Modelling regimes with Bayesian network mixtures Supplementary material

Marcus Bendtsen and Jose M. Peña

1 Parameter estimation

Since we cannot observe the hidden variables $H_{1:T}$, we cannot solve the parameter estimation problem by simply counting events from a dataset. Instead we must apply some method that can give us an approximate solution. The canonical way of parameter estimation in HMM is to use EM, and therefore we shall also adopt this technique for our GBN-HMMs.

As before, let $\mathbf{o}_{1:T}$ represent a sequence of observations over the variables $\mathbf{O}_{1:T}$ and let $h_{1:T} = \{h_1, h_2, ..., h_T\}$ represent a sequence of states. Let $\mathcal{H}$ represent the set of all state sequences $h_{1:T}$. The current parameters for our model are denoted Θ', and we seek parameters Θ that maximises the expected log-likelihood of the observed data. This expectation is expressed by $Q(\Theta, \Theta') = \sum_{h_{1:T} \in \mathcal{H}} p(\mathbf{o}_{1:T}, h_{1:T}|\Theta') \log p(\mathbf{o}_{1:T}, h_{1:T}|\Theta)$, thus it is the function Q that we wish to maximise.

We can substitute $p(\mathbf{o}_{1:T}, h_{1:T}|\Theta)$ in the Q function with our factorisation of the GBN-HMM, which gives us the expanded Q function in Equation 1. From this expansion we can see that the terms do not interact, thus we can maximise each term separately.

$$
\begin{aligned}
Q(\Theta, \Theta') = & \sum_{h_{1:T} \in \mathcal{H}} p(\mathbf{o}_{1:T}, h_{1:T}|\Theta') \log p(\mathbf{o}_{1:T}, h_{1:T}|\Theta) = \\
& \sum_{h_{1:T} \in \mathcal{H}} p(\mathbf{o}_{1:T}, h_{1:T}|\Theta') \log \pi_{h_1} \prod_{i=1}^{M} b^i_{h_1}(\mathbf{o}_1) \prod_{t=2}^{T} a_{h_{t-1}, h_t, z_{t-1}} \prod_{t=2}^{T} \prod_{i=1}^{M} b^i_{h_t}(\mathbf{o}_t) = \\
& \sum_{h_{1:T} \in \mathcal{H}} p(\mathbf{o}_{1:T}, h_{1:T}|\Theta') \log \pi_{h_1} + \\
& + \sum_{h_{1:T} \in \mathcal{H}} p(\mathbf{o}_{1:T}, h_{1:T}|\Theta') \sum_{t=2}^{T} \log a_{h_{t-1}, h_t, z_{t-1}} + \\
& + \sum_{h_{1:T} \in \mathcal{H}} p(\mathbf{o}_{1:T}, h_{1:T}|\Theta') \sum_{t=1}^{T} \sum_{i=1}^{M} \log b^i_{h_t}(\mathbf{o}_t)
\end{aligned}
\tag{1}
$$

1.1 Initial state distribution

The first term in Equation 1 can be seen as marginalising out all hidden state variables except the first. To show this we can look at a small example were $\mathcal{H} = \{\{1,1\},\{1,2\},\{2,1\},\{2,2\}\}$. If we expand the sum over these sequences we get Equation 2.

$$\begin{aligned}
&\sum_{h_{1:2}\in\mathcal{H}} p(\mathbf{o}_{1:2}, h_{1:2}|\Theta') \log \pi_{h_1} = \\
&\quad p(\mathbf{o}_{1:2}, h_1 = 1, h_2 = 1|\Theta') \log \pi_1 + p(\mathbf{o}_{1:2}, h_1 = 1, h_2 = 2|\Theta') \log \pi_1 + \\
&\quad p(\mathbf{o}_{1:2}, h_1 = 2, h_2 = 1|\Theta') \log \pi_2 + p(\mathbf{o}_{1:2}, h_1 = 2, h_2 = 2|\Theta') \log \pi_2 = \\
&\quad p(\mathbf{o}_{1:2}, h_1 = 1|\Theta') \log \pi_1 + p(\mathbf{o}_{1:2}, h_1 = 2|\Theta') \log \pi_2 = \\
&\quad \sum_{i=1}^{2} p(\mathbf{o}_{1:2}, h_1 = i|\Theta') \log \pi_i
\end{aligned} \tag{2}$$

Therefore, we can avoid summing over all possible sequences of hidden states and rewrite the first term of Equation 1 in the form of Equation 3.

$$\sum_{h_{1:T}\in\mathcal{H}} p(\mathbf{o}_{1:T}, h_{1:T}|\Theta') \log \pi_{h_1} = \sum_{i=1}^{N} p(\mathbf{o}_{1:T}, h_1 = i|\Theta') \log \pi_i \tag{3}$$

We wish to find the π_i that maximises this expression, under the constraint that $\sum_{i=1}^{N} \pi_i = 1$. In Equation 4 we therefore create a new function f with the addition of the Lagrange multiplier λ, using the aforementioned constraint.

$$f = \sum_{i=1}^{N} p(\mathbf{o}_{1:T}, h_1 = i|\Theta') \log \pi_i - \lambda \left(\sum_{i=1}^{N} \pi_i - 1 \right) \tag{4}$$

The derivative of f with respect to π_i is given in Equation 5.

$$\begin{aligned}
\frac{\partial f}{\partial \pi_i} =& \frac{\partial}{\partial \pi_i} \left(\sum_{i=1}^{N} p(\mathbf{o}_{1:T}, h_1 = i|\Theta') \log \pi_i - \lambda \left(\sum_{i=1}^{N} \pi_i - 1 \right) \right) = \\
=& \frac{\partial}{\partial \pi_i} \left(\sum_{i=1}^{N} p(\mathbf{o}_{1:T}, h_1 = i|\Theta') \log \pi_i - \sum_{i=1}^{N} \lambda \pi_i + \lambda \right) = \\
=& \frac{\partial}{\partial \pi_i} \left(\sum_{i=1}^{N} p(\mathbf{o}_{1:T}, h_1 = i|\Theta') \log \pi_i - \lambda \pi_i \right) + \cancelto{0}{\frac{\partial}{\partial \pi_i} \lambda}
\end{aligned} \tag{5}$$

From Equation 5 we can identify an expression for each partial derivative by realising that for, e.g. π_1, only the part of the sum concerning $i = 1$ will be nonzero in the partial derivative, thus we can remove all of the summands where $i \neq 1$. The partial derivatives are given in Equation 6.

$$\begin{aligned} \frac{\partial}{\partial \pi_1} \left(p(\mathbf{o}_{1:T}, h_1 = 1|\Theta') \log \pi_1 - \lambda \pi_1 \right) &= \frac{1}{\pi_1} p(\mathbf{o}_{1:T}, h_1 = 1|\Theta') - \lambda \\ &\vdots \\ \frac{\partial}{\partial \pi_N} \left(p(\mathbf{o}_{1:T}, h_1 = N|\Theta') \log \pi_N - \lambda \pi_N \right) &= \frac{1}{\pi_N} p(\mathbf{o}_{1:T}, h_1 = N|\Theta') - \lambda \end{aligned} \tag{6}$$

Any stationary point of the Lagrangian function is a stationary point of the original function, subject to the constraints. Therefore each partial derivative must be zero, thus from the last row of Equation 6 we can conclude that for $1 \leq i \leq N$ we must have $\frac{1}{\pi_i} p(\mathbf{o}_{1:T}, h_1 = i|\Theta') - \lambda = 0$. This implies that $p(\mathbf{o}_{1:T}, h_1 = i|\Theta')/\lambda = \pi_i$, and since $\sum_{i=1}^{N} \pi_i = 1$, we have that $\lambda = \sum_{i=1}^{N} p(\mathbf{o}_{1:T}, h_1 = i|\Theta') = p(\mathbf{o}_{1:T}|\Theta')$. Therefore, the choice of π_i that maximises the Q function is given by Equation 7.

$$\pi_i = \frac{p(\mathbf{o}_{1:T}, h_1 = i|\Theta')}{p(\mathbf{o}_{1:T}|\Theta')} \tag{7}$$

1.2 State transition distribution

The second term of Equation 1 can, in a similar way as the first term, be considered a marginalisation but for each time t rather than just $t = 1$. To show this we can expand our example from before, having $\mathcal{H} = \{ \{1,1,1\}, \{1,1,2\}, \{1,2,1\}, \{1,2,2\}, \{2,1,1\}, \{2,1,2\}, \{2,2,1\}, \{2,2,2\}\}$, gives us Equation 8.

$$
\begin{aligned}
&\sum_{h_{1:3}\in\mathcal{H}} p(\mathbf{o}_{1:3}, h_{1:3}|\Theta') \sum_{t=2}^{T} \log a_{h_{t-1}h_t z_{t-1}} = \\
&\quad p(\mathbf{o}_{1:3}, h_1=1, h_2=1, h_3=1|\Theta')(\log a_{11z_1} + \log a_{11z_2}) + \\
&\quad p(\mathbf{o}_{1:3}, h_1=1, h_2=1, h_3=2|\Theta')(\log a_{11z_1} + \log a_{12z_2}) + \\
&\quad p(\mathbf{o}_{1:3}, h_1=1, h_2=2, h_3=1|\Theta')(\log a_{12z_1} + \log a_{21z_2}) + \\
&\quad p(\mathbf{o}_{1:3}, h_1=1, h_2=2, h_3=2|\Theta')(\log a_{12z_1} + \log a_{22z_2}) + \\
&\quad p(\mathbf{o}_{1:3}, h_1=2, h_2=1, h_3=1|\Theta')(\log a_{21z_1} + \log a_{11z_2}) + \\
&\quad p(\mathbf{o}_{1:3}, h_1=2, h_2=1, h_3=2|\Theta')(\log a_{21z_1} + \log a_{12z_2}) + \\
&\quad p(\mathbf{o}_{1:3}, h_1=2, h_2=2, h_3=1|\Theta')(\log a_{22z_1} + \log a_{21z_2}) + \\
&\quad p(\mathbf{o}_{1:3}, h_1=2, h_2=2, h_3=2|\Theta')(\log a_{22z_1} + \log a_{22z_2}) = \\
&= \log a_{11z_1}(p(\mathbf{o}_{1:3}, h_1=1, h_2=1, h_3=1|\Theta') + p(\mathbf{o}_{1:3}, h_1=1, h_2=1, h_3=2|\Theta')) + \\
&\quad \log a_{11z_2}(p(\mathbf{o}_{1:3}, h_1=1, h_2=1, h_3=1|\Theta') + p(\mathbf{o}_{1:3}, h_1=2, h_2=1, h_3=1|\Theta')) + \\
&\quad \vdots \\
&\quad \log a_{22z_2}(p(\mathbf{o}_{1:3}, h_1=1, h_2=2, h_3=2|\Theta') + p(\mathbf{o}_{1:3}, h_1=2, h_2=2, h_3=2|\Theta')) = \\
&= \log a_{11z_1} p(\mathbf{o}_{1:3}, h_1=1, h_2=1|\Theta') + \log a_{11z_2} p(\mathbf{o}_{1:3}, h_2=1, h_3=1|\Theta') + \cdots \\
&\quad \cdots + \log a_{22z_2} p(\mathbf{o}_{1:3}, h_2=2, h_3=2|\Theta') = \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{t=2}^{3} p(\mathbf{o}_{1:3}, h_{t-1}=i, h_t=j|\Theta') \log a_{ijz_{t-1}}
\end{aligned}
\tag{8}
$$

Following this example we can again avoid summing over all possible sequences of hidden states, and rewrite the second term of Equation 1 in the form of Equation 9.

$$
\begin{aligned}
&\sum_{h_{1:T}\in\mathcal{H}} p(\mathbf{o}_{1:T}, h_{1:T}|\Theta') \sum_{t=2}^{T} \log a_{h_{t-1}h_t z_{t-1}} = \\
&\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=i, h_t=j) \log a_{ijz_{t-1}}
\end{aligned}
\tag{9}
$$

For each i in $(1 \le i \le N)$ and k in $(1 \le k \le S^Z)$ where S^Z represents the number of states of Z, it must be the case that $\sum_{j=1}^{N} a_{ijk} = 1$. We incorporate this constraint using Lagrange multipliers, and create the function f in Equation 10.

$$
f = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=i, h_t=j) \log a_{ijz_{t-1}} - \sum_{i=1}^{N}\sum_{k=1}^{S^Z} \lambda_{ik} \left(\sum_{j=1}^{N} a_{ijk} - 1 \right) \tag{10}
$$

The derivative of f with respect to a_{ijk} is given in Equation 11.

$$\begin{aligned}
\frac{\partial f}{\partial a_{ijk}} &= \frac{\partial}{\partial a_{ijk}} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1} = i, h_t = j) \log a_{ijz_{t-1}} - \right. \\
&\left. - \sum_{i=1}^{N} \sum_{k=1}^{S^z} \lambda_{ik} \left(\sum_{j=1}^{N} a_{ijk} - 1 \right) \right) = \\
&= \frac{\partial}{\partial a_{ijk}} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1} = i, h_t = j) \log a_{ijz_{t-1}} - \right. \\
&\left. - \sum_{i=1}^{N} \sum_{k=1}^{S^z} \sum_{j=1}^{N} \lambda_{ik} a_{ijk} + \sum_{i=1}^{N} \sum_{k=1}^{S^z} \lambda_{ik} \right) = \\
&= \frac{\partial}{\partial a_{ijk}} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1} = i, h_t = j) \log a_{ijz_{t-1}} - \right. \\
&\left. - \sum_{i=1}^{N} \sum_{k=1}^{S^z} \sum_{j=1}^{N} \lambda_{ik} a_{ijk} \right) + \cancelto{0}{\frac{\partial}{\partial a_{ijk}} \left(\sum_{i=1}^{N} \sum_{k=1}^{S^z} \lambda_{ik} \right)}
\end{aligned} \tag{11}$$

We can then write each partial derivative on its own by considering each combination of i, j and k separately, which follows from Equation 12.

$$\begin{aligned}
&\frac{\partial}{\partial a_{111}} \left(\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1} = 1, h_t = 1) \log a_{11z_{t-1}} - \lambda_{11} a_{111} \right) \\
&\qquad \vdots \\
&\frac{\partial}{\partial a_{NNS^z}} \left(\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1} = N, h_t = N) \log a_{NNz_{t-1}} - \lambda_{NN} a_{NNS^z} \right)
\end{aligned} \tag{12}$$

Expanding the sum in the partial derivative for a_{111} gives us the summands in Equation 13.

$$
\begin{aligned}
&\frac{\partial}{\partial a_{111}}\left(\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=1, h_t=1)\log a_{11 z_{t-1}} - \lambda_{11} a_{111}\right) = \\
&\quad \frac{\partial}{\partial a_{111}} p(\mathbf{o}_{1:T}, h_1=1, h_2=1)\log a_{11 z_1} + \\
&\quad + \frac{\partial}{\partial a_{111}} p(\mathbf{o}_{1:T}, h_1=1, h_2=1)\log a_{11 z_2} + \\
&\quad + \cdots + \\
&\quad + \frac{\partial}{\partial a_{111}} p(\mathbf{o}_{1:T}, h_{T-1}=1, h_T=1)\log a_{11 z_{T-1}} \\
&\quad - \frac{\partial}{\partial a_{111}} \lambda_{11} a_{111}
\end{aligned} \tag{13}
$$

From Equation 13 we can see that for t where $z_{t-1} \neq 1$, the term will be 0 after derivation with respect to a_{111}. Therefore, if we let $\delta(z_{t-1}=1)$ be one when $z_{t-1}=1$ and zero otherwise, we get Equation 14.

$$
\begin{aligned}
&\frac{\partial}{\partial a_{111}}\left(\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=1, h_t=1)\log a_{11 z_{t-1}} - \lambda_{11} a_{111}\right) = \\
&\frac{\partial}{\partial a_{111}}\left(\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=1, h_t=1)\log a_{11 z_{t-1}}\right) - \frac{\partial}{\partial a_{111}}\lambda_{11} a_{111} = \\
&\frac{1}{a_{111}}\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=1, h_t=1)\delta(z_{t-1}=1) - \lambda_{11}
\end{aligned} \tag{14}
$$

We therefore have that $a_{111} = \frac{\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=1, h_t=1)\delta(z_{t-1}=1)}{\lambda_{11}}$, and since $\sum_{j=1}^{N} a_{1j1} = 1$ we get $\lambda_{11} = \sum_{j=1}^{N}\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=1, h_t=j)\delta(z_{t-1}=1) = \sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=1)\delta(z_{t-1}=1)$. By generalising what we have done for a_{111} to any a_{ijk}, we find that the a_{ijk} that maximises the Q function is given by Equation 15.

$$
a_{ijk} = \frac{\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=i, h_t=j)\delta(z_{t-1}=k)}{\sum_{t=2}^{T} p(\mathbf{o}_{1:T}, h_{t-1}=i)\delta(z_{t-1}=k)} \tag{15}
$$

1.3 Observational model distribution

From the third term in Equation 1 we can see that it is possible to treat each observed variable separately. Therefore, we shall consider the case where $i=1$, from which the rest of the observational variables follow. As was the case with the initial state distribution and transition state distribution, we start by avoiding the summation over every possible hidden state

sequence via marginalisation. We revisit the example with $\mathcal{H} = \{\{1,1\},\{1,2\},\{2,1\},\{2,2\}\}$, and expand the summations in Equation 16.

$$
\begin{aligned}
&\sum_{h_{1:2}\in\mathcal{H}} p(\mathbf{o}_{1:2}, h_{1:2}|\Theta') \sum_{t=1}^{2} \log b^1_{h_t}(\mathbf{o}_t) = \\
&\quad p(\mathbf{o}_{1:2}, h_1 = 1, h_2 = 1|\Theta')(\log b^1_1(\mathbf{o}_1) + \log b^1_1(\mathbf{o}_2)) + \\
&+p(\mathbf{o}_{1:2}, h_1 = 1, h_2 = 2|\Theta')(\log b^1_1(\mathbf{o}_1) + \log b^1_2(\mathbf{o}_2)) + \\
&+p(\mathbf{o}_{1:2}, h_1 = 2, h_2 = 1|\Theta')(\log b^1_2(\mathbf{o}_1) + \log b^1_1(\mathbf{o}_2)) + \\
&+p(\mathbf{o}_{1:2}, h_1 = 2, h_2 = 2|\Theta')(\log b^1_2(\mathbf{o}_1) + \log b^1_2(\mathbf{o}_2)) = \\
&= \log b^1_1(\mathbf{o}_1)(p(\mathbf{o}_{1:2}, h_1 = 1, h_2 = 1|\Theta') + p(\mathbf{o}_{1:2}, h_1 = 1, h_2 = 2|\Theta')) + \\
&+ \log b^1_1(\mathbf{o}_2)(p(\mathbf{o}_{1:2}, h_1 = 1, h_2 = 1|\Theta') + p(\mathbf{o}_{1:2}, h_1 = 2, h_2 = 1|\Theta')) + \\
&+ \log b^1_2(\mathbf{o}_1)(p(\mathbf{o}_{1:2}, h_1 = 2, h_2 = 1|\Theta') + p(\mathbf{o}_{1:2}, h_1 = 2, h_2 = 2|\Theta')) + \\
&+ \log b^1_2(\mathbf{o}_2)(p(\mathbf{o}_{1:2}, h_1 = 1, h_2 = 2|\Theta') + p(\mathbf{o}_{1:2}, h_1 = 2, h_2 = 2|\Theta')) = \\
&= \log b^1_1(\mathbf{o}_1) p(\mathbf{o}_{1:2}, h_1 = 1|\Theta') + \log b^1_1(\mathbf{o}_2) p(\mathbf{o}_{1:2}, h_2 = 1|\Theta') + \\
&+ \log b^1_2(\mathbf{o}_1) p(\mathbf{o}_{1:2}, h_1 = 2|\Theta') + \log b^1_2(\mathbf{o}_2) p(\mathbf{o}_{1:2}, h_2 = 2|\Theta') = \\
&= \sum_{j=1}^{2}\sum_{t=1}^{2} p(\mathbf{o}_{1:2}, h_t = j|\Theta') \log b^1_j(\mathbf{o}_t)
\end{aligned}
\tag{16}
$$

Thus, we again avoid summing over every possible hidden state sequence, and therefore for observed variable $i = 1$ we have Equation 17.

$$
\sum_{h_{1:T}\in\mathcal{H}} \sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_{1:T}|\Theta') \log b^1_{h_t}(\mathbf{o}_t) = \sum_{j=1}^{N}\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = j|\Theta') \log b^1_j(\mathbf{o}_t) \tag{17}
$$

The observed variable under consideration has S^1 states, which we will index by l. For each hidden state $j \in [1, N]$ and parent configuration $k \in [1, K^j]$ of the observed variable (where K^j is the number of parent configurations for the observed variable under hidden state j), we know that summing over all S^1 states will result in unity. If we let b^1_{jkl} represent the probability of state l given hidden state j and parent configuration k, then we have the constraints that $\sum_{l=1}^{S^1} b^1_{jkl} = 1$.

As before, in Equation 18 we create a new function f by introducing Lagrange multipliers using the aforementioned constraints.

$$
f = \sum_{j=1}^{N}\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = j|\Theta') \log b^1_j(\mathbf{o}_t) - \sum_{j=1}^{N}\sum_{k=1}^{K^j} \lambda_{jk} \left(\sum_{l=1}^{S^1} b^1_{jkl} - 1 \right) \tag{18}
$$

The derivative of f is then given by Equation 19.

$$\begin{aligned}
\frac{\partial f}{\partial b^1_{jkl}} &= \frac{\partial}{\partial b^1_{jkl}} \left(\sum_{j=1}^{N} \sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = j|\Theta') \log b^1_j(\mathbf{o}_t) - \sum_{j=1}^{N} \sum_{k=1}^{K^j} \lambda_{jk} \left(\sum_{l=1}^{S^1} b^1_{jkl} - 1 \right) \right) = \\
&= \frac{\partial}{\partial b^1_{jkl}} \left(\sum_{j=1}^{N} \sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = j|\Theta') \log b^1_j(\mathbf{o}_t) - \sum_{j=1}^{N} \sum_{k=1}^{K^j} \sum_{l=1}^{S^1} \lambda_{jk} b^1_{jkl} + \right. \\
&\left. + \sum_{j=1}^{N} \sum_{k=1}^{K^j} \lambda_{jk} \right) = \\
&= \frac{\partial}{\partial b^1_{jkl}} \left(\sum_{j=1}^{N} \sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = j|\Theta') \log b^1_j(\mathbf{o}_t) - \sum_{j=1}^{N} \sum_{k=1}^{K^j} \sum_{l=1}^{S^1} \lambda_{jk} b^1_{jkl} \right) + \\
&+ \frac{\partial}{\partial b^1_{jkl}} \left(\sum_{j=1}^{N} \sum_{k=1}^{K^j} \lambda_{jk} \right) \to 0
\end{aligned} \tag{19}$$

From Equation 19 we can extract each partial derivative on its own in Equation 20.

$$\begin{aligned}
&\frac{\partial}{\partial b^1_{111}} \left(\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta') \log b^1_1(\mathbf{o}_t) - \lambda_{11} b^1_{111} \right) \\
&\vdots \\
&\frac{\partial}{\partial b^1_{NK^N S^1}} \left(\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = N|\Theta') \log b^1_N(\mathbf{o}_t) - \lambda_{NK^N} b^1_{NK^N S^1} \right)
\end{aligned} \tag{20}$$

We shall continue in Equation 21 by expanding the summation within the partial derivative with respect to ∂b^1_{111}, from which the rest will follow.

$$
\begin{aligned}
&\frac{\partial}{\partial b^1_{111}}\left(\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\log b^1_1(\mathbf{o}_t) - \lambda_{11} b^1_{111}\right) = \\
&= \frac{\partial}{\partial b^1_{111}} p(\mathbf{o}_{1:T}, h_1 = 1|\Theta')\log b^1_1(\mathbf{o}_1) + \\
&+ \frac{\partial}{\partial b^1_{111}} p(\mathbf{o}_{1:T}, h_2 = 1|\Theta')\log b^1_1(\mathbf{o}_2) + \\
&+ \cdots + \\
&+ \frac{\partial}{\partial b^1_{111}} p(\mathbf{o}_{1:T}, h_T = 1|\Theta')\log b^1_1(\mathbf{o}_T) - \\
&- \frac{\partial}{\partial b^1_{111}} \lambda_{11} b^1_{111}
\end{aligned}
\tag{21}
$$

By observing that each term $b^1_1(\mathbf{o}_t)$ in Equation 21 identifies the parameter b^1_{jkl} that is consistent with $\mathbf{o}_t$, then it follows that only terms $p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\log b^1_1(\mathbf{o}_t)$ in which the parameter identified is b^1_{111} contribute to the sum once the partial derivative is taken. If we let $\delta(\mathbf{o}_t, b^1_{111})$ be one when b^1_{111} is identified, and zero otherwise, we get Equation 22.

$$
\begin{aligned}
&\frac{\partial}{\partial b^1_{111}}\left(\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\log b^1_1(\mathbf{o}_t) - \lambda_{11} b^1_{111}\right) = \\
&\frac{1}{b^1_{111}}\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\delta(\mathbf{o}_t, b^1_{111}) - \lambda_{11}
\end{aligned}
\tag{22}
$$

The condition for stationary points is given by setting each partial derivative to zero, thus continuing the example we have $b^1_{111} = \frac{\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t=1|\Theta')\delta(\mathbf{o}_t, b^1_{111})}{\lambda_{11}}$. Summing over each state S^1 gives $\sum_{l=1}^{S} b^1_{11l} = \frac{\sum_{l=1}^{S}\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t=1|\Theta')\delta(\mathbf{o}_t, b^1_{11l})}{\lambda_{11}} = 1$. From here we can form an expression for λ_{11} and solve it according to Equation 23.

$$
\begin{aligned}
\lambda_{11} &= \sum_{l=1}^{S^1}\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\delta(\mathbf{o}_t, b^1_{11l}) = \\
&= \sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\delta(\mathbf{o}_t, b^1_{111}) + ... + \sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\delta(\mathbf{o}_t, b^1_{11S^1}) = \\
&= \sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\delta(\mathbf{o}_t, b^1_{11})
\end{aligned}
\tag{23}
$$

In the last row of Equation 23 we let $\delta(\mathbf{o}_t, b^i_{jk})$ be one only when $\mathbf{o}_t$ is consistent with parent configuration k for variable i under hidden state j, else it will be zero. It follows from the

fact that we are summing over all S^1 states that the observed variable takes, and since for each configuration of its parents it must take on a state, that we are essentially counting the number of times we are observing the parent state.

We therefore find that the value for parameter b^1_{111} that maximises the Q function is given by Equation 24 and in general for each observable variable and its respective parameters we have Equation 25.

$$b^1_{111} = \frac{\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\delta(\mathbf{o}_t, b^1_{111})}{\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = 1|\Theta')\delta(\mathbf{o}_t, b^1_{11})} \tag{24}$$

$$b^i_{jkl} = \frac{\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = j|\Theta')\delta(\mathbf{o}_t, b^i_{jkl})}{\sum_{t=1}^{T} p(\mathbf{o}_{1:T}, h_t = j|\Theta')\delta(\mathbf{o}_t, b^i_{jk})} \tag{25}$$

2 Inference

The first desired quantity that we shall consider will be $p(O_{1:T}, H_t|\Theta')$, which we shall refer to as $\gamma(H_T)$ and expand according to Equation 26.

$$\begin{aligned}\gamma(H_t) &= p(O_{1:T}, H_t|\Theta') = p(O_{1:t}, O_{t+1:T}, H_t|\Theta') = \\ &= p(O_{t+1:T}|\cancel{O_{1:t}}^{O_t}, H_t, \Theta')p(O_{1:t}, H_t|\Theta') \\ &= p(O_{t+1:T}|O_t, H_t, \Theta')p(O_{1:t}, H_t|\Theta')\end{aligned} \tag{26}$$

In the second row of Equation 26 we notice that we cannot cancel out all the observations in the first factor, as we have not conditioned on H_{t+1} and due to the Z variable we cannot assume independence.

The final two factors of Equation 26 can efficiently be computed. The joint probability of the state of H_t and the observations $O_{1:t}$ is given by Equation 27, and the conditional probability of the observations $O_{t+1:T}$ given the state of H_t and the observation O_t is given by Equation 28.

$$
\begin{aligned}
\alpha(H_t) &= p(H_t, O_{1:t}|\Theta') = \sum_{H_{t-1}} p(H_t, H_{t-1}, O_{1:t-1}, O_t|\Theta') \\
&= \sum_{H_{t-1}} p(O_t|H_t, \cancel{H_{t-1}}, \cancel{O_{1:t-1}}, \Theta') p(H_t, H_{t-1}, O_{1:t-1}|\Theta') \\
&= \sum_{H_{t-1}} p(O_t|H_t, \Theta') p(H_t|H_{t-1}, \cancel{O_{1:t-1}}^{O_{t-1}}, \Theta') p(H_{t-1}, O_{1:t-1}|\Theta') \\
&= p(O_t|H_t, \Theta') \sum_{H_{t-1}} p(H_t|H_{t-1}, O_{t-1}, \Theta') \alpha(H_{t-1}) \\
&= \prod_{i=1}^{M} b^i_{H_t}(O_t) \sum_{H_{t-1}} a_{H_{t-1} H_t Z_{t-1}} \alpha(H_{t-1})
\end{aligned}
\tag{27}
$$

$$
\begin{aligned}
\beta(H_t) &= p(O_{t+1:T}|H_t, O_t, \Theta') = \sum_{H_{t+1}} p(O_{t+1}, O_{t+2:T}, H_{t+1}|H_t, O_t, \Theta') \\
&= \sum_{H_{t+1}} p(O_{t+2:T}|H_{t+1}, O_{t+1}, \cancel{H_t}, \cancel{O_t}, \Theta') p(O_{t+1}, H_{t+1}|H_t, O_t, \Theta') \\
&= \sum_{H_{t+1}} p(O_{t+2:T}|H_{t+1}, O_{t+1}, \Theta') p(O_{t+1}|H_{t+1}, \cancel{H_t}, \cancel{O_t}, \Theta') p(H_{t+1}|H_t, O_t, \Theta') \\
&= \sum_{H_{t+1}} \beta(H_{t+1}) p(O_{t+1}|H_{t+1}, \Theta') p(H_{t+1}|H_t, O_t, \Theta') \\
&= \sum_{H_{t+1}} \beta(H_{t+1}) \prod_{i=1}^{M} b^i_{H_{t+1}}(O_{t+1}) a_{H_t H_{t+1} Z_t}
\end{aligned}
\tag{28}
$$

Using this new notation, we can write $\gamma(H_t) = \alpha(H_t)\beta(H_t)$. Noticing that $\sum_{i=1}^{N} \gamma(H_t = i) = p(O_{1:T}|\Theta')$, we can readily compute Equation 7 and Equation 25 since both α and β are expressed in known quantities (under the parameters Θ').

The second quantity that we need to be able to compute, in order to compute Equation 15, is $p(O_{1:T}, H_{t-1}, H_t|\Theta')$, which we shall refer to as $\xi(H_{t-1}, H_t)$. We expand this quantity in Equation 29, and express it in terms of known quantities.

$$
\begin{aligned}
&\xi(H_{t-1}, H_t) = \\
&= p(O_{1:T}, H_{t-1}, H_t|\Theta') = p(O_{1:t-1}, O_t, O_{t+1:T}, H_{t-1}, H_t|\Theta') = \\
&= p(O_{t+1:T}|\cancel{O_{1:t-1}}, O_t, \cancel{H_{t-1}}, H_t, \Theta')p(O_{1:t-1}, O_t, H_{t-1}, H_t|\Theta') \\
&= p(O_{t+1:T}|O_t, H_t, \Theta')p(O_t|\cancel{O_{1:t-1}}, \cancel{H_{t-1}}, H_t, \Theta')p(O_{1:t-1}, H_{t-1}, H_t|\Theta') \\
&= p(O_{t+1:T}|O_t, H_t, \Theta')p(O_t|H_t, \Theta')p(H_t|\cancel{O_{1:t-1}} \overset{O_{t-1}}{\rightarrow}, H_{t-1}, \Theta')p(O_{1:t-1}, H_{t-1}|\Theta') \\
&= p(O_{t+1:T}|O_t, H_t, \Theta')p(O_t|H_t, \Theta')p(H_t|O_{t-1}, H_{t-1}, \Theta')p(O_{1:t-1}, H_{t-1}|\Theta') = \\
&= \beta(H_t) \prod_{i=1}^{M} b^i_{H_t}(O_t) a_{H_{t-1}H_t Z_{t-1}} \alpha(H_{t-1})
\end{aligned}
\tag{29}
$$