehat{\theta}^{(n)}].$$

- So, if we could compute this expectation, maximize it with respect to θ, call the result $\widehat{\theta}^{(n+1)}$ and iterate, we can improve towards finding the θ that maximizes the likelihood (or at least not get worse). In other words, we can improve towards finding the MLE of θ.

  These expectation and maximization steps are precisely the EM algorithm!

The EM Algorithm for Mixture Densities

Assume that we have a random sample $X_1, X_2, \ldots, X_n$ is a random sample from the mixture density

$$f(x|\theta) = \sum_{j=1}^{N} p_i f_j(x|\theta_j).$$

Here, x has the same dimension as one of the X_i and θ is the parameter vector

$$\theta = (p_1, p_2, \ldots, p_N, \theta_1, \theta_2, \ldots, \theta_N).$$

Note that $\sum_{i=j}^{N} p_j = 1$ and each θ_i may be high-dimensional.

An X_i sampled from $f(x|\theta)$ may be interpreted as being sampled from one of the densities

$$f_1(x|\theta_1), f_2(x|\theta_2), \ldots, f_N(x|\theta_N)$$

with respective probabilities

$$p_1, p_2, \ldots, p_N.$$

- The likelihood is

$$L(\theta|X) = \prod_{i=1}^{n} f(X_i|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{N} p_j f_j(X_i|\theta).$$

- The log-likelihood is

$$\ell(\theta|X) = \ln L(X|\theta) = \sum_{i=1}^{N} \ln \left(\sum_{j=1}^{N} p_j f_j(X_i|\theta_j) \right).$$

- The log of the sum makes this difficult to work with. In order to make this problem more tractable, we will consider X as incomplete data and we will augment the data with additional variables $Y_1, Y_2, \ldots, Y_n$ which indicate the component that the corresponing X's come from.

  Specifically, for $k = 1, 2, \ldots, N$,if $Y_i = k$ then $X_i \sim f_k(x|\theta_k)$, and $p_k = P(Y_i = k)$.

- Now $L(X|\theta)$ is thought of an an incomplete likelihood and the complete likelihood is

$$\begin{aligned} L(\theta|X,Y) &= f(X,Y|\theta) = \textstyle\prod_{i=1}^n f(X_i,Y_i|\theta) \\ &= \textstyle\prod_{i=1}^n f(X_i|Y_i,\theta) f(Y_i|\theta). \end{aligned}$$

- Note that

$$f(y_i|\theta) = P(Y_i = y_i|\theta) = p_{y_i}$$

and that

$$f(x_i|y_i,\theta) = f_{y_i}(x_i|\theta_{y_i}).$$

- So,

$$\ell(\theta|x,y) = \ln L(\theta|x,y) = \sum_{i=1}^n \ln[p_{y_i f_{y_i}}(x_i|\theta_{y_i})].$$

- To apply the EM algorithm, we will start with some initial guesses ($k=0$) for the parameters:

$$\widehat{\theta}^{(k)} = (\widehat{p}_1^{(k)}, \widehat{p}_2^{(k)}, \ldots, \widehat{p}_N^{(k)}, \widehat{\theta}_1^{(k)}, \widehat{\theta}_2^{(k)}, \ldots, \widehat{\theta}_N^{(k)}).$$

- Then, we compute

$$f(y_i|x_i,\widehat{\theta}^{(k)}) = \frac{f(x_i|y_i,\widehat{\theta}^{(k)}) \cdot f(y_i|\widehat{\theta}^{(k)})}{f(x_i|\widehat{\theta}^{(k)})} = \frac{\widehat{p}_{y_i}^{(k)} f_{y_i}(x_i|\widehat{\theta}_{y_i^{(k)}})}{\sum_{j=1}^N \widehat{p}_j^{(k)} f_j(x_i|\widehat{\theta}^{(k)})}$$

for $y_i = 1,2,\ldots,N$.

- The EM algorithm "Q-function" is

$$\begin{aligned} Q(\theta|\widehat{\theta}^{(k)}) &= \mathsf{E}[\ln \underbrace{f(X,Y|\theta)}_{L(\theta|X,Y)} | X, \widehat{\theta}^{(k)}] \\ &= \textstyle\sum_y \ln L(\theta|X,y) \cdot f(y|X,\widehat{\theta}^{(k)}). \end{aligned}$$

Here,

$$f(y|X,\widehat{\theta}^{(k)}) = \prod_{i=1}^n f(y_i|x_i,\widehat{\theta}^{(k)}),$$

but we will ultimately expand our expression for $Q(\theta|\widehat{\theta}^{(k)})$ into terms involving the $f(y_i|x_i,\widehat{\theta}^{(k)})$ and not need to write this joint density down explicitly.

[TO BE CONTINUED...]