CS281A/Stat241A: Statistical Learning Theory

EM algorithm and Hidden Markov Models (10/21/04)

Lecturer: Michael I. Jordan *Scribes: Li Zhuang and Feng Zhou*

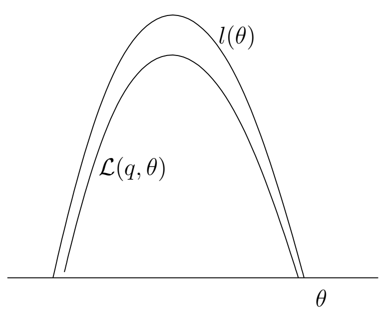
1 EM Algorithm

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: $\hat{\theta}_{\text{ML}} = \text{argmax}_\theta \, p(x|\theta) = \text{argmax}_\theta \sum_z p(x, z|\theta)$

First, we have:

$$\begin{aligned} l(\theta) = \log p(x|\theta) &= \log \sum_z p(x, z|\theta) = \log \sum_z q(z|x) \frac{p(x, z|\theta)}{q(z|x)} \\ &\geq \sum_z q(z|x) \log \frac{p(x, z|\theta)}{q(z|x)} \equiv \mathcal{L}(q, \theta) \quad \text{(according to Jensen's inequality)} \end{aligned}$$



EM is coordinate ascent on $\mathcal{L}$:

- E step: $q^{(t+1)} = \text{argmax}_q \, \mathcal{L}(q, \theta^{(t)})$
- M step: $\theta^{(t+1)} = \text{argmax}_\theta \, \mathcal{L}(q^{(t+1)}, \theta)$

More specifically,

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

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

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

$$\begin{aligned}
\widetilde{\mathcal{L}} &= \mathcal{L} + \lambda(1 - \sum_z q(z|x)) \\
\frac{\partial \widetilde{\mathcal{L}}}{\partial q(z|x)} &= \log p(x, z|\theta^{(t)}) - \log q(z|x) - 1 + \lambda \\
\text{(set=0)} \Rightarrow \ q(z|x) &= p(x, z|\theta^{(t+1)}) e^{\lambda - 1} \\
\text{(sum over } z\text{)} \Rightarrow \ 1 &= \widetilde{\lambda} \sum_z p(x, z|\theta^{(t)}) \quad (\widetilde{\lambda} \equiv e^{\lambda - 1}) \\
\Rightarrow \widetilde{\lambda} &= \frac{1}{p(x|\theta^{(t)})} \\
\Rightarrow q(z|x) &= p(z|x, \theta^{(t)})
\end{aligned}$$

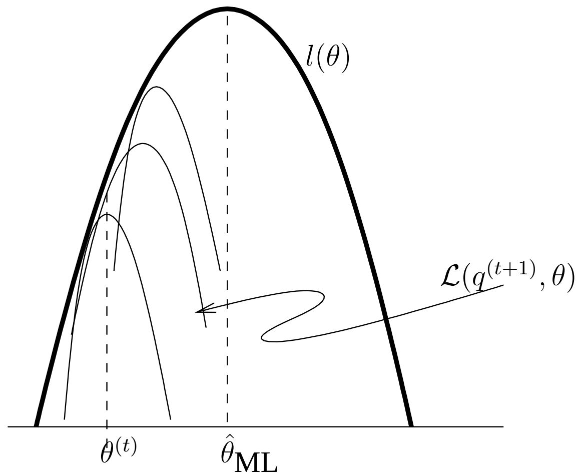
Now we plug $p(z|x, \theta^{(t)})$ into $\mathcal{L}(q, \theta^{(t)})$:

$$\begin{aligned}
\mathcal{L}(q, \theta^{(t)}) &= \mathcal{L}(p(z|x, \theta^{(t)}), \theta^{(t)}) \\
&= \sum_z p(z|x, \theta^{(t)}) \log p(x, z|\theta^{(t)}) - \sum_z p(z|x, \theta^{(t)}) \log p(z|x, \theta^{(t)}) \\
&= \sum_z p(z|x, \theta^{(t)}) \log \frac{p(x, z|\theta^{(t)})}{p(z|x, \theta^{(t)})} \\
&= \sum_z p(z|x, \theta^{(t)}) \log p(x|\theta^{(t)}) \\
&= \log p(x|\theta^{(t)}) \sum_z p(z|x, \theta^{(t)}) = \log p(x|\theta^{(t)}) \cdot 1 = \log p(x|\theta^{(t)})
\end{aligned}$$

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

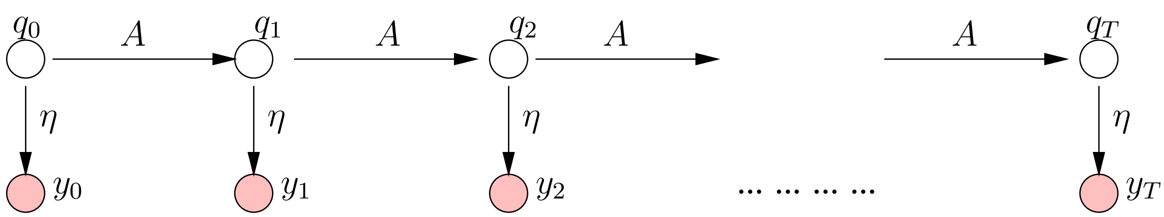
$$l(\theta^{(t+1)}) \leq l(\theta^{(t)}).$$

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:



where:

- $A_{ij} = p(q^j_{t+1} = 1|q^i_t = 1)$, where q^i (or q^j) means the i-th (or j-th) state. ($\mathbf{A}$)
- $p(y_t|q^i_t = 1)$ is the emission probability of the i-th state. ($\boldsymbol{\eta}$)
- $\pi_i = p(q^i_0 = 1)$. ($\boldsymbol{\pi}$)

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

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

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

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

where $\pi_{q_0} \triangleq \prod_i \pi_i^{q^i_0}$ and $a_{q_t,q_{t+1}} \triangleq \prod_{i,j} a_{ij}^{q^i_t q^j_{t+1}}$. Write expectations with respect to $p(q|y, \theta^{(t)})$ as $\langle \ \cdot \ \rangle_{(t)}$.

The expected complete log likelihood:

$$\langle l_c(\theta)\rangle_{(t)} = \left\langle \sum_i q_0^i \log \pi_i + \sum_{t=0}^{T-1} \sum_{i,j} q_t^i q_{t+1}^j \log a_{ij} + \sum_{t=0}^{T} \log p(y_t|q_t, \boldsymbol{\eta}) \right\rangle_{(t)}$$
$$= \sum_i \left\langle q_0^i \right\rangle_{(t)} \log \pi_i + \sum_{t=0}^{T-1} \left\langle q_t^i q_{t+1}^j \right\rangle_{(t)} \log a_{ij} + \sum_{t=0}^{T} \left\langle \log_p(y_t|q_t, \boldsymbol{\eta}) \right\rangle_{(t)}$$

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.

2. M step: maximize expected complete log likelihood.

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):