11-751: Speech recognition and understanding, Fall 2008

A proof for the convergence of Baum-Welch training

Teaching assistant: Wilson Tam (yct@cs.cmu.edu) *Due on:*

Assume that Λ is the current model and Λ' is the old model. We want to show that the likelihood of the observation O always increases for each training iteration until convergence is reached. In other words, we want to maximize the log likelihood ratio as follows:

$$
\begin{aligned}
\log \frac{p(O|\Lambda)}{p(O|\Lambda')} &= \sum_Q \log \frac{p(O|\Lambda) \cdot p(Q|O,\Lambda') \cdot p(Q|O,\Lambda)}{p(O|\Lambda') \cdot p(Q|O,\Lambda') \cdot p(Q|O,\Lambda)} && (1) \\
&= \sum_Q \log \frac{p(O,Q|\Lambda)}{p(O,Q|\Lambda')} \cdot \frac{p(Q|O,\Lambda')}{p(Q|O,\Lambda)} && (2) \\
&= \sum_Q p(Q|O,\Lambda') \log \frac{p(O,Q|\Lambda)}{p(O,Q|\Lambda')} \cdot \frac{p(Q|O,\Lambda')}{p(Q|O,\Lambda)} && (3) \\
&= \sum_Q p(Q|O,\Lambda') \log \frac{p(O,Q|\Lambda)}{p(O,Q|\Lambda')} + \sum_Q p(Q|O,\Lambda') \log \frac{p(Q|O,\Lambda')}{p(Q|O,\Lambda)} && (4) \\
&= L(O,\Lambda;\Lambda') + KL(p(Q|O,\Lambda')||p(Q|O,\Lambda)) && (5) \\
&\geq L(O,\Lambda;\Lambda') && (6)
\end{aligned}
$$

where KL() denotes the Kullback-Leibler divergence (or distance) between two probability distributions which must be non-negative. Q denotes the state sequence. Since the Baum-Welch training maximizes $L(O,\Lambda;\Lambda')$, the likelihood of the observation O is guaranteed to improve until $\Lambda = \Lambda'$.