# HW4

Gabriel ROMON

## Problem 14.4

Let us first state some context regarding Hidden Markov Models by following the exposition given in [1]. Given two measurable spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$, $Q$ a Markov kernel on $(X, \mathcal{X})$ and $G$ a Markov kernel from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$ we define the following transition kernel on the product space:

$$T((x,y), C) = \int 1_C(x,y) G(x', dy') Q(x, dx')$$

Together with the initial distribution $\nu \otimes G$, this defines a Markov chain $(X_k, Y_k)_{k \geq 0}$.
For the case considered in the problem, $(X, \mathcal{X}) = (\{1, \ldots, \kappa\}, \mathcal{P}(\{1, \ldots, \kappa\}))$, $(Y, \mathcal{Y}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and the model is *fully dominated*, meaning that both kernels have densities. Indeed, $\forall x \in X, Q(x, \cdot) \ll N_X$ where $N_X$ denotes the counting measure on $X$, with density $x' \mapsto \mathbb{P}_{xx'}$. In addition, $\forall x \in X, G(x, \cdot) \ll \lambda$ with density $y \mapsto f(y|x)$.
It follows that the transition kernel $T((x,y), \cdot)$ has density $(x', y') \mapsto t((x,y),(x',y')) = p(x,x')f(y'|x')$.
By a standard result on Markov chains, the joint distribution of $(X_0, Y_0, \ldots, Y_t, Y_t)$ is given by the equality

$$\begin{aligned}
E(g(X_0, Y_0, \ldots, Y_t, Y_t)) &= \int g(x_0, y_0, \ldots, x_t, y_t) \prod_{j=0}^{t-1} T((x_j, y_j), d(x_{j+1}, y_{j+1}))\, \nu \otimes G(d(x_0, y_0)) \\
&= \int \int \int g(x_0, y_0, \ldots, x_t, y_t) \prod_{j=0}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=0}^{t} f(y_j|x_j) d\lambda^{\otimes t+1}(y_{0:t}) dN_X^{\otimes t}(x_{1:t}) d\nu(x_0)
\end{aligned}$$

for every bounded measurable function $g$. Consequently, the joint distribution of $(X_1, Y_1, \ldots, Y_t, Y_t)$ is given by

$$E(g(X_1, Y_1, \ldots, X_t, Y_t)) = \int \int g(x_1, y_1, \ldots, x_t, y_t) \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1) d\lambda^{\otimes t}(y_{1:t}) dN_X^{\otimes t}(x_{1:t})$$

where $p(x_1) = \int \int \mathbb{P}_{x_0 x_1} f(y_0|x_0) d\lambda(y_0) d\nu(x_0) = \int \mathbb{P}_{x_0 x_1} d\nu(x_0)$.
This implies that the joint density of $(X_1, Y_1, \ldots, Y_t, Y_t)$ with respect to the measure $\lambda^{\otimes t} \otimes N_X^{\otimes t}$ is

$$(x_1, y_1, \ldots, x_t, y_t) \mapsto \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1)$$

We will use this result repeatedly in the problem.

(a) We are asked to compute the conditional densities of $X_1, \ldots, X_t$ given $Y_{1:t} = y_{1:t}$ and $X_1, \ldots, X_t$ given $Y_{1:t-1} = y_{1:t-1}$. It suffices to compute the ratio of the joint density and the relevant marginal density.

$$\begin{aligned}
p(x_{1:t}|y_{1:t}) &= \frac{\prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1)}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1) dN_X^{\otimes t}(x_{1:t})} \\
&= \frac{\prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1)}{\color{red}{\prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j) p(x_1)}} \frac{\color{red}{\prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j) p(x_1)}}{\color{blue}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j) p(x_1) dN_X^{\otimes t}(x_{1:t})}} \\
&\quad \times \frac{1}{\frac{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1) dN_X^{\otimes t}(x_{1:t})}{\color{blue}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j) p(x_1) dN_X^{\otimes t}(x_{1:t})}}} \\
&= f(y_t|x_t) p(x_{1:t}|y_{1:t-1}) \frac{1}{p(y_t|y_{1:t-1})}
\end{aligned}$$

The red term is the joint density of $(X_{1:t}, Y_{1:t-1})$ obtained by marginalizing with respect to $Y_t$, where we used $\int f(y_t|x_t) d\lambda(y_t) = 1$. The blue term is the joint density of $Y_{1:t-1}$ obtained by marginalizing with respect to $(X_{1:t}, Y_t)$.

$$\begin{aligned}
p(x_{1:t}|y_{1:t-1}) &= \frac{\prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j)p(x_1)}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j)p(x_1) dN_X^{\otimes t}(x_{1:t})} \\
&= \mathbb{P}_{x_{t-1}x_t} \frac{\prod_{j=1}^{t-2} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j)p(x_1)}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j)p(x_1) dN_X^{\otimes t}(x_{1:t})} \\
&= \mathbb{P}_{x_{t-1}x_t} \frac{p(x_{1:t-1}, y_{1:t-1})}{p(y_{1:t-1})} = \mathbb{P}_{x_{t-1}x_t} p(x_{1:t-1}|y_{1:t-1})
\end{aligned}$$

(b) Combining the first actualization equation with the second yields

$$p(x_{1:t}|y_{1:t}) = \frac{f(y_t|x_t)\mathbb{P}_{x_{t-1}x_t}}{p(y_t|y_{1:t-1})} p(x_{1:t-1}|y_{1:t-1})$$

To compute the filtering density, one needs to get a hold of $p(x_{1:t-1}|y_{1:t-1})$ and $p(y_t|y_{1:t-1})$. The second quantity is linked to the first one in the following way:

$$\begin{aligned}
p(y_t|y_{1:t-1}) &= \frac{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j)p(x_1) dN_X^{\otimes t}(x_{1:t})}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j)p(x_1) dN_X^{\otimes t}(x_{1:t})} \\
&= \int \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t) \frac{\prod_{j=1}^{t-2} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j)p(x_1)}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j)p(x_1) dN_X^{\otimes t}(x_{1:t})} dN_X^{\otimes t}(x_{1:t}) \\
&= \int \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t) p(x_{1:t-1}|y_{1:t-1}) dN_X^{\otimes t}(x_{1:t}) \\
&= \sum_{x_1,\ldots,x_t \in X^t} \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t) p(x_{1:t-1}|y_{1:t-1})
\end{aligned}$$

As a result, computing $p(y_t|y_{1:t-1})$ has the same complexity as computing $p(x_{1:t}|y_{1:t})$.

(c) From the actualization equations we already know $p(x_{1:t}|y_{1:t-1})$. $p(x_t|y_{1:t-1})$ can then be computed by a simple marginalization:

$$\begin{aligned}
p(x_t|y_{1:t-1}) &= \int p(x_t|y_{1:t-1}) dN_X^{\otimes t-1}(x_{1:t-1}) = \int \mathbb{P}_{x_{t-1}x_t} p(x_{1:t-1}|y_{1:t-1}) dN_X^{\otimes t-1}(x_{1:t-1}) \\
&= \sum_{x_{1:t-1} \in X^{t-1}} \mathbb{P}_{x_{t-1}x_t} p(x_{1:t-1}|y_{1:t-1})
\end{aligned}$$

# Problem 14.5

(a) • Computing $\alpha_1(i)$ boils down to finding the joint density of $(X_1, Y_1)$. Marginalizing like before,

$$\begin{aligned}
p(y_1, x_1) &= p(x_1)f(y_1|x_1) \int \int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \underbrace{\prod_{j=2}^{t} f(y_j|x_j) d\lambda^{\otimes t-1}(y_{2:t})}_{=1} dN_X^{\otimes t-1}(x_{2:t}) \\
&= p(x_1)f(y_1|x_1) \int \ldots \underbrace{\int \mathbb{P}_{x_{t-1}x_t} dN_X(x_t)}_{=1} \mathbb{P}_{x_{t-2}x_{t-1}} dN_X(x_{t-1}) \ldots dN_X(x_2) \\
&= p(x_1)f(y_1|x_1)
\end{aligned}$$

Hence $\alpha_1(i) = p(y_1, i) = p(i)f(y_1|i) = \pi_i f(y_1|X_1 = i)$ where the last equality enforces notations used in the statement of the problem.

- The recursion for $\alpha_{t+1}(j)$ follows from computing the joint density of $(X_{t+1}, Y_{1:t+1})$:

$$\begin{aligned}
p(y_{1:t+1}, x_{t+1}) &= f(y_{t+1}|x_{t+1}) \int \prod_{j=1}^{t} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1) dN_X^{\otimes t}(x_{1:t}) \\
&= f(y_{t+1}|x_{t+1}) \int \mathbb{P}_{x_t x_{t+1}} \int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1) dN_X^{\otimes t-1}(x_{1:t-1}) dN_X(x_t) \\
&= f(y_{t+1}|x_{t+1}) \int \mathbb{P}_{x_t x_{t+1}} p(y_{1:t}, x_t) dN_X(x_t) \\
&= f(y_{t+1}|x_{t+1}) \sum_{x_t \in X} \mathbb{P}_{x_t x_{t+1}} \alpha_t(x_t)
\end{aligned}$$

Hence

$$\alpha_{t+1}(j) = f(y_{t+1}|j) \sum_{x_t \in X} \mathbb{P}_{x_t\, j} \alpha_t(x_t) = f(y_{t+1}|X_{t+1} = j) \sum_{i=1}^{\kappa} \mathbb{P}_{ij}\ \alpha_t(i)$$

where the last equality enforces notations used in the statement of the problem.

- **There is a mistake in the statement of the problem (2nd edition of the book).** $\beta_t(i)$ should be defined as $p(y_{t+1:T}|x_t = i)$ (the conditioning sign is missing). Computing the ratio between the joint and marginal densities yields

$$\begin{aligned}
p(y_{t+1:T}|x_t) &= \frac{\int \prod_{j=1}^{T-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{T} f(y_j|x_j) p(x_1) dN_X^{\otimes T-1}(x_{1:T\setminus\{t\}}) d\lambda^{\otimes t}(y_{1:t})}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1) dN_X^{\otimes t-1}(x_{1:t-1}) d\lambda^{\otimes t}(y_{1:t})} \\
&= \frac{\int \prod_{j=1}^{T-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=t+1}^{T} f(y_j|x_j) p(x_1) dN_X^{\otimes T-1}(x_{1:T\setminus\{t\}})}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} p(x_1) dN_X^{\otimes t-1}(x_{1:t-1})} \\
&= \frac{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} p(x_1) dN_X^{\otimes t-1}(x_{1:t-1}) \cdot \int \prod_{j=t}^{T-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=t+1}^{T} f(y_j|x_j) dN_X^{\otimes T-t}(x_{t+1:T})}{\int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} p(x_1) dN_X^{\otimes t-1}(x_{1:t-1})} \\
&= \int \prod_{j=t}^{T-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=t+1}^{T} f(y_j|x_j) dN_X^{\otimes T-t}(x_{t+1:T}) \\
&= \int \mathbb{P}_{x_t x_{t+1}} f(y_{t+1}|x_{t+1}) \int \prod_{j=t+1}^{T-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=t+2}^{T} f(y_j|x_j) dN_X^{\otimes T-t-1}(x_{t+2:T})\, dN_X(x_{t+1})
\end{aligned}$$

The red term is identical to the blue one, except that $t$ is replaced by $t+1$. Since the blue term is $p(y_{t+1:T}|x_t)$, the red one is $p(y_{t+2:T}|x_{t+1})$, hence

$$\begin{aligned}
p(y_{t+1:T}|x_t) &= \int \mathbb{P}_{x_t x_{t+1}} f(y_{t+1}|x_{t+1}) p(y_{t+2:T}|x_{t+1}) dN_X(x_{t+1}) \\
&= \sum_{x_{t+1} \in X} \mathbb{P}_{x_t x_{t+1}} f(y_{t+1}|x_{t+1}) p(y_{t+2:T}|x_{t+1})
\end{aligned}$$

Consequently,

$$\begin{aligned}
\beta_t(i) &= \sum_{x_{t+1} \in X} \mathbb{P}_{i\, x_{t+1}} f(y_{t+1}|x_{t+1}) \beta_{t+1}(x_{t+1}) = \sum_{j=1}^{\kappa} \mathbb{P}_{ij} f(y_{t+1}|j) \beta_{t+1}(j) \\
&= \sum_{j=1}^{\kappa} \mathbb{P}_{ij}\ f(y_{t+1}|X_{t+1} = j) \beta_{t+1}(j)
\end{aligned}$$

where the last equality enforces notations used in the statement of the problem.

- Marginalizing yields

$$\begin{aligned}
p(x_t, y_{1:T}) &= \int \prod_{j=1}^{T-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{T} f(y_j|x_j) p(x_1) dN_X^{\otimes T-1}(x_{1:T\setminus\{t\}}) \\
&= \int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1) dN_X^{\otimes t-1}(x_{1:t-1}) \\
&\quad \times \int \prod_{j=t}^{T-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=t+1}^{T} f(y_j|x_j) dN_X^{\otimes T-t}(x_{t+1:T}) \\
&= p(x_t, y_{1:t})\ p(y_{t+1:T}|x_t)
\end{aligned}$$

The last equality follows from identification of each of the integrals: the first one is an obvious marginalization that yields $p(x_t, y_{1:t})$ and the second is the blue term computed a few moments ago, and we proved it is equal to $p(y_{t+1:T}|x_t)$. Besides,

$$p(y_{1:T}) = \int p(x_t, y_{1:T}) dN_X(x_t) = \int p(x_t, y_{1:t})\ p(y_{t+1:T}|x_t) dN_X(x_t)$$

Hence

$$p(x_t|y_{1:T}) = \frac{p(x_t, y_{1:t})\ p(y_{t+1:T}|x_t)}{\int p(x_t, y_{1:t})\ p(y_{t+1:T}|x_t) dN_X(x_t)} = \frac{\alpha_t(x_t)\beta_t(x_t)}{\sum_{x_t \in X} \alpha_t(x_t)\beta_t(x_t)}$$

and finally

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{\kappa} \alpha_t(j)\beta_t(j)}$$

(b) Let us assume first that all the $\alpha_t(i)$ and $\beta_t(i)$ have already been computed and stored in memory. We begin with a naive evaluation: for fixed $t$ and $i$, computing $\gamma_t(i)$ requires computing $\sum_{j=1}^{\kappa} \alpha_t(j)\beta_t(j)$, which takes $O(\kappa)$, hence a global time complexity of $O(T\kappa^2)$.
However, this is clearly inefficient since $\sum_{j=1}^{\kappa} \alpha_t(j)\beta_t(j)$ may be only evaluated once and stored in memory. For a given $t$, we compute all the $\alpha_t(i)\beta_t(i)$, then the sum $\sum_{j=1}^{\kappa} \alpha_t(j)\beta_t(j)$ and finally the ratios, yielding all the $\gamma_t(i)$ with a time complexity of $O(\kappa+\kappa+\kappa) = O(\kappa)$, hence a global complexity of $O(T\kappa)$.

For a more realistic evaluation we assume that the $\alpha_t(i)$ and $\beta_t(i)$ are not known and need to be computed beforehand. Since recursive programming results in a lot of repeated computations, it is inefficient and we turn to dynamic programming instead. Computing all the $\alpha_1(i)$ takes $O(\kappa)$. By storing $\alpha_t$ in memory, evaluating $\alpha_{t+1}(j)$ for each $j$ is $O(\kappa)$, hence computing $\alpha_{t+1}$ takes $O(\kappa^2)$, yielding a total time complexity of $O(\kappa) + \sum_{t=2}^{T} O(\kappa^2) = O(T\kappa^2)$ to compute $\alpha$. In a similar fashion, the evaluation of $\beta$ requires $O(T\kappa^2)$ in time. Finally, computing $\gamma$ from scratch has complexity $O(T\kappa^2) + O(T\kappa^2) + O(T\kappa) = O(T\kappa^2)$.

(c) Let us compute the joint density of $(X_t, X_{t+1}, Y_{1:T})$ by marginalizing:

$$\begin{aligned}
p(x_t, x_{t+1}, y_{1:T}) &= \int \prod_{j=1}^{T-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{T} f(y_j|x_j) p(x_1) dN_X^{\otimes T-2}(x_{1:T\setminus\{t,t+1\}}) \\
&= \int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1) dN_X^{\otimes t-1}(x_{1:t-1}) \\
&\quad \times \mathbb{P}_{x_t x_{t+1}} f(y_{t+1}|x_{t+1}) \\
&\quad \times \int \prod_{j=t+1}^{T-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=t+2}^{T} f(y_j|x_j) dN_X^{\otimes T-t-1}(x_{t+2:T}) \\
&= p(y_{1:t}, x_t)\mathbb{P}_{x_t x_{t+1}} f(y_{t+1}|x_{t+1}) p(y_{t+2:T}|x_{t+1})
\end{aligned}$$

In the second equality, the first integral is a straight marginalization that yields $p(y_{1:t}, x_t)$ and the last integral is the red term computed in (a), which was shown to be $p(y_{t+2:T}|x_{t+1})$. Therefore,

$$p(x_t, x_{t+1}|y_{1:T}) = \frac{p(y_{1:t}, x_t)\mathbb{P}_{x_t x_{t+1}} f(y_{t+1}|x_{t+1}) p(y_{t+2:T}|x_{t+1})}{\int p(y_{1:t}, x_t)\mathbb{P}_{x_t x_{t+1}} f(y_{t+1}|x_{t+1}) p(y_{t+2:T}|x_{t+1}) dN_X^{\otimes 2}(x_{t:t+1})}$$

Hence

$$\xi_t(i,j) = p(i,j|y_{1:T}) = \frac{p(y_{1:t},i)\mathbb{P}_{ij}f(y_{t+1}|j)p(y_{t+2:T}|j)}{\sum_{i=1}^{\kappa}\sum_{j=1}^{\kappa}p(y_{1:t},i)\mathbb{P}_{ij}f(y_{t+1}|j)p(y_{t+2:T}|j)}$$
$$= \frac{\alpha_t(i)\mathbb{P}_{ij}f(y_{t+1}|j)\beta_{t+1}(j)}{\sum_{i=1}^{\kappa}\sum_{j=1}^{\kappa}\alpha_t(i)\mathbb{P}_{ij}f(y_{t+1}|j)\beta_{t+1}(j)}$$

Computing $\alpha$ and $\beta$ has complexity $O(T\kappa^2)$ as shown before. For a fixed $t$, we compute all the numerators in $O(\kappa^2)$, then evaluate the sum $\sum_{i=1}^{\kappa}\sum_{j=1}^{\kappa}\alpha_t(i)\mathbb{P}_{ij}f(y_{t+1}|j)\beta_{t+1}(j)$ in $O(\kappa^2)$ and finally compute all the ratios also in $O(\kappa^2)$, yielding a total time complexity of $O(T(\kappa^2+\kappa^2+\kappa^2)) = O(T\kappa^2)$.

(d) Let $\alpha'_t(i) = \frac{\alpha_t(i)}{c_t}$ and note that

$$\frac{\alpha'_t(i)\beta_t(i)}{\sum_{j=1}^{\kappa}\alpha'_t(j)\beta_t(j)} = \frac{\frac{\alpha_t(i)}{c_t}\beta_t(i)}{\sum_{j=1}^{\kappa}\frac{\alpha_t(j)}{c_t}\beta_t(j)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{\kappa}\alpha_t(j)\beta_t(j)} = \gamma_t(i)$$

Hence the equalities for $\gamma$ are preserved.

# Problem 14.6

(a) For $s \geq 3$,

$$\begin{aligned}
p(x_s, x_{s-1}, y_{1:t}) &= \int \prod_{j=1}^{t-1}\mathbb{P}_{x_j x_{j+1}}\prod_{j=1}^{t}f(y_j|x_j)p(x_1)dN_X^{\otimes t-2}(x_{1:t\setminus\{s-1,s\}}) \\
&= \int \prod_{j=1}^{s-2}\mathbb{P}_{x_j x_{j+1}}\prod_{j=1}^{s-1}f(y_j|x_j)p(x_1)dN_X^{\otimes s-2}(x_{1:s-2}) \\
&\quad \times \mathbb{P}_{x_{s-1}x_s}f(y_s|x_s) \\
&\quad \times \int \prod_{j=s}^{t-1}\mathbb{P}_{x_j x_{j+1}}\prod_{j=s+1}^{t}f(y_j|x_j)dN_X^{\otimes t-s}(x_{s+1:t}) \\
&= p(x_{s-1}, y_{1:s-1}) \\
&\quad \times \mathbb{P}_{x_{s-1}x_s}f(y_s|x_s) \\
&\quad \times \int \prod_{j=s}^{t-1}\mathbb{P}_{x_j x_{j+1}}\prod_{j=s+1}^{t}f(y_j|x_j)dN_X^{\otimes t-s}(x_{s+1:t}) \qquad (*)
\end{aligned}$$

Since $p(x_{s-1}, y_{1:s-1})$ does not depend on $x_s$, it vanishes when computing the conditional density:

$$p(x_s|x_{s-1}, y_{1:t}) = \frac{\mathbb{P}_{x_{s-1}x_s}f(y_s|x_s)\int\prod_{j=s}^{t-1}\mathbb{P}_{x_j x_{j+1}}\prod_{j=s+1}^{t}f(y_j|x_j)dN_X^{\otimes t-s}(x_{s+1:t})}{\int\mathbb{P}_{x_{s-1}x_s}f(y_s|x_s)\int\prod_{j=s}^{t-1}\mathbb{P}_{x_j x_{j+1}}\prod_{j=s+1}^{t}f(y_j|x_j)dN_X^{\otimes t-s}(x_{s+1:t})dN_X(x_s)}$$

A similar computation yields

$$\begin{aligned}
p(x_s, x_{s-1}, y_{s:t}) &= p(x_{s-1}) \\
&\quad \times \mathbb{P}_{x_{s-1}x_s}f(y_s|x_s) \\
&\quad \times \int \prod_{j=s}^{t-1}\mathbb{P}_{x_j x_{j+1}}\prod_{j=s+1}^{t}f(y_j|x_j)dN_X^{\otimes t-s}(x_{s+1:t})
\end{aligned}$$

$p(x_{s-1})$ vanishes and we get

$$p(x_s|x_{s-1}, y_{s:t}) = \frac{\mathbb{P}_{x_{s-1}x_s}f(y_s|x_s)\int\prod_{j=s}^{t-1}\mathbb{P}_{x_j x_{j+1}}\prod_{j=s+1}^{t}f(y_j|x_j)dN_X^{\otimes t-s}(x_{s+1:t})}{\int\mathbb{P}_{x_{s-1}x_s}f(y_s|x_s)\int\prod_{j=s}^{t-1}\mathbb{P}_{x_j x_{j+1}}\prod_{j=s+1}^{t}f(y_j|x_j)dN_X^{\otimes t-s}(x_{s+1:t})dN_X(x_s)} = p(x_s|x_{s-1}, y_{1:t})$$

For $s = 2$ the equality is proved in a similar fashion.

(b) • Note that

$$p(x_t, x_{t-1}, y_{1:t}) = \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t) \int \prod_{j=1}^{t-2} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j) p(x_1) dN_X^{\otimes t-2}(x_{1:t-2})$$
$$= \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t) p(x_{t-1}, y_{1:t-1})$$

Hence $p(x_t|x_{t-1}, y_{1:t}) = \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t) \dfrac{p(x_{t-1}, y_{1:t-1})}{p(x_{t-1}, y_{1:t})}$ and we define an unnormalized version of this conditional density as $p_t^\star(x_t|x_{t-1}, y_{1:t}) := \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t)$.

• Let $2 \leq s \leq t-1$. From equation $(*)$ in (a) we have

$$p(x_s, x_{s-1}, y_{1:t}) = p(x_{s-1}, y_{1:s-1}) \mathbb{P}_{x_{s-1}x_s} f(y_s|x_s) \int \prod_{j=s}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=s+1}^{t} f(y_j|x_j) dN_X^{\otimes t-s}(x_{s+1:t})$$
$$= p(x_{s-1}, y_{1:s-1}) \mathbb{P}_{x_{s-1}x_s} f(y_s|x_s) \int \frac{p(x_{s+1}, x_s, y_{1:t})}{p(x_s, y_{1:s})} dN_X(x_{s+1})$$
$$= \mathbb{P}_{x_{s-1}x_s} f(y_s|x_s) \int p(x_{s+1}|x_s, y_{1:t}) \frac{p(x_s, y_{1:t})}{p(x_s, y_{1:s})} dN_X(x_{s+1}) p(x_{s-1}, y_{1:s-1})$$

Hence

$$p(x_s|x_{s-1}, y_{1:t}) = \mathbb{P}_{x_{s-1}x_s} f(y_s|x_s) \sum_{x_{s+1} \in X} \left[ p(x_{s+1}|x_s, y_{1:t}) \frac{p(x_s, y_{1:t})}{p(x_s, y_{1:s})} \right] \frac{p(x_{s-1}, y_{1:s-1})}{p(x_{s-1}, y_{1:t})}$$

and we let

$$p_s^\star(x_s|x_{s-1}, y_{1:t}) := \mathbb{P}_{x_{s-1}x_s} f(y_s|x_s) \sum_{x_{s+1} \in X} \left[ p(x_{s+1}|x_s, y_{1:t}) \frac{p(x_s, y_{1:t})}{p(x_s, y_{1:s})} \right]$$

• Note that

$$p(x_1, y_{1:t}) = p(x_1) f(y_1|x_1) \int \mathbb{P}_{x_1 x_2} f(y_2|x_2) \prod_{j=2}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=3}^{t} f(y_j|x_j) p(x_1) dN_X^{\otimes t-2}(x_{1:t-2})$$
$$= p(x_1) f(y_1|x_1) \int \frac{p(x_2, x_1, y_{1:t})}{p(x_1, y_1)} dN_X(x_2)$$

where we used $(*)$ in the second equality.
Hence $p(x_1|y_{1:t}) = p(x_1) f(y_1|x_1) \sum_{x_2 \in X} \left[ p(x_2|x_1, y_{1:t}) \dfrac{p(x_1, y_{1:t})}{p(x_1, y_1)} \right] \dfrac{1}{p(y_{1:t})}$ $\quad (**)$ and we let

$$p_1^\star(x_1|y_{1:t}) := p(x_1) f(y_1|x_1) \sum_{x_2 \in X} \left[ p(x_2|x_1, y_{1:t}) \frac{p(x_1, y_{1:t})}{p(x_1, y_1)} \right]$$

• Putting all these relations together we have the following recursion

$$\begin{cases} p_t^\star(x_t|x_{t-1}, y_{1:t}) = \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t) \\ p_s^\star(x_s|x_{s-1}, y_{1:t}) = \mathbb{P}_{x_{s-1}x_s} f(y_s|x_s) \displaystyle\sum_{x_{s+1} \in X} p_{s+1}^\star(x_{s+1}|x_s, y_{1:t}) \qquad 2 \leq s \leq t-1 \\ p_1^\star(x_1|y_{1:t}) = p(x_1) f(y_1|x_1) \displaystyle\sum_{x_2 \in X} p_2^\star(x_2|x_1, y_{1:t}) \end{cases}$$

We resort to dynamic programming again. We begin by computing and storing the $p_t^\star(x_t|x_{t-1}, y_{1:t})$ for every combination of $(x_t, x_{t-1})$, which takes $O(\kappa^2)$ time and $O(\kappa^2)$ memory. Let $s \leq t-1$ be fixed. We assume that the $p_{s+1}^\star(x_{s+1}|x_s, y_{1:t})$ have all been saved from the previous iteration, which has a memory cost of $O(\kappa^2)$. Let $x_s$ be fixed. We compute and store $\sum_{x_{s+1} \in X} p_{s+1}^\star(x_{s+1}|x_s, y_{1:t})$, and then for each $x_{s-1}$ we compute $\mathbb{P}_{x_{s-1}x_s} f(y_s|x_s) \sum_{x_{s+1} \in X} p_{s+1}^\star(x_{s+1}|x_s, y_{1:t})$, which yields the value of $p_s^\star(x_s|x_{s-1}, y_{1:t})$ in $O(\kappa)$ time. Doing this for each $x_s$ yields $p_s^\star(x_s|x_{s-1}, y_{1:t})$ for every combination of $(x_s, x_{s-1})$ in $O(\kappa^2)$ time. To complete this iteration, we erase all the $p_{s+1}^\star(x_{s+1}|x_s, y_{1:t})$ from memory and we store instead all the $p_s^\star(x_s|x_{s-1}, y_{1:t})$.
As a result, the global time complexity is $O(t\kappa^2)$ and the memory footprint is $O(\kappa^2)$.

(c) It is straightforward to prove that $p(x_{1:t}|y_{1:t}) = p(x_1|y_{1:t}) \prod_{s=1}^{t} p(x_s|x_{s-1}, y_{1:t})$.

Once the $p_s{}^\star$ have been computed, the normalized versions $p(x_s|x_{s-1}, y_{1:t})$ can be retrieved without increasing the computational complexity. Given the factorization of the joint distribution, we can sample from it by sequentially generating $x_1$ from $p(x_1|y_{1:t})$, $x_2$ from $p(x_2|x_1, y_{1:t})$ and so on. This process takes $O(t\kappa^2)$ in time.

(d) Using the formula for $p(x_{1:t}|y_{1:t})$ proved a bit later in (e), we have for $s \geq 3$

$$p(x_{1:s}|y_{1:s}) = \mathbb{P}_{x_{s-1}x_s} f(y_s|x_s) p(x_{1:s-1}|y_{1:s-1}) \frac{p(y_{1:s-1})}{p(y_{1:s})}$$

Hence

$$\begin{aligned}
\arg\max_{x_{1:t}} p(x_{1:t}|y_{1:t}) &= \arg\max_{x_t} \left[ \arg\max_{x_{1:t-1}} p(x_{1:t}|y_{1:t}) \right] \\
&= \arg\max_{x_t} \left[ \arg\max_{x_{1:t-1}} \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t) p(x_{1:t-1}|y_{1:t-1}) \right] \\
&= \arg\max_{x_t} \left[ \arg\max_{x_{t-1}} \left( \mathbb{P}_{x_{t-1}x_t} f(y_t|x_t) \arg\max_{x_{1:t-2}} p(x_{1:t-1}|y_{1:t-1}) \right) \right]
\end{aligned}$$

The blue terms show that computing the maximizer is amenable to dynamic programming. One starts by computing

$$\arg\max_{x_1} p(x_{1:2}|y_{1:2}) = \arg\max_{x_1} \mathbb{P}_{x_1x_2} f(y_1|x_1) p(x_1)$$

and then recurses all the way up to $\arg\max_{x_{1:t}} p(x_{1:t}|y_{1:t})$.

(e) Equation $(**)$ in (b) rewrites further as

$$p(x_1|y_{1:t}) = \frac{p_1^\star(x_1|y_{1:t})}{p(y_{1:t})}$$

Hence $p(y_{1:t}) = \sum_{x_1 \in X} p_1^\star(x_1|y_{1:t})$, which yields a representation of the observed likelihood.

As an aside,

$$\begin{aligned}
p(x_{1:t}|y_{1:t}) &= \frac{p(x_1) f(y_1|x_1) \sum_{x_2 \in X} p_2^\star(x_2|x_1, y_{1:t})}{\sum_{x_1 \in X} p_1^\star(x_1|y_{1:t})} \prod_{s=1}^{t-1} \frac{\mathbb{P}_{x_{s-1}x_s} f(y_s|x_s) \sum_{x_{s+1} \in X} p_{s+1}^\star(x_{s+1}|x_s, y_{1:t})}{\sum_{x_s \in X} p_s^\star(x_s|x_{s-1}, y_{1:t})} \\
&\quad \times \frac{\mathbb{P}_{x_{t-1}x_t} f(y_t|x_t)}{\sum_{x_t \in X} p_t^\star(x_t|x_{t-1}, y_{1:t})} \\
&= \frac{\prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1)}{\sum_{x_1 \in X} p_1^\star(x_1|y_{1:t})} \\
&= \frac{\prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j) p(x_1)}{p(y_{1:t})}
\end{aligned}$$

(f) As stated in (c), computing the conditional likelihood $p(x_{1:T}|y_{1:T})$ has complexity $O(T\kappa^2)$ in time.

## Problem 14.7

(a) The recursion for $\varphi_{t+1}(j)$ follows from computing the joint density of $(X_{t+1}, Y_{1:t})$:

$$
\begin{aligned}
p(x_{t+1}, y_{1:t}) &= \int\int \prod_{j=1}^{t} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t+1} f(y_j|x_j)p(x_1)d\lambda(y_{t+1})dN_X^{\otimes t}(x_{1:t}) \\
&= \int \prod_{j=1}^{t} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t} f(y_j|x_j)p(x_1)dN_X^{\otimes t}(x_{1:t}) \\
&= \int \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t) \int \prod_{j=1}^{t-1} \mathbb{P}_{x_j x_{j+1}} \prod_{j=1}^{t-1} f(y_j|x_j)p(x_1)dN_X^{\otimes t-1}(x_{1:t-1})\, dN_X(x_t) \\
&= \int \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)p(x_t, y_{1:t-1})dN_X(x_t) \\
&= p(y_{1:t-1}) \int \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)p(x_t|y_{1:t-1})dN_X(x_t) \qquad (*)
\end{aligned}
$$

$p(y_{1:t-1})$ vanishes when computing the conditional expectation:

$$
\begin{aligned}
p(x_{t+1}|y_{1:t}) &= \frac{\int \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)p(x_t|y_{1:t-1})dN_X(x_t)}{\int\int \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)p(x_t|y_{1:t-1})dN_X(x_t)dN_X(x_{t+1})} \\
&= \frac{\int \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)p(x_t|y_{1:t-1})dN_X(x_t)}{\int (\int \mathbb{P}_{x_t x_{t+1}} dN_X(x_{t+1})) f(y_t|x_t)p(x_t|y_{1:t-1})dN_X(x_t)} \\
&= \frac{\int \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)p(x_t|y_{1:t-1})dN_X(x_t)}{\int f(y_t|x_t)p(x_t|y_{1:t-1})dN_X(x_t)}
\end{aligned}
$$

With the notations used in the problem this turns into

$$
\varphi_{t+1}(x_{t+1}) = \frac{1}{c_t} \sum_{x_t \in X} \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)\varphi_t(x_t)
$$

Hence

$$
\varphi_{t+1}(j) = \frac{1}{c_t} \sum_{i=1}^{\kappa} \mathbb{P}_{ij} f(y_t|x_t = i)\varphi_t(i)
$$

Setting $\varphi_1(j) = p(x_1 = j)$ is just a convention that makes sense.

(b) Integrating $(*)$ in (a) with respect to $x_{t+1}$ yields $p(y_{1:t}) = p(y_{1:t-1}) \int f(y_t|x_t)p(x_t|y_{1:t-1})dN_X(x_t)$, hence

$$
p(y_t|y_{1:t-1}) = \int f(y_t|x_t)p(x_t|y_{1:t-1})dN_X(x_t) = \sum_{x_t \in X} f(y_t|x_t)p(x_t|y_{1:t-1})
$$

Besides, trivial telescoping gives $p(y_{1:t}) = p(y_1) \prod_{s=2}^{t} p(y_s|y_{1:s-1})$, thus

$$
\begin{aligned}
p(y_{1:t}) &= p(y_1) \prod_{s=2}^{t} \sum_{x_s \in X} f(y_s|x_s)p(x_s|y_{1:s-1}) \\
&= \sum_{x_1 \in X} f(y_1|x_1)p(x_1) \prod_{s=2}^{t} \sum_{x_s \in X} f(y_s|x_s)\varphi_s(x_s) \\
&= \prod_{s=1}^{t} \sum_{x_s \in X} f(y_s|x_s)\varphi_s(x_s)
\end{aligned}
$$

(c) We proved earlier that $p(y_{1:t}) = \sum_{x_1 \in X} p_1^\star(x_1|y_{1:t})$. Remember from question (b) in Problem 14.6 that computing the $p_s^\star$ has complexity $O(t\kappa^2)$ in time and $O(\kappa^2)$ in memory.

From the previous question we have $p(y_{1:t}) = \prod_{s=1}^{t} \sum_{x_s \in X} f(y_s|x_s)\varphi_s(x_s)$, which requires knowledge of the $\varphi_s$. These can be computed via dynamic programming: we begin by computing and storing all the $\varphi_1(x_1)$ which takes $O(\kappa)$ time and $O(\kappa)$ memory. Additionally we compute and store $\sum_{x_1 \in X} f(y_1|x_1)\varphi_1(x_1)$. Let $s \geq 2$ be fixed. We assume that all the $\varphi_{s-1}(x_{s-1})$ have been saved from the previous iteration and that $\prod_{r=1}^{s-1} \sum_{x_r \in X} f(y_r|x_r)\varphi_r(x_r)$ has also been saved, all of which has a memory cost of $O(\kappa)$. Using the forward equation, we compute all the $\varphi_s(x_s)$ in $O(\kappa^2)$ time, from which we get $\sum_{x_s \in X} f(y_s|x_s)\varphi_s(x_s)$ in $O(\kappa)$ time and then $\prod_{r=1}^{s} \sum_{x_r \in X} f(y_r|x_r)\varphi_r(x_r)$ in $O(1)$. To complete this iteration, we erase the $\varphi_{s-1}(x_{s-1})$ from memory and store the $\varphi_s(x_s)$instead.

As a result, computing $p(y_{1:t})$ takes $O(t\kappa^2)$ time and $O(\kappa)$ memory. Compared to the other method, this one has similar time complexity but demands less memory ($O(\kappa)$ versus $O(\kappa^2)$).

(d) Note that

$$\begin{aligned}
\nabla_\theta \log p(y_{1:t}) &= \nabla_\theta \left[ \sum_{s=1}^{t} \log \left( \sum_{x_s \in X} f(y_s|x_s)\varphi_s(x_s) \right) \right] \\
&= \sum_{s=1}^{t} \nabla_\theta \log \left( \sum_{x_s \in X} f(y_s|x_s)\varphi_s(x_s) \right) \\
&= \sum_{s=1}^{t} \frac{\sum_{x_s \in X} f(y_s|x_s)\nabla_\theta(\varphi_s(x_s)) + \varphi_s(x_s)\nabla_\theta(f(y_s|x_s))}{\sum_{x_s \in X} f(y_s|x_s)\varphi_s(x_s)} \\
&= \sum_{s=1}^{t} \frac{1}{c_s} \sum_{x_s \in X} f(y_s|x_s)\nabla_\theta(\varphi_s(x_s)) + \varphi_s(x_s)\nabla_\theta(f(y_s|x_s))
\end{aligned}$$

Using the forward equation,

$$\begin{aligned}
\nabla_\theta \varphi_{t+1}(x_{t+1}) = \nabla_\theta &\left[ \frac{1}{\sum_{x_t \in X} f(y_t|x_t)\varphi_t(x_t)} \sum_{x_t \in X} \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)\varphi_t(x_t) \right] \\
&= -\frac{1}{c_t^2} \sum_{x_t \in X} \left[ f(y_t|x_t)\nabla_\theta \varphi_t(x_t) + \varphi_t(x_t)\nabla_\theta f(y_t|x_t) \right] \sum_{x_t \in X} \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)\varphi_t(x_t) \\
&\quad + \frac{1}{c_t} \sum_{x_t \in X} \mathbb{P}_{x_t x_{t+1}} \left[ f(y_t|x_t)\nabla_\theta \varphi_t(x_t) + \varphi_t(x_t)\nabla_\theta f(y_t|x_t) \right] \\
&= -\frac{1}{c_t} \sum_{x_t \in X} \left[ f(y_t|x_t)\nabla_\theta \varphi_t(x_t) + \varphi_t(x_t)\nabla_\theta f(y_t|x_t) \right] \varphi_{t+1}(x_{t+1}) \\
&\quad + \frac{1}{c_t} \sum_{x_t \in X} \mathbb{P}_{x_t x_{t+1}} \left[ f(y_t|x_t)\nabla_\theta \varphi_t(x_t) + \varphi_t(x_t)\nabla_\theta f(y_t|x_t) \right] \\
&= \frac{1}{c_t} \sum_{x_t \in X} (\mathbb{P}_{x_t x_{t+1}} - \varphi_{t+1}(x_{t+1})) \left[ f(y_t|x_t)\nabla_\theta \varphi_t(x_t) + \varphi_t(x_t)\nabla_\theta f(y_t|x_t) \right]
\end{aligned}$$

(e) Note that

$$\begin{aligned}
\nabla_\eta \log p(y_{1:t}) &= \nabla_\eta \left[ \sum_{s=1}^{t} \log \left( \sum_{x_s \in X} f(y_s|x_s)\varphi_s(x_s) \right) \right] \\
&= \sum_{s=1}^{t} \nabla_\eta \log \left( \sum_{x_s \in X} f(y_s|x_s)\varphi_s(x_s) \right) \\
&= \sum_{s=1}^{t} \frac{1}{c_s} \sum_{x_s \in X} f(y_s|x_s)\nabla_\eta(\varphi_s(x_s))
\end{aligned}$$

Using the forward equation,

$$\begin{aligned}
\nabla_\eta \varphi_{t+1}(x_{t+1}) &= \nabla_\eta \left[ \frac{1}{\sum_{x_t \in X} f(y_t|x_t)\varphi_t(x_t)} \sum_{x_t \in X} \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)\varphi_t(x_t) \right] \\
&= -\frac{1}{c_t^2} \sum_{x_t \in X} \left[ f(y_t|x_t)\nabla_\eta \varphi_t(x_t) \right] \sum_{x_t \in X} \mathbb{P}_{x_t x_{t+1}} f(y_t|x_t)\varphi_t(x_t) \\
&\quad + \frac{1}{c_t} \sum_{x_t \in X} f(y_t|x_t) \left[ \mathbb{P}_{x_t x_{t+1}} \nabla_\eta \varphi_t(x_t) + \varphi_t(x_t)\nabla_\eta \mathbb{P}_{x_t x_{t+1}} \right] \\
&= -\frac{1}{c_t} \sum_{x_t \in X} \left[ f(y_t|x_t)\nabla_\eta \varphi_t(x_t) \right] \varphi_{t+1}(x_{t+1}) \\
&\quad + \frac{1}{c_t} \sum_{x_t \in X} f(y_t|x_t) \left[ \mathbb{P}_{x_t x_{t+1}} \nabla_\eta \varphi_t(x_t) + \varphi_t(x_t)\nabla_\eta \mathbb{P}_{x_t x_{t+1}} \right] \\
&= \frac{1}{c_t} \sum_{x_t \in X} f(y_t|x_t) \left[ \nabla_\eta \varphi_t(x_t)(\mathbb{P}_{x_t x_{t+1}} - \varphi_{t+1}(x_{t+1})) + \varphi_t(x_t)\nabla_\eta \mathbb{P}_{x_t x_{t+1}} \right]
\end{aligned}$$

(f) To compute $\nabla_\theta \log p(y_{1:t})$ we first compute and save all the $\varphi_s(x_s)$, which takes $O(t\kappa^2)$ in time and $O(\kappa)$ in memory. Next we compute the $\nabla_\theta(\varphi_s(x_s))$ via dynamic programming. For a fixed $s$, once all the $\nabla_\theta(\varphi_{s-1}(x_{s-1}))$ are computed, evaluating $\nabla_\theta(\varphi_s(x_s))$ for each $x_s$ with the forward equation takes $O(\kappa)$, so the summand $\frac{1}{c_s}\sum_{x_s \in X} f(y_s|x_s)\nabla_\theta(\varphi_s(x_s)) + \varphi_s(x_s)\nabla_\theta(f(y_s|x_s))$ is computed in $O(\kappa^2)$ time.

Therefore, the global time complexity to compute $\nabla_\theta \log p(y_{1:t})$ is $O(t\kappa^2 + t\kappa^2) = O(t\kappa^2)$. Similarly, computing $\nabla_\eta \log p(y_{1:t})$ takes $O(t\kappa^2)$ in time.

(g) Let $n$ be the dimension of $\theta$ and $m$ be that of $\eta$. To fulfill the next iteration in the computation of $\nabla_\theta(\varphi_s(x_s))$, one must save in memory the results of the previous iteration, that is to say $\nabla_\theta(\varphi_{s-1}(x_{s-1}))$ for each $x_{s-1}$. This costs $O(\kappa n)$. Therefore, computing $\nabla_\theta \log p(y_{1:t})$ and $\nabla_\eta \log p(y_{1:t})$ takes $O(\kappa(n + m)) = O(\kappa p)$ in memory.

## References

[1] Tobias Ryden Olivier Cappé, Eric Moulines. *Inference in Hidden Markov Models.* Springer Series in Statistics. Springer, 2005.