Baum-Welch for Markov Decision Processes

Raphaël Reynouard

September 26, 2022

1 Introduction

This document describes the Baum-Welch algorithm [1] for Markov Decision Processes.

2 Preliminaries

We define a Markov Decision Process (MDP) formally as follow:

Definition 2.1 (Markov Decision Process) *A MDP is a tuple $\langle S, \mathcal{L}, A, \pi, \{\tau^{(a)}\}_{a\in A}\rangle$ where:*

- *S is a non-empty set of states,*
- *$\mathcal{L}$ is a non-empty set of observations,*
- *A is a non-empty set of actions,*
- *$\pi := \mathcal{D}(S)$ is the initial distribution i.e. the model starts in state s with probability $\pi(s) := \pi_s$,*
- *$\tau^{(a)} : S \mapsto \mathcal{D}(\mathcal{L}\times S)$ is the transition function. The model moves from state s to s' generating ℓ when it receives action a with probability $\tau^{(a)}(s)(\ell, s') := \tau^{(a)}_{s,\ell,s'}$.*

A path is a sequence in $\mathbf{Paths} = (S \times A \times \mathcal{L})^* S$ representing a finite execution of a MDP $\mathcal{M}$, and a trace is a finite sequence in $\mathbf{Traces} = (A\times\mathcal{L})^*$ representing a finite execution of a MDP for which we cannot see the states.
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.
For $i \in \mathbb{N}_{>0}$, we define $X_i\colon \mathbf{Paths} \to S$, $Y_i\colon \mathbf{Paths} \to \mathcal{L}$, $A_i\colon \mathbf{Paths} \to A$, and $O_i\colon \mathbf{Paths} \to \mathbf{Traces}$ respectively as $X_i(\rho) = s_i$, $Y_i(\rho) = \ell_i$, $A_i(\rho) = a_i$, and $O_i(\rho) = (a_1, \ell_1)\cdots(a_i, \ell_i)$, where $\rho = (s_1, a_1, \ell_1)\cdots(s_n, a_n, \ell_n)s_{n+1}$.
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 $o = a_1\ell_1 \dots a_T\ell_T$. A such path is $o : \gamma := s_1 a_1 \ell_1 \dots s_T a_T \ell_T s_{T+1}$.

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:

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

Hence:

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

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 goes 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

$$\begin{aligned} \alpha_o(s,t) &= Pr^{\mathcal{M}}[Y_{1:t-1} = \ell_1 \,.\,.\, \ell_{t-1}, X_t = s | A_{1:t-1} = a_1 \,.\,.\, a_{t-1}] \,, \text{ and} \\ \beta_o(s,t) &= Pr^{\mathcal{M}}[Y_{t:T} = \ell_t \,.\,.\, \ell_T | X_t = s, A_{t:T} = a_t \,.\,.\, a_T] \,. \end{aligned}$$

These can be calculated according to the following recurrences

$$\alpha_o(s,t) = \begin{cases} \pi(s) & \text{if } t = 1 \\ \sum_{s' \in S} \alpha(s', t-1) \cdot \tau^{(a_t-1)}(s')(\ell_{t-1}, s) & \text{if } 1 < t \le T+1 \end{cases}$$

$$\beta_o(s,t) = \begin{cases} 1 & \text{if } t = T+1 \\ \sum_{s' \in S} \tau^{(a_t)}(s)(\ell_t, s') \cdot \beta(s', t+1) & \text{if } 1 \le t \le T \end{cases}$$

Thus:

$$\begin{aligned} \gamma_o(s,t) &= \frac{\alpha_o(s,t)\, \beta_o(s,t)}{\sum_{u \in S} \alpha_o(u,t) \beta_o(u,t)} \\ \xi_o(s,t)(s') &= \frac{\alpha_o(s,t) \cdot \tau^{(a_t)}(s)(\ell_t, s') \cdot \beta_o(s', t+1)}{\sum_{u \in S} \alpha_o(u,t) \beta_o(u,t)} \end{aligned}$$

3 Baum-Welch for MDP

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

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

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:

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

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

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

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 τ

Now, let focus on the $\tau^{(a)}_{s,\ell,s'}$'s:

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

Hence:

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

Furthermore:

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

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

$$\begin{aligned}
l_{\tau^a_s} &= \sum_{o \in \mathcal{O}} \sum_{u,\ell} \sum_{t=1}^{|o|} l(s_t = s, s_{t+1} = u, o; \mathcal{M}^{(n)}) \cdot 1_\ell(\ell_t) \cdot 1_a(a_t) && (7) \\
&= \sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, o; \mathcal{M}^{(n)}) \cdot 1_a(a_t) && (8)
\end{aligned}$$

And by plugging (8) into (5):

$$\begin{aligned}
\hat{\tau}^{(a)}_{s,\ell,s'} &= \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, s_{t+1} = s', o; \mathcal{M}^{(n)}) \cdot 1_\ell(\ell_t) \cdot 1_a(a_t)}{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, o; \mathcal{M}^{(n)}) \cdot 1_a(a_t)} \\
\hat{\tau}^{(a)}_{s,\ell,s'} &= \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, s_{t+1} = s' | o; \mathcal{M}^{(n)}) \cdot 1_\ell(\ell_t) \cdot 1_a(a_t)}{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s | o; \mathcal{M}^{(n)}) \cdot 1_a(a_t)}
\end{aligned}$$

Finally, using the previously defined coefficients:

$$\boxed{\hat{\tau}^{(a)}_{s,\ell,s'} = \frac{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} \xi_o(s,t)(s') \cdot 1_\ell(\ell_t) \cdot 1_a(a_t)}{\sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} \sum_{s \in S} \gamma_o(u,t) \cdot 1_a(a_t)}}$$

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.