Baum-Welch for Hidden Markov Models

Raphaël Reynouard

September 26, 2022

1 Introduction

This document describes the Baum-Welch algorithm [1] for Hidden Markov Models.

2 Preliminaries

We define a Hidden Markov Model (HMM) formally as follow:

Definition 2.1 (Hidden Markov Model) *A HMM is a tuple* $\langle S, \mathcal{L}, \pi, a, b \rangle$ *where:*

- S *is a set of states,*
- $\mathcal{L}$ *is a set of observations,*
- $\pi := \mathcal{D}(S)$ *is the initial distribution i.e. the model starts in state* s *with probability* $\pi(s) := \pi_s$*,*
- $a : S \mapsto \mathcal{D}(S)$ *is the transition function. The model moves from state* s *to* s' *with probability* $a(s)(s') := a_{s,s'}$*,*
- $b : S \mapsto \mathcal{D}(\mathcal{L})$ *is the generation function. In state* s*, the model generates* ℓ *with probability* $b(s)(\ell) := b_{s,\ell}$*.*

A path is a sequence in $\mathbf{Paths} = (S \times \mathcal{L})^* S$ representing a finite execution of a HMM $\mathcal{M}$, and a trace is a finite sequence in $\mathbf{Traces} = \mathcal{L}^*$ representing a finite execution of a HMM for which we cannot see the states.
For $i \in \mathbb{N}_{>0}$, we define $X_i \colon \mathbf{Paths} \to S$, $Y_i \colon \mathbf{Paths} \to \mathcal{L}$, and $O_i \colon \mathbf{Paths} \to \mathbf{Traces}$ respectively as $X_i(\rho) = s_i$, $Y_i(\rho) = \ell_i$, and $O_i(\rho) = \ell_1 \cdots \ell_i$, where $\rho = (s_1, \ell_1)(s_2, \ell_2) \cdots (s_n, \ell_n)s_{n+1}$ is a path.
We denote by $|\rho|$ the length of a path ρ, i.e. the number of observations in this path, and by $|o|$ the length of a trace o.
We denote by $\mathcal{D}(\Omega)$ the set of discrete probability distributions on Ω. The *Dirac distribution* concentrated at x is the distribution $1_x \in \mathcal{D}(\Omega)$ defined, for arbitrary $y \in \Omega$, as $1_x(y) = 1$ if $x = y$, 0 otherwise.

A path of length T can be built from a sequence $\gamma = s_1 \dots s_{T+1}$ of states and a trace $o = \ell_1 \dots \ell_T$. A such path is $o : \gamma := s_1 \ell_1 s_2 \ell_2 \dots s_T \ell_T s_{T+1}$.

We denote by $l(\rho; \mathcal{M})$ the likelihood of a path ρ under a model $\mathcal{M}$, and by $l(o; \mathcal{M})$ the likelihood of a trace o under a model $\mathcal{M}$. We have:

$$\begin{aligned}
l(\rho; \mathcal{M}) &= \pi_{s_1} \prod_{t=1}^{|\rho|} a(s_t)(s_{t+1}) \times b(s_t)(\ell_t) \\
l(o; \mathcal{M}) &= \sum_{\gamma \in S^{|o|}} l(o : \gamma; \mathcal{M})
\end{aligned}$$

Hence:

$$\ln l(\rho; \mathcal{M}) = \ln \pi_{s_1} + \sum_{t=1}^{|\rho|} \ln a(s_t)(s_{t+1}) + \sum_{t=1}^{|\rho|} \ln b(s_t)(\ell_t) \tag{1}$$

Now we define $\gamma_o \colon S \times \{1 \,.\,.\, T+1\} \to [0,1]$ and $\xi_o \colon S \times \{1 \,.\,.\, T\} \times S \to [0,1]$ as

$$\begin{aligned}
\gamma_o(s,t) &= Pr^{\mathcal{M}}[X_t = s | O_T = o] \,, \\
\xi_o(s,t)(s') &= Pr^{\mathcal{M}}[X_t = s, X_{t+1} = s' | O_T = o] \,.
\end{aligned}$$

Intuitively, $\gamma_o(s,t)$ is the likelihood of being in state s at the t-th steps, and $\xi_o(s,t)(s')$ is the likelihood that the t-th transition is from s to s'.

We define the forward and the backward functions $\alpha_o, \beta_o \colon S \times \{1 \,.\,.\, T+1\} \to [0,1]$ as

$$\begin{aligned}
\alpha_o(s,t) &= Pr^{\mathcal{M}}[Y_{1:t-1} = \ell_1 \,.\,.\, \ell_{t-1}, X_t = s] \,, \text{ and} \\
\beta_o(s,t) &= Pr^{\mathcal{M}}[Y_{t:T} = \ell_t \,.\,.\, \ell_T | X_t = s] \,.
\end{aligned}$$

These can be calculated according to the following recurrences

$$\begin{aligned}
\alpha_o(s,t) &= \begin{cases} \pi(s) & \text{if } t = 1 \\ \displaystyle\sum_{s' \in S} \alpha(s', t-1) \cdot a(s')(s) \cdot b(s')(\ell_{t-1}) & \text{if } 1 < t \leq T+1 \end{cases} \\
\beta_o(s,t) &= \begin{cases} 1 & \text{if } t = T+1 \\ \displaystyle b(s)(\ell_t) \sum_{s' \in S} a(s)(s') \cdot \beta(s', t+1) & \text{if } 1 \leq t \leq T \end{cases}
\end{aligned}$$

Thus:

$$\begin{aligned}
\gamma_o(s,t) &= \frac{\alpha_o(s,t)\, \beta_o(s,t)}{\sum_{u \in S} \alpha_o(u,t) \beta_o(u,t)} \\
\xi_o(s,t)(s') &= \frac{\alpha_o(s,t) \cdot a_{s,s'} \cdot b_{s,\ell} \cdot \beta_o(s', t+1)}{\sum_{u \in S} \alpha_o(u,t) \beta_o(u,t)}
\end{aligned}$$

3 Baum-Welch for HMM

On a given finite set $\mathcal{O}$ of traces, the Baum-Welch algorithm can be described as repeating the two following steps until convergence:

1. Compute $Q(\mathcal{M}', \mathcal{M}^{(n)}) = \sum_{\gamma} \sum_{o \in \mathcal{O}} \ln \left[l(o : \gamma; \mathcal{M}') \right] l(\gamma | o; \mathcal{M}^{(n)})$.
2. Set $\mathcal{M}^{(n+1)} = \arg\max_{\mathcal{M}'} Q(\mathcal{M}', \mathcal{M}^{(n)})$.

Let $\mathcal{M}^{(n)} = \langle S, \mathcal{L}, \pi, a, b \rangle$ and $\mathcal{M}' = \langle S, \mathcal{L}, \hat{\pi}, \hat{a}, \hat{b} \rangle$.

First, noting that $l(o : \gamma) = l(o) l(\gamma | o)$, we can write:

$$\begin{aligned}
\arg\max_{\mathcal{M}'} Q(\mathcal{M}', \mathcal{M}^{(n)}) &= \arg\max_{\mathcal{M}'} \sum_{o \in \mathcal{O}} \sum_{\gamma} \ln \left[l(o : \gamma; \mathcal{M}') \right] l(\gamma | o; \mathcal{M}^{(n)}) \\
&= \arg\max_{\mathcal{M}'} \sum_{o \in \mathcal{O}} \sum_{\gamma} \ln \left[l(o : \gamma; \mathcal{M}') \right] l(o : \gamma; \mathcal{M}^{(n)})
\end{aligned}$$

Plugging (1) into $Q(\mathcal{M}', \mathcal{M}^{(n)})$ we get:

$$\begin{aligned}
Q(\mathcal{M}', \mathcal{M}^{(n)}) &= \sum_{o \in \mathcal{O}} \sum_{\gamma} \ln \hat{\pi}_{s_1} \, l(o : \gamma; \mathcal{M}^{(n)}) \\
&+ \sum_{o \in \mathcal{O}} \sum_{\gamma} \sum_{t=1}^{|o|} \ln \hat{a}(s_t)(s_{t+1}) \, l(o : \gamma; \mathcal{M}^{(n)}) \\
&+ \sum_{o \in \mathcal{O}} \sum_{\gamma} \sum_{t=1}^{|o|} \ln \hat{b}(s_t)(\ell_t) \, l(o : \gamma; \mathcal{M}^{(n)})
\end{aligned}$$

Now we optimise with Lagrange multipliers (l_π, l_{a_s} and l_{b_s}). Let $L(\mathcal{M}', \mathcal{M}^{(n)})$ be the Lagrangian:

$$\begin{aligned}
L(\mathcal{M}', \mathcal{M}^{(n)}) &= Q(\mathcal{M}', \mathcal{M}^{(n)}) \\
&- l_\pi \left(\sum_{s \in S} \hat{\pi}_s - 1 \right) \\
&- \sum_{s \in S} l_{a_s} \left(\sum_{u} \hat{a}(s)(u) - 1 \right) \\
&- \sum_{s \in S} l_{b_s} \left(\sum_{\ell} \hat{b}(s)(\ell) - 1 \right)
\end{aligned}$$

3.1 Estimation of π

First, let focus on the π_s's:

$$\begin{aligned}
\frac{\partial \hat{L}(\mathcal{M}', \mathcal{M}^{(n)})}{\partial \hat{\pi}_s} &= \frac{\partial Q(\mathcal{M}', \mathcal{M}^{(n)})}{\partial \hat{\pi}_s} - l_\pi = 0 \\
&= \frac{\partial}{\partial \hat{\pi}_s} \left(\sum_{\gamma} \sum_{o \in \mathcal{O}} \ln \hat{\pi}(s_1) l(o : \gamma; \mathcal{M}^{(n)}) \right) - l_\pi = 0 \\
&= \frac{\partial}{\partial \hat{\pi}_s} \left(\sum_{s'} \sum_{o \in \mathcal{O}} \ln \hat{\pi}(s') l(s_1 = s', o; \mathcal{M}^{(n)}) \right) - l_\pi = 0 \\
&= \sum_{o \in \mathcal{O}} \frac{l(s_1 = s, o; \mathcal{M}^{(n)})}{\hat{\pi}_s} - l_\pi = 0
\end{aligned}$$

Hence:

$$\hat{\pi}_s = \sum_{o \in \mathcal{O}} \frac{l(s_1 = s, o; \mathcal{M}^{(n)})}{l_\pi} \tag{2}$$

Furthermore:

$$\frac{\partial \hat{L}(\mathcal{M}', \mathcal{M}^{(n)})}{\partial l_\pi} = - \left(\sum_{s \in S} \hat{\pi}_s - 1 \right) = 0 \tag{3}$$

By plugging (2) into (3) we get:

$$l_\pi = \sum_{o \in \mathcal{O}} \sum_{s'} l(s_1 = s', o; \mathcal{M}^{(n)}) \tag{4}$$

And by plugging (4) into (2):

$$\begin{aligned}
\hat{\pi}_s &= \frac{\sum_{o \in \mathcal{O}} l(s_1 = s, o; \mathcal{M}^{(n)})}{\sum_{o \in \mathcal{O}} \sum_{s'} l(s_1 = s', o; \mathcal{M}^{(n)})} \\
\hat{\pi}_s &= \frac{\sum_{o \in \mathcal{O}} l(s_1 = s | o; \mathcal{M}^{(n)})}{\sum_{o \in \mathcal{O}} \sum_{s'} l(s_1 = s' | o; \mathcal{M}^{(n)})}
\end{aligned}$$

Finally, using the previously defined coefficients:

$$\boxed{\hat{\pi}_s = \frac{\sum_{o \in \mathcal{O}} \gamma_o(s, 0)}{\sum_{o \in \mathcal{O}} \sum_{s' \in S} \gamma_o(s', 0)}}$$

3.2 Estimation of a

Now, let focus on the $a_{s,s'}$'s:

$$\begin{aligned}
\frac{\partial L(\mathcal{M}', \mathcal{M}^{(n)})}{\partial \hat{a}_{s,s'}} &= \frac{\partial}{\partial \hat{a}_{s,s'}} \left(\sum_{\gamma} \sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} \ln[\hat{a}_{s_t, s'_{t+1}}] l(o : \gamma; \mathcal{M}^{(n)}) \right) - l_{a_s} = 0 \\
&= \frac{\partial}{\partial \hat{a}_{s,s'}} \left(\sum_{o \in \mathcal{O}} \sum_{u,u' \in S} \sum_{t=1}^{|o|} \ln[\hat{a}_{u,u'}] l(s_t = u, s_{t+1} = u', o; \mathcal{M}^{(n)}) \right) - l_{a_s} = 0 \\
&= \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, s_{t+1} = s', o; \mathcal{M}^{(n)})}{\hat{a}_{s,s'}} - l_{a_s} = 0
\end{aligned}$$

Hence:

$$\hat{a}_{s,s'} = \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, s_{t+1} = s', o; \mathcal{M}^{(n)})}{l_{a_s}} \tag{5}$$

Furthermore:

$$\frac{\partial L(\mathcal{M}', \mathcal{M}^{(n)})}{\partial l_{a_s}} = -\left(\sum_{u} \hat{a}_{s,u} - 1 \right) = 0 \tag{6}$$

By plugging (5) into (6) we get:

$$l_{a_s} = \sum_{o \in \mathcal{O}} \sum_{u} \sum_{t=1}^{|o|} l(s_t = s, s_{t+1} = u, o; \mathcal{M}^{(n)}) \tag{7}$$

And by plugging (7) into (5):

$$\hat{a}_{s,s'} = \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, s_{t+1} = s', o; \mathcal{M}^{(n)})}{\sum_{o \in \mathcal{O}} \sum_{u} \sum_{t=1}^{|o|} l(s_t = s, s_{t+1} = u, o; \mathcal{M}^{(n)})}$$

$$\hat{a}_{s,s'} = \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, s_{t+1} = s' | o; \mathcal{M}^{(n)})}{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s | o; \mathcal{M}^{(n)})}$$

Finally, using the previously defined coefficients:

$$\boxed{\hat{a}_{s,s'} = \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} \xi_o(s,t)(s')}{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} \gamma_o(s,t)}}$$

3.3 Estimation of b

Now, let focus on the $b_{s,\ell}$'s:

$$\begin{aligned}
\frac{\partial L(\mathcal{M}', \mathcal{M}^{(n)})}{\partial \hat{b}_{s,\ell}} &= \frac{\partial}{\partial \hat{b}_{s,\ell}} \left(\sum_{\gamma} \sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} \ln[\hat{b}_{s_t,\ell_t}] l(o : \gamma; \mathcal{M}^{(n)}) \right) - l_{b_s} = 0 \\
&= \frac{\partial}{\partial \hat{b}_{s,\ell}} \left(\sum_{o \in \mathcal{O}} \sum_{u \in S} \sum_{t=1}^{|o|} \ln[\hat{b}_{u,\ell_t}] l(s_t = u; \mathcal{M}^{(n)}) \right) - l_{b_s} = 0 \\
&= \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, o; \mathcal{M}^{(n)}) \cdot 1_\ell(\ell_t)}{\hat{b}_{s,\ell}} - l_{b_s} = 0
\end{aligned}$$

Hence:

$$\hat{b}_{s,\ell} = \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, o; \mathcal{M}^{(n)}) \cdot 1_\ell(\ell_t)}{l_{b_s}} \tag{8}$$

Furthermore:

$$\frac{\partial L(\mathcal{M}', \mathcal{M}^{(n)})}{\partial l_{b_s}} = -\left(\sum_{\ell} \hat{b}_{s,\ell} - 1 \right) = 0 \tag{9}$$

By plugging (8) into (9) we get:

$$l_{b_s} = \sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, o; \mathcal{M}^{(n)}) \tag{10}$$

And by plugging (10) into (8):

$$\hat{b}_{s,\ell} = \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} 1_\ell(\ell_t) \cdot l(s_t = s, o; \mathcal{M}^{(n)})}{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, o; \mathcal{M}^{(n)})}$$

$$\hat{b}_{s,\ell} = \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} 1_\ell(\ell_t) \cdot l(s_t = s | o; \mathcal{M}^{(n)})}{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s | o; \mathcal{M}^{(n)})}$$

Finally, using the previously defined coefficients:

$$\boxed{\hat{b}_{s,\ell} = \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} 1_\ell(\ell_t) \cdot \gamma_o(s, t)}{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} \gamma_o(s, t)}}$$

References

[1] L. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains,” 1970.