Supplement:
A Spectral Algorithm for Inference in Hidden semi-Markov Models

Igor Melnyk [*] and Arindam Banerjee[*]

[*]Department of Computer Science and Engineering, University of Minnesota, Minneapolis,
{melnyk, banerjee}@cs.umn.edu

A More on Tensor Algebra

In this section, we provide more discussion on the basic tensor operations used in the paper.

Recall that if we let $\underset{m_1,\ldots,m_N}{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{I_{m_1}\times\cdots\times I_{m_N}}$, $\underset{q_1,\ldots,q_L,r_1,\ldots,r_M}{\boldsymbol{\mathcal{Y}}} \in \mathbb{R}^{I_{q_1}\times\cdots\times I_{q_L}\times I_{r_1}\times\cdots\times I_{r_M}}$ then the tensor multiplication is defined as:

$$\underset{p_1,\ldots,p_K,r_1,\ldots,r_M}{\boldsymbol{\mathcal{Z}}} = \underset{p_1,\ldots,p_K,q_1,\ldots,q_L}{\boldsymbol{\mathcal{X}}} \times_{q_1,\ldots,q_L} \underset{q_1,\ldots,q_L,r_1,\ldots,r_M}{\boldsymbol{\mathcal{Y}}}$$

where $\underset{p_1,\ldots,p_K,r_1,\ldots,r_M}{\boldsymbol{\mathcal{Z}}} \in \mathbb{R}^{I_{p_1}\times\cdots\times I_{p_K}\times I_{r_1}\times\cdots\times I_{r_M}}$. Observe that in the above we can flatten the tensors $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ in multiple different ways as long as the matrix multiplication remains valid. For example, we could assign the multiplication modes in both tensors to columns, in this case the matrix product becomes $\mathbf{Z} = \mathbf{X}\mathbf{Y}^T$. Alternatively, the tensor $\boldsymbol{\mathcal{Y}}$ could be matrisized with the multiplication modes corresponding to rows, resulting in the product $\mathbf{Z} = \mathbf{X}\mathbf{Y}$.

In a series of tensor multiplications the order is irrelevant as long as the multiplication is performed along the matching modes:

$$\underset{sp}{\boldsymbol{\mathcal{X}}} \times_s \left(\underset{tr}{\boldsymbol{\mathcal{Y}}} \times_r \underset{rs}{\boldsymbol{\mathcal{Z}}}\right) = \left(\underset{sp}{\boldsymbol{\mathcal{X}}} \times_s \underset{rs}{\boldsymbol{\mathcal{Z}}}\right) \times_r \underset{tr}{\boldsymbol{\mathcal{Y}}}$$

If we let the matrisized tensors to be $\mathbf{X} \in \mathbb{R}^{I_p\times I_s}$, $\mathbf{Y} \in \mathbb{R}^{I_t\times I_r}$ and $\mathbf{Z} \in \mathbb{R}^{I_r\times I_s}$, then the above can be verified to be true since

$$\mathbf{X}\left(\mathbf{Y}\mathbf{Z}\right) = \left(\mathbf{X}\mathbf{Z}^T\right)\mathbf{Y}^T$$

Note that in many places we will drop the multiplication subscripts. The implied modes of multiplication can then be inferred from the subscripts of the tensors. Specifically, when two tensors are multiplied, we first check their modes and then multiply along the modes which are common to both of them. For example, in the product $\underset{pqr}{\boldsymbol{\mathcal{X}}} \times \underset{qsr}{\boldsymbol{\mathcal{Y}}}$ the implied multiplication is performed along the common modes, i.e., q and r.

Tensor inverse $\boldsymbol{\mathcal{X}}^{-1}$ is defined with respect to a certain subset of modes:

$$\underset{p_1,\ldots,p_K,q_1,\ldots,q_L}{\boldsymbol{\mathcal{X}}} \times_{q_1,\ldots,q_L} \underset{p_1,\ldots,p_K,q_1,\ldots,q_L}{\boldsymbol{\mathcal{X}}^{-1}} = \underset{p_1,\ldots,p_K,p_1,\ldots,p_K}{\boldsymbol{\mathcal{I}}}$$

where the inversion is performed with respect to the modes $q_1,\ldots,q_L$. Observe that in the above, tensor $\underset{p_1,\ldots,p_K,p_1,\ldots,p_K}{\boldsymbol{\mathcal{I}}}$ has duplicate modes and denotes an identity tensor, whose elements are everywhere zero, except at $\boldsymbol{\mathcal{I}}(i_1,\ldots,i_K,i_1,\ldots,i_K) = 1$. In general, if a tensor has duplicate modes, the corresponding sub-tensor can be interpreted as a hyper-diagonal. For example, if for a tensor $\underset{pq}{\boldsymbol{\mathcal{X}}}$ we construct a tensor $\underset{pppq}{\overline{\boldsymbol{\mathcal{X}}}}$, which has its mode p duplicated three times, then for a fixed index i, the sub-tensor $\overline{\boldsymbol{\mathcal{X}}}(:,:,:,i)$ is a hypercube with elements $\boldsymbol{\mathcal{X}}(:,i)$ on the diagonal.

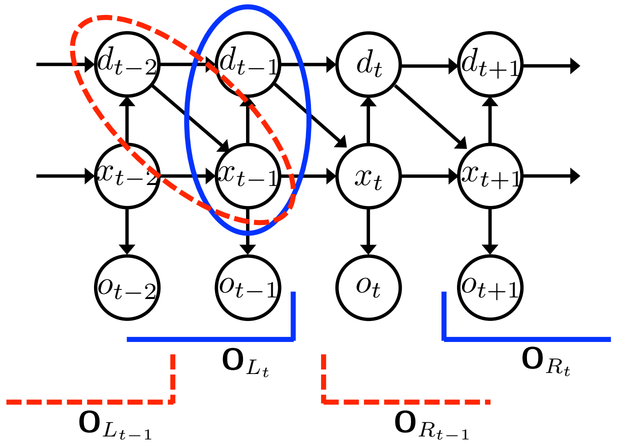

Figure 1: Conditional independence in HSMM. The figure depicts two sets of relationships: $\mathbf{O}_{L_t}$ and $\mathbf{O}_{R_t}$ are independent conditioned on $x_{t-1}d_{t-1}$, similarly, $\mathbf{O}_{L_{t-1}}$ and $\mathbf{O}_{R_{t-1}}$ are conditionally independent given $x_{t-1}d_{t-2}$. We defined $\mathbf{O}_{L_t} = \{\ldots, o_{t-2}, o_{t-1}\}$ and $\mathbf{O}_{R_t} = \{o_{t+1}, o_{t+2}, \ldots\}$.

Mode duplication enables us to multiply several tensors along the same mode. For example, if we need to multiply tensors $\underset{sp}{\mathcal{X}}$, $\underset{pr}{\mathcal{Y}}$ and $\underset{tp}{\mathcal{Z}}$ along the mode p, then a simple product of the form

$$\underset{sp}{\mathcal{X}} \times_p \underset{pr}{\mathcal{Y}} \times_p \underset{tp}{\mathcal{Z}}$$

cannot be done since any product of two tensors along the mode p would eliminate it, preventing any further multiplications. In general, if there are N multiplications along the specific mode, then there are must be cumulatively $2N$ modes in the participating tensors. In our example, we might duplicate the mode p in, say, tensor $\mathcal{Z}$ to have

$$\underset{sp}{\mathcal{X}} \times_p \left(\underset{pr}{\mathcal{Y}} \times_p \underset{tpp}{\mathcal{Z}} \right)$$

To reduce clutter, in some places we do not explicitly show the duplicated variables in the subscripts; the implied mode numerosity will be evident from the context or explicitly stated in cases when there is a confusion. For example, the notation for the identity tensor becomes $\underset{p_1,\ldots,p_K}{\mathcal{I}}$.

B Estimation of Observable Tensors

In the main paper we expressed the joint probability of the observed variables in the form:

$$\underset{o_1,\ldots,o_T}{\mathcal{P}} = \prod_t \underset{\mathbf{O}_{R_{t-1}}\mathbf{O}_{R_t}}{\tilde{\mathcal{D}}} \times_{\mathbf{O}_{R_t}} \left(\underset{\mathbf{O}_{R_t} o_t \mathbf{O}_{R_t}}{\tilde{\mathcal{X}}} \times_{o_t} \underset{o_t o_t}{\tilde{\mathcal{O}}} \right) \tag{B.1}$$

and derived the observable form for tensor $\underset{\mathbf{O}_{R_{t-1}}\mathbf{O}_{R_t}}{\tilde{\mathcal{D}}}$:

$$\underset{\mathbf{O}_{R_{t-1}}\mathbf{O}_{R_t}}{\tilde{\mathcal{D}}} = \underset{\mathbf{O}_{L_{t-1}}\mathbf{O}_{R_{t-1}}}{\mathcal{M}^{-1}} \times_{\mathbf{O}_{L_{t-1}}} \underset{\mathbf{O}_{L_{t-1}}\mathbf{O}_{R_t}}{\mathcal{M}} \tag{B.2}$$

In this Section, we present the omitted derivations for $\underset{\mathbf{O}_{R_t} o_t \mathbf{O}_{R_t}}{\tilde{\mathcal{X}}}$ and $\underset{o_t o_t}{\tilde{\mathcal{O}}}$.

B.1 Computation of Tensor $\underset{\mathbf{o}_{R_t}o_t\mathbf{o}_{R_t}}{\tilde{\mathcal{X}}}$

The form of this tensor was established to be:

$$\underset{\mathbf{o}_{R_t}o_t\mathbf{o}_{R_t}}{\tilde{\mathcal{X}}} = \underset{\mathbf{o}_{R_t}|x_{t-1}d_{t-1}}{\mathcal{F}^{-1}} \times_{x_{t-1}d_{t-1}} \left(\underset{x_t x_t|x_{t-1}d_{t-1}d_{t-1}}{\mathcal{X}} \times_{x_t} \underset{o_t|x_t}{\mathcal{F}} \right) \times_{x_t d_{t-1}} \underset{\mathbf{o}_{R_t}|x_t d_{t-1}}{\mathcal{F}} \tag{B.3}$$

Consider the following conditional independence relationship (see Figure 1):

$$\underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}}{\mathcal{M}} = \underset{\mathbf{o}_{L_t}|x_{t-1}d_{t-1}}{\mathcal{F}} \times_{x_{t-1}d_{t-1}} \underset{\mathbf{o}_{R_t}|x_{t-1}d_{t-1}}{\mathcal{F}} \times_{x_{t-1}d_{t-1}} \underset{x_{t-1}d_{t-1}}{\mathcal{K}}, \tag{B.4}$$

where $\underset{x_{t-1}d_{t-1}}{\mathcal{K}} = \underset{x_{t-1}d_{t-1}x_{t-1}d_{t-1}}{\mathcal{K}}$ and we omitted the duplicated modes.

Express the inverse of tensor $\underset{\mathbf{o}_{R_t}|x_{t-1}d_{t-1}}{\mathcal{F}}$ from the above equation

$$\underset{\mathbf{o}_{R_t}|x_{t-1}d_{t-1}}{\mathcal{F}^{-1}} = \underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}}{\mathcal{M}^{-1}} \times_{\mathbf{o}_{L_t}} \underset{\mathbf{o}_{L_t}|x_{t-1}d_{t-1}}{\mathcal{F}} \times_{x_{t-1}d_{t-1}} \underset{x_{t-1}d_{t-1}}{\mathcal{K}}$$

where tensor $\underset{\mathbf{o}_{R_t}|x_{t-1}d_{t-1}}{\mathcal{F}}$ is inverted with respect to mode $\mathbf{o}_{R_t}$, while $\underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}}{\mathcal{M}}$ is inverted with respect to mode $\mathbf{o}_{L_t}$. Substituting back to (B.3), we get

$$\underset{\mathbf{o}_{R_t}o_t\mathbf{o}_{R_t}}{\tilde{\mathcal{X}}} = \underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}}{\mathcal{M}^{-1}} \times_{\mathbf{o}_{L_t}} \underset{\mathbf{o}_{L_t}|x_{t-1}d_{t-1}}{\mathcal{F}} \times_{x_{t-1}d_{t-1}} \underset{x_{t-1}d_{t-1}}{\mathcal{K}} \times_{x_{t-1}d_{t-1}} \left(\underset{x_t x_t|x_{t-1}d_{t-1}d_{t-1}}{\mathcal{X}} \times_{x_t} \underset{o_t|x_t}{\mathcal{F}} \right) \times_{x_t d_{t-1}} \underset{\mathbf{o}_{R_t}|x_t d_{t-1}}{\mathcal{F}}$$

Multiplying together the last five factors, we obtain

$$\underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}o_t}{\mathcal{M}} = \underset{\mathbf{o}_{L_t}|x_{t-1}d_{t-1}}{\mathcal{F}} \times_{x_{t-1}d_{t-1}} \underset{x_{t-1}d_{t-1}}{\mathcal{K}} \times_{x_{t-1}d_{t-1}} \left(\underset{x_t x_t|x_{t-1}d_{t-1}d_{t-1}}{\mathcal{X}} \times_{x_t} \underset{o_t|x_t}{\mathcal{F}} \right) \times_{x_t d_{t-1}} \underset{\mathbf{o}_{R_t}|x_t d_{t-1}}{\mathcal{F}}$$

Finally, (B.3) can now be written as

$$\underset{\mathbf{o}_{R_t}o_t\mathbf{o}_{R_t}}{\tilde{\mathcal{X}}} = \underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}}{\mathcal{M}^{-1}} \times_{\mathbf{o}_{L_t}} \underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}o_t}{\mathcal{M}} \tag{B.5}$$

where the right hand side can now be estimated directly from data, without the need of the model parameters.

B.2 Computation of Tensor $\underset{o_t o_t}{\tilde{\mathcal{O}}}$

Finally, we consider the tensor

$$\underset{o_t o_t}{\tilde{\mathcal{O}}} = \underset{o_t|x_t}{\mathcal{F}^{-1}} \times_{x_t} \underset{o_t|x_t}{\mathcal{O}} \tag{B.6}$$

The conditional independence relationship can take the form

$$\underset{o_t o_{t+1}}{\mathcal{M}} = \underset{o_t|x_t}{\mathcal{F}} \times_{x_t} \underset{o_{t+1}|x_t}{\mathcal{F}} \times_{x_t} \underset{x_t x_t}{\mathcal{K}} \tag{B.7}$$

Expressing the inverse of $\underset{o_t|x_t}{\mathcal{F}}$

$$\underset{o_t|x_t}{\mathcal{F}^{-1}} = \underset{o_t o_{t+1}}{\mathcal{M}^{-1}} \times_{o_{t+1}} \underset{o_{t+1}|x_t}{\mathcal{F}} \times_{x_t} \underset{x_t x_t}{\mathcal{K}}$$

and substituting in (B.6), we get

$$\begin{aligned} \underset{o_t o_t}{\tilde{\mathcal{O}}} &= \underset{o_t o_{t+1}}{\mathcal{M}^{-1}} \times_{o_{t+1}} \underset{o_{t+1}|x_t}{\mathcal{F}} \times_{x_t} \underset{x_t}{\mathcal{K}} \times_{x_t} \underset{o_t|x_t}{\mathcal{O}} \\ &= \underset{o_t o_{t+1}}{\mathcal{M}^{-1}} \times_{o_{t+1}} \underset{o_t o_{t+1}}{\mathcal{M}} \end{aligned} \tag{B.8}$$

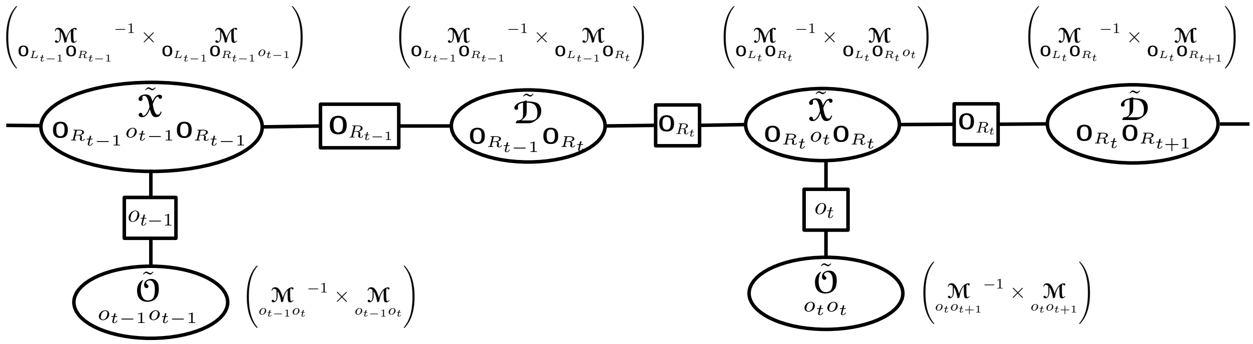
$\left(\underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_{t-1}}}{\mathcal{M}}^{-1} \times \underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_{t-1}}o_{t-1}}{\mathcal{M}}\right)$
$\left(\underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_{t-1}}}{\mathcal{M}}^{-1} \times \underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_t}}{\mathcal{M}}\right)$
$\left(\underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}}{\mathcal{M}}^{-1} \times \underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}o_t}{\mathcal{M}}\right)$
$\left(\underset{\mathbf{o}_{L_t}\mathbf{o}_{R_t}}{\mathcal{M}}^{-1} \times \underset{\mathbf{o}_{L_t}\mathbf{o}_{R_{t+1}}}{\mathcal{M}}\right)$
$\underset{\mathbf{o}_{R_{t-1}}o_{t-1}\mathbf{o}_{R_{t-1}}}{\tilde{\mathcal{X}}}$
$\mathbf{o}_{R_{t-1}}$
$\underset{\mathbf{o}_{R_{t-1}}\mathbf{o}_{R_t}}{\tilde{\mathcal{D}}}$
$\mathbf{o}_{R_t}$
$\underset{\mathbf{o}_{R_t}o_t\mathbf{o}_{R_t}}{\tilde{\mathcal{X}}}$
$\mathbf{o}_{R_t}$
$\underset{\mathbf{o}_{R_t}\mathbf{o}_{R_{t+1}}}{\tilde{\mathcal{D}}}$
o_{t-1}
o_t
$\underset{o_{t-1}o_{t-1}}{\tilde{\mathcal{O}}}$
$\left(\underset{o_{t-1}o_t}{\mathcal{M}}^{-1} \times \underset{o_{t-1}o_t}{\mathcal{M}}\right)$
$\underset{o_t o_t}{\tilde{\mathcal{O}}}$
$\left(\underset{o_t o_{t+1}}{\mathcal{M}}^{-1} \times \underset{o_t o_{t+1}}{\mathcal{M}}\right)$

Figure 2: Graphical representation of the HSMM spectral algorithm for inference in Algorithm 1. The cliques and separators are now defined in terms of the tensors, which are defined with respect to the observed data. The expressions in the parenthesis show the observable representation of the corresponding tensors.

Algorithm 1 Basic Spectral Algorithm for HSMM inference

```
Input: Training sequences: $\mathbf{S}^i = \{o_1^i, \dots, o_{T_i}^i\}, i = 1, \dots, N$.
Testing sequence: $\mathbf{S}^{test} = \{o_1^{test}, \dots, o_T^{test}\}$.
Output: $p(\mathbf{S}^{test})$

Training phase:
for all $t$ do
    Estimate $\underset{\mathbf{o}_{R_{t-1}}\mathbf{o}_{R_t}}{\tilde{\mathcal{D}}}$, $\underset{\mathbf{o}_{R_t}o_t\mathbf{o}_{R_t}}{\tilde{\mathcal{X}}}$ and $\underset{o_t o_t}{\tilde{\mathcal{O}}}$ from data $\{\mathbf{S}^1, \dots, \mathbf{S}^N\}$ using equations (B.2), (B.5) and (B.8).
end for

Testing phase:
$p(\mathbf{S}^{test}) = 1$
for $t = T$ down to $t = 1$ do
    $p(\mathbf{S}^{test}) = p(\mathbf{S}^{test}) \times \underset{\mathbf{o}_{R_{t-1}}\mathbf{o}_{R_t}}{\tilde{\mathcal{D}}} \times_{\mathbf{o}_{R_t}} \left( \underset{\mathbf{o}_{R_t}o_t\mathbf{o}_{R_t}}{\tilde{\mathcal{X}}} \times_{o_t} \underset{o_t o_t}{\tilde{\mathcal{O}}}\Big|_{o_t = o_t^{test}} \right)$
end for
```

C Spectral Algorithm

In this Section we present additional details for the derived spectral algorithm in its basic form as well as the form, which improves the algorithm's accuracy and reduces its complexity.

C.1 Basic Version

Using (B.2), (B.5) and (B.8) in (B.1) we can obtain the spectral algorithm to compute $\underset{o_1,\dots,o_T}{\mathcal{P}}$ entirely using the observed variables and Algorithm 1 shows its basic version. Figure 2 shows the graphical representation of this algorithm in terms of the transformed junction tree.

The notation $\underset{o_t o_t}{\tilde{\mathcal{O}}}\Big|_{o_t = o_t^{test}}$ means that based on the value of the tth symbol in testing sequence, we slice the tensor $\underset{o_t o_t}{\tilde{\mathcal{O}}}$ along the element o_t^{test} in the dimension o_t. For example, if $\underset{o_t o_t}{\tilde{\mathcal{O}}} \in \mathbb{R}^{10\times10}$ and $o_t^{test} = 3$ then $\underset{o_t o_t}{\tilde{\mathcal{O}}}\Big|_{o_t = o_t^{test}} \in \mathbb{R}^{10\times1}$, a third column in the original matrix.

Analyzing (B.2), (B.5) and (B.8), we see that the computational complexity of the training phase of the algorithm is determined by the tensor inverses and multiplications. For example, if in (B.2) we denote $|\mathbf{o}_R| = |\mathbf{o}_L| = \ell$, then $\underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_{t-1}}}{\mathbf{M}} \in \mathbb{R}^{n_o^\ell \times n_o^\ell}$ and $\underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_t}}{\mathbf{M}} \in \mathbb{R}^{n_o^\ell \times n_o^\ell}$. The computational complexity of the multiplications and inversions would then be $\mathcal{O}(n_o^{3\ell})$. Performing this computations for all t and assuming that the length of training and testing sequences is T, would result in $\mathcal{O}\left(n_o^{3\ell}T\right)$. Additionally, there will be a cost of $\mathcal{O}\left(\ell NT\right)$ to estimate the sample moments $\boldsymbol{\mathfrak{M}}$, which is based on counting the co-occurrences of certain observable symbols. Here N is the number of training sequences.

In the testing phase of the algorithm, we perform a series of tensor multiplications with the cost of $\mathcal{O}(n_o^{3\ell}T)$. Thus, the overall cost of Algorithm 1 is then $\mathcal{O}\left((n_o^{3\ell}+\ell N)T\right)$.

In the following section we show how to improve the accuracy and efficiency of the basic spectral Algorithm 1. The idea is to estimate only three tensors $\tilde{\boldsymbol{\mathfrak{X}}}$, $\tilde{\boldsymbol{\mathfrak{D}}}$ and $\tilde{\boldsymbol{\mathfrak{O}}}$ in the batch, by averaging across all t.

C.2 Efficient Version

We show the details for computing the tensors $\tilde{\boldsymbol{\mathfrak{D}}}$ in the batch form. The derivations for other tensors $\tilde{\boldsymbol{\mathfrak{X}}}$ and $\tilde{\boldsymbol{\mathfrak{O}}}$ can be computed in a similar manner. Recall from (B.2) the form of $\underset{\mathbf{o}_{R_{t-1}}\mathbf{o}_{R_t}}{\tilde{\boldsymbol{\mathfrak{D}}}}$, and consider the following structure:

$$\left(\sum_t \underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_{t-1}}}{\boldsymbol{\mathfrak{M}}}\right)^{-1} \times_{\mathbf{o}_L} \left(\sum_t \underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_t}}{\boldsymbol{\mathfrak{M}}}\right) \tag{C.1}$$

where $\mathbf{o}_L$ denotes a generic mode of the averaged tensor $\boldsymbol{\mathfrak{M}}$, corresponding to $\mathbf{o}_{L_{t-1}}$ for all t. Note that in practice, instead of summation, we use averaging to avoid numerical overflow problems. It is equivalent to the considered expression, since the term $\frac{1}{T}$ then cancels out.

Since

$$\underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_{t-1}}}{\boldsymbol{\mathfrak{M}}} = \underset{\mathbf{o}_{L_{t-1}}|x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{F}}} \times_{x_{t-1}d_{t-2}} \underset{\mathbf{o}_{R_{t-1}}|x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{F}}} \times_{x_{t-1}d_{t-2}} \underset{x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{K}}} \tag{C.2}$$

the first term inside brackets can be rewritten as:

$$\begin{aligned}
&\sum_t \underset{\mathbf{o}_{L_{t-1}}|x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{F}}} \times_{x_{t-1}d_{t-2}} \underset{\mathbf{o}_{R_{t-1}}|x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{F}}} \times_{x_{t-1}d_{t-2}} \underset{x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{K}}}\\
&\quad = \sum_t \underset{\mathbf{o}_{R_{t-1}}|x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{F}}} \times_{x_{t-1}d_{t-2}} \underset{\mathbf{o}_{L_{t-1}}x_{t-1}d_{t-2}}{\overline{\boldsymbol{\mathfrak{F}}}}\\
&\quad = \underset{\mathbf{o}_{R_2}|x_2d_1}{\boldsymbol{\mathfrak{F}}} \times \left(\sum_t \underset{\mathbf{o}_{L_{t-1}}x_{t-1}d_{t-2}}{\overline{\boldsymbol{\mathfrak{F}}}}\right)
\end{aligned} \tag{C.3}$$

where in the second line we combined the two factors, i.e., $\underset{\mathbf{o}_{L_{t-1}}x_{t-1}d_{t-2}}{\overline{\boldsymbol{\mathfrak{F}}}} = \underset{\mathbf{o}_{L_{t-1}}|x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{F}}} \times_{x_{t-1}d_{t-2}} \underset{x_{t-1}d_{t-2}x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{K}}}$ and in the third line we used the homogeneity property of HSMM, i.e., the fact that $\underset{\mathbf{o}_{R_{t-1}}|x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{F}}}$ does not depend on time stamp t, and extracted one of the common factors. Note that the term $\underset{\mathbf{o}_{L_{t-1}}x_{t-1}d_{t-2}}{\overline{\boldsymbol{\mathfrak{F}}}}$, on the other hand, does depend on t since the factor $\underset{x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{K}}}$ changes as the time stamp t changes.

Similarly, since

$$\underset{\mathbf{o}_{L_{t-1}}\mathbf{o}_{R_t}}{\boldsymbol{\mathfrak{M}}} = \underset{\mathbf{o}_{L_{t-1}}|x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{F}}} \times_{x_{t-1}d_{t-2}} \underset{x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{K}}} \times_{x_{t-1}d_{t-2}} \underset{d_{t-1}|x_{t-1}x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{D}}} \times_{x_{t-1}d_{t-1}} \underset{\mathbf{o}_{R_t}|x_{t-1}d_{t-1}}{\boldsymbol{\mathfrak{F}}} \tag{C.4}$$

rewrite the second term in (C.1) as

$$\begin{aligned}
&\sum_t \underset{\mathbf{o}_{L_{t-1}}|x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{F}}} \times_{x_{t-1}d_{t-2}} \underset{x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{K}}} \times_{x_{t-1}d_{t-2}} \underset{d_{t-1}|x_{t-1}x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{D}}} \times_{x_{t-1}d_{t-1}} \underset{\mathbf{o}_{R_t}|x_{t-1}d_{t-1}}{\boldsymbol{\mathfrak{F}}}\\
&\quad = \sum_t \underset{\mathbf{o}_{L_{t-1}}x_{t-1}d_{t-2}}{\overline{\boldsymbol{\mathfrak{F}}}} \times_{x_{t-1}d_{t-2}} \underset{d_{t-1}|x_{t-1}x_{t-1}d_{t-2}}{\boldsymbol{\mathfrak{D}}} \times_{x_{t-1}d_{t-1}} \underset{\mathbf{o}_{R_t}|x_{t-1}d_{t-1}}{\boldsymbol{\mathfrak{F}}}\\
&\quad = \left(\sum_t \underset{\mathbf{o}_{L_{t-1}}x_{t-1}d_{t-2}}{\overline{\boldsymbol{\mathfrak{F}}}}\right) \times \underset{d_2|x_2x_2d_1}{\boldsymbol{\mathfrak{D}}} \times_{x_2d_2} \underset{\mathbf{o}_{R_3}|x_2d_2}{\boldsymbol{\mathfrak{F}}}
\end{aligned} \tag{C.5}$$

Algorithm 2 Efficient Spectral Algorithm for HSMM inference

```
**Input:** Training sequences: $\mathbf{S}^i = \{o^i_1, \dots, o^i_{T_i}\}, i = 1, \dots, N$.
Testing sequence: $\mathbf{S}^{test} = \{o^{test}_1, \dots, o^{test}_T\}$.
**Output:** $p(\mathbf{S}^{test})$

**Training phase:**
Estimate $\tilde{\mathfrak{D}}, \tilde{\mathfrak{X}}$ and $\tilde{\mathfrak{O}}$ from data $\{\mathbf{S}^1, \dots, \mathbf{S}^N\}$ using equations (C.8), (C.9) and (C.10).

**Testing phase:**
$p(\mathbf{S}^{test}) = 1$
**for** $i = T$ **down to** $i = 1$ **do**
    $p(\mathbf{S}^{test}) = p(\mathbf{S}^{test}) \times \tilde{\mathfrak{D}} \times \left(\tilde{\mathfrak{X}} \times \tilde{\mathfrak{O}}|_{o=o_i^{test}}\right)$
**end for**
```

where we used the transformations similar as in (C.3). Now if we multiply the inverse of (C.3) with (C.5), we get

$$\begin{aligned}
&\underset{\mathsf{\mathbf{O}}_{R_2}|x_2d_1}{\mathfrak{F}^{-1}} \times \left(\sum_t \underset{\mathsf{\mathbf{O}}_{L_{t-1}}x_{t-1}d_{t-2}}{\overline{\mathfrak{F}}}\right)^{-1} \times \left(\sum_t \underset{\mathsf{\mathbf{O}}_{L_{t-1}}x_{t-1}d_{t-2}}{\overline{\mathfrak{F}}}\right) \times \underset{d_2|x_2x_2d_1}{\mathfrak{D}} \times \underset{\mathsf{\mathbf{O}}_{R_3}|x_2d_2}{\mathfrak{F}} \qquad \text{(C.6)}\\
&= \underset{\mathsf{\mathbf{O}}_{R_2}|x_2d_1}{\mathfrak{F}^{-1}} \times_{x_2d_1} \underset{d_2|x_2x_2d_1}{\mathfrak{D}} \times_{x_2d_2} \underset{\mathsf{\mathbf{O}}_{R_3}|x_2d_2}{\mathfrak{F}}\\
&= \underset{\mathsf{\mathbf{O}}_{R_2}\mathsf{\mathbf{O}}_{R_3}}{\tilde{\mathfrak{D}}} = \underset{\mathsf{\mathbf{O}}_{R_{t-1}}\mathsf{\mathbf{O}}_{R_t}}{\tilde{\mathfrak{D}}} \qquad \text{(C.7)}
\end{aligned}$$

where in (C.6) we used the fact that the order in which tensors are multiplied is irrelevant and also the fact that the terms in parenthesis are invertible. This is due to the fact that the set of observations $\mathsf{\mathbf{O}}_{L_{t-1}}$ for all t was selected so as to make each of the summand invertible; the selection of $\mathsf{\mathbf{O}}_{L_{t-1}}$ was discussed in Section 6 in the main paper and we will provide mode details in this supplement in Section D. Moreover, in (C.7) we used the definition of $\underset{\mathsf{\mathbf{O}}_{R_{t-1}}\mathsf{\mathbf{O}}_{R_t}}{\tilde{\mathfrak{D}}}$

$$\underset{\mathsf{\mathbf{O}}_{R_{t-1}}\mathsf{\mathbf{O}}_{R_t}}{\tilde{\mathfrak{D}}} = \underset{\mathsf{\mathbf{O}}_{R_{t-1}}|x_{t-1}d_{t-2}}{\mathfrak{F}^{-1}} \times \underset{d_{t-1}|x_{t-1}d_{t-2}}{\mathfrak{D}} \times \underset{\mathsf{\mathbf{O}}_{R_t}|x_{t-1}d_{t-1}}{\mathfrak{F}}$$

together with the homogeneity property of HSMM.
Therefore, we can conclude that the batch form of the tensor takes the form:

$$\tilde{\mathfrak{D}} = \left(\sum_t \underset{\mathsf{\mathbf{O}}_{L_{t-1}}\mathsf{\mathbf{O}}_{R_{t-1}}}{\mathfrak{M}}\right)^{-1} \times_{\mathsf{\mathbf{O}}_L} \left(\sum_t \underset{\mathsf{\mathbf{O}}_{L_{t-1}}\mathsf{\mathbf{O}}_{R_t}}{\mathfrak{M}}\right). \qquad \text{(C.8)}$$

Similar derivations can be carried out to obtain the rest of the tensors in the batch form:

$$\tilde{\mathfrak{X}} = \left(\sum_t \underset{\mathsf{\mathbf{O}}_{L_t}\mathsf{\mathbf{O}}_{R_t}}{\mathfrak{M}}\right)^{-1} \times_{\mathsf{\mathbf{O}}_L} \left(\sum_t \underset{\mathsf{\mathbf{O}}_{L_t}\mathsf{\mathbf{O}}_{R_t}o_t}{\mathfrak{M}}\right) \qquad \text{(C.9)}$$

$$\tilde{\mathfrak{O}} = \left(\sum_t \underset{o_to_{t+1}}{\mathfrak{M}}\right)^{-1} \times_o \left(\sum_t \underset{o_to_{t+1}}{\mathfrak{M}}\right). \qquad \text{(C.10)}$$

where in the last expression the mode o corresponds to the mode $o_{t_{t+1}}$ after averaging of tensor $\underset{o_to_{t+1}}{\mathfrak{M}}$ for all t.

Analyzing (C.8), (C.9) and (C.10), we see that the computational complexity of the training phase of the algorithm is $\mathcal{O}\left((n_o^{2\ell} + \ell N)T\right)$, mainly determined by the tensor additions and the estimation of the sample moments $\mathfrak{M}$. The number of inverses and multiplications is now fixed and independent of sequence length

T. The computational complexity of the testing phase is $\mathcal{O}(n_o^{3\ell}T)$, which is the same as for Algorithm 1. Thus, the overall cost of Algorithm 2 is $\mathcal{O}\left((n_o^{3\ell}+\ell N)T\right)$.

Note that although not proved, we observed in practice that such a batch tensor computation significantly improves the accuracy of the resulting spectral algorithm. In part, this is due to the fact that we now use more data to estimate the tensors as compared to the original form (B.1). The estimates obtained in this form have lower variance, which in turn ensures that the inverses we compute in (C.8), (C.9) and (C.10) are more stable and accurate.

D Proof of Theorem 5.1

In this Section we prove the main result of the main paper, i.e., Theorem 5.1., stated below for references:

Theorem D.1 *Let the number of observations be* $|\mathbf{O}_{R_{t-1}}| = \ell$ *and define the set of indices* $\mathcal{S} = \left\{\max\left[t,\ t+(n_d-1)-(n_x^i-1)\right]\ \mid i=0,\ldots,\ell-1\right\}$, *such that* $\mathbf{O}_{R_{t-1}} = \{o_k | k \in \mathcal{S}\}$ *then the rank of tensor* $\underset{\mathbf{O}_{R_{x_{t-1}}}|x_{t-1}d_{t-2}}{\mathcal{F}}$ *is* $\min[n_x^\ell,\ n_x n_d]$.

In the following, we will study the rank structure of tensor $\underset{\mathbf{O}_{R_{t+1}}|x_t d_t}{\mathcal{F}}$ instead. This specific choice was only done to ensure the compactness in our notations, however the HSMM homogeneity property enables us to transfer this result for tensors for any t.

Also note that

$$\underset{\mathbf{O}_{R_{t+1}}|x_t d_t}{\mathcal{F}} = \underset{\mathbf{O}_{R_{t-1}}|x_{t-2}d_{t-2}}{\mathcal{F}} = \underset{\mathbf{O}_{R_{t-1}}|x_{t-1}d_{t-2}}{\mathcal{F}} \times_{x_{t-1}d_{t-2}} \underset{x_{t-1}d_{t-2}|x_{t-2}d_{t-2}}{\mathcal{X}}$$

where the first equality is due to the homogeneity property of the model and in the second equality we embedded the HSMM transition matrix into tensor $\underset{x_{t-1}d_{t-2}|x_{t-2}d_{t-2}}{\mathcal{X}}$ with mode d_{t-2} duplicated. It can be shown that the matricized tensor $\underset{x_{t-1}d_{t-2}|x_{t-1}d_{t-2}}{\mathbf{X}} \in \mathbb{R}^{n_x n_d \times n_x n_d}$ has rank $n_x n_d$. Therefore, the rank structure of $\underset{\mathbf{O}_{R_{t+1}}|x_t d_t}{\mathcal{F}}$ will determine the rank structure of $\underset{\mathbf{O}_{R_{t-1}}|x_{t-1}d_{t-2}}{\mathcal{F}}$, so they are equivalent in this case.

For references, we also repeat below the assumptions we made in the main paper about HSMM parameters:

Assumptions D.2
1. *$\mathcal{X}$ is full rank and has non-zero probability of visiting any state from any other state.*
2. *D has a non-zero probability of any duration in any state.*
3. *O is full column rank and, as a consequence, $n_x \leq n_o$.*

D.1 Rank Structure of Tensor $\underset{\mathbf{O}_{R_{t+1}}|x_t d_t}{\mathcal{F}}$

Define by $\mathbf{X}_{R_{t+1}} = \{x_{t+2}, x_{t+3}, \ldots\}$, the sequence of hidden states corresponding to $\mathbf{O}_{R_{t+1}} = \{o_{t+2}, o_{t+3}, \ldots\}$. Then using conditional independence property of HSMM in Figure 3, namely, that the variables $\mathbf{O}_{R_{t+1}}$ and $x_t d_t$ are independent given $\mathbf{X}_{R_{t+1}}$, we can write:

$$\underset{\mathbf{O}_{R_{t+1}}|x_t d_t}{\mathcal{F}} = \underset{\mathbf{O}_{R_{t+1}}|\mathbf{X}_{R_{t+1}}}{\mathfrak{Q}} \times_{\mathbf{X}_{R_{t+1}}} \underset{\mathbf{X}_{R_{t+1}}|x_t d_t}{\mathfrak{T}} \tag{D.1}$$

for some tensors $\mathfrak{Q}$ and $\mathfrak{T}$, representing the appropriate probability distributions.

Denoting $\ell = |\mathbf{O}_{R_{t+1}}| = |\mathbf{X}_{R_{t+1}}|$, it can be verified, that the matrisized form of $\mathfrak{Q}$ in (D.1) can be written as $\mathbf{Q} = \otimes_\ell O \in \mathbb{R}^{n_o^\ell \times n_x^\ell}$, a Kronecker product of the observation matrix O with itself ℓ times. According to statement 3 in Assumptions D.2, $rank(O) = n_x$ and $n_x \leq n_o$, and using the rank property of the Kronecker product, we infer that $rank(\mathbf{Q}) = n_x^\ell$.

Combining the above conclusion with the fact that the matrisized form of the other two tensors in (D.1) is $\mathbf{F} \in \mathbb{R}^{n_o^\ell \times n_x n_d}$ and $\mathbf{T} \in \mathbb{R}^{n_x^\ell \times n_x n_d}$, to ensure invertibility of $\mathcal{F}$, we need to select a set of variables $\mathbf{X}_{R_{t+1}}$

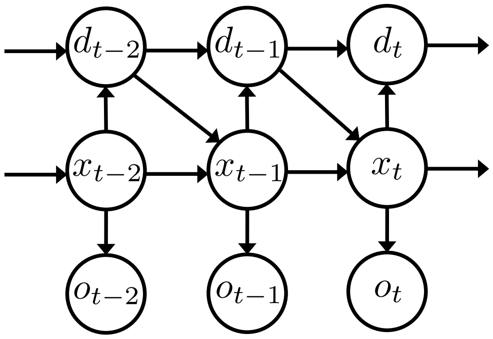

Figure 3: Hidden semi-Markov Model (HSMM).

so that $rank\Big(\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}\Big) = n_x n_d$ with the condition that $n_x^\ell \geq n_x n_d$. Thus, the problem of the analysis of the rank structure of tensor $\underset{\mathbf{o}_{R_{t+1}}|x_t d_t}{\boldsymbol{\mathcal{F}}}$ was translated to the problem of rank structure of matrix $\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$. In what follows, we assume that $\mathbf{x}_{R_{t+1}} = \{x_{t+2}, \ldots, x_{t+\ell}\}$ are sequential and so we would be interested in determining ℓ which makes $rank\Big(\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}\Big) = n_x n_d$. Later, the sequential assumption will be removed and we show how to select such variables in a more efficient, non-sequential way.

D.1.1 Computation of Factor T

In order to study the rank structure of $\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$ we will have to understand the mechanism how this matrix is constructed and how the rank changes as the size of $\mathbf{x}_{R_{t+1}}$ increases. We start by considering the following conditional independence relationships from the model in Figure 3:

$$p(x_{t+3}, x_{t+2}|x_{t+1}, d_{t+1}) = \sum_{d_{t+2}} p(x_{t+3}|x_{t+2}, d_{t+2})\, \underbrace{p(d_{t+2}|x_{t+2}, d_{t+1}) p(x_{t+2}|x_{t+1}, d_{t+1})} \tag{D.2}$$

$$p(x_{t+3}, x_{t+2}, x_{t+1}|x_t, d_t) = \sum_{d_{t+1}} p(x_{t+3}, x_{t+2}|x_{t+1}, d_{t+1})\, \underbrace{p(d_{t+1}|x_{t+1}, d_t) p(x_{t+1}|x_t, d_t)}. \tag{D.3}$$

Using the model's homogeneity property, we see that the quantity underlined in (D.2) is the same as the one in (D.3). Moreover, equation (D.2) can then be thought of as transforming $p(x_{t+1}|x_t, d_t)$ into $p(x_{t+2}, x_{t+1}|x_t, d_t)$, while the expression in (D.3) is, in effect, transforms $p(x_{t+2}, x_{t+1}|x_t, d_t)$ into $p(x_{t+3}, x_{t+2}, x_{t+1}|x_t, d_t)$. Thus (D.2) and (D.3) encode the following chain of transformations:

$$p(x_{t+1}|x_t, d_t) \to p(x_{t+2}, x_{t+1}|x_t, d_t) \to p(x_{t+3}, x_{t+2}, x_{t+1}|x_t, d_t).$$

Based on the above considerations, we can rewrite (D.2) and (D.3) in the tensor form as follows:

$$\underset{x_{t+3}, x_{t+2}|x_{t+1}, d_{t+1}}{\boldsymbol{\mathcal{T}}} = \underset{x_{t+3}, x_{t+2}|x_{t+2}, d_{t+2}}{\boldsymbol{\mathcal{T}}} \times_{x_{t+2} d_{t+2}} \underset{x_{t+2}, d_{t+2}|x_{t+1} d_{t+1}}{\boldsymbol{\mathcal{V}}} \tag{D.4}$$

$$\underset{x_{t+3}, x_{t+2}, x_{t+1}|x_t, d_t}{\boldsymbol{\mathcal{T}}} = \underset{x_{t+3}, x_{t+2}, x_{t+1}|x_{t+1}, d_{t+1}}{\boldsymbol{\mathcal{T}}} \times_{x_{t+1} d_{d+1}} \underset{x_{t+1}, d_{t+1}|x_t d_t}{\boldsymbol{\mathcal{V}}}, \tag{D.5}$$

where $\underset{x_{t+2}, d_{t+2}|x_{t+1}, d_{t+1}}{\boldsymbol{\mathcal{V}}} = \underset{x_{t+1}, d_{t+1}|x_t, d_t}{\boldsymbol{\mathcal{V}}} = \underset{x_{t+1}, d_{t+1}|x_{t+1}, d_t}{\boldsymbol{\mathcal{D}}} \times_{x_{t+1} d_t} \underset{x_{t+1}, d_t|x_t, d_t}{\boldsymbol{\mathcal{X}}}$. The homogeneity property allows us to rewrite the above as

$$\underset{x_{t+2}, x_{t+1}|x_t, d_t}{\boldsymbol{\mathcal{T}}} = \underset{x_{t+1}, x_t|x_t, d_t}{\boldsymbol{\mathcal{T}}} \times \boldsymbol{\mathcal{V}} \tag{D.6}$$

$$\underset{x_{t+3}, x_{t+2}, x_{t+1}, x_{t+1}|x_t, d_t}{\boldsymbol{\mathcal{T}}} = \underset{x_{t+2}, x_{t+1}|x_t, d_t}{\boldsymbol{\mathcal{T}}} \times \boldsymbol{\mathcal{V}}. \tag{D.7}$$

Our next step is to represent the above tensor equations in the matrix form. First, consider tensor $\boldsymbol{\mathcal{V}}$, its matricized form can be written as:

$$\mathbf{V} = \underset{x_{t+1},d_{t+1}|x_{t+1},d_t}{\mathbf{D}} \quad \underset{x_{t+1},d_t|x_t,d_t}{\mathbf{X}} \tag{D.8}$$

where $\underset{x_{t+1},d_{t+1}|x_{t+1},d_t}{\mathbf{D}} \in \mathbb{R}^{n_x n_d \times n_x n_d}$ and $\underset{x_{t+1},d_t|x_t,d_t}{\mathbf{X}} \in \mathbb{R}^{n_x n_d \times n_x n_d}$. Next, consider the equations (D.6) and (D.7), its matrix version is of the form:

$$\underset{x_{t+2},x_{t+1}|x_t,d_t}{\mathbf{T}} = \underset{x_{t+1},x_t|x_t,d_t}{\mathbf{T}} \quad \mathbf{V} \tag{D.9}$$

$$\underset{x_{t+3},x_{t+2},x_{t+1}|x_t,d_t}{\mathbf{T}} = \underset{x_{t+2},x_{t+1},x_t|x_t,d_t}{\mathbf{T}} \quad \mathbf{V}, \tag{D.10}$$

here $\underset{x_{t+1},x_t|x_t,d_t}{\mathbf{T}} \in \mathbb{R}^{n_x^2 \times n_x n_d}$, $\underset{x_{t+2},x_{t+1}|x_t,d_t}{\mathbf{T}} \in \mathbb{R}^{n_x^2 \times n_x n_d}$, and similarly $\underset{x_{t+2},x_{t+1},x_t|x_t,d_t}{\mathbf{T}} \in \mathbb{R}^{n_x^3 \times n_x n_d}$, and matrix $\underset{x_{t+3},x_{t+2},x_t|x_t,d_t}{\mathbf{T}} \in \mathbb{R}^{n_x^3 \times n_x n_d}$.

Summarizing the above derivations, we can describe the following algorithmic approach for analyzing $\underset{\mathsf{X}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$ as $\mathsf{X}_{R_{t+1}}$ increases. We begin with $\underset{x_{t+1}|x_t,d_t}{\mathbf{T}} = [\mathcal{X}\ \mathbf{I}\ \cdots\ \mathbf{I}] \in \mathbb{R}^{n_x \times n_x n_d}$, where the first block $\mathcal{X} \in \mathbb{R}^{n_x \times n_x}$ corresponds to $d_t = 1$, and the subsequent $(n_d - 1)$ blocks of $\mathbf{I} \in \mathbb{R}^{n_x \times n_x}$ correspond to $d_t > 1$ for which $x_{t+1} = x_t$. We then use (D.9) to get $\underset{x_{t+2},x_{t+1}|x_t,d_t}{\mathbf{T}}$. However, notice that in (D.9) the matrix $\underset{x_{t+1},x_t|x_t,d_t}{\mathbf{T}}$ has a duplicated mode x_t, therefore, we need to restructure $\underset{x_{t+1}|x_t,d_t}{\mathbf{T}}$, which can be accomplished with:

$$\underset{x_{t+1},x_t|x_t,d_t}{\mathbf{T}'} = \underset{x_{t+1}|x_t,d_t}{\mathbf{T}} \odot \mathbf{E},$$

where $\mathbf{E} = [\mathbf{I}\ \cdots\ \mathbf{I}] \in \mathbb{R}^{n_x \times n_x n_d}$ and $\odot$ denotes a Khatri-Rao product (row-wise Kronecker product)[1]. Then, we use (D.10) to transform $\underset{x_{t+2},x_{t+1}|x_t,d_t}{\mathbf{T}}$ into $\underset{x_{t+3},x_{t+2},x_{t+1}|x_t,d_t}{\mathbf{T}}$ where, again a preliminary step is to restructure the matrix as follows:

$$\underset{x_{t+2},x_{t+1},x_t|x_t,d_t}{\mathbf{T}'} = \underset{x_{t+2},x_{t+1}|x_t,d_t}{\mathbf{T}} \odot \mathbf{E}.$$

Algorithm 3 summarizes the above constructions for a general case.
In the next section, we will provide analysis of the Algorithm 3 and specifically study the rank structure of matrix $\mathbf{T}$. To understand the analysis, it is important to know how the structure of matrix $\mathbf{T}$ evolves across iterations. For this, we present in Figure 4 a schematic description of a few steps of the algorithm.

D.1.2 Analysis of Algorithm 3

In this Section our goal is to analyze the Algorithm 3 and study how the rank of matrix $\underset{\mathsf{X}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$ evolves across iterations. First, we state the main result of this analysis:

Theorem D.3 *The rank of the output matrix* $\underset{\mathsf{X}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$ *in Algorithm 3 is* $\min(\ell n_x, n_x n_d)$.

Applying now Theorem D.3 to equation (D.1) in matrix form

$$\underset{\mathsf{O}_{R_{t+1}}|x_t d_t}{\mathbf{F}} = \underset{\mathsf{O}_{R_{t+1}}|\mathsf{X}_{R_{t+1}}}{\mathbf{Q}} \times \underset{\mathsf{X}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$$

where $rank(\mathbf{Q}) = n_x^\ell$ we can now conclude the following result:

[1] Let $\mathbf{P} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ and $\mathbf{Q} \in \mathbb{R}^{k \times n}$ then $\mathbf{P} \odot \mathbf{Q} = \begin{bmatrix} \mathbf{p}_1 \otimes \mathbf{Q} \\ \mathbf{p}_2 \otimes \mathbf{Q} \\ \vdots \\ \mathbf{p}_n \otimes \mathbf{Q} \end{bmatrix} \in \mathbb{R}^{mk \times n}$, where $\otimes$ is a Kronecker product.

Algorithm 3 Computation of $\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$

```
**Input:** $p(d_t|x_t, d_{t-1})$ and $p(x_t|x_{t-1}, d_{t-1})$ - duration and transition distributions, $\ell$ - number of steps
**Initialization:**
    $p(x_{t+1}|x_t, d_t) \rightarrow \underset{x_{t+1}|x_t,d_t}{\mathbf{T}}$
    $p(d_{t+1}|x_{t+1}, d_t) \rightarrow \underset{x_{t+1},d_{t+1}|x_{t+1},d_t}{\mathbf{D}}$
    $p(x_{t+1}|x_t, d_t) \rightarrow \underset{x_{t+1},d_t|x_t,d_t}{\mathbf{X}}$
    $\mathbf{V} = \underset{x_{t+1},d_{t+1}|x_{t+1},d_t}{\mathbf{D}} \underset{x_{t+1},d_t|x_t,d_t}{\mathbf{X}}, \quad \mathbf{E} = [\mathbf{I} \cdots \mathbf{I}]$

**for** $i = 1$ **to** $\ell - 1$ **do**
    $\underset{x_{t+i},\,\dots\,,x_{t+1},x_t|x_t,d_t}{\mathbf{T}'} = \underset{x_{t+i},\,\dots\,,x_{t+1}|x_t,d_t}{\mathbf{T}} \odot \mathbf{E}$    (D.11)
    $\underset{x_{t+i+1},\,\dots\,,x_{t+2},x_{t+1}|x_t,d_t}{\mathbf{T}} = \underset{x_{t+i},\,\dots\,,x_{t+1},x_t|x_t,d_t}{\mathbf{T}'} \mathbf{V}$    (D.12)
**end for**
```

Corollary D.4 *To achieve the full column rank for* $\underset{\mathbf{o}_{R_{t+1}}|x_t d_t}{\mathbf{F}} \in \mathbb{R}^{n_o^\ell \times n_x n_d}$*, i.e. to ensure that the rank of tensor* $\underset{\mathbf{o}_{R_{t+1}}|x_t d_t}{\mathfrak{F}}$ *is* $n_x n_d$*, the number of observations* ℓ *in* $\mathbf{O}_{R_{t+1}} = \{o_{t+2}, o_{t+3}, \dots, o_{t+\ell}\}$ *must be equal to the maximum state persistence i.e.,* $\ell = n_d$*.*

Before we prove Theorem D.3, we will establish certain auxiliary results.

Lemma D.5 *Let* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *be a matrix with no all-zero columns then* $rank\,(\mathbf{I} \odot \mathbf{A}) = rank\,(\mathbf{A} \odot \mathbf{I}) = n$*, where* $\odot$ *denotes Khatri-Rao product and* $\mathbf{I} \in \mathbb{R}^{n \times n}$*.*

Proof Let $\mathbf{K} = (\mathbf{I} \odot \mathbf{A}) \in \mathbb{R}^{mn \times n}$. By definition of Khatri-Rao product, $\mathbf{K}(:, j) = \mathbf{e}_j \otimes \mathbf{A}(:, j)$, for $j = 1, \dots, n$, which consists of zeros, except for rows $(j-1)m + 1, \dots, (j-1)m + m$, containing the column $\mathbf{A}(:, j)$. Here $\otimes$ denotes Kronecker product and $\mathbf{e}_j$ is everywehre zero except for position j which is 1. As long as there is no all-zero columns in $\mathbf{A}$, each column of $\mathbf{K}$ is independent of each other and therefore the rank is n. Moreover, since the matrix $\mathbf{A} \odot \mathbf{I}$ is a row-permuted version of $\mathbf{A} \odot \mathbf{I}$, their ranks are the same. ▌

Lemma D.6 *Define a block-row matrix* $\mathbf{M} = [\mathbf{A}_1\ \mathbf{A}_2\ \cdots\ \mathbf{A}_k] \in \mathbb{R}^{m \times kn}$*, where each* $\mathbf{A}_i \in \mathbb{R}^{m \times n}$*. Define by* $r_j,\ j = 1, \dots, n$ *the rank of matrix* $[\mathbf{A}_1(:, j)\ \cdots\ \mathbf{A}_k(:, j)]$ *composed of* j*th columns of* $\mathbf{A}$*'s, and let* $\mathbf{E} = [\mathbf{I}\ \mathbf{I}\ \cdots\ \mathbf{I}] \in \mathbb{R}^{n \times kn}$*, where* $\mathbf{I} \in \mathbb{R}^{n \times n}$*. Then the rank of matrix* $\mathbf{W} = \mathbf{M} \odot \mathbf{E} \in \mathbb{R}^{mn \times kn}$*, obtained using a Khatri-Rao product, is* $\min(mn, \sum_j r_j)$*.*

Proof First note that $\mathbf{M} \odot \mathbf{E}$ and $\mathbf{E} \odot \mathbf{M}$ are row permuted version of each other, so they have the same rank. Therefore, consider $\mathbf{W}' = \mathbf{E} \odot \mathbf{M} = [\mathbf{I} \odot \mathbf{A}_1 \cdots \mathbf{I} \odot \mathbf{A}_k]$. Also, note that $\mathbf{e}_j \otimes [\mathbf{A}_1(:, j) \cdots\ \mathbf{A}_k(:, j)]$, $j = 1, \dots, n$ is a matrix which consists of zeros except for rows $(j-1)m + 1, \dots, (j-1)m + m$ where it contains the columns $[\mathbf{A}_1(:, j)\ \cdots\ \mathbf{A}_k(:, j)]$. The rank of these columns is r_j and all other columns in $\mathbf{W}$ are independent of them due to the structure of the Khatri-Rao product. Therefore, each set of such columns adds r_j to the total rank. Since the overall rank of $\mathbf{W}$ cannot exceed either the number of rows or columns, we conclude that $rank(\mathbf{W}) = \min(mn, \sum_j r_j)$. ▌

Lemma D.7 *Let* $V = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ *be a set of independent vectors. Define* $\mathbf{u} = \sum_{i=1}^n c_i \mathbf{v}_i$*, where coefficients* $c_i \neq 0, i = 1, \dots, n$*. Define* U *to be a strict subset of* V*, i.e.,* $U \subset V$*, then a set of vectors* $\mathbf{u} \cup U$ *is independent.*

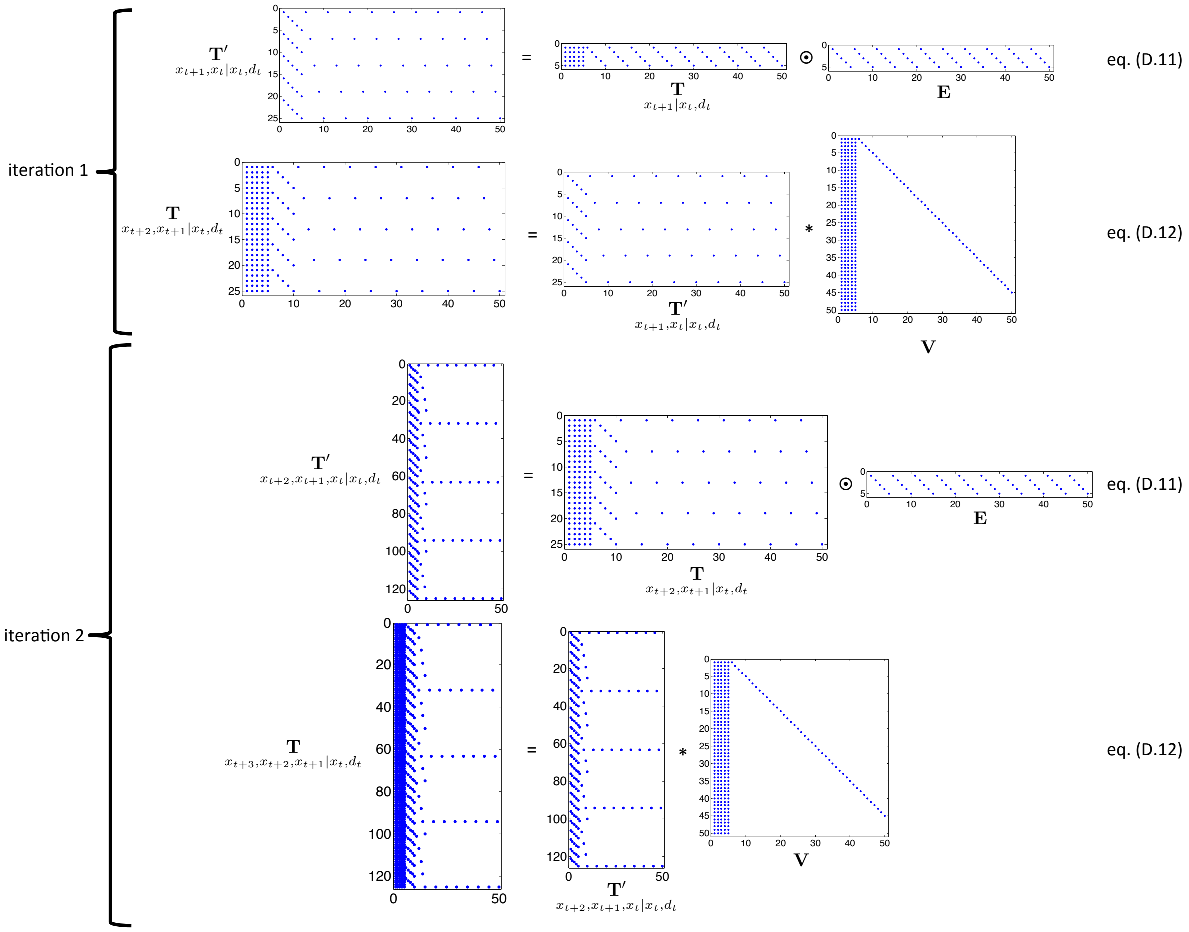
iteration 1
$\mathbf{T}'_{x_{t+1},x_t|x_t,d_t}$
=
$\mathbf{T}_{x_{t+1}|x_t,d_t}$
⊙
$\mathbf{E}$
eq. (D.11)
$\mathbf{T}_{x_{t+2},x_{t+1}|x_t,d_t}$
=
$\mathbf{T}'_{x_{t+1},x_t|x_t,d_t}$
*
$\mathbf{V}$
eq. (D.12)
iteration 2
$\mathbf{T}'_{x_{t+2},x_{t+1},x_t|x_t,d_t}$
=
$\mathbf{T}_{x_{t+2},x_{t+1}|x_t,d_t}$
⊙
$\mathbf{E}$
eq. (D.11)
$\mathbf{T}_{x_{t+3},x_{t+2},x_{t+1}|x_t,d_t}$
=
$\mathbf{T}'_{x_{t+2},x_{t+1},x_t|x_t,d_t}$
*
$\mathbf{V}$
eq. (D.12)

Figure 4: Schematic representation of Algorithm 3. This example illustrates the HSMM with $n_x = 5$ and $n_d = 10$. The non-zero matrix elements are displayed as dots.

Proof Define $\{1, \ldots, n\} = \alpha \cup \bar{\alpha}$, where α denotes a subset of indices for vectors corresponding to U. Then we can write $\mathbf{u} = \sum_{i:i\in\alpha} c_i \mathbf{v}_i + \sum_{j:j\in\bar{\alpha}} c_j \mathbf{v}_j$.

Assuming the opposite, i.e., $\mathbf{u} \cup U$ are dependent, we can write $k_0\mathbf{u} + \sum_{i:i\in\alpha} k_i \mathbf{v}_i = 0$ where $k_0 \neq 0$ and some of $k_i, i \in \alpha$ are also must be non-zero. Substituting the definition of $\mathbf{u}$ and rearranging the terms, we get:

$$k_0 \sum_{i:i\in\alpha} (c_i + k_i)\mathbf{v}_i + k_0 \sum_{j:j\in\bar{\alpha}} c_j \mathbf{v}_j = 0$$

Since $c_j \neq 0, j \in \bar{\alpha}$, the above equation claims the linear dependence of vectors in V, which is a contradiction of our assumption and so $\mathbf{u} \cup U$ are independent. ▮

We now ready to analyze Algorithm 3. It can be verified that (D.8) is of the form:

$$\mathbf{V} = \left[\begin{array}{c|ccc} \Psi & \mathbf{I} & & \\ & & \ddots & \\ & & & \mathbf{I} \\ \hline & \mathbf{0} & \cdots & \mathbf{0} \end{array}\right] \in \mathbb{R}^{n_x n_d \;\times\; n_x n_d} \qquad \text{where} \quad \Psi = \begin{bmatrix} \operatorname{diag}\left[D(1,:)\right]\mathcal{X} \\ \operatorname{diag}\left[D(2,:)\right]\mathcal{X} \\ \vdots \\ \operatorname{diag}\left[D(n_d,:)\right]\mathcal{X} \end{bmatrix} \in \mathbb{R}^{n_x n_d \;\times\; n_x} \tag{D.13}$$

where diag $[D(i,:)]$ is the diagonal matrix with ith row from D on the diagonal. Note that we can also write $\Psi = (D \odot \mathbf{I})\,\mathcal{X}$. Observe that the rank of $\mathbf{V}$ is $n_x n_d$ because the $n_x(n_d-1) \times n_x(n_d-1)$ block diagonal matrix delineated in (D.13) and the last $n_x \times n_x$ block matrix $\text{diag}\,[D(n_d,:)]\,\mathcal{X}$ in Ψ together comprising $n_x n_d$ independent columns of $\mathbf{V}$. Note that $\text{diag}\,[D(n_d,:)]\,\mathcal{X}$ has rank n_x because $\mathcal{X}$ is full rank and $D(n_d,:)$ is non-zero, which follows from statements 1 and 2 in Assumptions D.2). As a side note observe that the requirement to have $D(n_d,:)$ non-zero implies that there is a non-zero probability of maximum state persistence.

In analyzing the Algorithm 3, it would be useful to denote the matrices at iteration i in (D.11) and (D.12) as

$$\begin{aligned}
\underset{x_{t+i},\ \dots\ ,x_{t+1}|x_t,d_t}{\mathbf{T}} &= [\mathbf{A}_1^{(i)}\ \cdots\ \mathbf{A}_{n_d}^{(i)}] \\
\underset{x_{t+i},\ \dots\ ,x_{t+1},x_t|x_t,d_t}{\mathbf{T}'} &= [\mathbf{B}_1^{(i)}\ \cdots\ \mathbf{B}_{n_d}^{(i)}] \\
\underset{x_{t+i+1},\dots,x_{t+2},x_{t+1}|x_t,d_t}{\mathbf{T}} &= [\mathbf{C}_1^{(i)}\ \cdots\ \mathbf{C}_{n_d}^{(i)}].
\end{aligned}$$

Moreover, utilizing the structure of matrix $\mathbf{V}$ from (D.13), the operations involved in step (D.12) are as follows:

$$\Big[\mathbf{C}_1^{(i)}\ \ \mathbf{C}_2^{(i)}\ \ \mathbf{C}_3^{(i)}\ \ \cdots\ \mathbf{C}_{n_d}^{(i)}\Big] = \Big[[\mathbf{B}_1^{(i)}\ \cdots\ \mathbf{B}_{n_d}^{(i)}]\Psi\ \ \ \mathbf{B}_1^{(i)}\ \ \mathbf{B}_2^{(i)}\ \ \cdots\ \ \mathbf{B}_{n_d-1}^{(i)}\Big]. \tag{D.14}$$

With the above information we can now present the proof of Theorem D.3:

Proof At the start of the algorithm, we have $\underset{x_{t+1}|x_t,d_t}{\mathbf{T}} = [\mathcal{X}\ \mathbf{I}\ \cdots\ \mathbf{I}] = [\mathbf{A}_1^{(1)} \cdots \mathbf{A}_{n_d}^{(1)}]$, which has rank n_x. The rank of matrix $\Big[\mathbf{A}_1^{(1)}(:,l) \cdots \mathbf{A}_{n_d}^{(1)}(:,l)\Big]$ for $l = 1,\dots,n_x$ is $r_l = 2$ since among all the columns only two of them are independent. Therefore, according to Lemma D.6, the result of operations in (D.11), has rank $\sum_l r_l = 2n_x$. Moreover, we note that since $[\mathbf{B}_1^{(1)}\ \mathbf{B}_2^{(1)}\ \cdots\ \mathbf{B}_{n_d}^{(1)}] = [\mathcal{X}\odot\mathbf{I}\ \ \mathbf{I}\odot\mathbf{I}\ \cdots\ \mathbf{I}\odot\mathbf{I}]$, it can be seen that its $2n_x$ independent vectors can be formed by the columns $[\mathbf{B}_1^{(1)}\ \mathbf{B}_2^{(1)}]$, so that the rank of $\Big[\mathbf{B}_1^{(1)}(:,l) \cdots \mathbf{B}_{n_d}^{(1)}(:,l)\Big]$ for $l = 1,\dots,n_x$ is 2.

Next, since the rank of $\mathbf{V}$ is $n_x n_d$, the operations in (D.12) produce matrix $[\mathbf{C}_1^{(1)}\ \mathbf{C}_2^{(1)}\ \cdots\ \mathbf{C}_{n_d}^{(1)}]$ with the rank still being $2n_x$. Moreover, the columns of $\mathbf{C}_1^{(1)}$ are linearly dependent on the rest of the columns, $[\mathbf{C}_2^{(1)}\ \cdots\ \mathbf{C}_{n_d}^{(1)}]$, due to (D.14). However, the rank of $\Big[\mathbf{C}_1^{(1)}(:,l) \cdots \mathbf{C}_{n_d}^{(1)}(:,l)\Big]$ is now $r_l = 3$ for $l = 1,\dots,n_x$. To understand this, note that

$$\begin{aligned}
[\mathbf{B}_1^{(1)}\ \ \mathbf{B}_2^{(1)}\ \cdots\ \mathbf{B}_{n_d}^{1}] &= [\mathcal{X}\odot\mathbf{I}\ \ \mathbf{I}\odot\mathbf{I}\ \cdots\ \mathbf{I}\odot\mathbf{I}] \\
[\mathbf{C}_1^{(1)}\ \ \mathbf{C}_2^{(1)}\ \ \mathbf{C}_3^{(1)}\ \cdots\ \mathbf{C}_{n_d}^{(1)}] &= [\mathbf{C}_1^{(1)}\ \ \mathcal{X}\odot\mathbf{I}\ \ \mathbf{I}\odot\mathbf{I}\ \cdots\ \mathbf{I}\odot\mathbf{I}],
\end{aligned}$$

where, according to (D.14), $\mathbf{C}_1^{(1)} = [\mathbf{B}_1^{(1)} \cdots \mathbf{B}_{n_d}^{(1)}]\Psi$. As we established before, the rank of $\Big[\mathbf{C}_2^{(1)}(:,l) \cdots \mathbf{C}_{n_d}^{(1)}(:,l)\Big] = \Big[\mathbf{B}_1^{(1)}(:,l) \cdots \mathbf{B}_{n_d-1}^{(1)}(:,l)\Big]$ is $r_l = 2$. Moreover, it can also be checked that $\mathbf{C}_1^{(1)}(:,l)$ is independent of $\Big[\mathbf{C}_2^{(1)}(:,l) \cdots \mathbf{C}_{n_d}^{(1)}(:,l)\Big]$ due to Lemma D.7. Clearly, then the cumulative rank of $\Big[\mathbf{C}_1^{(1)}(:,l) \cdots \mathbf{C}_{n_d}^{(1)}(:,l)\Big]$ is 3 for $l = 1,\dots,n_x$.

To generalize, if at the iteration i the rank of $\Big[\mathbf{A}_1^{(i)} \cdots \mathbf{A}_{n_d}^{(i)}\Big]$ is in_x while the rank of $\Big[\mathbf{A}_1^{(i)}(:,l) \cdots \mathbf{A}_{n_d}^{(i)}(:,l)\Big]$ is $(i+1)$, then the operations in step (D.11) produce $\Big[\mathbf{B}_1^{(i)} \cdots \mathbf{B}_{n_d}^{(i)}\Big]$ having rank $(i+1)n_x$ due to Lemma D.6. The step in (D.12) keeps the rank of $\Big[\mathbf{C}_1^{(i)} \cdots \mathbf{C}_{n_d}^{(i)}\Big]$ at $(i+1)n_x$ due to the full rank structure of $\mathbf{V}$. At the same time, this step increases the rank of $\Big[\mathbf{C}_1^{(i)}(:,l) \cdots \mathbf{C}_{n_d}^{(i)}(:,l)\Big]$ to $(i+2)$ due to Lemma D.7, i.e., independence of $\mathbf{C}_1^{(i)}(:,l)$ from $\Big[\mathbf{C}_2^{(i)}(:,l) \cdots \mathbf{C}_{n_d}^{(i)}(:,l)\Big]$ with the latter having the rank $(i+1)$.

Therefore, each iteration increases the rank of matrix $\mathbf{T}$ by n_x and so after $2 \le \ell \le n_d$ steps the rank of the resulting matrix $\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$ is ℓn_x.

```
**Algorithm 4** Efficient computation of $\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$

**Input:** $p(d_t|x_t, d_{t-1})$ and $p(x_t|x_{t-1}, d_{t-1})$ - duration and transition distributions, $\ell$ - number of steps
**Initialization:**

            $p(x_{t+1}|x_t, d_t) \rightarrow \underset{x_{t+1}|x_t,d_t}{\mathbf{T}}$
            $p(d_{t+1}|x_{t+1}, d_t) \rightarrow \underset{x_{t+1},d_{t+1}|x_{t+1},d_t}{\mathbf{D}}$
            $p(x_{t+1}|x_t, d_t) \rightarrow \underset{x_{t+1},d_t|x_t,d_t}{\mathbf{X}}$
            $\mathbf{V} = \underset{x_{t+1},d_{t+1}|x_{t+1},d_t}{\mathbf{D}} \; \underset{x_{t+1},d_t|x_t,d_t}{\mathbf{X}}, \quad \mathbf{E} = [\mathbf{I} \cdots \mathbf{I}]$

$c = 1$
**for** $i = 1$ **to** $\ell - 1$ **do**

            $\mathbf{T} = \mathbf{T} \;\; \mathbf{V}$                                        (D.15)

    **if** $i == (n_x)^c - 1$ **or** $i == \ell - 1$ **do**

            $\mathbf{T} = \mathbf{T} \odot \; \mathbf{E}$                                        (D.16)

        $c = c + 1$
    **end if**
**end for**
```

Note that if $\ell = 1$ then the Algorithm 3 is not executed and returns the trivial $\underset{x_{t+1}|x_t,d_t}{\mathbf{T}}$ with rank n_x. On the other hand, if $\ell > n_d$ then the rank of $\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$ is $n_x n_d$ since this is the number of columns in that matrix and so is the maximum achievable rank. ∎

D.2 Efficient Computation of Factor T

In Corollary D.4 we established that the required number of observations in $\mathbf{O}_{R_{t+1}} = \{o_{t+2}, o_{t+3}, \ldots, o_{t+\ell}\}$ is $\ell = n_d$. Therefore, the sizes of the estimated quantities $\tilde{\mathfrak{D}} \in \mathbb{R}^{n_o^{n_d} \times n_o^{n_d}}$ and $\tilde{\mathfrak{X}} \in \mathbb{R}^{n_o^{n_d} \times n_o^{n_d} \times n_o}$ in the Algorithm 2 will have exponential dependency on n_d. When maximum state persistence is large, the estimation of such quantity becomes impractical. Fortunately, we can modify Algorithm 3 to significantly reduce the number of observations. The idea is to apply the step (D.12) multiple times in-between the applications of step (D.11). Recall that in the previous construction we established that we needed $\ell = n_d$ consecutive observations, e.g., $\mathbf{O}_{R_{t+1}} = \{o_{t+2}, \ldots, o_{t+\ell}\}$. In contrast, in the proposed approach, every time we add an observation, say $o_{t+\tau}$, we skip certain number δ of time steps before adding another observation $o_{t+\tau+\delta}$, so that the observations are non-consecutive. As we illustrate next, the span of these non-consecutive observations is still n_d but the number of them is only logarithmic in n_d. The proposed approach still achieves the full rank structure of $\underset{\mathbf{O}_{R_{t+1}}|x_t d_t}{\mathbf{F}}$ but with smaller number of data points. The Algorithm 4, which is a simple modification of the Algorithm 3, summarizes the above procedure.
The following result establishes the rank structure of the matrix $\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$ in the output of the Algorithm 4.

Theorem D.8 *The rank of the output matrix* $\underset{\mathbf{x}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$ *in Algorithm 4 is* $\min(n_x^\ell, n_x n_d)$.

Note that based on the above theorem, Algorithm 4 increases the rank at every step exponentially rather than linearly. In order for $\underset{\mathsf{X}_{R_{t+1}}|x_t d_t}{\mathbf{T}}$ to achieve the rank $n_x n_d$ we will now require $\ell = \lceil 1 + \frac{\log n_d}{\log n_x} \rceil$ observations, since we need to ensure $n_x^\ell = n_x n_d$. Observe that the span of the selected observations is still n_d, while the number of the observations is only logarithmic in n_d. The following Corollary summarizes the above conclusions.

Corollary D.9 *To achieve the full column rank for* $\underset{\mathsf{O}_{R_{t+1}}|x_t d_t}{\mathbf{F}} \in \mathbb{R}^{n_o^\ell \times n_x n_d}$*, i.e. to ensure that the rank of tensor* $\underset{\mathsf{O}_{R_{t+1}}|x_t d_t}{\mathfrak{F}}$ *is* $n_x n_d$*, the number of observations* ℓ *in* $\mathsf{O}_{R_{t+1}}$ *must be equal to* $\ell = \lceil 1 + \frac{\log n_d}{\log n_x} \rceil$*, since we need to ensure* $n_x^\ell = n_x n_d$*.*

Before we prove Theorem D.8, it is instructive to visualize the progress of Algorithm 4. Figure 5 shows a schematic description of a few steps of the algorithm.
We are now ready to present the proof of Theorem D.8.

Proof For the proof, we refer back to Algorithm 3 and the proof of Theorem D.3. Recall, that at iteration $i = 1$, the result of step (D.11) is a matrix $[\mathbf{B}_1^{(1)} \cdots \mathbf{B}_{n_d}^{(1)}] \in \mathbb{R}^{n_x^2 \times n_x n_d}$, whose rank is $2n_x$, since $\left[\mathbf{A}_1^{(1)}(:,l) \cdots \mathbf{A}_{n_d}^{(1)}(:,l)\right] = [\mathcal{X}\ \mathbf{I} \cdots \mathbf{I}] \in \mathbb{R}^{n_x \times n_x n_d}$ for $l = 1, \ldots, n_x$ had two independent columns. Then, the transformations in step (D.12) produced $\left[\mathbf{C}_1^{(1)}(:,l) \cdots \mathbf{C}_{n_d}^{(1)}(:,l)\right]$ for $l = 1, \ldots, n_x$ with rank $3n_x$.

Note that if $n_x > 2$ then $\left[\mathbf{A}_1^{(1)}(:,l) \cdots \mathbf{A}_{n_d}^{(1)}(:,l)\right]$ potentially can have a rank up to n_x, while in Algorithm 3 we only have it equal to 2. It turns out that if we apply step (D.12) multiple times and use Lemma D.7, we can increase the rank of $\left[\mathbf{C}_1^{(1)}(:,l) \cdots \mathbf{C}_{n_d}^{(1)}(:,l)\right]$ for $l = 1, \ldots, n_x$ to n_x.

Specifically, consider the step (D.15). Then at iteration $i = 1$ we have $[\mathbf{A}_1^{(1)} \cdots \mathbf{A}_{n_d}^{(1)}] = [\mathbf{B}_1^{(1)} \cdots \mathbf{B}_{n_d}^{(1)}]$ and for $l = 1, \ldots, n_x$ the two independent columns are $\left[\mathbf{B}_1^{(1)}(:,l)\ \ \mathbf{B}_2^{(1)}(:,l)\right] = [\mathcal{X}(:,l)\ \ \mathbf{I}(:,l)]$. The result of step (D.15) gives us then three independent columns

$$\left[\mathbf{C}_1^{(1)}(:,l)\ \ \mathbf{C}_2^{(1)}(:,l)\ \ \mathbf{C}_3^{(1)}(:,l)\right] = \left[\mathbf{C}_1^{(1)}(:,l)\ \ \mathcal{X}(:,l)\ \ \mathbf{I}(:,l)\right]$$

where $\mathbf{C}_1^{(1)} = [\mathcal{X}\ \mathbf{I}\ \cdots\ \mathbf{I}]\Psi$. The independence follows from Lemma D.7. The repeated application of step (D.15) one more time gives four independent columns

$$\left[\mathbf{C}_1^{(2)}(:,l)\ \ \mathbf{C}_2^{(2)}(:,l)\ \ \mathbf{C}_3^{(2)}(:,l)\ \ \mathbf{C}_4^{(2)}(:,l)\right] = \left[\mathbf{C}_1^{(2)}(:,l)\ \ \mathbf{C}_1^{(1)}(:,l)\ \ \mathcal{X}(:,l)\ \ \mathbf{I}(:,l)\right]$$

where $\mathbf{C}_1^{(2)} = [\mathbf{C}_1^{(1)} \cdots \mathbf{C}_{n_d}^{(1)}]\Psi$. Observe that since the number of rows is n_x, we can increase the rank at most up to n_x. Therefore, if in the beginning we had *two* independent columns and we want to get n_x independent columns, we will need to apply the step (D.15) $n_x - 2$ times, so we will have $[\mathbf{C}_1^{(n_x-2)}(:,l)\ \cdots\ \mathbf{C}_{n_d}^{(n_x-2)}(:,l)]$ with rank n_x.

If we now apply step (D.16) it will give us $[\mathbf{A}_1^{(1)}\ \cdots\ \mathbf{A}_{n_d}^{(1)}] \in \mathbb{R}^{n_x^2 \times n_x n_d}$ with rank n_x^2 due to Lemma D.6. Continuing in this manner, we can again repeatedly apply the step (D.15) to create a matrix with a rank at most n_x^2, since there are n_x^2 rows and assuming that $n_x n_d \geq n_x^2$. The number of times we need to apply (D.15) is now $n_x^2 - n_x$ since we need to go from n_x to n_x^2 independent columns.

In general, the step (D.15) needs to be applied $n_x^c - n_x^{c-1}$, in order to obtain n_x^c independent columns. The application of step (D.16) then creates $\mathbf{T}$ with rank n_x^{c+1}. Note, that since $\mathbf{T}$ has $n_x n_d$ columns, the maximum achievable rank is $n_x n_d$. ▮

Observe that the above proof also provided the method for selecting the non-sequential observations $\mathsf{X}_{R_{t+1}}$. Specifically, since the set of observations $\mathsf{X}_{R_{t+1}} = \{o_{t+2}, \ldots\}$ must start from observation o_{t+2} and $|\mathsf{X}_{R_{t+1}}| = \ell$, we denote $s = t + 2$. Then, ith added observation is $o_{s+(n_d-1)-(n_x^i-1)}$ for $i = 0, \ldots, \ell - 2$ and the ℓth observation is $o_s = o_{t+2}$. For tensor $\underset{\mathsf{O}_{R_{t+1}}|x_t d_t}{\mathfrak{F}}$ to achieve rank $n_x n_d$ we need to add $\ell = \lceil 1 + \frac{\log n_d}{\log n_x} \rceil$ observations.

Theorem D.8 together with Corollary D.9 now proves the main result of the paper (Theorem 5.1).

E Initial and Final Parts of HSMM

In this Section we present the derivations for the initial and final steps of HSMM, which were omitted from the main text. Specifically, this amounts to computing the factor $\mathcal{X}$ for two parts of the model, corresponding to $\mathbb{X}_{root}$ and $\mathbb{X}_T$ in Figs. 6 and 7. The derivations for all other parts of HSMM were presented in the main text and this supplement.

To begin, recall the expression for the joint likelihood of the observed sequence:

$$\underset{o_1,\ldots,o_T}{\mathcal{P}} = \prod_t \underset{d_{t-1}|x_{t-1}d_{t-2}}{\mathcal{D}} \times_{x_{t-1}d_{t-1}} \left(\underset{x_t|x_{t-1}d_{t-1}}{\mathcal{X}} \times_{x_t} \underset{o_t|x_t}{\mathcal{O}} \right)$$

and rewrite the above expression by keeping only the initial and final factors:

$$\begin{aligned} \underset{o_1,\ldots,o_T}{\mathcal{P}} = & \left(\underset{o_1|x_1}{\mathcal{O}} \times_{x_1} \left(\underset{x_2x_2|x_1d_1}{\mathcal{X}} \times_{x_2} \underset{o_2|x_2}{\mathcal{O}} \right) \right) \times_{x_2d_1} \underset{d_2|x_2x_2d_1}{\mathcal{D}} \times \cdots \\ & \cdots \times \underset{d_{T-1}|x_{T-1}x_{T-1}d_{T-2}}{\mathcal{D}} \times_{x_{T-1}d_{T-1}} \left(\underset{x_T|x_{T-1}d_{T-1}}{\mathcal{X}} \times_{x_T} \underset{o_T|x_T}{\mathcal{O}} \right) \end{aligned} \tag{E.1}$$

Introduce the identity tensors into (E.1), regroup the terms and extract the factors $\mathcal{X}$:

$$\underset{\omega_{x_1}\omega_{x_2}\omega_{x_2d_1}}{\tilde{\mathcal{X}}} = \underset{\omega_{x_1}|x_1}{\mathcal{F}} \times_{x_1} \left(\underset{x_2x_2|x_1d_1}{\mathcal{X}} \times_{x_2} \underset{\omega_{x_2}|x_2}{\mathcal{F}} \right) \times_{x_2d_1} \underset{\omega_{x_2d_1}|x_2d_1}{\mathcal{F}} \tag{E.2}$$

$$\underset{\omega_{x_{T-1}d_{T-1}}\omega_{x_T}}{\tilde{\mathcal{X}}} = \underset{\omega_{x_{T-1}d_{T-1}}|x_{T-1}d_{T-1}}{\mathcal{F}^{-1}} \times_{x_{T-1}d_{T-1}} \left(\underset{x_T|x_{T-1}d_{T-1}}{\mathcal{X}} \times_{x_T} \underset{\omega_{x_T}|x_T}{\mathcal{F}} \right) \tag{E.3}$$

Defining the observable sets $\omega_{x_1} = o_1$, $\omega_{x_2} = o_2$ and $\omega_{x_2d_1} = \mathsf{\mathbf{O}}_{R_3}$ we can rewrite (E.2) as follows:

$$\underset{o_1o_2\mathsf{\mathbf{O}}_{R_3}}{\tilde{\mathcal{X}}} = \underset{o_1|x_1}{\mathcal{F}} \times_{x_1} \left(\underset{x_2x_2|x_1d_1}{\mathcal{X}} \times_{x_2} \underset{o_2|x_2}{\mathcal{F}} \right) \times_{x_2d_1} \underset{\mathsf{\mathbf{O}}_{R_3}|x_2d_1}{\mathcal{F}} \tag{E.4}$$

Note that since all the factors participating in (E.4) are valid probability distributions, the resulting factor, i.e., $\underset{o_1o_2\mathsf{\mathbf{O}}_{R_3}}{\tilde{\mathcal{X}}}$ is also a valid probability distribution, so it can be estimated directly from data. This is in contrast to the derivations we made for other parts of the model, where we had to perform additional transformations such as, for example in (B.2), in order to bring to the form, which could be estimated from the data samples.

In order to estimate (E.3), we compare it to the similar factor we considered in the main paper:

$$\underset{\omega_{x_{t-1}d_{t-1}}\omega_{x_t}\omega_{x_td_{t-1}}}{\tilde{\mathcal{X}}} = \underset{\omega_{x_{t-1}d_{t-1}}|x_{t-1}d_{t-1}}{\mathcal{F}^{-1}} \times_{x_{t-1}d_{t-1}} \left(\underset{x_tx_t|x_{t-1}x_{t-1}d_{t-1}}{\mathcal{X}} \times_{x_t} \underset{\omega_{x_t}|x_t}{\mathcal{F}} \right) \times_{x_td_{t-1}} \underset{\omega_{x_td_{t-1}}|x_td_{t-1}}{\mathcal{F}} \tag{E.5}$$

Observe that the last factor $\underset{\omega_{x_td_{t-1}}|x_td_{t-1}}{\mathcal{F}}$ in (E.5) is a conditional probability distribution, which has the following marginalization property

$$\underset{\omega_{x_td_{t-1}}|x_td_{t-1}}{\mathcal{F}} \times_{\omega_{x_td_{t-1}}} \underset{\omega_{x_td_{t-1}}}{\mathbf{1}} = \underset{x_td_{t-1}}{\mathbf{1}} \tag{E.6}$$

where $\mathbf{1}$ is the tensor, which has all elements equal to 1. The above can also be written in the scalar notations, $\sum_{\omega_{x_td_{t-1}}} p(\omega_{x_td_{t-1}}|x_td_{t-1}) = 1$ for each value of x_td_{t-1}. Therefore, if we apply (E.6) to (E.5), we get $\underset{\omega_{x_{t-1}d_{t-1}}\omega_{x_t}}{\tilde{\mathcal{X}}}$, which is the time-shifted version of $\underset{\omega_{x_{T-1}d_{T-1}}\omega_{x_T}}{\tilde{\mathcal{X}}}$. Therefore, to compute (E.3), we estimate the tensor in (B.5), i.e.,

$$\underset{\mathsf{\mathbf{O}}_{R_t}o_t\mathsf{\mathbf{O}}_{R_t}}{\tilde{\mathcal{X}}} = \underset{\mathsf{\mathbf{O}}_{L_t}\mathsf{\mathbf{O}}_{R_t}}{\mathcal{M}^{-1}} \times_{\mathsf{\mathbf{O}}_{L_t}} \underset{\mathsf{\mathbf{O}}_{L_t}\mathsf{\mathbf{O}}_{R_t}o_t}{\mathcal{M}}$$

and marginalize out the right set of modes, corresponding to $\mathbf{o}_{R_t}$. Alternatively, we can use the batch estimate

$$\tilde{\mathfrak{X}} = \left(\sum_t \mathfrak{M}_{\mathbf{o}_{L_t} \mathbf{o}_{R_t}} \right)^{-1} \times_{\mathbf{o}_L} \left(\sum_t \mathfrak{M}_{\mathbf{o}_{L_t} \mathbf{o}_{R_t} o_t} \right)$$

and similarly perform the marginalization.

This concludes our derivations.

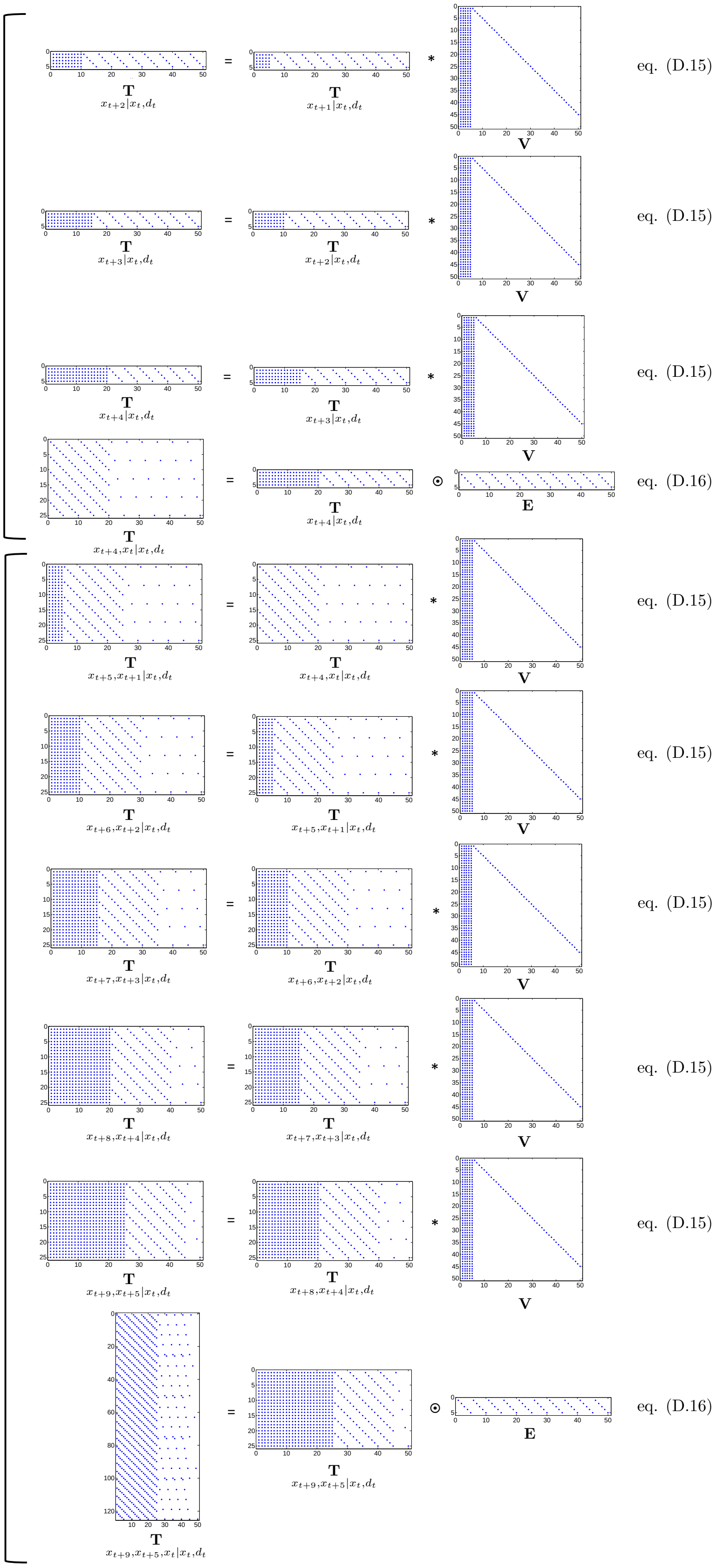
$\mathbf{T}_{x_{t+2}|x_t,d_t}$ = $\mathbf{T}_{x_{t+1}|x_t,d_t}$ * $\mathbf{V}$ eq. (D.15)
$\mathbf{T}_{x_{t+3}|x_t,d_t}$ = $\mathbf{T}_{x_{t+2}|x_t,d_t}$ * $\mathbf{V}$ eq. (D.15)
$\mathbf{T}_{x_{t+4}|x_t,d_t}$ = $\mathbf{T}_{x_{t+3}|x_t,d_t}$ * $\mathbf{V}$ eq. (D.15)
$\mathbf{T}_{x_{t+4},x_t|x_t,d_t}$ = $\mathbf{T}_{x_{t+4}|x_t,d_t}$ ⊙ $\mathbf{E}$ eq. (D.16)
$\mathbf{T}_{x_{t+5},x_{t+1}|x_t,d_t}$ = $\mathbf{T}_{x_{t+4},x_t|x_t,d_t}$ * $\mathbf{V}$ eq. (D.15)
$\mathbf{T}_{x_{t+6},x_{t+2}|x_t,d_t}$ = $\mathbf{T}_{x_{t+5},x_{t+1}|x_t,d_t}$ * $\mathbf{V}$ eq. (D.15)
$\mathbf{T}_{x_{t+7},x_{t+3}|x_t,d_t}$ = $\mathbf{T}_{x_{t+6},x_{t+2}|x_t,d_t}$ * $\mathbf{V}$ eq. (D.15)
$\mathbf{T}_{x_{t+8},x_{t+4}|x_t,d_t}$ = $\mathbf{T}_{x_{t+7},x_{t+3}|x_t,d_t}$ * $\mathbf{V}$ eq. (D.15)
$\mathbf{T}_{x_{t+9},x_{t+5}|x_t,d_t}$ = $\mathbf{T}_{x_{t+8},x_{t+4}|x_t,d_t}$ * $\mathbf{V}$ eq. (D.15)
$\mathbf{T}_{x_{t+9},x_{t+5},x_t|x_t,d_t}$ = $\mathbf{T}_{x_{t+9},x_{t+5}|x_t,d_t}$ ⊙ $\mathbf{E}$ eq. (D.16)
0 10 20 30 40 50
0 5 10 15 20 25
0 5 10 15 20 25 30 35 40 45 50
0 20 40 60 80 100 120

Figure 5: Schematic representation of Algorithm 4. This example illustrates the HSMM with $n_x = 5$ and $n_d = 10$. The non-zero matrix elements are displayed as dots.

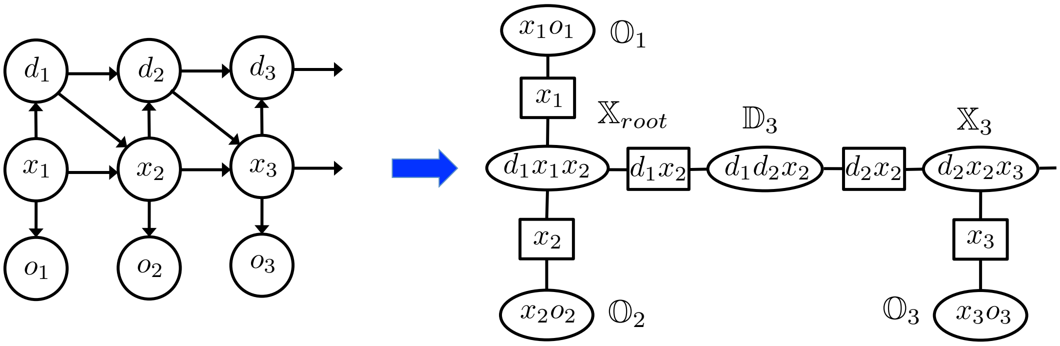

Figure 6: Part of HSMM corresponding to the initial time stamps and the related part of junction tree.

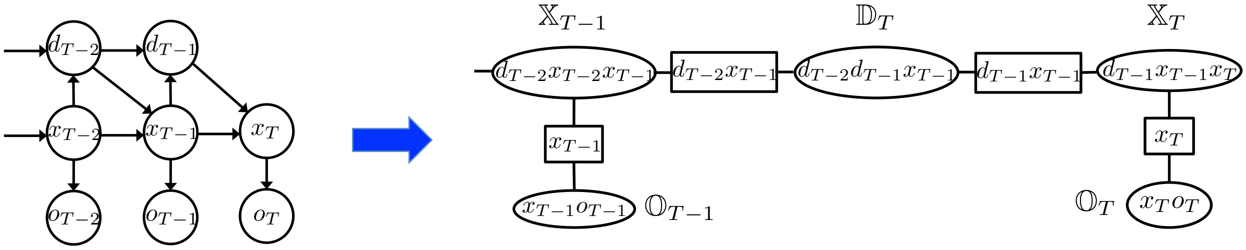

Figure 7: Part of HSMM corresponding to the final time stamps and the related part of junction tree.