Expectation Maximization for Hidden Markov Models:
Derived from First Principles

Jesse Hoey

December 15, 2000

Abstract

Expectation Maximization is a popular technique for deriving MAP estimation of model parameters. A successful application is learning parameters of hidden Markov models (HMMs). This note derives the Baum-Welsh learning algorithm for HMMs from first principles.

1 Introduction

A hidden Markov model is a Bayesian network as shown (unrolled) in Figure 1. A set of hidden states X_t generates a set of observed variables Z_t at each of a set of times $t = 0, ..., T$. We refer to the set of states at all times as $X = \{X_0, ..., X_T\}$ and $Z = \{Z_0, ..., Z_T\}$. We assume for now that both the observations and the hidden variables are discrete valued with N_z and N_x values each. At time t, Z_t, is assumed conditionally independent of all other variables at all other times given its parent X_t, and the usual Markovian independence assumption holds between the X_ts. The parameters of the model are threefold. First, the *transition probabilities*, $\Theta_{xij} = P(x_{t,i}|x_{t-1,j})$, give the probability that the hidden state at time t will be $X_t = i$, given that the state at the previous time

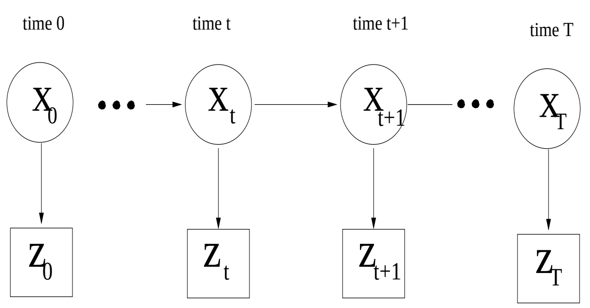

Figure 1: Hidden Markov model as a dynamic Bayesian network.

step $t-1$ was $X_t = j$. Second, the *starting probabilities*, $\Pi_i = P(X_{0,i})$, are the probabilities that the system starts in state $X_0 = i$. Third, the *emission probabilities*, $\Theta_{zij} = P(Z_{t,i}|X_{t,j})$, are the probability that observation $Z_t = i$ will be made while in state $X_t = j$. We refer to the set of all parameters as $\Theta = \{\Theta_X, \Theta_Z, Pi\}$.

Hidden Markov models have been extensively studied in the context of speech recognition [5], along with many other applications in areas such as vision [2, 4]. The focus of this note is the learning the maximum *a-posteriori* estimates of the parameters, Θ, given by the values which maximize $P(Z, \Theta)$. The standard technique is an application of the expectation-maximization (EM) algorithm of [1], which results in an efficient recursive estimation technique known as the Baum-Welsh or forward-backward training procedure. This note will provide a simple derivation of this particular application of the EM algorithm. The first part, (Section 2), borrows heavily from the derivation in [3], as concisely described by Thomas Minka [1], and gives a general derivation of the expectation and maximization steps which can be applied to any model. This is followed in Section 3 by a derviation of the forward-backward equations assuming multinomial distributions and Dirichlet priors.

2 EM derivation

2.1 General EM

We wish to find the values of Θ which maximize

$$f(\Theta) = P(Z, \Theta) = \int_X P(Z, X, \Theta),$$

where the integral is over the values of the *hidden* variables X in the HMM. We lower bound the value of $f(\Theta)$ with a function $q(X)$:

$$f(\Theta) = \int_X \frac{P(Z, X, \Theta)}{q(X)} q(X) \geq \prod_X \left(\frac{P(Z, X, \Theta)}{q(X)} \right)^{q(X)}, \tag{1}$$

where the inequality is given by Jensen's inequality. Taking logarithms we get

$$G(\Theta, q) = \int_X q(X) \log P(Z, X, \Theta) - q(X) \log q(X). \tag{2}$$

Then, at the current guess for Θ, Θ', we choose q to maximize G, so that g touches f at Θ'. We use the constraint that $\int_X q(X) = 1$ and get

$$q(X) = \frac{P(Z, X, \Theta)}{\int_X P(Z, X, \Theta)} = \frac{P(Z, X, \Theta)}{P(Z, \Theta)} = P(X|Z, \Theta)$$

The idea is then to

[1] See his excellent set of tutorial notes at http://www-white.media.mit.edu/ tpminka/papers/tutorial.html. The EM tutorial is at ftp://vismod.www.media.mit.edu/pub/tpminka/papers/minka-em-tut.ps.gz

1. Choose $q(X)$ to maximize the bound at the current guess, Θ'. This just means getting $P(X|Z, \Theta)$ and is the "E" step of EM.

2. Maximize the bound over Θ. This means maximizing 2 with the $q(X)$ derived in the "E" step. That is, we maximize

$$\int_X P(X|Z, \Theta') \log P(Z, X, \Theta). \tag{3}$$

2.2 Multinomial-Dirichlet EM

This section will specialize the above procedure for probability distributions in the multinomial family, with Dirichlet conjugate priors. We show the derivation for Θ_{Xij} in what follows. The other parameters can be similarily derived. Maximization of equation 3 can be performed by noting that there is a constraint on Θ_{Xij},

$$\sum_i \Theta_{Xij} = 1,$$

which means that we want to solve

$$\frac{\partial}{\partial \Theta_{Xij}} \left[\int_X P(X|Z, \Theta') \log P(Z, X, \Theta) - \lambda(\sum_i \Theta_{Xij} - 1) \right] = 0$$

which is

$$\int_X P(X|Z, \Theta') \frac{\partial}{\partial \Theta_{Xij}} \left[\log P(Z, X|\Theta) P(\Theta) \right] = \lambda. \tag{4}$$

Assuming that the likelihood function of the completed data is given by a multinomial distribution,

$$P(Z, X|\Theta') = \prod_{ij} \Theta_{Xij}^{N_{Xij}} \Theta_{Zij}^{N_{Zij}} \Pi_i^{N_i},$$

were N_{Xij} is the number of times $X_t = i$ when $X_{t-1} = j$ in the (completed) data. Assuming further that the priors are given by a Dirichlet distribution

$$P(\Theta) = P(\Theta_X)P(\Theta_Z)P(\Pi) = Dir(\Theta|\alpha_{Xij})Dir(\Theta_Z|\alpha_{Zij})Dir(\Pi|\alpha_i),$$

then equation 4 becomes

$$\begin{aligned}\int_X P(X|Z, \Theta') \frac{\partial}{\partial \Theta_{Xij}} [\sum_{ij} (N_{Xij} + \alpha_{Xij}) \log \Theta_{Xij} &+ \\ \sum_{ij} (N_{Zij} + \alpha_{Zij}) \log \Theta_{Zij} &+ \\ \sum_i (N_i + \alpha_i) \log \Pi_i] &= \lambda.\end{aligned}$$

And therefore,

$$\int_X P(X|Z,\Theta')\frac{N_{Xij}+\alpha_{Xij}}{\Theta_{Xij}} = \lambda$$

$$\lambda = \sum_i \int_X P(X|Z,\Theta')$$

$$\begin{aligned}
\Theta_{Xij} &= \frac{\int_X P(X|Z,\Theta')N_{Xij}+\alpha_{Xij}}{\sum_i \int_X P(X|Z,\Theta')N_{Xij}+\alpha_{Xij}} \\
&= \frac{\int_X P(X|Z,\Theta')N_{Xij}+\alpha_{Xij}}{\sum_i \int_X P(X|Z,\Theta')N_{Xij}+\alpha_{Xij}} \\
&= \frac{\alpha_{Xij}+E_{P(X|Z,\Theta')}(N_{Xij})}{\sum_i \alpha_{Xij}+E_{P(X|Z,\Theta')}(N_{Xij})}
\end{aligned}$$

So we see that the parameters can be updates by simply taking the expected counts, which form the *sufficient statistics* for the multinomial distribution.

3 Baum-Welsh Derivation

Our goal is then to find the expectations $E_{P(X|Z,\Theta')}(N_{Xij})$, which is

$$\begin{aligned}
E_{P(X|Z,\Theta')}(N_{Xij}) &= \int_X P(X|Z,\Theta')N_{Xij} \\
&= \int_{X_0...X_T} P(X_0...X_T|Z_0...Z_T,\Theta')\sum_t \delta(X_{t,i})\delta(X_{t-1,j}) \quad (5) \\
& \quad (6)
\end{aligned}$$

where

$$\delta(X_{t,i}) = \begin{cases} 1 & \text{if } X_t = i \\ 0 & \text{otherwise} \end{cases}$$

The sum over t can be taken outside the integrations in (6), and the integrations over $X_0...X_{t-2}, X_{t+1}...X_T$ can be immediately performed, each giving factors of 1. Further, the integrations over X_{t-1} and X_t can be performed, since the δ-functions simply pick out a particular value of these varibles: $X_{t-1,j}$ and $X_{t,i}$. Thus, we are left with:

$$E_{P(X|Z,\Theta')}(N_{Xij}) = \sum_{t=1}^{T} P(X_{t,i}, X_{t-1,j}|Z,\Theta), \tag{7}$$

the expected number of times $X_t = i$ and $X_{t-1} = j$ given the data, $Z = Z_0...Z_T$, and the current model parameters, Θ. Factoring the term in the sum by splitting the data Z up into two sets $Z_0...Z_{t-1}$ and $Z_t...Z_T$ gives the Baum-Welsh equations for updating the parameters of the HMM using expectation maximization (we leave out the Θ upon which every term is conditioned)

$$\begin{aligned}
&P(X_{t,i}, X_{t-1,j}|Z,\Theta)\\
&= P(Z_t...Z_T|X_{t,i}, X_{t-1,j}, Z_0...Z_{t-1})P(X_{t,i}, X_{t-1,j}, Z_0...Z_{t-1})\\
&= P(Z_t...Z_T|X_{t,i}X_{t-1,j})P(X_{t,i}|X_{t-1,j}Z_0...Z_{t-1})P(X_{t-1,j}|Z_0...Z_{t-1})\\
&= P(Z_t|X_{t,i}X_{t-1,j}Z_{t+1}...Z_T)P(Z_{t+1}...Z_T|X_{t,i}X_{t-1,j})\\
&\qquad P(X_{t,i}|X_{t-1,j})\frac{P(X_{t-1,j}Z_0...Z_{t-1})}{P(Z_0...Z_{t-1})}\\
&= P(Z_t|X_{t,i})P(Z_{t+1}...Z_T|X_{t,i})P(X_{t,i}|X_{t-1,j})P(X_{t-1,j}Z_0...Z_{t-1})\\
&= \Theta_{Z_t*i}\beta_{t,i}\Theta_{Xij}\alpha_{t-1,j}
\end{aligned}$$

which is exactly the (unnormalized) equation (37) from [5]. We use Θ_{Z_t*i}, where the $*$ represents the value of the observation Z_t. We have left to evaluate the α and β terms (the forward and backward variables, respectively). The alpha term is simply the joint probability of $X_t = j$ and all the observations prior to time t. It can be expanded recursively by summing over the previous states, X_{t-1}, as follows:

$$\begin{aligned}
\alpha_{t,j} &= P(X_{t,j}Z_0...Z_t)\\
&= P(Z_t|X_{t,j}Z_0...Z_{t-1})P(X_{t,j}Z_0...Z_{t-1})\\
&= P(Z_t|X_{t,j})\sum_k P(X_{t,j}X_{t-1,k}Z_0...Z_{t-1})\\
&= P(Z_t|X_{t,j})\sum_k P(X_{t,j}|X_{t-1,k})P(X_{t-1,k}Z_0...Z_{t-1})\\
&= \Theta_{Z_t*j}\sum_k \Theta_{Xjk}\alpha_{t-1,k}
\end{aligned}$$

Similarly, the beta term is the probability of all observations posterior to time t, given the state at time t, X_t. It can be expanded recursively by summing over all next states, X_{t+1} as follows:

$$\begin{aligned}
\beta_{t,i} &= P(Z_{t+1}...Z_T|X_{t,i})\\
&= \sum_k P(Z_{t+1}|X_{t+1,k})P(Z_{t+2}...Z_T|X_{t+1,k})P(X_{t+1,k}|X_{t,i})\\
&= \sum_k \Theta_{Z_{t+1}*k}\beta_{t+1,k}\Theta_{Xki}
\end{aligned}$$

The terms α_0 and β_T must be evaluated separately. The β_T is initialized evenly, and the $\alpha_{0,i} = \Theta_{Z0i}\Pi_i$.

4 Baum-Welsh for POMDPs

Consider further that we not only have evidence conditioned on (or generated by) the hidden states, X, but that there is evidence which the hidden states are

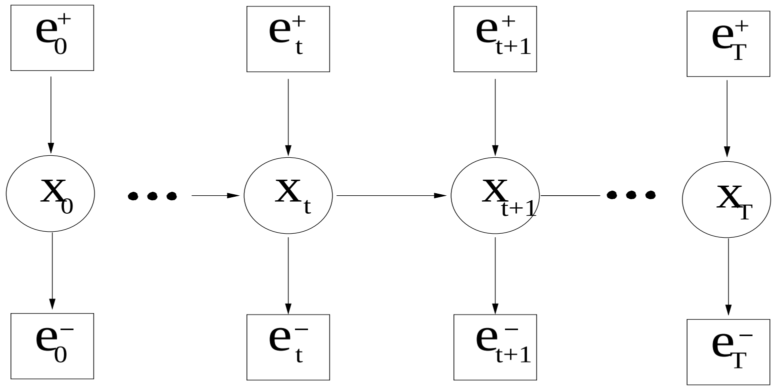

Figure 2: POMDP as a dynamic Bayesian network.

conditioned on, as shown in Figure 2. This type of model is called a partially observable Markov decision process if the evidence e_t^+ are actions, a, taken from a set of possible actions, $\mathcal{A}$. Now the transition probabilities, Θ_X are replaced by conditional probabilities for the hidden states given the actions and the previous states, $\Theta_{Xijk} = P(X_{t,i}|e^+_{t,k}X_{t-1,j})$. Now we are interested in examining the probability distributions conditioned on both sets of evidence e^+ and e^-. We refer to the set of both evidences as simply $e = e_0^+...e_T^+, e_0^-...e_T^-$, and $e_t = e_t^+, e_t^-$.

$$\begin{aligned}
&P(X_{t,i}, X_{t-1,j}|e,\Theta)\\
&\quad= P(e_t...e_T|X_{t,i},X_{t-1,j},e_0...e_{t-1})P(X_{t,i},X_{t-1,j}|e_0...e_{t-1})\\
&\quad= P(e_t...e_T|X_{t,i}X_{t-1,j})P(X_{t-1,j}|e_0...e_{t-1})P(X_{t,i}|X_{t-1,j}e_0...e_{t-1})\\
&\quad= P(e_t|X_{t,i}X_{t-1,j}e_{t+1}...e_T)P(e_{t+1}...e_T|X_{t,i}X_{t-1,j})\\
&\qquad\quad \frac{P(X_{t-1,j}e_0...e_{t-1})}{P(e_0...e_{t-1})}P(X_{t,i}|X_{t-1,j})\\
&\quad= P(e_t^-e_t^+|X_{t,i}X_{t-1,j})P(e_{t+1}...e_T|X_{t,i})P(X_{t-1,j}e_0...e_{t-1})P(X_{t,i}|X_{t-1,j})\\
&\quad= P(e_t^-|X_{t,i}e_t^+)P(e_t^+|X_{t,i}X_{t-1,j})\beta_{t,i}\alpha_{t-1,j}\\
&\qquad\quad \left[\sum_k P(X_{t,i}|e^+_{t,k}X_{t-1,j})P(e^+_{t,k}|X_{t-1,j})\right]\\
&\quad= P(e_t^-|X_{t,i})P(X_{t,i}|e_t^+X_{t-1,j})P(e_t^+|X_{t-1,j})\beta_{t,i}\alpha_{t-1,j}\\
&\qquad\quad \left[\sum_k \Theta_{Xijk}P(e^+_{t,k}|X_{t-1,j})\right]\\
&\quad= \Theta_{e_t^-*i}\Theta_{Xij*}P(e_t^+|X_{t-1,j})\beta_{t,i}\alpha_{t-1,j}\sum_k \Theta_{Xijk}P(e^+_{t,k}|X_{t-1,j})
\end{aligned}$$

The new term which shows up is the $P(e_t^+|X_{t-1,j})$, which is referred to as the *policy* in a POMDP. With a deterministic policy, this term simply fixes the value of e_t^+ as the action taken. However, in a more general case, the policy is probabilistic, and the 'actions' e^+ are not fixed. We can evaluate the alpha and beta terms much as before, but they now include the policy term as well. The alpha term is:

$$
\begin{aligned}
\alpha_{t,j} &= P(X_{t,j}e_0...e_t) \\
&= P(e_t|X_{t,j}e_0...e_{t-1})P(X_{t,j}e_0...e_{t-1}) \\
&= P(e_t^+e_t^-|X_{t,j})\sum_k P(X_{t,j}X_{t-1,k}e_0...e_{t-1}) \\
&= P(e_t^-|X_{t,j})P(e_t^+|X_{t,j})\sum_k P(X_{t,j}|X_{t-1,k}e_0...e_{t-1})P(X_{t-1,k}e_0...e_{t-1}) \\
&= \Theta_{e_t^-*j}\sum_k \Theta_{Xjk*}\alpha_{t-1,k}
\end{aligned}
$$

The beta term gives:

$$
\begin{aligned}
\beta_{t,i} &= P(e_{t+1}...e_T|X_{t,i}) \\
&= \sum_k P(e_{t+1}^-|X_{t+1,k})P(e_{t+2}...e_T|X_{t+1,k})P(X_{t+1,k}|e_{t+1}^+X_{t,i})P(e_{t+1}^+|X_{t,i}) \\
&= \sum_k \Theta_{e_{t+1}^-*k}\beta_{t+1,k}\Theta_{Xk*i}P(e_{t+1}^+|X_{t,i})
\end{aligned}
$$

The new term which shows up is the $P(e_t^+|X_{t-1,j})$, which is referred to as the *policy* in a POMDP. With a deterministic policy, this term simply fixes the value of e_t^+ as the action taken.

References

[1] A.P. Dempster, N.M.Laird, and D.B. Rubin. Maximum likelihood from incomplete data using the EM algorithm. *Journal of the Royal Statistical Society*, 39(B), 1977.

[2] Jesse Hoey and James J. Little. Representation and recognition of complex human motion. In *Proc. IEEE CVPR*, Hilton Head, SC, June 2000.

[3] Stephen P. Luttrell. Partitioned mixture distribution: an adaptive Bayesian network for low-level vision processing. *IEEE Proc. on Vision, Image and Signal Processing*, 141(4):251–260, 1994.

[4] Carlos Morimoto, Yaser Yacoob, and Larry Davis. Recognition of head gestures using hidden Markov models. In *Proceeding of ICPR*, pages 461–465, Austria, 1996.

[5] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–296, February 1989.