Recall that the goal is to maximize the likelihood (or log likelihood) of the observed data:

$$\log(P_Y(y;\theta)).$$

Sometimes, when the log likelihood is difficult to maximize, a missing (or latent) variable helps the computations. Consider Z to be the missing/latent variable. By using the law of total probability (and more):

$$\begin{aligned}
\log(P_Y(y;\theta)) &= \log\left(\sum_z P_{Y,Z}(y,z;\theta)\right) \\
&= \log\left(\sum_z q_{Z,\theta^i}(z)\frac{P_{Y,Z}(y,z;\theta)}{q_{Z,\theta^i}(z)}\right) \\
&= \log\left(E_{q_{Z,\theta^i}}\left[\frac{P_{Y,Z}(y,z;\theta)}{q_{Z,\theta^i}(z)}\right]\right) \\
&\geq E_{q_{Z,\theta^i}}\left[\log\left(\frac{P_{Y,Z}(y,z;\theta)}{q_{Z,\theta^i}(z)}\right)\right] \text{ by Jensen's Inequality} \qquad (1)
\end{aligned}$$

Equation (1) creates a lower bound on the quantity we want to maximize. For ease of computation, the focus will be on the right side of equation (1) during the *maximization step*.

$$\begin{aligned}
E_{q_{Z,\theta^i}}\left[\log\left(\frac{P_{Y,Z}(y,z;\theta)}{q_{Z,\theta^i}(z)}\right)\right] &= E_{q_{Z,\theta^i}}[\log(P_{Y,Z}(y,z;\theta))] - E_{q_{Z,\theta^i}}[\log(q_{Z,\theta^i}(z))] \qquad (2) \\
&= Q(\theta|\theta^i) - E_{q_{Z,\theta^i}}[\log(q_{Z,\theta^i}(z))] \qquad (3)
\end{aligned}$$

Because the goal is to maximize the likelihood with respect to θ, only the first term on the right side of equation (3) is relevant. That is, if Z is known (for a given value of θ^i), then the maximization of the likelihood simplifies to:

The M-Step

$$\begin{aligned}
\hat{\theta} &\leftarrow \text{argmax}_\theta E_{q_{Z,\theta^i}}[\log(P_{Y,Z}(y,z;\theta))] \\
\hat{\theta} &\leftarrow \text{argmax}_\theta \ Q(\theta|\theta^i)
\end{aligned}$$

But unfortunately, we don't typically know the value of Z, or really, $q_{Z,\theta^i}(z)$.

The E-Step

In the *expectation step* we aim to find $q_{Z,\theta^i}(z)$ that will optimize the likelihood. Recall the quantity above that we hope to maximize:

$$\begin{aligned}
E_{q_{Z,\theta^i}}\left[\log\left(\frac{P_{Y,Z}(y,z;\theta)}{q_{Z,\theta^i}(z)}\right)\right] &= E_{q_{Z,\theta^i}}\left[\log\left(\frac{P_Y(y;\theta)P_{Z|Y}(z|y;\theta)}{q_{Z,\theta^i}(z)}\right)\right] \\
&= \log(P_Y(y;\theta)) - E_{q_{Z,\theta^i}}\left[\log\left(\frac{q_{Z,\theta^i}(z)}{P_{Z|Y}(z|y;\theta)}\right)\right]
\end{aligned}$$

It turns out that $\frac{q_{Z,\theta^i}(z)}{P_{Z|Y}(z|y;\theta)}$ is always greater than 1 (it is called the Kullback-Leibler divergence), so $\log(\frac{q_{Z,\theta^i}(z)}{P_{Z|Y}(z|y;\theta)})$ is always greater than zero. In order to make it as small as possible (i.e., to maximize the righthand size), we want the ratio to be as close to one as possible.

$$q_{Z,\theta^i}(z) \approx P_{Z|Y}(z|y;\theta)$$

That is, the value of $q_{Z,\theta^i}(z)$ that maximizes the likelihood, is $P_{Z|Y}(z|y;\theta)$. So, what did we do? We wanted to maximize the likelihood of the given data. Because it was difficult to do directly, we found an algorithm that would iterate between maximizing the likelihood with respect to θ when $q_{Z,\theta^i}(z)$ is known, and then solving for $q_{Z,\theta^i}(z)$ when θ is known.

Does it work?

The EM Algorithm seems like an intuitive way to go back and forth between parameter estimation and estimation of missing information. However, how can we show that it actually converges to a maximum of some kind?

$$\begin{aligned}
\log(P_Y(y;\theta)) &= \log(P_{Y,Z}(y,z;\theta)) - \log(P_{Z|Y}(z|y;\theta)) \text{ cond prob, rearranged} \\
E_{q_{Z,\theta^i}}[\log(P_Y(y;\theta))] &= E_{q_{Z,\theta^i}}[\log(P_{Y,Z}(y,z;\theta))] - E_{q_{Z,\theta^i}}[\log(P_{Z|Y}(z|y;\theta))] \\
\log(P_Y(y;\theta)) &= Q(\theta|\theta^i) + H(\theta|\theta^i)
\end{aligned}$$

which holds for any value of θ, including θ^i.

$$\log(P_Y(y;\theta^i)) = Q(\theta^i|\theta^i) + H(\theta^i|\theta^i)$$

Subtracting the two previous equations gives:

$$\log(P_Y(y;\theta)) - \log(P_Y(y;\theta^i)) = Q(\theta|\theta^i) - Q(\theta^i|\theta^i) + H(\theta|\theta^i) - H(\theta^i|\theta^i)$$

And Gibbs' inequality tells us that $H(\theta|\theta^i) \geq H(\theta^i|\theta^i)$. So we can conclude that:

$$\log(P_Y(y;\theta)) - \log(P_Y(y;\theta^i)) \geq Q(\theta|\theta^i) - Q(\theta^i|\theta^i)$$

If θ makes $Q(\theta|\theta^i)$ bigger than $Q(\theta^i|\theta^i)$, then $\log(P_Y(y;\theta))$ cannot go lower than $\log(P_Y(y;\theta^i))$.

EM Algorithm

Algorithm 1 EM Algorithm

Take initial guesses for the parameters, $i = 0$.
for $i = 1, 2, 3, \ldots$ **do**
Expectation Step: compute the probabilities of each possible value of Z, given θ. Use them to estimate the complete data likelihood as a function of θ.

$$\begin{aligned}
q_{Z,\theta^{i-1}} &\leftarrow p_{Z|Y}(z|y;\theta^{i-1}) \\
Q(\theta|\theta^{i-1}) &= E_{q_{Z,\theta^{i-1}}}[P_{Y,Z}(y,z;\theta)]
\end{aligned}$$

Maximization Step: compute the values of the parameters by maximizing the likelihood with the distribution of z known (that is, under the probability distribution of Z given above).

$$\theta^i \leftarrow \text{argmax}_\theta E_{q_{Z,\theta^{i-1}}}[P_{Y,Z}(y,z;\theta)]$$

Alternatively:

$$\theta^i \leftarrow \text{argmax}_\theta\ Q(\theta|\theta^{i-1})$$

if $\theta^i \approx \theta^{i-1}$ **then**
return θ^i.
end if
end for