Suppose we have an estimation problem in which we have a training set $\{x^{(1)}, \ldots, x^{(m)}\}$ consisting of m independent examples. We wish to fit the parameters of a model $p(x,z)$ to the data, where the likelihood is given by:

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{m} \log p(x;\theta) \\ &= \sum_{i=1}^{m} \log \sum_{z} p(x,z;\theta) \end{aligned}$$

Here, the $z^{(i)}$'s are the latent random variables; and it is often the case that if the $z^{(i)}$'s were observed, then maximum likelihood estimation would be easy.

In such a setting, the **EM Algorithm** gives an efficient method for maximum likelihood estimation.

$z^{(i)} \sim Q_i(z^{(i)})$

$$
\begin{aligned}
\sum_i \log p(x^{(i)};\theta) &= \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta) & (1)\\
&= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} & (2)\\
&\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} & (3)
\end{aligned}
$$

下限提高,
极大似然值也就高了

$E[f(x)] \geqslant f(E[x])$ if f() is a convex function

$f(E[x]) \geqslant E[f(x)]$ if f() is a concave function

→ $log(E[x]) \geqslant E[log(x)]$

→ $$f\left(\mathrm{E}_{z^{(i)}\sim Q_i}\left[\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]\right) \geq \mathrm{E}_{z^{(i)}\sim Q_i}\left[f\left(\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right)\right]$$

There're many possible choices for the $Q^{(i)}$'s . Which should we choose to maximize the likelihood?

$$\mathrm{E}_{z^{(i)} \sim Q_i}\left[f\left(\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right)\right] = \mathrm{E}_{z^{(i)} \sim Q_i}\left[f(z^{(i)})\right]$$

To make the bound tight for a particular value of θ, we need for the step involving Jensen's inequality in our derivation above to hold with equality. For this to be true, we know it is sufficient that that the expectation be taken over a "constant"-valued random variable.

In order to obtain the maximum value, we need $f(z^{(1)}) = f(z^{(2)}) = \cdots = f(z^{(n)})$

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = c$$ Which implies: $Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)}; \theta)$

Since $\sum_z Q_i(z^{(i)}) = 1$ and $Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)}; \theta)$ we can yield:

$$Q_i(z^{(i)}) = \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z; \theta)} = \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} = p(z^{(i)} | x^{(i)}; \theta)$$

From the above equation, we can find that the best Q is the posterior probability given current θ.

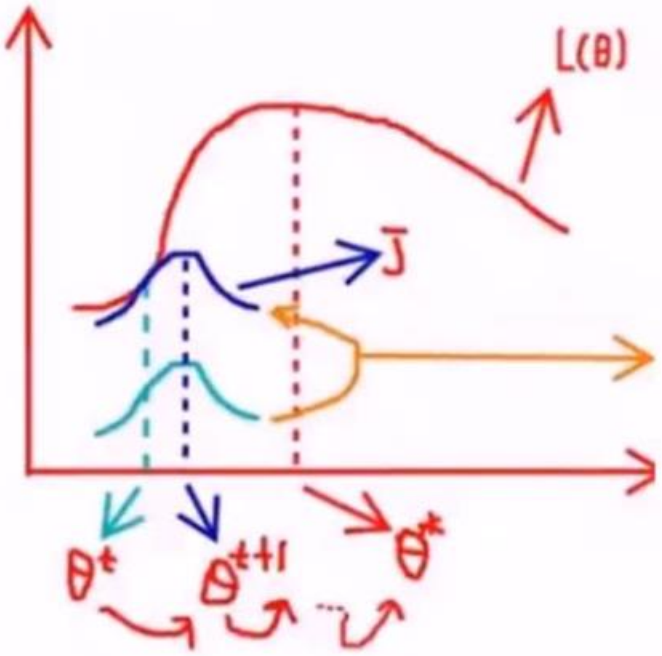
The **EM Algorithm**: Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta)$$

(M-step) Set

$$\theta := \arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \quad \}$$



Repeat until convergence: {

(E-step) For each i, j, set

$$w_j^{(i)} = p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

(M-step) Update the parameters:

$$\phi_j = \frac{1}{m} \sum_{i=1}^m w_j^{(i)}$$

$$\mu_j = \frac{\sum_{i=1}^m w_j^{(i)} x^{(i)}}{\sum_{i=1}^m w_j^{(i)}}$$

$$\Sigma_j = \frac{\sum_{i=1}^m w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^m} w_j^{(i)}$$

}