1 EM Algorithm

The expectation-maximization (EM) algorithm provides a way to compute maximum likelihood parameter estimates for models with missing data (i.e. latent variables). We have:

- \( x \) - observed variables
- \( z \) - latent variables
- model: \( p(x, z|\theta) \)
- goal: \( \hat{\theta}_{\text{ML}} = \arg\max_{\theta} p(x|\theta) = \arg\max_{\theta} \sum_z p(x, z|\theta) \)

First, we have:

\[
\ell(\theta) = \log p(x|\theta) = \log \sum_z p(x, z|\theta) = \log \sum_z q(z|x) \frac{p(x, z|\theta)}{q(z|x)} \\
\geq \sum_z q(z|x) \log \frac{p(x, z|\theta)}{q(z|x)} \equiv \mathcal{L}(q, \theta) \quad \text{(according to Jensen’s inequality)}
\]

EM is coordinate ascent on \( \mathcal{L} \):

- E step: \( q^{(t+1)} = \arg\max_q \mathcal{L}(q, \theta^{(t)}) \)
- M step: \( \theta^{(t+1)} = \arg\max_{\theta} \mathcal{L}(q^{(t+1)}, \theta) \)
EM algorithm and Hidden Markov Models

More specifically,

- E step: \( q^{(t+1)}(z|x) = p(z|x, \theta^{(t)}) \)

- M step: maximize expected complete log likelihood.

We need to show that the E step can leverage the equivalence above \( q^{(t+1)}(z|x) = p(z|x, \theta^{(t)}) \). First,

\[
L(q, \theta^{(t)}) = \sum_z q(z|x) \log p(x, z|\theta^{(t)}) - \sum_z q(z|x) \log q(z|x).
\]

Then, compute the maximum \( \frac{\partial L}{\partial q(z|x)} \):

\[
\widetilde{L} = L + \lambda(1 - \sum_z q(z|x))
\]

\[
\frac{\partial \widetilde{L}}{\partial q(z|x)} = \log p(x, z|\theta^{(t)}) - \log q(z|x) - 1 + \lambda \]

(set=0) \( q(z|x) = p(x, z|\theta^{(t)}) e^\lambda - 1 \)

(sum over z) \( \Rightarrow 1 = \tilde{\lambda} \sum_z p(x, z|\theta^{(t)}) \) \( (\tilde{\lambda} \equiv e^\lambda - 1) \)

\[ \Rightarrow \tilde{\lambda} = \frac{1}{p(x|\theta^{(t)})} \]

\[ \Rightarrow q(z|x) = p(z|x, \theta^{(t)}) \]

Now we plug \( p(z|x, \theta^{(t)}) \) into \( L(q, \theta^{(t)}) \):

\[
L(q, \theta^{(t)}) = L(p(z|x, \theta^{(t)}), \theta^{(t)})
\]

\[
= \sum_z p(z|x, \theta^{(t)}) \log p(x, z|\theta^{(t)}) - \sum_z p(z|x, \theta^{(t)}) \log p(z|x, \theta^{(t)})
\]

\[
= \sum_z p(z|x, \theta^{(t)}) \log \frac{p(x, z|\theta^{(t)})}{p(z|x, \theta^{(t)})}
\]

\[
= \sum_z p(z|x, \theta^{(t)}) \log p(x|\theta^{(t)})
\]

\[
= \log p(x|\theta^{(t)}) \sum_z p(z|x, \theta^{(t)}) = \log p(x|\theta^{(t)}) \cdot 1 = \log p(x|\theta^{(t)})
\]

The equation above shows that there is no gap at \( \theta^{(t)} \) based on \( q^{(t+1)} \). That is,

\[ l(\theta^{(t+1)}) \leq l(\theta^{(t)}) \]

This is also shown in the following figure:
2 Hidden Markov Models (HMMs)

The graphic model of the HHM is shown in the following graph:

\[ \begin{array}{c}
q_0 \xrightarrow{A} q_1 \xrightarrow{A} q_2 \ldots \xrightarrow{A} q_T \\
y_0 \xrightarrow{\eta} y_1 \xrightarrow{\eta} y_2 \ldots \xrightarrow{\eta} y_T
\end{array} \]

where:

- \( A_{ij} = p(q_{t+1} = j | q_t = i) \), where \( q^i \) (or \( q^j \)) means the \( i \)-th (or \( j \)-th) state. (A)
- \( p(y_t | q_t = 1) \) is the emission probability of the \( i \)-th state. (\( \eta \))
- \( \pi_i = p(q_0 = i) \). (\( \pi \))

Thus the parameters are \( \theta = (A, \pi, \eta) \). Our goal, given data \( D = \{y_{0,n}, y_{1,n}, \ldots, y_{T,n}\}_{n=1}^N \), is to find \( \hat{\theta}_{ML} \).

Now we do ML estimation of \( \hat{\theta} \) using the EM algorithm:

1. Write out complete log likelihood \((N = 1)\):

\[
\ell_c(\theta) = \log [p(q, y | \theta)] = \log \left[ \pi_{q_0} \prod_{t=0}^{T-1} a_{q_t, q_{t+1}} \prod_{t=0}^{T} p(y_t | q_t, \eta) \right],
\]

where \( \pi_{q_0} \triangleq \prod_i \pi_i^{q_0} \) and \( a_{q_t, q_{t+1}} \triangleq \prod_{i,j} a_{ij}^{q_t q_{t+1}} \). Write expectations with respect to \( p(q | y, \theta^{(t)}) \) as

\[ \langle \cdot \rangle_{(t)}. \]
The expected complete log likelihood:

\[
\langle l_{c}(\theta) \rangle_{(t)} = \left\langle \sum_{i} q_{0}^{i} \log \pi_{i} + \sum_{t=0}^{T-1} \sum_{i,j} q_{t}^{i} q_{t+1}^{j} \log a_{ij} + \sum_{t=0}^{T} \log p(y_{t}|q_{t}, \eta) \right\rangle_{(t)}
\]

\[
= \sum_{i} \left\langle q_{0}^{i} \right\rangle_{(t)} \log \pi_{i} + \sum_{t=0}^{T-1} \left\langle q_{t}^{i} q_{t+1}^{j} \right\rangle_{(t)} \log a_{ij} + \sum_{t=0}^{T} \left\langle \log p(y_{t}|q_{t}, \eta) \right\rangle_{(t)}
\]

Note that \( \left\langle q_{0}^{i} \right\rangle_{(t)} \) and \( \left\langle q_{t}^{i} q_{t+1}^{j} \right\rangle_{(t)} \) are expected sufficiency statistics.


Some notes about the parameters. For example, the stochastic automaton, when the number of states \( K = 3 \), then

\[ q_{t} \in \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\} . \]

The values in the \( A \) matrix (state transition matrix) are the edge weights in the following state transition graph (where outgoing edges for a single state must sum to one):

![Diagram of state transition graph]