Hidden Markov Models - A tutorial

Mylene Haslehner

November 2020

Contents

1 Introduction

The goal of this tutorial is to reformulate and to complete the demonstrations of the three problems of Hidden Markov Models (HMM) in [Rabiner(1989)] by missing steps of a mathematically fully comprehensive demonstration, since to us, important small calculation steps were not mentioned and may hinder many novice readers - like us - of fully understanding the topic. Our intention was to better understand the mathematical basis of this important field of Hidden Markov Models.

Hidden Markov Models describe a stochastic process of an unknown state $(x_1, ..., x_T)$, that is regularly observed. The observations $(z_1, ..., z_T)$ are a probabilistic function of the states, each z_t being a function of x_t for $t \in [1, ..., T]$.

HMM have a vast field of applications originating in signal theory, that goes from speech recognition to gene sequencing, over weather or satellite trajectory forecasting.

We consider that $x_t \in [1, ..., N]$ if the set of states is finite of size $N \in \mathbf{N}^*$, and that $z_t \in [1, ..., M]$ for $M \in \mathbf{N}^*$. The general framing of a HMM is based on initial knowledge of its model parameters, $\lambda := (\pi_i, a_{ij}, b(z_t))$,

$$\begin{aligned} \pi_i &:= p[x_1 = i] \\ a_{ij} &:= p[x_{t+1} = j | x_t = i] \\ b_i(k) &:= p[z_t = k | x_t = i], \end{aligned} \tag{1}$$

for $i, j \in [1, N]$, $k \in [1, M]$, where π is the prior (the probability distribution of the first state at time $t = 1$), a_{ij} is the transition probability and $b(k)$ is the 'emission probability'. Since we consider a Markov process, a_{ij} is independent of time.

Three characteristic problems have been formulated in the context of HMM by [Rabiner(1989)], to which we add a fourth problem, that describes prediction in the future. Let us denote by $p(x) \underset{x}{\rightarrow} \max$ the value of x for which $p(x)$ reaches a maximum.

- Problem 1: Calculate $p[z_1, ..., z_T|\lambda]$.
- Problem 2: $p[\mathbf{x}, \mathbf{z}] \underset{\mathbf{x}}{\rightarrow} \max$, given λ.
- Problem 3: $p[\lambda] := p[\mathbf{z}|\lambda] \underset{\lambda}{\rightarrow} \max$.
- Problem 4: predict $p(x_{T+1}|(z_1, ..., z_T), (x_1, ..., x_T))$,
  the estimate of the state vector at time $T + 1$.

According to [Rabiner(1989)], Problem 1 is an 'evaluation problem' that enables to choose the best model among different competing models that best matches the observations: Calculate the probability that the observation sequence was produced by that model. Problem 2 finds the 'optimal' state sequence given model parameters and the observation sequence, eg in speech recognition. Problem 3 allows to train the model parameters in order to optimally adapt the model to observed training data. This problem allows to create the best model for real phenomena. Problem 4 (not derived here) is a prediction problem. It calculates the estimated, most likely, forecast state given the model, and current and past states and observations. Problem 4 is usually solved eg by using Kalman filter and particle filters.

2 Problem 1 (Forward Algorithm)

The problem consists of calculating the probability of a sequence of observations $z_1, ..., z_T$, given the model parameters λ as defined above in (1),

$$p[z_1, ..., z_T|\lambda]. \tag{2}$$

The demonstration starts as follows. We have

$$\begin{aligned}
p[z_1, ..., z_t, x_t] &= p[z_1, ...z_{t-1}, z_t|x_t]p[x_t] \\
&= p[z_1, ..., z_{t-1}|x_t]p[z_t|x_t]p[x_t] \\
&= p[z_t|x_t]p[z_1, ..., z_{t-1}, x_t] \\
&= p[z_t|x_t] \sum_{x_{t-1}} p[z_1, ..., z_{t-1}, x_{t-1}, x_t] \\
&= p[z_t|x_t] \sum_{x_{t-1}} p[z_1, ..., z_{t-1}, x_t|x_{t-1}]p[x_{t-1}] \\
&= p[z_t|x_t] \sum_{x_{t-1}} p[z_1, ..., z_{t-1}|x_{t-1}]p[x_t|x_{t-1}]p[x_{t-1}] \\
&= p[z_t|x_t] \sum_{x_{t-1}} p[z_1, ..., z_{t-1}, x_{t-1}]p[x_t|x_{t-1}],
\end{aligned} \tag{3}$$

where we have conditioned on x_t using Bayes theorem and applied the conditional independence of the observations $(z_1, ..., z_t)$ in the first two steps, then grouped $z_1, ..., z_{t-1}$ and x_t into a joint probability and summed over the variable x_{t-1} in the next two steps. Finally, in the last three steps, we have again conditioned on x_{t-1} using Bayes, applied the conditional independence of $z_1, ..., z_{t-1}$ and x_t, and grouped $z_1, ..., z_{t-1}$ and x_{t-1} into a joint probability.

Let $\alpha_t := p[z_1, ..., z_t, x_t]$. For each time step t, α_t is a function of x_t. We have

$$\alpha_t = p[z_t|x_t] \sum_{x_{t-1}} \alpha_{t-1} p[x_t|x_{t-1}]. \tag{4}$$

In this way, we can calculate recursively $p[z_1, ..., z_t, x_t]$ using previous time steps $t - 1$. This is called forward algorithm.

Finally, in order to calculate $p[z_1, ..., z_t]$, it is sufficient to sum $p[z_1, ..., z_t, x_t]$ over x_t:

$$p[z_1, ..., z_t] = \sum_{x_t} \alpha_t, \tag{5}$$

where

$$\alpha_t = p[z_t|x_t] \sum_{x_{t-1}} \alpha_{t-1} p[x_t|x_{t-1}] \tag{6}$$

and

$$\alpha_1 = p[z_1, x_1]. \tag{7}$$

The probability of the observation sequence can now be calculated recursively.

3 Problem 2 (Viterbi)

This problem consists of maximizing the likelihood of a sequence of states and of observations $\mathbf{z}$ with respect to the states

$$p[\mathbf{x}, \mathbf{z}] \underset{\mathbf{x}}{\rightarrow} \max. \tag{8}$$

This problem is equivalent to minimizing the log of the likelihood

$$\begin{aligned}
& -\log p[\mathbf{x}, \mathbf{z}] \underset{\mathbf{x}}{\rightarrow} \min & (9)\\
\Leftrightarrow & -\log p[\mathbf{z}|\mathbf{x}]p[\mathbf{x}] \underset{\mathbf{x}}{\rightarrow} \min & (10)\\
\Leftrightarrow & -\log p[\mathbf{z}|\mathbf{x}] - \log p[\mathbf{x}] \underset{\mathbf{x}}{\rightarrow} \min & (11)\\
\Leftrightarrow & -\log \prod_t p[\mathbf{z}_t|\mathbf{x}_t] - \log \prod_t p[\mathbf{x}_{t+1}|\mathbf{x}_t] \underset{\mathbf{x}_t}{\longrightarrow} \min & (12)\\
\Leftrightarrow & \sum_t \left(-\log p[\mathbf{z}_t|\mathbf{x}_t] - \log p[\mathbf{x}_{t+1}|\mathbf{x}_t]\right) \underset{\mathbf{x}_t}{\longrightarrow} \min & (13)\\
\Leftrightarrow & \sum_t \gamma_t \underset{\mathbf{x}_t}{\longrightarrow} \min & (14)\\
\Leftrightarrow & \lambda_T \underset{\mathbf{x}_t}{\longrightarrow} \min, & (15)
\end{aligned}$$

where

$$\gamma_t := -\log p[\mathbf{z}_t|\mathbf{x}_t] - \log p[\mathbf{x}_{t+1}|\mathbf{x}_t], \tag{16}$$

$$\lambda_T := \sum_{t=0}^{T-1} \gamma_t. \tag{17}$$

λ_T can be interpreted as a path. The task is to find the shortest path by iteration through a graph. The minimization is done for each time step [Forney(1973)].

Let $\tilde{x}_T := (\tilde{x}_0, ..., \tilde{x}_T)$ be the shortest path segment until time T.
Define $\Gamma(x_T) := \lambda_T(\tilde{x}_T)$ as the length λ of the shortest path segment $\tilde{x}_T$

- Initialize: For t = 0, $\Gamma(\mathbf{x}_0) = 0$ and $\tilde{x}_0 = x_0$.

- For each t > 1, find $\Gamma(\mathbf{x}_{t+1}) := \min_{\mathbf{x}_t}(\Gamma(\mathbf{x}_t) + \gamma_t))$.
  Store $\Gamma(\mathbf{x}_{t+1})$ and $\tilde{x}_{t+1}$

Replace $t + 1$ by t and repeat the procedure until $t = T - 1$. Note that the path segments are M-dimensional vectors, so this procedure needs to be done separately for each of the M coordinates of the path-vector.

In particular, we have

$$
\begin{aligned}
\Gamma(\mathbf{x}_1) &= \min_{\mathbf{x}_0}(\Gamma(\mathbf{x}_0) + \gamma_1)) \\
&= \min_{\mathbf{x}_0}(-\log p[\mathbf{z}_0|\mathbf{x}_0] - \log p[\mathbf{x}_1|\mathbf{x}_0]) \\
\Gamma(\mathbf{x}_2) &= \min_{\mathbf{x}_1}(\Gamma(\mathbf{x}_1) + \gamma_2)) \\
&= \min_{\mathbf{x}_1}(\Gamma(\mathbf{x}_1) - \log p[\mathbf{z}_1|\mathbf{x}_1] - \log p[\mathbf{x}_2|\mathbf{x}_1]) \\
&\dots \\
\Gamma(\mathbf{x}_T) &= \min_{\mathbf{x}_{T-1}}(\Gamma(\mathbf{x}_{T-1}) + \gamma_T)) \\
&= \min_{\mathbf{x}_{T-1}}(\Gamma(\mathbf{x}_{T-1}) - \log p[\mathbf{z}_{T-1}|\mathbf{x}_{T-1}] - \log p[\mathbf{x}_T|\mathbf{x}_{T-1}]).
\end{aligned}
\tag{18}
$$

4 Problem 3 (Baum-Welch)

The third problem consists of calculating the model parameters that maximize the probability of a sequence of observations:

$$p[\lambda] := p[z|\lambda] \underset{\lambda}{\rightarrow} \max. \tag{19}$$

Since it is mathematically unfeasible to find the maximum of this likelihood, we try to look for a $\bar{\lambda}$ that just does the work of increasing it, i.e. find a $\bar{\lambda}$ such that

$$p[\bar{\lambda}] \geq p[\lambda], \tag{20}$$

or, equivalently,

$$
\begin{aligned}
&\frac{p[\bar{\lambda}]}{p[\lambda]} \geq 1 \\
\Leftrightarrow &\log\left(\frac{p[\bar{\lambda}]}{p[\lambda]}\right) \geq 0,
\end{aligned}
\tag{21}
$$

Using the measure-theoretical definition of probabilities,

$$p[\lambda] = \int_z p(z, \lambda) d\mu(z) \tag{22}$$

(where μ is a non-negative measure and $\mu(z) = 1$) and applying Hoelder inequality to the concave function $\log z$, [Baum(1972)] showed that

$$
\begin{aligned}
&\log\left(\frac{p[\bar\lambda]}{p[\lambda]}\right) \geq 0 \\
\Leftrightarrow &\log \frac{\int_z p[z,\bar\lambda] d\mu(z)}{P[\lambda]} \\
= &\log \int_z p[z,\bar\lambda] \frac{d\mu(z)}{P[\lambda]} \\
= &\log \int_z \frac{p[z,\bar\lambda]}{p[z,\lambda]} \left[\frac{p[z,\lambda] d\mu(z)}{P[\lambda]}\right] \geq \int_z \log \frac{p[z,\bar\lambda]}{p[z,\lambda]} \left[\frac{p[z,\lambda] d\mu(z)}{p[\lambda]}\right] \\
&= \frac{1}{p[\lambda]} \int_z \log\left(\frac{p[z,\bar\lambda]}{p[z,\lambda]}\right) p[z,\lambda] d\mu(z) \\
&= \frac{1}{P[\lambda]} \left(Q[\lambda,\bar\lambda] - Q[\lambda,\lambda]\right) \geq 0,
\end{aligned}
\tag{23}
$$

where the function

$$
Q[\lambda,\bar\lambda] := \int_z p[z,\lambda] \log p[z,\bar\lambda] d\mu(z) \tag{24}
$$

is Baum's auxiliary function Q ([Baum(1972)]). Note that, if the set of states is discrete, the integrands become sums, $\int_z f(z) d\mu(z) = \sum_z f(z)$.

This means that increasing the likelihood can be achieved by increasing the auxiliary function for a well-chosen $\bar\lambda$:

$$
\text{If} \quad Q(\lambda,\bar\lambda) \geq Q(\lambda,\lambda) \Rightarrow P[\bar\lambda] \geq P[\lambda]. \tag{25}
$$

Such a $\bar\lambda$ can be found by maximizing Q with respect to $\bar\lambda$.

Recalling expression (1), the model parameters are $\bar\lambda = (\pi_i, a_{ij}, b_j(k))$, where

$$
\begin{aligned}
\pi_i &:= p[x_0 = i] \\
a_{ij} &:= p[x_{t+1} = j | x_t = i] \\
b_j(k) &:= p[z_{t+1} | x_{t+1} = j, x_t = i].
\end{aligned}
\tag{26}
$$

Let us rewrite Q in discrete form:

$$
\begin{aligned}
Q[\lambda,\bar\lambda] &= \sum_x p[x,\lambda] \log p[x,\bar\lambda] \\
&= \sum_{x=(x_0,x_1,\dots,x_T)} p[x,\lambda] \log \prod_t p[x_t,\bar\lambda_t] \\
&= \sum_{x_0=1}^{N} \dots \sum_{x_T=1}^{N} p[x,\lambda] \left(\log p[x_0,\bar\lambda] + \sum_{t>0} \log p[x_{t+1}|x_t,\bar\lambda] + \sum_{t>0} \log p[z_{t+1}|x_{t+1},x_t,\bar\lambda]\right) \\
&= \sum_{x_0=1}^{N} \dots \sum_{x_T=1}^{N} p[x,\lambda] \left(\log \bar\pi_{x_0} + \sum_{t>0} \log \bar a_{ij}(t) + \sum_{t>0} \log \bar b_j(t)\right)
\end{aligned}
\tag{27}
$$

With the constraints

$$\sum_{x_0=1}^{N} \bar{\pi}_{x_0} = 1, \qquad \sum_{j=x_{t+1}=1}^{N} \bar{a}_{ij} = 1, \quad \sum_{z_t=1}^{N} \bar{b}_j(t) = 1, \tag{28}$$

we maximize the function

$$L(\bar{\pi}_{x_0}, \bar{a}_{ij}, \bar{b}_j(t), \mu_1, \boldsymbol{\mu_2}, \boldsymbol{\mu_3}) := Q(\lambda, \bar{\pi}_{x_0}, \bar{a}_{ij}, \bar{b}_j(t))$$
$$-\mu_1 \left(\sum_{x_0=1}^{N} \bar{\pi}_{x_0} - 1 \right) - \sum_{i=x_t=1}^{N} \mu_{2i} \left(\sum_{j=x_{t+1}=1}^{N} \bar{a}_{ij} - 1 \right) - \sum_{j=x_{t+1}=1}^{N} \mu_{3j} \left(\sum_{z_t=1}^{N} \bar{b}_j(t) - 1 \right)$$

using Lagrange multipliers. In order to find $\bar{\pi}_{x_0=i}$, it is sufficient to calculate the partial derivative, with respect to $\bar{\pi}_0$ and to μ_1, of

$$\begin{aligned} L(\bar{\pi}_{x_0}, \mu_1) &= \sum_{x=(x_0,\ldots,x_T)} p[x, \lambda] \log \bar{\pi}_{x_0} - \mu_1 \left(\sum_{x_0=1}^{N} \bar{\pi}_{x_0} - 1 \right) \\ &= \sum_{x_0} \cdots \sum_{x_T} p[x_0, x_1, ..., x_T, \lambda] \log \bar{\pi}_{x_0} - \mu_1 \left(\sum_{x_0=1}^{N} \bar{\pi}_{x_0} - 1 \right). \end{aligned} \tag{29}$$

Let $x_0 = i$. We have

$$\frac{\partial}{\partial \bar{\pi}_{x_0=i}} L(\bar{\pi}_{x_0}, \mu_1) = 0 \Leftrightarrow \sum_{x_1} \cdots \sum_{x_T} p[x_0 = i, x_1, ..., x_T, \lambda] \frac{1}{\bar{\pi}_i} = \mu_1, \tag{30}$$

$$\frac{\partial}{\partial \mu_1} L(\bar{\pi}_{x_0}, \mu_1) = 0 \Leftrightarrow \sum_{i=1}^{N} \bar{\pi}_i = 1. \tag{31}$$

From equation (30), we find

$$\bar{\pi}_i = \frac{\sum_{x_1} \cdots \sum_{x_T} p[x_0 = i, x_1, ..., x_T, \lambda]}{\mu_1}, \tag{32}$$

which we substitute into condition (31). We get

$$\sum_{i=1}^{N} \bar{\pi}_i = \sum_{x_0=i=1}^{N} \frac{\sum_{x_1} \cdots \sum_{x_T} p[x_0 = i, x_1, ..., x_T, \lambda]}{\mu_1} = 1 \tag{33}$$

$$\Leftrightarrow \mu_1 = \sum_{x_0=i=1}^{N} \sum_{x_1} \cdots \sum_{x_T} p[x_0 = i, x_1, ..., x_T, \lambda] \tag{34}$$

$$\Leftrightarrow \mu_1 = \sum_{x} p[x, \lambda]. \tag{35}$$

Hence, equation (47) becomes

$$\bar{\pi}_i = \frac{\sum_{x_1} \cdots \sum_{x_T} p[x_0 = i, x_1, ..., x_T, \lambda]}{\sum_{x=(x_0,\ldots,x_T)} p[x, \lambda]}. \tag{36}$$

An analogous calculation for $L(\bar{a}_{ij}, \boldsymbol{\mu_2})$ and for $L(\bar{b}_j(t), \boldsymbol{\mu_3})$ leads to the parameters $\bar{a}_{ij}$ and $\bar{b}_j(t)$.

Let $x_t = i$ and $x_{t+1} = j$. We have

$$\frac{\partial}{\partial \bar{a}_{i,j}} L(\bar{a}_{i,j}, \mu_{\mathbf{2}}) = 0 \Leftrightarrow \tag{37}$$

$$\frac{\partial}{\partial \bar{a}_{i,j}} \left(\sum_{x=(x_1 \ldots, x_T)} p[x, \lambda] \sum_{t=1}^{T} \log \bar{a}_{x_t, x_{t+1}} - \sum_{i=x_t=1}^{N} \boldsymbol{\mu}_{2i} \left(\sum_{j=x_{t+1}=1}^{N} \bar{a}_{ij} - 1 \right) \right)$$

$$= \frac{\partial}{\partial \bar{a}_{x_t=i, x_{t+1}=j}} \left(\sum_{x=(x_1 \ldots, x_T)} p[x, \lambda] \left(\log \bar{a}_{x_1, x_2} + \log \bar{a}_{x_2, x_3} + \ldots + \log \bar{a}_{x_{T-1}, x_T} \right) \right. \tag{38}$$

$$\left. - \sum_{i=x_t=1}^{N} \boldsymbol{\mu}_{2i} \left(\sum_{j=x_{t+1}=1}^{N} \bar{a}_{ij} - 1 \right) \right)$$

$$= \sum_{t=1}^{T} p[x_1, \ldots, x_t = i, x_{t+1} = j, ..x_T, \lambda] \frac{1}{\bar{a}_{x_t=i, x_{t+1}=j}} - \mu_{2i} = 0 \tag{39}$$

$$\Leftrightarrow \quad \bar{a}_{ij} = \sum_{t=1}^{T} p[x_1, \ldots, x_t = i, x_{t+1} = j, ..x_T, \lambda] \frac{1}{\mu_{2i}} \tag{40}$$

With the condition (28) on a_{ij}, we have

$$\sum_{j=x_{t+1}=1}^{N} \sum_{t=1}^{T} p[x_1, \ldots, x_t = i, x_{t+1} = j, ..x_T, \lambda] \frac{1}{\mu_{2i}} = 1 \tag{41}$$

$$\Leftrightarrow \quad \sum_{j=1}^{N} \sum_{t=1}^{T} p[x_1, \ldots, x_t = i, x_{t+1} = j, ..x_T, \lambda] = \mu_{2i}. \tag{42}$$

Substituting μ_{2i} into (40) leads to

$$\bar{a}_{ij} = \sum_{t=1}^{T} p[x_1, \ldots, x_t = i, x_{t+1} = j, ..x_T, \lambda] \frac{1}{\mu_{2i}} \tag{43}$$

$$= \frac{\sum_{t=1}^{T} p[x_1, \ldots, x_t = i, x_{t+1} = j, ..x_T, \lambda]}{\sum_{t=1}^{T} \sum_{j=1}^{N} p[x_1, \ldots, x_t = i, x_{t+1} = j, ..x_T, \lambda]}. \tag{44}$$

Hence, we have

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} p[x_t = i, x_{t+1} = j, \lambda]}{\sum_{t=1}^{T} \sum_{j=1}^{N} p[x_t = i, x_{t+1} = j, \lambda]}. \tag{45}$$

Finally, we derive $\bar{b}_j(z_t)$ in a similar way. Let $z_{t+1} := k$ and $x_{t+1} := j$ (it wouldn't make sense to define an observation earlier). We have

$$\frac{\partial}{\partial \bar{b}_j(k)} L(\bar{b}_j, \boldsymbol{\mu_3}) = 0 \Leftrightarrow \tag{46}$$

$$\frac{\partial}{\partial \bar{b}_j(k)} \left(\sum_{t>0} \sum_{x=(x_1 \ldots, x_T)} \delta_{k,z_{t+1}} p[x,\lambda] \log \bar{b}_j(k) - \sum_{j=x_{t+1}=1}^{N} \boldsymbol{\mu}_{3j} \left(\sum_{k=1}^{M} \bar{b}_j(k) - 1 \right) \right)$$

$$= \sum_{t>0} \delta_{k,z_{t+1}} p[x,\lambda] \frac{1}{\bar{b}_j(k)} = \mu_{3j} \tag{47}$$

$$\Leftrightarrow \quad \bar{b}_j(k) = \sum_{t>0} \delta_{k,z_{t+1}} p[x,\lambda] \frac{1}{\mu_{3j}}. \tag{48}$$

With the condition (28), we have

$$\sum_{z_{t+1}=k=1}^{M} \bar{b}_j(k) = 1 \tag{49}$$

$$\Leftrightarrow \sum_{k=1}^{M} \sum_{t>0} \delta_{k,z_{t+1}} p[x,\lambda] \frac{1}{\mu_{3j}} = 1 \tag{50}$$

$$\Leftrightarrow \quad \mu_{3j} = \sum_{t>0} p[x,\lambda] \tag{51}$$

in virtue of

$$\sum_{k=1}^{M} \delta_{k,z_{t+1}} = 1. \tag{52}$$

Finally,

$$\bar{b}_j(k) = \sum_{t>0} \delta_{k,z_{t+1}} p[x,\lambda] \frac{1}{\mu_{3j}} = \sum_{t>0} \delta_{k,z_{t+1}} p[x,\lambda] \frac{1}{\sum_{t>0} p[x,\lambda]} \tag{53}$$

$$\Leftrightarrow \quad \bar{b}_j(k) = \frac{\sum_{t>0|z_{t+1}=k} p[x_{t+1}=j,\lambda]}{\sum_{t>0} p[x_{t+1}=j,\lambda]}. \tag{54}$$

To summarize, the parameter $\bar{\lambda} := (\bar{\pi}, \bar{a}_{ij}, \bar{b}_j(z_{t+1=k}))$ that maximizes the Q function is given by

$$\bar{\pi}_i = \frac{\sum_{x_1} \ldots \sum_{x_T} p[x_0 = i, x_1, \ldots, x_T, \lambda]}{\sum_{x=(x_0,\ldots,x_T)} p[x,\lambda]} \tag{55}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} p[x_t = i, x_{t+1} = j, \lambda]}{\sum_{t=1}^{T} \sum_{j=1}^{N} p[x_t = i, x_{t+1} = j, \lambda]} \tag{56}$$

$$\bar{b}_j(k) = \frac{\sum_{t>0|z_{t+1}=k} p[x_{t+1}=j,\lambda]}{\sum_{t>0} p[x_{t+1}=j,\lambda]}. \tag{57}$$

For the purpose of the derivation of the time-recursive formula for these parameters, let us reformulate the above expressions using Bayes theorem, and by conditioning on $\mathbf{z}$ and λ into:

$$
\begin{aligned}
\bar{\pi}_i &= \frac{\sum_{x=(x_1,...,x_T)} p[x_0 = i, x_1, ..., x_T|\mathbf{z}, \lambda]}{\sum_{x=(x_0,...,x_T)} p[x|\mathbf{z}, \lambda]} \\
&= \sum_{x=(x_1,...,x_T)} p[x_0 = i, x_1, ..., x_T|\mathbf{z}, \lambda] \\
&= \sum_{x_1=j}^{N} p[x_0 = i, x_1 = j|\mathbf{z}, \lambda] \\
\bar{a}_{ij} &= \frac{\sum_{t=1}^{T} p[x_t = i, x_{t+1} = j|\mathbf{z}, \lambda]}{\sum_{t=1}^{T} \sum_{j=1}^{N} p[x_t = i, x_{t+1} = j|\mathbf{z}, \lambda]} \\
\bar{b}_j(k) &= \frac{\sum_{t>0|z_{t+1}=k} p[x_{t+1} = j|\mathbf{z}, \lambda]}{\sum_{t>0} p[x_{t+1} = j|\mathbf{z}, \lambda]} \\
&= \frac{\sum_{t>0|z_{t+1}=k} \sum_i p[x_t = i, x_{t+1} = j|\mathbf{z}, \lambda]}{\sum_{t>0} \sum_i p[x_t = i, x_{t+1} = j|\mathbf{z}, \lambda]}. \qquad (58)
\end{aligned}
$$

Note that for $\bar{\pi}_i$, we were able to use the fact that, after conditioning on z and λ, the denominator sums up to one, since in general $\sum_x p(x, y) = p(y)$ for any x and y.

Let us show that the model parameters $\bar{\pi}_i$, $\bar{a}_{ij}$, $\bar{b}_j$ can be calculated inductively. Note that all parameters depend on the quantity $p[x_t = i, x_{t+1} = j|\mathbf{z}, \lambda]$, which will be reformulated below. For the sake of ease, let us for now omit to write the indices i and j, as well as λ, and make transformations on $p[x_t, x_{t+1}|\mathbf{z}]$.

Let us split the observation sequence into three components, $(z_1, ..., z_t)$, z_{t+1}, $(z_{t+2}, ..., z_T)$. $\mathbf{z_1} := (z_1, ..., z_t)$, $\mathbf{z_2} := (z_{t+2}, ..., z_T)$.

We have

$$
\begin{aligned}
p[x_t, x_{t+1}|\mathbf{z}] &= \frac{p[x_t, x_{t+1}, \mathbf{z}]}{p[\mathbf{z}]} \\
&= \frac{p[\mathbf{z}|x_t, x_{t+1}]p[x_t, x_{t+1}]}{p[\mathbf{z}]} \\
&= \frac{p[\mathbf{z_1}, z_{t+1}, \mathbf{z_2}|x_t, x_{t+1}]p[x_t, x_{t+1}]}{p[\mathbf{z}]} \\
&= \frac{p[\mathbf{z_1}|x_t, x_{t+1}]p[z_{t+1}|x_t, x_{t+1}]p[\mathbf{z_2}|x_t, x_{t+1}]p[x_t, x_{t+1}]}{p[\mathbf{z}]} \qquad (59) \\
&= \frac{p[\mathbf{z_1}|x_t]p[z_{t+1}|x_{t+1}]p[\mathbf{z_2}|x_{t+1}]P[x_t, x_{t+1}]}{p[\mathbf{z}]} \\
&= \frac{p[\mathbf{z_1}|x_t]p[x_t]p[z_{t+1}|x_{t+1}]p[\mathbf{z_2}|x_{t+1}]p[x_{t+1}|x_t]}{p[\mathbf{z}]} \\
&= \frac{p[\mathbf{z_1}, x_t]p[z_{t+1}|x_{t+1}]p[\mathbf{z_2}|x_{t+1}]p[x_{t+1}|x_t]}{p[\mathbf{z}]}
\end{aligned}
$$

Reintroducing the indices in the notation for a_{ij} and $b_j(k)$ as in (1), and defining α_t and β_{t+1} as follows,

$$a_{ij} := p[x_{t+1} = j|x_t = i] \qquad (60)$$

$$b_j(k) := p[z_{t+1} = k|x_{t+1} = j] \qquad (61)$$

$$\alpha_t := p[\mathbf{z_1}, x_t] \qquad (62)$$

$$\beta_{t+1}(j) := p[\mathbf{z_2}|x_{t+1} = j] \qquad (63)$$

Hence,

$$
\begin{aligned}
p[x_t = i, x_{t+1} = j|\mathbf{z}] &= \frac{p[x_t, x_{t+1}, \mathbf{z}]}{p[\mathbf{z}]} \\
&= \frac{p[x_t = i, x_{t+1} = j, \mathbf{z}]}{\sum_i \sum_j p[x_t = i, x_{t+1} = j|\mathbf{z}]} \qquad (64) \\
&= \frac{\alpha_t a_{ij} b_j(k) \beta_{t+1}(j)}{\sum_i \sum_j \alpha_t a_{ij} b_j(k) \beta_{t+1}(j)}.
\end{aligned}
$$

Finally, if we set

$$\xi_t(i, j) := \frac{\alpha_t a_{ij} b_j(k) \beta_{t+1}(j)}{\sum_i \sum_j \alpha_t a_{ij} b_j(k) \beta_{t+1}(j)}, \qquad (65)$$

the model parameters are

$$\begin{aligned}
\bar{\pi}_i &= \sum_{j=1}^{N} p[x_0 = i, x_1 = j | \mathbf{z}, \lambda] = \sum_j \xi_0, (i, j), \\
\bar{a}_{ij} &= \frac{\sum_{t=1}^{T} \xi_t(i, j)}{\sum_{t=1}^{T} \sum_j \xi_t(i, j)}, \\
\bar{b}_j(k) &= \frac{\sum_{t>0|z_{t+1}=k} \sum_i \xi_t(i, j)}{\sum_{t>0} \sum_i \xi_t(i, j)}.
\end{aligned} \tag{66}$$

References

[Baum(1972)] Baum L. E., An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes (Inequalities, 1972)

[Bishop(2006)] Bishop C. M., Pattern Recognition and Machine Learning (Springer 2006)

[Forney(1973)] Forney G. D., The Viterbi Algorithm (IEEE, 1973)

[Rabiner(1989)] Rabiner L. R., A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition (IEEE, 1989)