Derivation of Baum-Welch Algorithm for Hidden Markov Models

Stephen Tu

1 Introduction

This short document goes through the derivation of the Baum-Welch algorithm for learning model parameters of a hidden markov model (HMM). For more generality, we treat the multiple observations case. Note that Baum-Welch is simply an instantiation of the more general Expectation-Maximization (EM) algorithm.

2 Setup

Let us consider discrete (categorical) HMMs of length $T$ (each observation sequence is $T$ observations long). Let the space of observations be $X = \{1, 2, ..., N\}$, and let the space of underlying states be $Z = \{1, 2, ..., M\}$. An HMM $\theta = (\pi, A, B)$ is parameterized by the initial state matrix $\pi$, the state transition matrix $A$, and the emission matrix $B$; $\pi_i = P(z_1 = i)$, $A_{ij} = P(z_{t+1} = j | z_t = i)$, and $B_i(j) = P(x_t = j | z_t = i)$. See [1] for a more detailed treatment of HMMs.

We study the problem of learning the parameterization of $\theta$ from a dataset of $D$ observations. Let $X = (X^{(1)}, ..., X^{(D)})$, where each $X^{(i)} = (x_1^{(i)}, x_2^{(i)}, ..., x_T^{(i)})$. We assume each observation is drawn iid. The learning problem is non-trivial because we are not given the latent variables $Z^{(i)}$ for each $X^{(i)}$, otherwise we could directly compute $\theta^* = \arg\max_\theta P(X, Z; \theta)$. Without $Z$, the naive solution would be to directly compute $\theta^* = \arg\max_\theta \sum_{z \in Z} P(X, z; \theta)$. This is not tractable, since there are $DT^M$ different values of $z$ to try.

3 Baum-Welch

Baum-Welch is an iterative procedure for estimating $\theta^*$ from only $X$. It works by maximizing a proxy to the log-likelihood, and updating the current model to be closer to the optimal model. Each iteration of Baum-Welch is guaranteed to increase the log-likelihood of the data. But of course, convergence to the optimal solution is not guaranteed.

Baum-Welch can be described simply as repeating the following steps until convergence:

1. Compute $Q(\theta, \theta^*) = \sum_{z \in Z} \log [P(X, z; \theta)] P(z|X; \theta^*)$.

2. Set $\theta^{s+1} = \arg\max_\theta Q(\theta, \theta^*)$.

Without justifying why this works, the rest of this document will focus on deriving the necessary update steps to run this algorithm. First, noting that $P(z, X) = P(X) P(z|X)$, we can write

$$\arg\max_\theta \sum_{z \in Z} \log [P(X, z; \theta)] P(z|X; \theta^*) = \arg\max_\theta \sum_{z \in Z} \log [P(X, z; \theta)] P(z, X; \theta^*)$$

since $P(X)$ is not affected by choice of $\theta$. Now $P(z, X; \theta)$ is easy to write down

$$P(z, X; \theta) = \prod_{d=1}^D \left( \pi_{z_1^{(d)}} B_{z_1^{(d)}}(x_1^{(d)}) \prod_{t=2}^T A_{z_{t-1}^{(d)} z_t^{(d)}} B_{z_t^{(d)}}(x_t^{(d)}) \right)$$
Taking the log gives us

$$\log P(z, X; \theta) = \sum_{d=1}^{D} \left[ \log \pi_{z_1}^{(d)} + \sum_{t=2}^{T} \log A_{z_{t-1}, z_t}^{(d)} + \sum_{t=1}^{T} \log B_{z_t}^{(d)}(x_t^{(d)}) \right]$$

Plugging this into \( \hat{Q}(\theta, \theta^*) \), we get

$$\hat{Q}(\theta, \theta^*) = \sum_{z \in \mathcal{Z}} \sum_{d=1}^{D} \log \pi_{z_1}^{(d)} P(z, X; \theta^*) + \sum_{z \in \mathcal{Z}} \sum_{d=1}^{D} \sum_{t=2}^{T} \log A_{z_{t-1}, z_t}^{(d)} P(z, X; \theta^*) + \sum_{z \in \mathcal{Z}} \sum_{d=1}^{D} \sum_{t=1}^{T} \log B_{z_t}^{(d)}(x_t^{(d)}) P(z, X; \theta^*)$$

This is a nice form which we can optimize analytically with Lagrange multipliers. We need Lagrange multipliers because we have equality constraints which come from requiring that \( \pi, A_i \), and \( B_i(\cdot) \) form valid probability distributions. Let \( \hat{L}(\theta, \theta^*) \) be the Lagrangian

$$\hat{L}(\theta, \theta^*) = \hat{Q}(\theta, \theta^*) - \lambda \left( \sum_{i=1}^{M} \pi_i - 1 \right) - \sum_{i=1}^{M} \lambda A_i \left( \sum_{j=1}^{M} A_{ij} - 1 \right) - \sum_{i=1}^{M} \lambda B_i \left( \sum_{j=1}^{N} B_i(j) - 1 \right)$$

First let us focus on the \( \pi_i \)'s

$$\frac{\partial \hat{L}(\theta, \theta^*)}{\partial \pi_i} = \frac{\partial}{\partial \pi_i} \left( \sum_{z \in \mathcal{Z}} \sum_{d=1}^{D} \log \pi_{z_1}^{(d)} P(z, X; \theta^*) \right) - \lambda = 0$$

$$= \frac{\partial}{\partial \pi_i} \left( \sum_{j=1}^{M} \sum_{d=1}^{D} \log \pi_j P(z_1^{(d)} = j, X; \theta^*) \right) - \lambda = 0$$

$$= \sum_{d=1}^{D} \frac{P(z_1^{(d)} = i, X; \theta^*)}{\pi_i} - \lambda = 0$$

$$\frac{\partial \hat{L}(\theta, \theta^*)}{\partial \lambda} = - \left( \sum_{i=1}^{M} \pi_i - 1 \right) = 0$$

The second step is simply the result of marginalizing out, for each \( d \), all \( z_1^{(d)} \) and \( z_{t \neq d} \) for all \( t \). We use this style of trick extensive throughout the remainder of the document. Some algebra yields

$$\pi_i = \frac{\sum_{d=1}^{D} P(z_1^{(d)} = i, X; \theta^*)}{\sum_{d=1}^{D} \sum_{j=1}^{M} P(z_1^{(d)} = j, X; \theta^*)} = \frac{\sum_{d=1}^{D} P(z_1^{(d)} = i, X; \theta^*)}{\sum_{d=1}^{D} \sum_{j=1}^{M} P(z_1^{(d)} = j, X; \theta^*)} = \frac{\sum_{d=1}^{D} P(z_1^{(d)} = i, X; \theta^*)}{\sum_{d=1}^{D} P(X; \theta^*)} = \frac{\sum_{d=1}^{D} P(X; \theta^*) P(z_1^{(d)} = i | X; \theta^*)}{\sum_{d=1}^{D} P(X; \theta^*)} = \frac{1}{D} \sum_{d=1}^{D} P(z_1^{(d)} = i | X; \theta^*) = \frac{1}{D} \sum_{d=1}^{D} P(z_1^{(d)} = i | X^{(d)}; \theta^*)$$

2
We now follow a similar process for the $A_{ij}$’s.

\[
\frac{\partial \hat{L}(\theta, \theta^*)}{\partial A_{ij}} = \frac{\partial}{\partial A_{ij}} \left( \sum_{z \in Z} \sum_{d=1}^{D} \sum_{t=2}^{T} \log A_{z_{t-1} \rightarrow z_t} P(z, \mathcal{X}; \theta^*) \right) - \lambda_{A_i} = 0
\]

\[
= \frac{\partial}{\partial A_{ij}} \left( \sum_{j=1}^{M} \sum_{d=1}^{D} \sum_{t=2}^{T} \log A_{j,k} P(z_{t-1}^{(d)} = j, z_t^{(d)} = k; \mathcal{X}; \theta^*) \right) - \lambda_{A_i} = 0
\]

\[
= \sum_{d=1}^{D} \sum_{t=2}^{T} \frac{P(z_{t-1}^{(d)} = i, z_t^{(d)} = j; \mathcal{X}; \theta^*)}{A_{ij}} - \lambda_{A_i} = 0
\]

This yields

\[
A_{ij} = \frac{\sum_{d=1}^{D} \sum_{t=2}^{T} P(z_{t-1}^{(d)} = i, z_t^{(d)} = j; \mathcal{X}; \theta^*)}{\sum_{j=1}^{M} \sum_{d=1}^{D} \sum_{t=2}^{T} P(z_{t-1}^{(d)} = i, z_t^{(d)} = j; \mathcal{X}; \theta^*)}
\]

\[
= \frac{\sum_{d=1}^{D} \sum_{t=2}^{T} P(\mathcal{X}; \theta^*) P(z_{t-1}^{(d)} = i, z_t^{(d)} = j| \mathcal{X}; \theta^*)}{\sum_{d=1}^{D} \sum_{t=2}^{T} P(\mathcal{X}; \theta^*) P(z_{t-1}^{(d)} = i| \mathcal{X}; \theta^*)}
\]

\[
= \frac{\sum_{d=1}^{D} \sum_{t=2}^{T} P(z_{t-1}^{(d)} = i, z_t^{(d)} = j| \mathcal{X}^{(d)}; \theta^*)}{\sum_{d=1}^{D} \sum_{t=2}^{T} P(z_{t-1}^{(d)} = i| \mathcal{X}^{(d)}; \theta^*)}
\]

The final thing is the $B_i(j)$’s, which are slightly trickier. Let $I(x)$ denote an indicator function which is 1 if $x$ is true, 0 otherwise.

\[
\frac{\partial \hat{L}(\theta, \theta^*)}{\partial B_i(j)} = \frac{\partial}{\partial B_i(j)} \left( \sum_{z \in Z} \sum_{d=1}^{D} \sum_{t=1}^{T} \log B_{z_t \rightarrow z_t^{(d)}} (x_t^{(d)}) P(z, \mathcal{X}; \theta^*) \right) - \lambda_{B_i} = 0
\]

\[
= \frac{\partial}{\partial B_i(j)} \left( \sum_{i=1}^{N} \sum_{d=1}^{D} \sum_{t=1}^{T} \log B_i \left( x_t^{(d)} \right) P(z_t^{(d)} = i, \mathcal{X}; \theta^*) \right) - \lambda_{B_i} = 0
\]

\[
= \sum_{d=1}^{D} \sum_{t=1}^{T} \frac{P(z_t^{(d)} = i, \mathcal{X}; \theta^*) I(x_t^{(d)} = j)}{B_i(j)} - \lambda_{B_i} = 0
\]

\[
\frac{\partial \hat{L}(\theta, \theta^*)}{\partial \lambda_{B_i}} = - \left( \sum_{j=1}^{N} B_i(j) - 1 \right) = 0
\]
This should come as no surprise by now

\[
B_s(j) = \frac{\sum_{d=1}^{D} \sum_{t=1}^{T} P(z_{t}^{(d)} = i, X; \theta^{*}) I(x_{t}^{(d)} = j)}{\sum_{j=1}^{N} \sum_{d=1}^{D} \sum_{t=1}^{T} P(z_{t}^{(d)} = i, X; \theta^{*}) I(x_{t}^{(d)} = j)}
\]

\[
= \frac{\sum_{d=1}^{D} \sum_{t=1}^{T} P(z_{t}^{(d)} = i, X; \theta^{*}) I(x_{t}^{(d)} = j)}{\sum_{d=1}^{D} \sum_{t=1}^{T} P(z_{t}^{(d)} = i, X; \theta^{*})}
\]

\[
= \frac{\sum_{d=1}^{D} \sum_{t=1}^{T} P(z_{t}^{(d)} = i|X^{(d)}; \theta^{*}) I(x_{t}^{(d)} = j)}{\sum_{d=1}^{D} \sum_{t=1}^{T} P(z_{t}^{(d)} = i|X^{(d)}; \theta^{*})}
\]

To summarize, the update steps are

\[
\pi_{i}^{(s+1)} = \frac{1}{D} \sum_{d=1}^{D} P(z_{1}^{(d)} = i|X^{(d)}; \theta^{*})
\]

\[
A_{i,j}^{(s+1)} = \frac{\sum_{d=1}^{D} \sum_{t=2}^{T} P(z_{t-1}^{(d)} = i, z_{t}^{(d)} = j|X^{(d)}; \theta^{*})}{\sum_{d=1}^{D} \sum_{t=2}^{T} P(z_{t-1}^{(d)} = i|X^{(d)}; \theta^{*})}
\]

\[
B_{i}^{(s+1)}(j) = \frac{\sum_{d=1}^{D} \sum_{t=1}^{T} P(z_{t}^{(d)} = i|X^{(d)}; \theta^{*}) I(x_{t}^{(d)} = j)}{\sum_{d=1}^{D} \sum_{t=1}^{T} P(z_{t}^{(d)} = i|X^{(d)}; \theta^{*})}
\]

Note that \( P(z_{t}|X; \theta) \) and \( P(z_{t-1}, z_{t}|X; \theta) \) are both quantities which can be computed efficiently for HMMs by the forward-backwards algorithm. Once again, see [1] for more details.

**References**