20.1 Hidden Markov Models: The training problem and EM Algorithm

Given $\theta$, a series of observation:
Compute $\lambda = (A, B, \pi)$ to maximize $P(\theta | \lambda)$

20.1.1 Maximum likelihood

Consider a dataset, $D$, and $m$ samples from $D$ drawn independently. Then,

$$D = P(X|\theta), \theta = \text{vector of parameters}$$

$$P(D|\theta) = \prod_{k=1}^{m} P(x_k|\theta)$$

Definition 20.1 The Maximum Likelihood Estimator is the value, $\hat{\theta}$, that maximizes $P(D|\theta)$

- Analytically, it is easier to use the log of the value. Because log is monotonically increasing, the $\hat{\theta}$ that maximizes this is also the maximum likelihood. That is, $L(\theta) = \ln P(D|\theta)$ and $\hat{\theta} = \max L(\theta)$

If $P(D|\theta)$ is differentiable function of $\theta$, then

$$\theta = \{\theta_1...\theta_n\},$$

$$\nabla_\theta = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_n} \end{pmatrix}$$

From the above, we can then gather that $\hat{\theta}$ occurs where there are inflection points of $L$. That is, $\nabla_\theta(L) = 0$

20.1.2 Q function

We can use the principle of maximum likelihood to allow for the learning of parameter governing the distribution at this point, of which some data may have missing features. If there is no missing features, then...
finding \( \hat{\theta} \) will satisfy in finding the best parameters.
The basic idea of the EM algorithm is to iteratively estimate the likelihood, given the data that is present.

**Definition 20.2** Each data point \( X_k \) can be divided into \( \{X_{k_g}, X_{k_b}\} \), where those points \( g \) are "good" or present data, and \( b \) is "bad" or missing data.

We can define the Q-function:
\[
Q(\theta, \theta^i) = \mathbb{E}_{\mathcal{D}_b} [\ln P(\mathcal{D}_b|\theta)|\mathcal{D}, \theta^i]
\]
This is taking the expected value over the missing features, assuming that the parameters in \( \theta^i \) are true.

- The parameter vector, \( \theta^i \) is the current best estimate of the full distribution \( \mathcal{D} \)
- Given \( \theta \), we can calculate the likelihood of \( \mathcal{D}_b \), marginalized with respect to the current best distribution
- Different estimates \( \rightarrow \) different likelihood
- Expectation-minimization will select the best candidate \( \theta \), which we will call \( \theta^{i+1} \)

### 20.1.3 EM algorithm

**Result:** \( \hat{\theta} = \theta^{i+1} \)

initialize \( \theta^0, \epsilon, i = 0; \)
while \( Q(\theta^{i+1}, \theta^i) - Q(\theta^i, \theta^{i-1}) \leq \epsilon \) do

\[
\text{Compute } Q(\theta, \theta^i)
\]
\[
\theta^{i+1} = \max Q(\theta, \theta^i)
\]
end

**Algorithm 1:** EM algorithm

The EM algorithm guarantees that the log likelihood of the good data with the bad data marginalized will increase linearly. This is not the same as finding the values of the bad data or the complete distribution.