Quick Refresher on HMM and LDS

Shoichiro Yamanishi

May 10, 2020

Abstract

This is a personal notes as my own memory aid on Hidden Markov Models and Linear Dynamical Systems. Specifically the following topics.

- Baum-Welch EM algorithm
- Viterbi algorithm
- Kalman Filter
- Rauch-Tung-Striebel smoother and EM algorithm

Chapter 13 of PRML[2] Chap 13 is an excellent source for HMM (Baum-Welch, Viterbi) and Kalman Filter as in $p(\boldsymbol{z}_n|\boldsymbol{z}1, \cdots, \boldsymbol{z}_n)$, but not so good for Kalman smoother (RTS smoother) as in $p(\boldsymbol{z}_n|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N)$. Especially the derivation of $p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N)$, which is required for EM-algorithm, is a bit shaky between (13.103) and (13.104). For deriving RTS smoother, I used an excellent course notes [4] from Professor Särkkä of Aalto Univ. Also, Chap 24 of Barber [1] contains comprehensive materials for LDS, but it is a bit difficult to understand and I personally do not like the style of notations.

1 Baum-Welch Algorithm

1.1 HMM Model formation

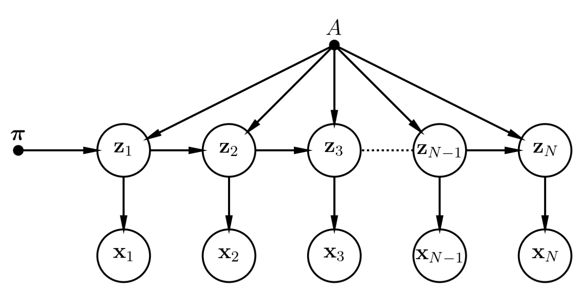

Figure 1: Parameter Reduction

Let $\boldsymbol{x}_i$ be the variables observed, and $\boldsymbol{z}_i$ be one-of-K latent variables. And let the joint distribution be:

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N) = p(\boldsymbol{z}_1|\boldsymbol{\pi}) \prod_{n=1}^{N} p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}) \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n) \tag{1}$$

as a hidden Markov model. Please note that $(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$ are not i.i.d. We parameterize $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})$ called *transition probability* as $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, A)$, where A is a $K \times K$ matrix, and each row A_i represent a multinomial distribution, i.e., $\sum_{j=1}^{K} A_{i,j} = 1$, and $A_{i,j} \geq 0$. This means A_i represents the distribution of $\boldsymbol{z}_n$, if $[\boldsymbol{z}_{n-1}]_i = 1$ (i is chosen for $\boldsymbol{z}_{n-1}$.) So,

$$p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, A) = \prod_{j=1}^{K} \prod_{i=1}^{K} A_{i,j}^{[\boldsymbol{z}_{n-1}]_i [\boldsymbol{z}_n]_j} \tag{2}$$

and for the initial distribution,

$$p(\boldsymbol{z}_1|\boldsymbol{\pi}) = \prod_{j=1}^{K} \pi_j^{[\boldsymbol{z}_1]_j} \tag{3}$$

We also parameterize $p(\boldsymbol{x}_n|\boldsymbol{z}_n)$ called *emission probability* as

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\phi}) = \prod_{j=1}^{K} p(\boldsymbol{x}_n|\phi_j)^{[\boldsymbol{z}_n]_j} \tag{4}$$

This can be a Gaussian mixture. Then we aggregate the parameters A and $\boldsymbol{\phi}$ into $\boldsymbol{\theta}$ The parameterized joint distribution is:

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N|\boldsymbol{\theta}) = p(\boldsymbol{z}_1|\boldsymbol{\pi}) \prod_{n=1}^{N} p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, A) \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\phi}) \tag{5}$$

1.2 Maximum Likelihood with EM-algorithm

The maximum likelihodd estimate for the parameter $\boldsymbol{\theta}$ given the observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$ is:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \{p(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N|\boldsymbol{\theta})\} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \sum_{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N} p(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N|\boldsymbol{\theta}) \right\} \tag{6}$$

The summation over $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N$ on RHS is intractable, and we need to use EM-framework. For that we need to form the following function of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ derived from ELBO.

$$
\begin{aligned}
Q(\boldsymbol{\theta}_{old}, \boldsymbol{\theta}) &= \mathbb{E}_{(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N)\sim p(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N,\boldsymbol{\theta}_{old})}\left[\ln p(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N,\boldsymbol{\theta})\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{K}[\boldsymbol{z}_1]_k \ln \pi_k + \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{i=1}^{K}[\boldsymbol{z}_{n-1}]_i[\boldsymbol{z}_n]_j \ln A_{i,j} + \sum_{n=1}^{N}\sum_{i=1}^{K}[\boldsymbol{z}_{n-1}]_i \ln p(\boldsymbol{x}_n|\phi_k)\right] \\
&= \sum_{k=1}^{K}\mathbb{E}\left[[\boldsymbol{z}_1]_k\right] \ln \pi_k + \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{i=1}^{K}\mathbb{E}\left[[\boldsymbol{z}_{n-1}]_i[\boldsymbol{z}_n]_j\right] \ln A_{i,j} + \sum_{n=1}^{N}\sum_{i=1}^{K}\mathbb{E}\left[[\boldsymbol{z}_n]_i\right] \ln p(\boldsymbol{x}_n|\phi_k)
\end{aligned} \tag{7}
$$

We introduce two types of marginal distributions,

$$
\begin{aligned}
\boldsymbol{\gamma}(\boldsymbol{z}_i) &= p(\boldsymbol{z}_i|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N,\boldsymbol{\theta}_{old}), \in \mathcal{R}^K \\
\boldsymbol{\xi}(\boldsymbol{z}_i,\boldsymbol{z}_{i+1}) &= p(\boldsymbol{z}_i,\boldsymbol{z}_{i+1}|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N,\boldsymbol{\theta}_{old}) \in \mathcal{R}^{K\times K}
\end{aligned} \tag{8}
$$

as in (13.13) and (13.14) of PRML [2]. Then we can express the expectations above as follows.

$$
\begin{aligned}
\mathbb{E}\left[[\boldsymbol{z}_i]_k\right] &= \sum_{k=1}^{K}[\boldsymbol{\gamma}(\boldsymbol{z}_i)]_k[\boldsymbol{z}_i]_k \\
\mathbb{E}\left[[\boldsymbol{z}_i]_k[\boldsymbol{z}_{i+1}]_m\right] &= \sum_{k=1}^{K}\sum_{m=1}^{K}[\boldsymbol{\xi}(\boldsymbol{z}_i,\boldsymbol{z}_{i+1})]_{km}[\boldsymbol{z}_i]_k[\boldsymbol{z}_{i+1}]_m
\end{aligned} \tag{9}
$$

Please note $p(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N,\boldsymbol{\theta}_{old})$ is not cleanly factorized, i.e. Marginalization is required for $\boldsymbol{\gamma}(\boldsymbol{z}_i)$ and $\boldsymbol{\xi}(\boldsymbol{z}_i,\boldsymbol{z}_{i+1})$. This is where the sum-product belief propagaion along the chain comes to the rescue.

First we factorize $\boldsymbol{\gamma}(\boldsymbol{z}_n)$ conditioned on $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N$. using Baye's rule and the conditional independence.

$$
\begin{aligned}
\boldsymbol{\gamma}(\boldsymbol{z}_n) &= p(\boldsymbol{z}_n|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N) \\
&= \frac{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N|\boldsymbol{z}_n)p(\boldsymbol{z}_n)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)} \\
&= \frac{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1},\cdots,\boldsymbol{x}_N|\boldsymbol{z}_n)p(\boldsymbol{z}_n)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)} \\
&= \frac{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n,\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1},\cdots,\boldsymbol{x}_N|\boldsymbol{z}_n)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)} \\
&= \frac{\boldsymbol{\alpha}(\boldsymbol{z}_n)\boldsymbol{\beta}(\boldsymbol{z}_n)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)}
\end{aligned} \tag{10}
$$

where $\boldsymbol{\alpha}(\boldsymbol{z}_n) = p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n,\boldsymbol{z}_n)$ and $\boldsymbol{\beta}(\boldsymbol{z}_n) = p(\boldsymbol{x}_{n+1},\cdots,\boldsymbol{x}_N|\boldsymbol{z}_n)$.

In the same way, we factorize $\boldsymbol{\xi}(\boldsymbol{z}_i, \boldsymbol{z}_{i+1})$ conditioned on $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$.

$$
\begin{aligned}
\boldsymbol{\xi}(\boldsymbol{z}_n, \boldsymbol{z}_{n+1}) &= p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1} | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) \\
&= \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_n, \boldsymbol{z}_{n+1})}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)} \\
&= \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N | \boldsymbol{z}_n, \boldsymbol{z}_{n+1}) p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1})}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)} \\
&= \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n | \boldsymbol{z}_n, \boldsymbol{z}_{n+1}) p(\boldsymbol{x}_{n+1} | \boldsymbol{z}_n, \boldsymbol{z}_{n+1}) p(\boldsymbol{x}_{n+2}, \cdots, \boldsymbol{x}_N | \boldsymbol{z}_n, \boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1} | \boldsymbol{z}_n) p(\boldsymbol{z}_n)}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)} \\
&= \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n | \boldsymbol{z}_n) p(\boldsymbol{x}_{n+1} | \boldsymbol{z}_{n+1}) p(\boldsymbol{x}_{n+2}, \cdots, \boldsymbol{x}_N | \boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1} | \boldsymbol{z}_n) p(\boldsymbol{z}_n)}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)} \\
&= \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n | \boldsymbol{z}_n) p(\boldsymbol{z}_n) p(\boldsymbol{x}_{n+1} | \boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1} | \boldsymbol{z}_n) p(\boldsymbol{x}_{n+2}, \cdots, \boldsymbol{x}_N | \boldsymbol{z}_{n+1})}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)} \\
&= \frac{\boldsymbol{\alpha}(\boldsymbol{z}_n) p(\boldsymbol{x}_{n+1} | \boldsymbol{z}_{n+1}) p(\boldsymbol{z}_{n+1} | \boldsymbol{z}_n) \boldsymbol{\beta}(\boldsymbol{z}_{n+1})}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)}
\end{aligned}
\tag{11}
$$

1.3 Belief Propagation (Sum-Product Algorithm over a Chain

Next we define $\boldsymbol{\alpha}(\boldsymbol{z}_n)$ in the forward propagation, and $\boldsymbol{\beta}(\boldsymbol{z}_n)$ in the backward propagation as follows.

$$
\begin{aligned}
\boldsymbol{\alpha}(\boldsymbol{z}_1) &= p(\boldsymbol{x}_1, \boldsymbol{z}_1) = p(\boldsymbol{x}_1 | \boldsymbol{z}_1) p(\boldsymbol{z}_1) \\
\boldsymbol{\alpha}(\boldsymbol{z}_n) &= p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n, \boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n | \boldsymbol{z}_n) p(\boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_n | \boldsymbol{z}_n) p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1} | \boldsymbol{z}_n) p(\boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_n | \boldsymbol{z}_n) p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}, \boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_n | \boldsymbol{z}_n) \sum_{k=1}^{K} p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}, [\boldsymbol{z}_{n-1}]_k, \boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_n | \boldsymbol{z}_n) \sum_{k=1}^{K} p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}, \boldsymbol{z}_n \mid [\boldsymbol{z}_{n-1}]_k) p([\boldsymbol{z}_{n-1}]_k) \\
&= p(\boldsymbol{x}_n | \boldsymbol{z}_n) \sum_{k=1}^{K} p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1} \mid [\boldsymbol{z}_{n-1}]_k) p(\boldsymbol{z}_n \mid [\boldsymbol{z}_{n-1}]_k) p([\boldsymbol{z}_{n-1}]_k) \\
&= p(\boldsymbol{x}_n | \boldsymbol{z}_n) \sum_{k=1}^{K} p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1} \mid [\boldsymbol{z}_{n-1}]_k) p([\boldsymbol{z}_{n-1}]_k) p(\boldsymbol{z}_n \mid [\boldsymbol{z}_{n-1}]_k) \\
&= p(\boldsymbol{x}_n | \boldsymbol{z}_n) \sum_{k=1}^{K} [\boldsymbol{\alpha}(\boldsymbol{z}_{n-1})]_k p(\boldsymbol{z}_n \mid [\boldsymbol{z}_{n-1}]_k)
\end{aligned}
\tag{12}
$$

This is a forward propagation summing over $\boldsymbol{z}_{n-1}$. Please note $\boldsymbol{\alpha}(\boldsymbol{z}_n)$ is not a normalized probability distribution. In the same way,

$$
\begin{aligned}
\boldsymbol{\beta}(\boldsymbol{z}_N) &= \mathbf{1} \\
\boldsymbol{\beta}(\boldsymbol{z}_n) &= p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N | \boldsymbol{z}_n) \\
&= \frac{p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_n)}{p(\boldsymbol{z}_n)} \\
&= \frac{\sum_{k=1}^K p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_n, [\boldsymbol{z}_{n+1}]_k)}{p(\boldsymbol{z}_n)} \\
&= \frac{\sum_{k=1}^K p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_n \mid [\boldsymbol{z}_{n+1}]_k) p([\boldsymbol{z}_{n+1}]_k)}{p(\boldsymbol{z}_n)} \\
&= \frac{\sum_{k=1}^K p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N \mid [\boldsymbol{z}_{n+1}]_k) p(\boldsymbol{z}_n \mid [\boldsymbol{z}_{n+1}]_k) p([\boldsymbol{z}_{n+1}]_k)}{p(\boldsymbol{z}_n)} \\
&= \frac{\sum_{k=1}^K p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N \mid [\boldsymbol{z}_{n+1}]_k) p(\boldsymbol{z}_n, [\boldsymbol{z}_{n+1}]_k)}{p(\boldsymbol{z}_n)} \\
&= \frac{\sum_{k=1}^K p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N \mid [\boldsymbol{z}_{n+1}]_k) p([\boldsymbol{z}_{n+1}]_k | \boldsymbol{z}_n) p(\boldsymbol{z}_n)}{p(\boldsymbol{z}_n)} \\
&= \frac{p(\boldsymbol{z}_n) \sum_{k=1}^K p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N \mid [\boldsymbol{z}_{n+1}]_k) p([\boldsymbol{z}_{n+1}]_k | \boldsymbol{z}_n)}{p(\boldsymbol{z}_n)} \\
&= \sum_{k=1}^K p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N \mid [\boldsymbol{z}_{n+1}]_k) p([\boldsymbol{z}_{n+1}]_k | \boldsymbol{z}_n) \\
&= \sum_{k=1}^K p(\boldsymbol{x}_{n+1} \mid [\boldsymbol{z}_{n+1}]_k) p(\boldsymbol{x}_{n+2}, \cdots, \boldsymbol{x}_N \mid [\boldsymbol{z}_{n+1}]_k) p([\boldsymbol{z}_{n+1}]_k | \boldsymbol{z}_n) \\
&= \sum_{k=1}^K \boldsymbol{\beta}(\boldsymbol{z}_{n+1}) p(\boldsymbol{x}_{n+1} \mid [\boldsymbol{z}_{n+1}]_k) p([\boldsymbol{z}_{n+1}]_k | \boldsymbol{z}_n),
\end{aligned}
\tag{13}
$$

This is a backward propagation summing over $\boldsymbol{z}_{n+1}$. Please note that the derivation of $\beta(\boldsymbol{z}_n)$ in PRML [2] in pp 621 is wrong between the 2nd and 3rd lines.

After finding $\alpha(\boldsymbol{z}_n)$ and $\beta(\boldsymbol{z}_n)$, we can calculate $\boldsymbol{\gamma}(\boldsymbol{z}_n)$ and $\boldsymbol{\xi}(\boldsymbol{z}_n, \boldsymbol{z}_{n+1})$, and then $\mathbb{E}\left[[\boldsymbol{z}_i]_k\right]$ and $\mathbb{E}\left[[\boldsymbol{z}_i]_k[\boldsymbol{z}_{i+1}]_m\right]$. This corresponds to the E-step.

Then at the M-step, we can find $\boldsymbol{\theta}^*$ such that $\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \{Q(\boldsymbol{\theta}_{old}, \boldsymbol{\theta})\}$. Usually this is achieved by $\nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}_{old}, \boldsymbol{\theta}) = 0$.

1.4 Scaling α and β

As stated above, $\alpha(\boldsymbol{z}_n)$ and $\beta(\boldsymbol{z}_n)$ are not normalized and it can lead to numerical issues through the recursion. To avoid those issues, we normalize $\alpha(\boldsymbol{z}_n)$ and we increase the value of $\beta(\boldsymbol{z}_n)$ as follows.

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n) = \frac{\boldsymbol{\alpha}(\boldsymbol{z}_n)}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)} = \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n, \boldsymbol{z}_n)}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)} = p(\boldsymbol{z}_n|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) \tag{14}$$

$$\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n) = \frac{\boldsymbol{\beta}(\boldsymbol{z}_n)}{p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)} = \frac{p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N|\boldsymbol{z}_n)}{p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)} \tag{15}$$

Please note that $\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)$ is a proper pribability density function, but $\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n)$ is not. Next, we calculate $\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n)$ recursively with belief propagation. For that, use a a tool as a factor as follows.

$$\begin{aligned} c_1 &= p(\boldsymbol{x}_1) \\ c_n &= p(\boldsymbol{x}_n|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}) \end{aligned} \tag{16}$$

It has the following characteristics.

$$\begin{aligned} \prod_{i=1}^{n} c_i &= p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1, \boldsymbol{x}_2)\cdots p(\boldsymbol{x}_n|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}) \\ &= p(\boldsymbol{x}_1)\frac{p(\boldsymbol{x}_2, \boldsymbol{x}_1)}{p(\boldsymbol{x}_1)}\frac{p(\boldsymbol{x}_3, \boldsymbol{x}_2, \boldsymbol{x}_1)}{p(\boldsymbol{x}_1, \boldsymbol{x}_2)}\cdots\frac{p(\boldsymbol{x}_n, \boldsymbol{x}_{n-1}, \cdots, \boldsymbol{x}_1)}{p(\boldsymbol{x}_{n-1}, \cdots, \boldsymbol{x}_1)} \\ &= p(\boldsymbol{x}_n, \boldsymbol{x}_{n-1}, \cdots, \boldsymbol{x}_1) \end{aligned} \tag{17}$$

$$\begin{aligned} \prod_{i=n+1}^{N} c_i &= p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)p(\boldsymbol{x}_{n+2}|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n+1})\cdots p(\boldsymbol{x}_N|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N-1}) \\ &= \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n+1})}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)}\frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n+2})}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n+1})}\cdots\frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N-1})} \\ &= \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)}{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)} \\ &= p(\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) \end{aligned} \tag{18}$$

Then analogous to the equatation 12, the recursive definision of $\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)$ will be

$$
\begin{aligned}
&p(\boldsymbol{x}_n|\boldsymbol{z}_n)\sum_{k=1}^{K}[\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_{n-1})]_k p(\boldsymbol{z}_n \mid [\boldsymbol{z}_{n-1}]_k) \\
&= p(\boldsymbol{x}_n|\boldsymbol{z}_n)\sum_{k=1}^{K}\left[\frac{\boldsymbol{\alpha}(\boldsymbol{z}_{n-1})}{p(\boldsymbol{x}_n,\cdots,\boldsymbol{x}_{n-1})}\right]_k p(\boldsymbol{z}_n \mid [\boldsymbol{z}_{n-1}]_k) \\
&= \frac{1}{p(\boldsymbol{x}_n,\cdots,\boldsymbol{x}_{n-1})}p(\boldsymbol{x}_n|\boldsymbol{z}_n)\sum_{k=1}^{K}[\boldsymbol{\alpha}(\boldsymbol{z}_{n-1})]_k\, p(\boldsymbol{z}_1 \mid [\boldsymbol{z}_{n-1}]_k) \\
&= \frac{\boldsymbol{\alpha}(\boldsymbol{z}_n)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n-1})} \\
&= \frac{\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n-1})} \\
&= \hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)p(\boldsymbol{x}_n|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n-1}) \\
&= c_n\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)
\end{aligned}
\tag{19}
$$

Since $\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)$ is a normalized distribution, we can calculate c_n as the partition function of RHS of 19 as follows.

$$
c_n = \sum_{k=1}^{K}[\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)p(\boldsymbol{x}_n|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n-1})]_k \tag{20}
$$

In the same way, for $\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n)$, analogous to 13,

$$
\begin{aligned}
&\sum_{k=1}^{K}[\hat{\boldsymbol{\beta}}(\boldsymbol{z}_{n+1})]_k p(\boldsymbol{x}_{n+1}|[\boldsymbol{z}_{n+1}]_k)[p([\boldsymbol{z}_{n+1} \mid [\boldsymbol{z}_n)]_k \\
&= \frac{\sum_{k=1}^{K}[\boldsymbol{\beta}(\boldsymbol{z}_{n+1})]_k p(\boldsymbol{x}_{n+1}|[\boldsymbol{z}_{n+1}]_k)[p([\boldsymbol{z}_{n+1} \mid [\boldsymbol{z}_n)]_k}{p(\boldsymbol{x}_{n+2},\cdots,\boldsymbol{x}_N|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n+1})} \\
&= \frac{\boldsymbol{\beta}(\boldsymbol{z}_n)}{p(\boldsymbol{x}_{n+2},\cdots,\boldsymbol{x}_N|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n+1})} \\
&= \frac{p(\boldsymbol{x}_{n+1},\cdots,\boldsymbol{x}_N|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)}{p(\boldsymbol{x}_{n+2},\cdots,\boldsymbol{x}_N|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n+1})}\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n) \\
&= \frac{\frac{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)}}{\frac{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n+1})}}\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n) \\
&= \frac{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n+1})}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)}\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n) \\
&= p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n) \\
&= c_{n+1}\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n)
\end{aligned}
\tag{21}
$$

Then

$$
\begin{aligned}
\boldsymbol{\gamma}(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N) &= \frac{\boldsymbol{\alpha}(\boldsymbol{z}_n)\boldsymbol{\beta}(\boldsymbol{z}_n)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)} \\
&= \frac{\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1},\cdots,\boldsymbol{x}_N|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)} \\
&= \frac{\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n)p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)} \\
&= \hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)\hat{\boldsymbol{\beta}}(\boldsymbol{z}_n)
\end{aligned} \tag{22}
$$

and

$$
\begin{aligned}
\boldsymbol{\xi}(\boldsymbol{z}_n,\boldsymbol{z}_{n+1}) = p(\boldsymbol{z}_n,\boldsymbol{z}_{n+1}|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N) &= \frac{\boldsymbol{\alpha}(\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)\boldsymbol{\beta}(\boldsymbol{z}_{n+1})}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)} \\
&= \frac{\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)\hat{\boldsymbol{\beta}}(\boldsymbol{z}_{n+1})p(\boldsymbol{x}_{n+2},\cdots,\boldsymbol{x}_N|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n+1})}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)} \\
&= \frac{\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)\hat{\boldsymbol{\beta}}(\boldsymbol{z}_{n+1})p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n+1})} \\
&= \frac{\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)\hat{\boldsymbol{\beta}}(\boldsymbol{z}_{n+1})}{p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n+1})} \\
&= \frac{\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)\hat{\boldsymbol{\beta}}(\boldsymbol{z}_{n+1})}{p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)} \\
&= \frac{\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)p(\boldsymbol{x}_{n+1}|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1}|\boldsymbol{z}_n)\hat{\boldsymbol{\beta}}(\boldsymbol{z}_{n+1})}{c_{n+1}}
\end{aligned} \tag{23}
$$

2 Viterbi Algorithm

Let's restate the chain model expressed by equation below.

$$
p(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N,\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_N|\boldsymbol{\theta}) = p(\boldsymbol{z}_1|\boldsymbol{\pi})\prod_{n=1}^{N}p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1},A)\prod_{n=1}^{N}p(\boldsymbol{x}_n|\boldsymbol{z}_n,\boldsymbol{\phi}) \tag{24}
$$

Assume $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$ are observed and $\boldsymbol{\theta}$ is fixed. The Viterbi algorithm is used to find the set of $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_N$ that maximizes the conditional distribution, i.e.,

$$\begin{aligned}
\boldsymbol{z}_1^*, \boldsymbol{z}_2^*, \cdots, \boldsymbol{z}_N^* &= \underset{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N}{\operatorname{argmax}} \{p(\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N, \boldsymbol{\theta})\} \\
&= \underset{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N}{\operatorname{argmax}} \{p(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N | \boldsymbol{\theta})\} \\
&= \underset{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N}{\operatorname{argmax}} \{\ln p(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N | \boldsymbol{\theta})\} \\
&= \underset{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N}{\operatorname{argmax}} \left\{ \ln p(\boldsymbol{z}_1 | \boldsymbol{\pi}) + \sum_{n=1}^{N} \ln p(\boldsymbol{z}_n | \boldsymbol{z}_{n-1}, A) + \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n | \boldsymbol{z}_n, \boldsymbol{\phi}) \right\} \\
&= \underset{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N}{\operatorname{argmax}} \left\{ \sum_{k=1}^{K} [\boldsymbol{z}_1]_k \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{i=1}^{K} [\boldsymbol{z}_{n-1}]_i [\boldsymbol{z}_n]_j \ln A_{i,j} + \sum_{n=1}^{N} \sum_{i=1}^{K} [\boldsymbol{z}_n]_i \ln p(\boldsymbol{x}_n | \phi_k) \right\}
\end{aligned} \tag{25}$$

This can be considered to be a MAP estimate with the non-informative prior over $\boldsymbol{z}_n$. The above maximization is recursively defined using the property of the discrete variable $\boldsymbol{z}_n$ and distributive property of *max* operation as follows.

$$\begin{aligned}
p(\boldsymbol{z}_1^*, \boldsymbol{z}_2^*, \cdots, \boldsymbol{z}_N^*) = & \max_{k_N \text{ over } \boldsymbol{z}_N} \{\ln(\boldsymbol{x}_N | \boldsymbol{\phi}_{k_N}) + \ln A_{k_{N-1}, k_N} + \\
& \max_{k_{N-1} \text{ over } \boldsymbol{z}_{N-1}} \{ \\
& \cdots \\
& \max_{k_2 \text{ over } \boldsymbol{z}_2} \{\ln p(\boldsymbol{x}_2 | \boldsymbol{\phi}_{k_2}) + \ln A_{k_2, k_3} + \\
& \max_{k_1 \text{ over } \boldsymbol{z}_1} \{\ln p(\boldsymbol{x}_1 | \boldsymbol{\phi}_{k_1}) + \ln A_{k_1, k_2} + \ln \pi_{k_1}\} \\
& \} \cdots \} \}
\end{aligned} \tag{26}$$

And $\boldsymbol{z}_1^*, \boldsymbol{z}_2^*, \cdots, \boldsymbol{z}_N^*$ are retrieved by back tracking.

3 Kalman Filter

3.1 Model Formation : Liniear Dynamical System

The underlying graphical model is the same as the HMM and the joint probability distribution is expressed by equation 1. The difference is that $\boldsymbol{z}_n$ and $\boldsymbol{x}_n$ are continuous and they all follow Gaussian distribution as follows.

$$\begin{aligned}
p(\boldsymbol{z}_1) &= \mathcal{N}(\boldsymbol{z}_1 | \boldsymbol{\mu}_0, P_0) \\
p(\boldsymbol{z}_1 | \boldsymbol{z}_{n-1}) &= \mathcal{N}(\boldsymbol{z}_n | A\boldsymbol{z}_{n-1}, \Gamma) \\
p(\boldsymbol{x}_n | \boldsymbol{z}_n) &= \mathcal{N}(\boldsymbol{x}_n | C\boldsymbol{z}_n, \Sigma)
\end{aligned} \tag{27}$$

3.2 Forward Propagation : Kalman Filter

In this section we mainly follow Section 13.3 of PRML [2], with significantly filling the gaps between the equations. The kalman filter is defined as $p(\boldsymbol{z}_n | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$.

This can be recursively defined as in equation 12 and 19 as follows. Let

$$p(\boldsymbol{z}_n|\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n) = \hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n) = \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, V_n),$$

then

$$c_n\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n) = p(\boldsymbol{x}_n|\boldsymbol{z}_n)\int \hat{\alpha}(\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})d\boldsymbol{z}_{n-1} \tag{28}$$

where $c_n = p(\boldsymbol{x}_n|\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{n-1})$.

Now define $\boldsymbol{\mu}_n$ and V_n recursively. In order to do that, we use the following identities repeatedly. First, two identies in the closed form among joint normal distributions. Let

$$\begin{aligned} p(\boldsymbol{v}) &= \mathcal{N}(\boldsymbol{v}|\boldsymbol{\mu},\Lambda^{-1}) \\ p(\boldsymbol{w}|\boldsymbol{v}) &= \mathcal{N}(\boldsymbol{w}|M\boldsymbol{v}+\boldsymbol{b}, L^{-1}) \end{aligned} \tag{29}$$

then

$$\begin{aligned} p(\boldsymbol{w}) &= \int p(\boldsymbol{w}|\boldsymbol{v})p(\boldsymbol{v})d\boldsymbol{v} \\ &= \int \mathcal{N}(\boldsymbol{w}|M\boldsymbol{v}+\boldsymbol{b}, L^{-1})\mathcal{N}(\boldsymbol{v}|\boldsymbol{\mu},\Lambda^{-1})d\boldsymbol{v} \\ &= \mathcal{N}(\boldsymbol{w}|M\boldsymbol{\mu}+\boldsymbol{b}, L^{-1}+M\Lambda^{-1}M^T) \end{aligned} \tag{30}$$

and

$$\begin{aligned} p(\boldsymbol{v}|\boldsymbol{w}) &= \int \mathcal{N}(\boldsymbol{v}|Q(M^TL(\boldsymbol{w}-\boldsymbol{b})+\Lambda\boldsymbol{\mu}), Q) \\ Q &= (\Lambda + M^TLM)^{-1} \end{aligned} \tag{31}$$

Let $\boldsymbol{\mu}_{\boldsymbol{vw}}$ and $\mathrm{Cov}_{\boldsymbol{vw}}$ be the mean and the covariance of the joint distribution $p(\boldsymbol{v},\boldsymbol{w})$.

$$\boldsymbol{\mu}_{\boldsymbol{vw}} = \begin{bmatrix} \boldsymbol{\mu} \\ M\boldsymbol{\mu}+\boldsymbol{b} \end{bmatrix} \tag{32}$$

$$\mathrm{Cov}_{\boldsymbol{vw}} = \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1}M^T \\ M\Lambda^{-1} & L^{-1}+M\Lambda^{-1}M^T \end{bmatrix} \tag{33}$$

and

$$p(\boldsymbol{v},\boldsymbol{w}) = \mathcal{N}\Big(\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} \Big| \begin{bmatrix} \boldsymbol{\mu} \\ M\boldsymbol{\mu}+\boldsymbol{b} \end{bmatrix}, \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1}M^T \\ M\Lambda^{-1} & L^{-1}+M\Lambda^{-1}M^T \end{bmatrix}\Big) \tag{34}$$

Next, two identities of matrix inverse,

$$(P^{-1}+B^TR^{-1}B)^{-1}B^TR^{-1} = PB^T(BPB^T+R)^{-1} \tag{35}$$

$$(A+BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D+CA^{-1}B)^{-1}CA^{-1} \tag{36}$$

Equation 36 is called the Woodbury identity.

Now, we solve the integration in equation 28.

$$\begin{aligned} \int \hat{\alpha}(\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})d\boldsymbol{z}_{n-1} &= \int \mathcal{N}(\boldsymbol{z}_n|A\boldsymbol{z}_{n-1},\Gamma)\mathcal{N}(\boldsymbol{z}_{n-1}|\mu_{n-1},V_{n-1})d\boldsymbol{z}_{n-1} \\ &= \mathcal{N}(\boldsymbol{z}_n|A\boldsymbol{\mu}_{n-1},\Gamma+AV_{n-1}A^T) \\ &= \mathcal{N}(\boldsymbol{z}_n|A\boldsymbol{\mu}_{n-1},P_{n-1}) \\ &= p(\boldsymbol{z}_n|\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{n-1}) \end{aligned} \tag{37}$$

where $P_{n-1} = \Gamma + AV_{n-1}A^T$ and used identity 30. Then we solve the RHS of equation 28.

$$
\begin{aligned}
c_n\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n) &= p(\boldsymbol{x}_n|\boldsymbol{z}_n)\int \hat{\alpha}(\boldsymbol{z}_{n-1})p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})d\boldsymbol{z}_{n-1} \\
&= p(\boldsymbol{x}_n|\boldsymbol{z}_n)\mathcal{N}(\boldsymbol{z}_n|A\boldsymbol{\mu}_{n-1}, P_{n-1}) \\
&= \mathcal{N}(\boldsymbol{x}_n|C\boldsymbol{z}_n, \Sigma)\mathcal{N}(\boldsymbol{z}_n|A\boldsymbol{\mu}_{n-1}, P_{n-1}) \\
&= p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1})
\end{aligned} \tag{38}
$$

then we solve $\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n)$ as follows.

$$
\begin{aligned}
\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n) &= \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1})}{c_n} \\
&= \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1})}{p(\boldsymbol{x}_n|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1})} \\
&= p(\boldsymbol{z}_n|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}, \boldsymbol{x}_n)
\end{aligned} \tag{39}
$$

Using $\mathcal{N}(\boldsymbol{x}_n|C\boldsymbol{z}_n, \Sigma)$, $\mathcal{N}(\boldsymbol{z}_n|A\boldsymbol{\mu}_{n-1}, P_{n-1})$, and identity 31,

$$
\begin{aligned}
\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_n) &= \mathcal{N}(\boldsymbol{z}_n|V_n(C^T\Sigma^{-1}\boldsymbol{x}_n + P_{n-1}^{-1}A\boldsymbol{\mu}_{n-1}), V_n) \\
&= \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, V_n)
\end{aligned} \tag{40}
$$

where

$$
\begin{aligned}
V_n &= (P_{n-1}^{-1} + C^T\Sigma^{-1}C)^{-1} \\
&= P_{n-1} - P_{n-1}C^T(\Sigma + CP_{n-1}C^T)^{-1}CP_{n-1} \quad \text{(from Eq. 36)}
\end{aligned} \tag{41}
$$

and

$$
\begin{aligned}
\boldsymbol{\mu}_n &= V_n(C^T\Sigma^{-1}\boldsymbol{x}_n + P_{n-1}^{-1}A\boldsymbol{\mu}_{n-1}) \\
&= V_nC^T\Sigma^{-1}\boldsymbol{x}_n + V_nP_{n-1}^{-1}A\boldsymbol{\mu}_{n-1} \\
&= (P_{n-1}^{-1} + C^T\Sigma^{-1}C)^{-1}C^T\Sigma^{-1}\boldsymbol{x}_n + V_nP_{n-1}^{-1}A\boldsymbol{\mu}_{n-1} \\
&= P_{n-1}C^T(\Sigma - CP_{n-1}C^T)^{-1}\boldsymbol{x}_n + V_nP_{n-1}^{-1}A\boldsymbol{\mu}_{n-1} \text{ from Eq.35} \\
&= K_n\boldsymbol{x}_n + V_nP_{n-1}^{-1}A\boldsymbol{\mu}_{n-1}
\end{aligned} \tag{42}
$$

where $K_n = P_{n-1}C^T(\Sigma - CP_{n-1}C^T)^{-1}$, and it is called **Kalman Gain**. Using K_n we redefine V_n as follows.

$$
\begin{aligned}
V_n &= P_{n-1} - P_{n-1}C^T(\Sigma + CP_{n-1}C^T)^{-1}CP_{n-1} \\
&= P_{n-1} - K_nCP_{n-1} \\
&= (I - K_nC)P_{n-1}
\end{aligned} \tag{43}
$$

Then we continue the derivation of $\boldsymbol{\mu}_n$,

$$\begin{aligned}
\boldsymbol{\mu}_n &= K_n\boldsymbol{x}_n + V_nP_{n-1}A\boldsymbol{\mu}_{n-1} \\
&= K_n\boldsymbol{x}_n + (P_{n-1}^{-1} + C^T\Sigma^{-1}C)^{-1}P_{n-1}^{-1}A\boldsymbol{\mu}_{n-1} \\
&= K_n\boldsymbol{x}_n + \left(P_{n-1} - P_{n-1}C^T(\Sigma + CP_{n-1}C^T)^{-1}CP_{n-1}\right)P_{n-1}^{-1}A\boldsymbol{\mu}_{n-1} \quad \text{(from Eq. 36)} \\
&= K_n\boldsymbol{x}_n + \left(P_{n-1} - K_nCP_{n-1}\right)P_{n-1}^{-1}A\boldsymbol{\mu}_{n-1} \\
&= K_n\boldsymbol{x}_n + \left(I - K_nC\right)A\boldsymbol{\mu}_{n-1} \\
&= A\boldsymbol{\mu}_{n-1} + K_n(\boldsymbol{x}_n - CA\boldsymbol{\mu}_{n-1})
\end{aligned} \tag{44}$$

The first term of equation 44 is the prediction, and the second term can be considered a correction after observing $\boldsymbol{x}_n$.

c_n is defined as follows.

$$\begin{aligned}
c_n &= p(\boldsymbol{x}_n|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}) \\
&= \mathcal{N}(\boldsymbol{x}_n|CA\boldsymbol{\mu}_{n-1}, CP_{n-1}C^T + \Sigma)
\end{aligned} \tag{45}$$

where we have used the identify 30 for the marginal distribution.

The initial condition will be defined as follows.

$$\begin{aligned}
c_1\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_1) &= p(\boldsymbol{z}_1)(\boldsymbol{x}_1|\boldsymbol{z}_1) \\
&= \mathcal{N}(\boldsymbol{x}_1|C\boldsymbol{z}_1, \Sigma)\mathcal{N}(\boldsymbol{z}_1|\boldsymbol{\mu}_0, P_0)
\end{aligned} \tag{46}$$

Using $\mathcal{N}(\boldsymbol{x}_1|C\boldsymbol{z}_1, \Sigma)$, $\mathcal{N}(\boldsymbol{z}_1|\boldsymbol{\mu}_0, P_0)$, and identity 31,

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}(\boldsymbol{z}_1) &= \mathcal{N}(\boldsymbol{z}_1|V_1(C^T\Sigma^{-1}\boldsymbol{x}_1 + P_0^{-1}\boldsymbol{\mu}_0), V_1) \\
&= \mathcal{N}(\boldsymbol{z}_1|\boldsymbol{\mu}_1, V_1)
\end{aligned} \tag{47}$$

where

$$\begin{aligned}
V_1 &= (P_0^{-1} + C^T\Sigma^{-1}C)^{-1} \\
&= P_0 - P_0C^T(\Sigma + CP_0C^T)^{-1}CP_0 \quad \text{(from Eq. 36)} \\
&= P_0 - K_1CP_0 \\
&= (I - K_1C)P_0
\end{aligned} \tag{48}$$

and

$$\begin{aligned}
\boldsymbol{\mu}_1 &= V_1(C^T\Sigma^{-1}\boldsymbol{x}_1 + P_0^{-1}\boldsymbol{\mu}_0) \\
&= V_1C^T\Sigma^{-1}\boldsymbol{x}_1 + V_1P_0^{-1}\boldsymbol{\mu}_0 \\
&= (P_0^{-1} + C^T\Sigma^{-1}C)^{-1}C^T\Sigma^{-1}\boldsymbol{x}_1 + (I - K_1C)P_0P_0^{-1}\boldsymbol{\mu}_0 \\
&= P_0C^T(\Sigma - CP_0C^T)^{-1}\boldsymbol{x}_1 + \boldsymbol{\mu}_0 - K_1C\boldsymbol{\mu}_0 \\
&= K_1\boldsymbol{x}_1 + \boldsymbol{\mu}_0 - K_1C\boldsymbol{\mu}_0 \\
&= \boldsymbol{\mu}_0 + K_1(\boldsymbol{x}_1 - C\boldsymbol{\mu}_0)
\end{aligned} \tag{49}$$

4 Rauch-Tung-Striebel Smoother

Here we mainly follow the descirption of [4], which is more elegant and readable than PRML[2]. PRML tries to use HMM's α,β notations, but it messes up at the very end between (13.103) and (13.104) when deriving $p(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$.

In the previous chapter we dealt with $p(\boldsymbol{z}_n | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$ with the forward propagation. In this chapter we deal with $p(\boldsymbol{z}_n | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$, i.e., conditioned on all the samples. We also derive $p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1} | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$, which is required for EM-like algorithms.

We follow the process in [4]. We assume $p(\boldsymbol{z}_n | \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ is available. First we get $p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1} | \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ from $p(\boldsymbol{z}_{n+1} | \boldsymbol{z}_n)$ and $p(\boldsymbol{z}_n | \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$. Then we get $p(\boldsymbol{z}_n | \boldsymbol{z}_{n+1}, \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ which is equivalent to $p(\boldsymbol{z}_n | \boldsymbol{z}_{n+1}, \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ as $\boldsymbol{z}_{n+1}$ blocks $\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N$. See figure 2

From $p(\boldsymbol{z}_n | \boldsymbol{z}_{n+1}, \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ and $p(\boldsymbol{z}_{n+1} | \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$, which is assumed to be available, we get $p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1} | \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$. By marginalizing it, we obtain $p(\boldsymbol{z}_n | \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$.

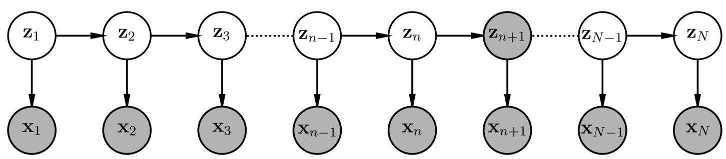

Figure 2: $\boldsymbol{z}_{n+1}$ blocking $\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N$ from $\boldsymbol{x}_n$

By definition,

$$\begin{aligned} p(\boldsymbol{z}_{n+1} | \boldsymbol{z}_n) &= \mathcal{N}(\boldsymbol{z}_{n+1} | A\boldsymbol{z}_n, \Gamma) \\ p(\boldsymbol{z}_n | \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) &= \mathcal{N}(\boldsymbol{z}_{n+1} | \boldsymbol{\mu}_n, V_n) \end{aligned} \tag{50}$$

Then using identity 31,

$$\begin{aligned} p(\boldsymbol{z}_n | \boldsymbol{z}_{n+1}, \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) =& p(\boldsymbol{z}_n | \boldsymbol{z}_{n+1}, \boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) \\ =& \mathcal{N}(\boldsymbol{z}_n | \boldsymbol{\mu}'_n, V'_n) \end{aligned} \tag{51}$$

where

$$\begin{aligned} V'_n &= \left(V_n^{-1} + A^T \Gamma^{-1} A\right)^{-1} \\ &= V_n - V_n A^T \left(\Gamma + A V_n A^T\right)^{-1} A V_n \\ &= V_n - K'_n A V_n \\ &= (I - K'_n A) V_n \end{aligned} \tag{52}$$

and

$$\begin{aligned} K'_n &= V_n A^T \left(\Gamma + A V_n A^T\right)^{-1} \\ &= (V_n^{-1} + A \Gamma^{-1} A)^{-1} A^T \Gamma^{-1} \quad \text{from identity35} \end{aligned} \tag{53}$$

and

$$\begin{aligned}\boldsymbol{\mu}'_n &= V'_n \left(A^T\Gamma^{-1}\boldsymbol{z}_{n+1} + V_n^{-1}\boldsymbol{\mu}_n\right) \\ &= (V_n^{-1} + A^T\Gamma^{-1}A)^{-1}A^T\Gamma^{-1}\boldsymbol{z}_{n+1} + (I - K'_nA)V_nV_n^{-1}\boldsymbol{\mu}_n \\ &= K'_n\boldsymbol{z}_{n+1} + (I - K'_nA)\boldsymbol{\mu}_n \quad \text{from equation53}\end{aligned} \tag{54}$$

Now we are ready to derive $p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1}|\boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ and $p(\boldsymbol{z}_n|\boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$. Let $p(\boldsymbol{z}_n|\boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) = \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n^\gamma, V_n^\gamma)$. From the equation 34,

$$\begin{aligned}p(\boldsymbol{z}_{n+1}, \boldsymbol{z}_n|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) &= p(\boldsymbol{z}_n|\boldsymbol{z}_{n+1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)p(\boldsymbol{z}_{n+1}|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N) \\ &= \mathcal{N}(\boldsymbol{z}_n|K'_n\boldsymbol{z}_{n+1} + (I - K'_nA)\boldsymbol{\mu}_n, V'_n)\mathcal{N}(\boldsymbol{z}_{n+1}|\boldsymbol{\mu}_{n+1}^\gamma, V_{n+1}^\gamma) \\ &= \mathcal{N}\Bigg(\begin{bmatrix} \boldsymbol{z}_{n+1} \\ \boldsymbol{z}_n \end{bmatrix} \Bigg| \begin{bmatrix} \boldsymbol{\mu}_{n+1}^\gamma \\ K'_n\boldsymbol{\mu}_{n+1}^\gamma + (I - K'_nA)\boldsymbol{\mu}_n \end{bmatrix}, \begin{bmatrix} V_{n+1}^\gamma & V_{n+1}^\gamma {K'_n}^T \\ K'_nV_{n+1}^\gamma & V'_n + K'_nV_{n+1}^\gamma K'_n \end{bmatrix} \Bigg)\end{aligned} \tag{55}$$

and

$$\begin{aligned}p(\boldsymbol{z}_n|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) &= \mathcal{N}\left(\boldsymbol{z}_n|K'_n\boldsymbol{\mu}_{n+1}^\gamma + (I - K'_nA)\boldsymbol{\mu}_n,\ V'_n + K'_nV_{n+1}^\gamma K'_n\right) \\ &= \mathcal{N}\left(\boldsymbol{z}_n|\boldsymbol{\mu}_n^\gamma, V_n^\gamma\right)\end{aligned} \tag{56}$$

with the initial condition

$$\begin{aligned}p(\boldsymbol{z}_N|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) &= \mathcal{N}\left(\boldsymbol{z}_N|\boldsymbol{\mu}_N^\gamma, V_N^\gamma\right) \\ &= \mathcal{N}\left(\boldsymbol{z}_N|\boldsymbol{\mu}_N, V_N\right)\end{aligned} \tag{57}$$

For the EM algorithm, we need $\mathbb{E}[\boldsymbol{z}_n]$, $\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_{n-1}^T]$, and $\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^T]$ for the given $\boldsymbol{\mu}_0$, P_0, A, Γ, C, and Σ. for the E-step. They are given by

$$\begin{aligned}\mathbb{E}[\boldsymbol{z}_n] &= \boldsymbol{\mu}_n^\gamma \\ \mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^T] &= V_n^\gamma + \boldsymbol{\mu}_n^\gamma {\boldsymbol{\mu}^\gamma}_n^T \\ \mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_{n-1}^T] &= V_n^\gamma {K'}_{n-1}^T + \boldsymbol{\mu}_n^\gamma {\boldsymbol{\mu}^\gamma}_{n-1}^T\end{aligned} \tag{58}$$

For the M-step, which maximizes $\boldsymbol{\mu}_0$, P_0, A, Γ, C, and Σ, please see pp 642, 643, 13.3.2 of PRML [2].

References

[1] David Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2011.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer, 1 edition, 2007.

[3] Simon J. D. Prince. *Computer Vision: Models, Learning, and Inference.* Cambridge University Press, 2012.

[4] Simmo Säkkä. Beyesian estimation of time-varying systems, lecture 7: Optimal smoothing aalto university. https://users.aalto.fi/ ssarkka/course_k2011/pdf/handout7.pdf, 03 2011. Accessed: 2020-03-30.