A Baum-Welch extension to learn Gaussian observation Hidden Markov Models

Raphaël Reynouard

September 26, 2022

1 Introduction

This document describes an extension of the Baum-Welch algorithm [1] for Gaussian observation Hidden Markov Models.

2 Preliminaries

We define a Gaussian observation Hidden Markov Model (GoHMM) as a hidden markov model that generates an observation $\ell \in \mathbb{R}$ ($\mathbb{R}$ denotes the set of real numbers) each time it is in a state, according to a Gaussian distribution. The parameters of the Gaussian distribution depend on the current state of the model. Formally:

Definition 2.1 (Gaussian observation Hidden Markov Model) *A GoHMM is a tuple $\langle S, \pi, a, \{\theta_s\}_{s\in S}\rangle$ where:*

- *S is a set of states,*
- *$\pi := \mathcal{D}(S)$ 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 $a(s)(s') := a_{s,s'}$,*
- *$\theta_s = \{\mu_s, \sigma_s\}$ are the parameters used by the gaussian distribution to generate the observation while in state s.*

We denote by $b(s)(\ell)$ (or shortly $b_{s,\ell}$) the likelihood that the model generates $\ell \in \mathbb{R}$ while in state s:

$$b(s)(\ell) = \frac{1}{\sigma_s\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\mu_s-\ell}{\sigma_s}\right)^2}$$

A path is a sequence in $\mathbf{Paths} = (S \times \mathbb{R})^* S$ representing a finite execution of a GoHMM $\mathcal{M}$, and a trace is a finite sequence in $\mathbf{Traces} = \mathbb{R}^*$ representing

a finite execution of a GoHMM 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 $X_i \colon \mathbf{Paths} \to S$, $Y_i \colon \mathbf{Paths} \to \mathbb{R}$, 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 $\mathcal{D}(\Omega)$ 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:

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

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

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

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

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

These can be calculated according to the following recurrences

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

$$\beta_o(s,t) = \begin{cases} 1 & \text{if } t = T+1 \\ \sum_{s' \in S} a(s)(s') \cdot b(s)(\ell_t) \cdot \beta(s', t+1) & \text{if } 1 \leq t \leq T \end{cases}$$

Thus:

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

3 Baum-Welch for GoHMM

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 $\mathcal{M}^{(n)} = \langle S, \pi, a, \{\theta_s\}_{s\in S}\rangle$ and $\mathcal{M}' = \langle S, \hat{\pi}, \hat{a}, \{\hat{\theta}_s\}_{s\in S}\rangle$.

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:

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

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

$$L(\mathcal{M}', \mathcal{M}^{(n)}) = Q(\mathcal{M}', \mathcal{M}^{(n)}) - l_\pi\left(\sum_{s\in S}\hat{\pi}_s - 1\right) - \sum_{s\in S} l_{a_s}\left(\sum_{s'}\hat{a}(s)(s') - 1\right)$$

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):

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

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:

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

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

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

$$
= \sum_{o \in \mathcal{O}} \sum_{t=1}^{|o|} l(s_t = s, o; \mathcal{M}^{(n)}) \tag{8}
$$

And by plugging (8) into (5):

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

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

Finally, using the previously defined coefficients:

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

3.3 Estimation of b

3.3.1 Estimation of μ

First, notice that:

$$\ln b(s)(\ell) = \ln\left(\frac{1}{\sigma_s\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\ell-\mu_s)^2}{\sigma_s^2}}\right) \tag{9}$$

$$= -\ln(\sigma_s\sqrt{2\pi}) - \frac{1}{2}\frac{(l-\mu_s)^2}{\sigma_s^2} \tag{10}$$

$$= -\ln(\sigma_s\sqrt{2\pi}) - \frac{\ell^2}{2\sigma_s^2} + \frac{\ell\mu_s}{\sigma_s^2} - \frac{\mu_s^2}{2\sigma_s^2} \tag{11}$$

$$\frac{\partial \ln b(s)(\ell)}{\partial \mu_s} = \frac{\ell}{\sigma_s^2} - \frac{\mu_s}{\sigma_s^2}$$

As usual:

$$\begin{aligned}\frac{\partial L(\mathcal{M}', \mathcal{M}^{(n)})}{\partial \hat{\mu}_s} &= \frac{\partial}{\partial \hat{\mu}_s}\left(\sum_{\gamma}\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} \ln \hat{b}_{s_t,\ell_t} l(o:\gamma;\mathcal{M}^{(n)})\right) = 0\\ &= \frac{\partial}{\partial \hat{\mu}_s}\left(\sum_{o\in\mathcal{O}}\sum_{u\in S}\sum_{t=1}^{|o|} \ln \hat{b}_{s,\ell_t}\, l(s_t=u,o;\mathcal{M}^{(n)})\right) = 0\\ &= \sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} \frac{\ell_t\, l(s_t=s,o;\mathcal{M}^{(n)})}{\hat{\sigma}_s^2} - \frac{\hat{\mu}_s l(s_t=s,o;\mathcal{M}^{(n)})}{\hat{\sigma}_s^2} = 0\end{aligned}$$

Hence we have:

$$\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} \frac{\hat{\mu}_s l(s_t=s,o;\mathcal{M}^{(n)})}{\hat{\sigma}_s^2} = \sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} \frac{\ell_t l(s_t=s,o;\mathcal{M}^{(n)})}{\hat{\sigma}_s^2}$$

So:

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

$$\hat{\mu}_s = \frac{\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} \ell_t l(s_t=s|o;\mathcal{M}^{(n)})}{\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} l(s_t=s|o;\mathcal{M}^{(n)})}$$

Finally, using the previously defined coefficients:

$$\boxed{\hat{\mu}_s = \frac{\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} \ell_t \gamma_o(s,t)}{\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|}\sum_{s'\in S} \gamma_o(s',t)}}$$

3.3.2 Estimation of σ_s

From (11) we have:

$$\frac{\partial \ln b(s)(\ell)}{\partial \sigma_s} = -\frac{1}{\sigma_s} + \frac{\ell^2}{\sigma_s^3} - \frac{2\ell\mu_s}{\sigma_s^3} + \frac{\mu_s^2}{\sigma_s^3} \tag{12}$$

$$= \frac{1}{\sigma_s^3}(-\sigma_s^2 + \ell^2 - 2\ell\mu_s + \mu_s^2) \tag{13}$$

$$= \frac{1}{\sigma_s^3}\left((\ell - \mu_s)^2 - \sigma_s^2\right) \tag{14}$$

As usual:

$$\frac{\partial L(\mathcal{M}', \mathcal{M}^{(n)})}{\partial \hat{\sigma}_s} = \frac{\partial}{\partial \hat{\sigma}_s}\left(\sum_{\gamma}\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} \ln \hat{b}_{s_t,\ell_t} l(o:\gamma;\mathcal{M}^{(n)})\right) = 0$$

$$= \frac{\partial}{\partial \hat{\sigma}_s}\left(\sum_{o\in\mathcal{O}}\sum_{s}\sum_{t=1}^{|o|} \ln \hat{b}_{s,\ell_t} l(s_t = s, o;\mathcal{M}^{(n)})\right) = 0$$

And, from (14):

$$\frac{\partial L(\mathcal{M}', \mathcal{M}^{(n)})}{\partial \hat{\sigma}_s} = \frac{1}{\hat{\sigma}_s^3}\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} l(s_t = s, o;\mathcal{M}^{(n)})\left((\ell_t - \hat{\mu}_s)^2 - \hat{\sigma}_s^2\right)$$

Hence we have:

$$\frac{1}{\hat{\sigma}_s^3}\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} l(s_t = s, o;\mathcal{M}^{(n)})(\ell_t - \hat{\mu}_s)^2 = \frac{1}{\hat{\sigma}_s}\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} l(s_t = s, o;\mathcal{M}^{(n)})$$

Then, by isolating $\hat{\sigma}_s$ on the left side:

$$\hat{\sigma}_s^2 = \frac{\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} l(s_t = s, o;\mathcal{M}^{(n)})(\ell_t - \hat{\mu}_s)^2}{\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} l(s_t = s, o;\mathcal{M}^{(n)})}$$

$$\hat{\sigma}_s^2 = \frac{\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} l(s_t = s|o;\mathcal{M}^{(n)})(\ell_t - \hat{\mu}_s)^2}{\sum_{o\in\mathcal{O}}\sum_{t=1}^{|o|} l(s_t = s|o;\mathcal{M}^{(n)})}$$

Finally, using the previously defined coefficients:

$$\hat{\sigma}_s = \frac{\sum_{o\in\mathcal{O}} \sum_{t=1}^{|o|} (\ell_t - \hat{\mu}_s)^2 \gamma_o(s,t)}{\sum_{o\in\mathcal{O}} \sum_{t=1}^{|o|} \sum_{s'\in S} \gamma_o(s',t)}$$

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.