More on the EM Algorithm

The Expectation Maximization algorithm is a general purpose method for finding the MLE in a model with hidden variables. It does not require committing to any particular model. It consists of two steps:

- **E-step**: "fill in" the latent variables using the posterior ("expectation")
- **M-step**: maximize the expected Complete Log Likelihood with respect to the parameters

The variables used are

\[
D = \{x_1, \ldots, x_N\} \text{ are the observed data} \\
Z \quad \text{are the hidden random variables} \\
\Theta \quad \text{are the model parameters}
\]

The goal is to find parameters that maximize the Complete Log Likelihood:

\[
\hat{\theta} = \arg \max_{\theta} \log p(X, Z|\theta) = \arg \max_{\theta} \left[ \log p(Z|\theta) + \log p(X, Z|\theta) \right]
\]

Complete Log Likelihood

In the latent variable setting,

\[
\hat{\theta} = \arg \max_{\theta} \log \sum_z p(z|\theta)p(X|z, \theta)
\]

Jensen’s Inequality

If \( \lambda \in (0, 1) \) and we have a convex function \( f \),

\[
\lambda f(x) + (1 - \lambda) f(y) \geq f(\lambda x + (1 - \lambda) y)
\]

We can generalize this to expectation with the formula

\[
E[f(X)] \geq f(E[X])
\]

This applies for a convex \( f \), if \( f \) is concave we simply flip the inequality.

**EM Objective Function**

From before, we have

\[
\log p(X|\theta) = \log \sum_z p(z|\theta)p(X|z, \theta)
\]

\[
= \log \sum_z p(z|\theta)p(X|z, \theta) \frac{q(z)}{q(z)}
\]

1
for some distribution $q(z)$ over the latent variables. Using the definition $E[f(X)] = \sum_x p(x)f(x)$, we have

$$\log p(X|\theta) = \log E_q \left[ \frac{p(Z|\theta)p(X|Z,\theta)}{q(Z)} \right]$$

Now we apply Jensen’s Inequality, noting that the log function is concave:

$$\log p(X|\theta) \geq E_q \left[ \frac{p(Z|\theta)p(X|Z,\theta)}{q(Z)} \right] = E_q \left[ \log p(Z|\theta) \right] + E_q \left[ \log p(X|Z,\theta) \right] - E_q \left[ \log q(Z) \right] = \mathcal{L}(\theta; q)$$

which is the EM objective function.

**Coordinate Ascent**

EM proceeds by coordinate ascent. For instance, at iteration $t$, we start with $q^{(t)}$ and $\theta^{(t)}$:

- **E-step**: $q^{(t+1)} = \arg \max_q \mathcal{L}(q, \theta^{(t)}) = p(Z|X)$, which is the posterior
- **M-step**: $\theta^{(t+1)} = \arg \max_{\theta} \mathcal{L}(q^{(t+1)}, \theta)$

**Why is $q$ optimal? Are we maximizing $\mathcal{L}$?**

From before,

$$\mathcal{L}(q, \theta) = E_q \left[ \log p(X, Z|\theta) \right] - E_q \left[ \log q(Z) \right]$$

Because the second term is constant with respect to $\theta$, it will not affect our optimization. Thus, we are only concerned with the first part of $\mathcal{L}$, which is the expected complete log likelihood.

Claim: when $q = p(Z|X,\theta)$ is the posterior, $\mathcal{L}(q, \theta)$ is optimized with respect to $q$.

$$\mathcal{L}(q, \theta) = \sum_z q(z) \log \frac{p(z,X|\theta)}{q(z)} = \sum_z p(z|X,\theta) \log \frac{p(z,X|\theta)}{p(z|X)}$$

$$\mathcal{L}(p(Z|X,\theta), \theta) = \sum_z p(z|X,\theta) \log \frac{p(X,z|\theta)}{p(z|X,\theta)}$$

$$= \sum_z p(z|X,\theta) \log \frac{p(X,z|\theta)p(X)}{p(X,z)} \iff p(Z|X,\theta) = \frac{p(Z,X|\theta)}{p(X|\theta)}$$

$$= \sum_z p(z|X) \log p(X|\theta)$$

$$= p(X|\theta)$$

Because $\mathcal{L}$ is a bound on the likelihood of the data, and because $\log p(X|\theta)$ actually is the likelihood, this $q$ cannot bound the likelihood any more tightly.