1 HMM

1.1 What is Hidden Markov Chain

1. Discrete-state Markov Chain with hidden state $z_t \in \{1, 2, ..., K\}$
2. Observation model $p(x_t|z_t)$

Joint distribution of the hidden states and observations over window $1, 2, ..., T$:

$$\begin{aligned} p(z_{1:T}, x_{1:T}) &= p(z_{1:T})p(x_{1:T}|z_{1:T}) \\ &= \Big[p(z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})\Big]\Big[\prod_{t=1}^{T} p(x_t|z_t)\Big] \end{aligned} \tag{1}$$

HMM inference: $p(z_{1:T}|x_{1:T})$, observation $\rightarrow$ hidden state; data $\rightarrow$ parameters.

1.2 Inference Problems

1. Filtering: $p(z_t|x_{1:t})$, online; recursively as data stream in.
2. Smoothing: $p(z_t|x_{1:T})$, offline; condition on past and future(whole dataset) $\rightarrow$ reduce uncertainty.
3. MAP: $\arg\max_{z_{1:T}} p(z_{1:T}|x_{1:T})$; viterbi decoding.
4. Fixed lag smoothing: $p(z_{t-l}|x_{1:t}), l > 0$ is called the lag. This gives better performance than filtering, but incurs a slight delay. Knowing more observation to filtering.
5. Prediction: $p(z_{t+h}|x_{1:t}), h > 0$; predict future hidden state by past observation.

$$p(z_{t+h}|x_{1:t}) = \sum_{z_{t+h-1}} ... \sum_{z_{t+1}} p(z_{t+h}|z_{t+h-1})...p(z_{t+1}|z_t)p(z_t|x_{1:t}) \tag{2}$$

6. Prediction for future observation: $p(x_{t+h}|x_{1:t})$; predict future observation by past observation.

$$p(x_{t+h}|x_{1:t}) = \sum_{z_{t+h}} p(x_{t+h}|z_{t+h})p(z_{t+h}|x_{1:t}) \tag{3}$$

7. Posterior samples: $z_{1:T} \sim p(z_{1:T}|x_{1:T})$;
8. Probability of evidence: $p(x_{1:T}) = \sum_{z_{1:T}} p(z_{1:T}, x_{1:T})$; evidence $\rightarrow$ data.

1.3 Filtered Marginal $\alpha_t = p(z_t|x_{1:t})$

Forward Algorithm, Predict-Update Circle.

1. Predict:

$$p(z_t = j|x_{1:t-1}) = \sum_i p(z_t = j|z_{t-1} = i)p(z_{t-1} = i|x_{1:t-1}) \tag{4}$$

2. Update:

$$\begin{aligned}
p(z_t = j|x_{1:t}) &= p(z_t = j|x_t, x_{1:t-1}) \\
&= \frac{p(z_t = j, x_t, x_{1:t-1})}{p(x_t, x_{1:t-1})} \\
&= \frac{p(x_t|z_t = j, x_{1:t-1})p(z_t = j, x_{1:t-1})}{p(x_t, x_{1:t-1})} \\
&= \frac{p(x_t|z_t = j)p(z_t = j|x_{1:t-1})p(x_{1:t-1})}{p(x_t, x_{1:t-1})} \\
&= \frac{p(x_t|z_t = j)p(z_t = j|x_{1:t-1})}{p(x_t|x_{1:t-1})} \\
&= \frac{p(x_t|z_t = j)p(z_t = j|x_{1:t-1})}{\sum_j p(x_t, z_t = j|x_{1:t-1})} \\
&= \frac{p(x_t|z_t = j)p(z_t = j|x_{1:t-1})}{\sum_j p(x_t|z_t = j, x_{1:t-1})p(z_t = j|x_{1:t-1})} \\
&= \frac{p(x_t|z_t = j)p(z_t = j|x_{1:t-1})}{\sum_j p(x_t|z_t = j)p(z_t = j|x_{1:t-1})}
\end{aligned} \tag{5}$$

3. Matrix-Vector

   Local evidence at time t:

$$\psi_t = p(x_t|z_t) \in R_t^K \tag{6}$$

   Transition matrix:

$$\Psi \in R_t^{KK} \tag{7}$$

   Predict:

$$\begin{aligned}
p(z_t = j|x_{1:t-1}) &= \sum_i p(z_t = j|z_{t-1} = i)p(z_{t-1} = i|x_{1:t-1}) \\
&= \sum_i \Psi(i, j)\alpha_{t-1}(i)
\end{aligned} \tag{8}$$

Update:

$$
\begin{aligned}
p(z_t = j|x_{1:t}) &= \frac{p(x_t|z_t = j)p(z_t = j|x_{1:t-1})}{\sum_j p(x_t|z_t = j)p(z_t = j|x_{1:t-1})} \\
&= \frac{\psi_t(j)\sum_i \Psi(i,j)\alpha_{t-1}(i)}{\sum_j \psi_t(j)\sum_i \Psi(i,j)\alpha_{t-1}(i)} \\
&\propto \psi_t(j)\sum_i \Psi(i,j)\alpha_{t-1}(i) \\
&= normalize\Big(\psi_t(j)\sum_i \Psi(i,j)\alpha_{t-1}(i)\Big)
\end{aligned} \tag{9}
$$

$$
\alpha_t = normalize\Big(\psi_t \odot (\Psi^T \alpha_{t-1})\Big) \tag{10}
$$

$$
\alpha_1 = normalize\Big(\psi_1 \odot \pi\Big) \tag{11}
$$

1.4 Smoothed Marginal $p(z_t|x_{1:T})$

Forwards-Backwards Algorithm, using offline inference. We introduce the conditional likelihood of future evidence given that the hidden state at time t, $\beta_t(j) = p(x_{t+1:T}|z_t = j)$

1. Future evidence(backward algorithm):

$$
\beta_t(j) = p(x_{t+1:T}|z_t = j) \tag{12}
$$

$$
\begin{aligned}
\beta_{t-1}(i) &= p(x_{t:T}|z_{t-1} = i) \\
&= \sum_j p(z_t = j, x_t, x_{t+1:T}|z_{t-1} = i) \\
&= \sum_j \frac{p(x_t, x_{t+1}, z_t = j, z_{t-1} = i)}{p(z_{z_{t-1}=i})} \\
&= \sum_j p(x_t, x_{t+1:T}|z_t = j, z_{t-1} = i)p(z_t = j|z_{t-1} = i) \\
&= \sum_j p(x_t, x_{t+1:T}|z_t = j)p(z_t = j|z_{t-1} = i) \\
&= \sum_j p(x_t|z_t = j)p(x_{t+1:T}|x_t, z_t = j)p(z_t = j|z_{t-1} = i) \\
&= \sum_j \psi_t(j)\beta_t(j)\Psi(i,j)
\end{aligned} \tag{13}
$$

$$
\beta_{t-1} = \Psi(\psi_t \odot \beta_t) \tag{14}
$$

$$
\beta_T(i) = p(x_{T+1:T}|z_T = i) = 1 \tag{15}
$$

2. Smoothed posterior marginal(forward-backward algorithm):

$$\begin{aligned}
\gamma_t(j) &= p(z_t = j|x_{1:T}) \\
&= p(z_t = j|x_{1:t}, x_{t+1:T}) \\
&= \frac{p(z_t = j, x_{1:t}, x_{t+1:T})}{p(x_{1:T})} \\
&= \frac{p(z_t = j, x_{1:t})p(x_{t+1:T}|z_t = j, x_{1:t})}{p(x_{1:T})} \\
&= \frac{p(z_t = j|x_{1:t})p(x_{1:t})p(x_{t+1:T}|z_t = j)}{p(x_{1:T})} \\
&\propto p(z_t = j|x_{1:t})p(x_{t+1:T}|z_t = j) \\
&= normalize\Big(\alpha_t(j)\beta_t(j)\Big)
\end{aligned} \tag{16}$$

1.5 Two-slice smoothed marginal $p(z_t = i, z_{t+1} = j|x_{1:T})$

1. Two-slice smoothed marginal:

$$\begin{aligned}
\xi_{t,t+1}(i,j) &= p(z_t = i, z_{t+1} = j|x_{1:T}) \\
&= p(z_t = i|x_{1:T})p(z_{t+1} = j|z_t = i, x_{1:T}) \\
&= \gamma_t(i)p(z_{t+1} = j|z_t = i, x_{1:T}) \\
&= \frac{\gamma_t(i)p(z_{t+1} = j, z_t = i, x_{1:T})}{p(z_t = i, x_{1:T})} \\
&= \frac{\gamma_t(i)p(x_{1:T}|z_{t+1} = j, z_t = i)p(z_{t+1} = j|z_t = i)p(z_t = i)}{p(z_t = i, x_{1:T})} \\
&= \frac{\gamma_t(i)p(x_{t+1:T}|z_{t+1} = j)p(z_{t+1} = j|z_t = i)}{p(x_{1:T}|z_t = i)} \\
&= \frac{\gamma_t(i)p(x_{t+1}, x_{t+2:T}|z_{t+1} = j)p(z_{t+1} = j|z_t = i)}{p(x_{1:T}|z_t = i)} \\
&= \frac{\gamma_t(i)p(x_{t+1}|z_{t+1} = j)p(x_{t+2:T}|z_{t+1} = j)p(z_{t+1} = j|z_t = i)}{p(x_{1:T}|z_t = i)} \\
&= \frac{\alpha_t(i)\beta_t(i)\psi_{t+1}(j)\beta_{t+1}(j)\Psi(i,j)}{p(x_{t+1:T}|z_t = i)} \\
&= \frac{\alpha_t(i)\beta_t(i)_{t+1}(j)\beta_{t+1}(j)\Psi(i,j)}{\beta_t(i)} \\
&= \alpha_t(i)\psi_{t+1}(j)\beta_{t+1}(j)\Psi(i,j)
\end{aligned} \tag{17}$$

$$\xi_{t,t+1} = \Psi(i,j) \odot \Big(\alpha_t(\psi_{t+1} \odot \beta_{t+1})^T\Big) \tag{18}$$

1.6 HMM parameter estimation

1. Parameter set:

$$\theta = (\pi, A, B) \tag{19}$$

Initial distribution π

$$\pi(i) = p(z_1 = i) \tag{20}$$

Transition matrix A

$$A(i, j) = p(z_{t+1} = j | z_t = i) \tag{21}$$

Class-conditional densities B (local evidence)

$$B(j, l) = p(x_t = l | z_t = j) \tag{22}$$

2. Full data observed (we know all $x_{1:T}$ and $z_{1:T}$):

$$p(z_{1:T}|\theta) = \prod_{j=1}^{K} (\pi(j))^{I(x_1=j)} \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} (A(j,k))^{I(z_{t-1}=j, z_t=k)} \tag{23}$$

For N iid multiple sequences:

$$D = \{[z_{i,1}, ...z_{i,T_i}]\}_{i=1}^{N} \tag{24}$$

The likelihood is:

$$\begin{aligned} \log(D|A, \pi) &= \sum_{i=1}^{N} \log(p(z_{i,1:T}|\theta)) \\ &= \sum_{j=1}^{K} N_j \log(\pi(j)) + \sum_{j=1}^{K} \sum_{k=1}^{K} N_{jk} \log(A(j,k)) \end{aligned} \tag{25}$$

$$N_j = \sum_{i=1}^{N} I(z_{i,1} = j) \tag{26}$$

$$N_{jk} = \sum_{i=1}^{N} \sum_{t=2}^{T_i} I(z_{i,t-1} = j, z_{i,t} = k) \tag{27}$$

Then we get MLE of π and A:

$$\hat{\pi}(j) = \frac{N_j}{\sum_{j=1}^{K} N_j} \tag{28}$$

$$\hat{A}(j,k) = \frac{N_{jk}}{\sum_{k=1}^{K} N_{jk}} \tag{29}$$

Why $\sum_{k=1}^{K}$? Because we fixed $z_{t-1} = j$.

The MLE for B:

$$\hat{B}(j,l) = \frac{N_{jl}^{x}}{\sum_{l=1}^{L} N_{jl}^{x}} \tag{30}$$

$$N_{jl}^{x} = \sum_{i=1}^{N} \sum_{t=1}^{T} I(x_{i,t} = l, z_{i,t} = j) \tag{31}$$

3. When z_t are not observed $\rightarrow$ EM algorithm.

   E-step:

   The expected log-likelihood:

$$
\begin{aligned}
Q(\theta, \theta^{old}) &= \sum_{j=1}^{K} E[N_j] \log(\pi(j)) + \sum_{j=1}^{K} \sum_{k=1}^{K} E[N_{jk}] \log(A(j,k)) \\
&+ \sum_{l=1}^{L} \sum_{j=1}^{K} E[N_{jl}^x] \log(B(j,l))
\end{aligned} \tag{32}
$$

$$
\begin{aligned}
E[N_j] &= \sum_{i=1}^{N} p(z_{i,1} = j | [x_{i,1} ... x_{i,T}], \theta^{old}) \\
&= \sum_{i=1}^{N} \gamma_{i,1}(j)
\end{aligned} \tag{33}
$$

$$
\begin{aligned}
E[N_{jk}] &= \sum_{i=1}^{N} \sum_{t=2}^{T_i} p(z_{i,t-1} = j, z_{i,t} = k | [x_{i,1} ... x_{i,T}], \theta^{old}) \\
&= \sum_{i=1}^{N} \sum_{t=2}^{T_i} \xi_{t-1,t}(j,k)
\end{aligned} \tag{34}
$$

$$
E[N_{jl}^x] = \sum_{i=1}^{N} \sum_{t=1}^{T_i} p(x_{i,t} = l, z_{i,t} = j | [x_{i,1} ... x_{i,T}], \theta^{old}) \tag{35}
$$

   $E[N_{jl}^x]$ is the likelihood of $x_{i,t}$ and $z_{i,t}$, we can compute this with the relationship between $x_{i,t}$ and $z_{i,t}$. For example, $x_{i,t}$ is the gaussian mixture of $z_{i,t}$.

   M-step:

$$
\hat{\pi}(j) = \frac{E[N_j]}{N} \tag{36}
$$

$$
\hat{A}(j,k) = \frac{E[N_{jk}]}{\sum_{k=1}^{K} E[N_{jk}]} \tag{37}
$$

   $\hat{B}(j,l)$ could be the solution of $\nabla_B \sum_{l=1}^{L} \sum_{j=1}^{K} E[N_{jl}^x] \log(B(j,l)) = 0$