ECE276A: Sensing & Estimation in Robotics
Lecture 18: Hidden Markov Models

Instructor:

Nikolay Atanasov: natanasov@ucsd.edu

Teaching Assistants:

Qiaojun Feng: qif007@eng.ucsd.edu

Tianyu Wang: tiw161@eng.ucsd.edu

Ibrahim Akbar: iakbar@eng.ucsd.edu

You-Yi Jau: yjau@eng.ucsd.edu

Harshini Rajachander: hrajacha@eng.ucsd.edu

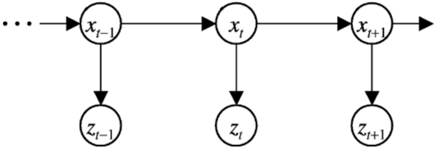
UC San Diego

JACOBS SCHOOL OF ENGINEERING

Electrical and Computer Engineering

Hidden Markov Model (HMM)

- Same **graphical model** as Bayes filter
- Discrete states: $x_t \in \{1, \ldots, N\}$



- Observations can be either:
  - Discrete: $z_t \in \{1, \ldots, M\}$
  - Continuous: $z_t \in \mathbb{R}^m$
- **Prior**: $\pi \in [0,1]^N$ with $\pi(i) := \mathbb{P}(x_0 = i)$
- **Motion model**: due to the Markov assumption, $p_f(x_{t+1} \mid x_t)$ can be specified by a transition matrix $A \in \mathbb{R}^{N \times N}$ where $A(i,j) = p_f(i \mid x_t = j)$
- **Observation model**:
  - Discrete: $B \in \mathbb{R}^{M \times N}$ such that $B(i,j) = p_h(i \mid x_t = j)$
  - Continuous: $p_h(z_t \mid x_t = j) = \phi(z_t; \mu_j, \Sigma_j)$
- **Model parameters**: $\theta := (\pi, A, B)$ or $\theta := (\pi, A, \{\mu_j, \Sigma_j\}_{j=1}^N)$

The Three Main HMM Problems

P1 Given an observation sequence $z_{0:T}$ and model parameters $\theta := (\pi, A, B)$, how do we efficiently compute the likelihood $p_\theta(z_{0:T})$ of the observation sequence?

P2 Given an observation sequence $z_{0:T}$ and model parameters $\theta := (\pi, A, B)$, how do we choose a corresponding state sequence $x_{0:T}$ which best "explains" the observations?

P3 How do we adjust the model parameters $\theta := (\pi, A, B)$ to maximize $p_\theta(z_{0:T})$?

Forward-backward Procedure

- **Smoothing for HMMs**: given $\theta := (\pi, A, B)$ and $z_{0:T}$, compute the observation likelihood $p_\theta(z_{0:T})$

- **Joint probability density function**:

$$p_\theta(x_{0:T}, z_{0:T}) = \underbrace{\pi(x_0)}_{\text{prior}} \prod_{t=0}^{T} \underbrace{B(z_t, x_t)}_{\text{observation model}} \prod_{t=1}^{T} \underbrace{A(x_t, x_{t-1})}_{\text{motion model}}$$

- **Idea**: marginalize $x_{0:T}$ from the joint pdf to obtain the observation likelihood:

$$p_\theta(z_{0:T}) = \sum_{x_{0:T} \in \{1,\ldots,N\}^{T+1}} p_\theta(x_{0:T}, z_{0:T})$$

  - Summing over all possible $x_{0:T}$ requires $O(N^T)$ operations
  - Fortunately, $p_\theta(z_{0:T})$ can be computed recursively using $O(N^2 T)$ operations via the **forward-backward procedure**

Forward Procedure

- Define: $\alpha_t(i) := p(z_{0:t}, x_t = i)$
- Initialize: $\alpha_0(i) = p(z_0 = j, x_0 = i) = B(j, i)\pi(i)$
- Induction:

$$\begin{aligned}
\alpha_{t+1}(i) &= p(z_{0:t+1}, x_{t+1} = i) \\
&= \sum_{j=1}^{N} p(z_{0:t}, x_t = j) p_f(i \mid x_t = j) p_h(z_{t+1} \mid x_{t+1} = i) \\
&= \underbrace{B(z_{t+1}, i)}_{\text{Update}} \underbrace{\sum_{j=1}^{N} A(i, j)\alpha_t(j)}_{\text{Predict}}
\end{aligned}$$

- Termination: $p(z_{0:T}) = \sum_{i=1}^{N} p(z_{0:T}, x_T = i) = \sum_{i=1}^{N} \alpha_T(i)$
- Complexity: $O(N^2 T)$
  - N times for each state, perform N multiplications in the sum over T time periods

Backward Procedure

- Define: $\beta_t(i) := p(z_{t+1:T} \mid x_t = i)$
- Initialize: $\beta_T(i) = 1$
- Induction:

$$\begin{aligned}\beta_t(i) &= p(z_{t+1:T} \mid x_t = i) \\ &= \sum_{j=1}^{N} p_h(z_{t+1} \mid x_{t+1} = j) p_f(j \mid x_t = i) p(z_{t+2:T} \mid x_{t+1} = j) \\ &= \sum_{j=1}^{N} B(z_{t+1}, j) A(j, i) \beta_{t+1}(j)\end{aligned}$$

- Termination: $p(z_{0:T}) = \sum_{i=1}^{N} p(z_{0:T}, x_0 = i)\pi(i) = \sum_{i=1}^{N} \beta_0(i)\pi(i)$
- Complexity: $O(N^2 T)$
  - N times for each state, perform N multiplications in the sum over T time periods

Inference in HMMs

- **Forward Procedure** (Filtering): computes marginals online using only the available observations:

$$p(x_t = i \mid z_{0:t}) = \frac{p(x_t = i, z_{0:t})}{p(z_{0:t})} = \frac{\alpha_t(i)}{\sum_j \alpha_t(j)}$$

- **Forward-Backward Procedure** (Smoothing): computes marginals using the entire observation sequence:

$$\gamma_t(i) := p(x_t = i \mid z_{0:T}) = \frac{p(x_t = i, z_{0:T})}{p(z_{0:T})} = \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)}$$

- **Viterbi Decoding**: computes the most-likely explanation (state sequence) of the observations:

$$x^*_{0:T} = \arg\max_{x_{0:T}} p(x_{0:T}, z_{0:T})$$

Pair of States Density

- **Pair of States Density**:

$$\xi_t(i,j) := p(x_t = j, x_{t+1} = i \mid z_{0:T}) = \frac{\alpha_t(j)A(i,j)B(z_{t+1},i)\beta_{t+1}(i)}{\sum_{i',j'} \alpha_t(j')A(i',j')B(z_{t+1},i')\beta_{t+1}(i')}$$

- The joint pdf between a pair of states x_t and x_{t+1} conditioned on the complete observation sequence $z_{0:T}$ (smoothing) is:

$$\begin{aligned}
\xi_t(i,j) &:= p(x_t = j, x_{t+1} = i \mid z_{0:T}) \propto p(x_t = j, x_{t+1} = i, z_{0:T}) \\
&\overset{\text{Conditional}}{\underset{\text{Probability}}{=\!=\!=}} p(z_{0:T} \mid x_t = j, x_{t+1} = i) p(x_{t+1} = i \mid x_t = j) p(x_t = j) \\
&\overset{\text{Markov}}{\underset{\text{Assumption}}{=\!=\!=}} \underbrace{p(z_{0:t} \mid x_t = j) p(x_t = j)}_{\alpha_t(j)} A(i,j) \underbrace{p_h(z_{t+1} \mid x_{t+1} = i)}_{B(z_{t+1},i)} \underbrace{p(z_{t+2:T} \mid x_{t+1} = i)}_{\beta_{t+1}(i)} \\
&= \alpha_t(j)A(i,j)B(z_{t+1},i)\beta_{t+1}(i)
\end{aligned}$$

Viterbi Decoding

$$\delta_t(i) := \max_{x_{0:t-1}} p(x_{0:t-1}, x_t = i, z_{0:t})$$

Likelihood of the observed sequence with the most likely state assignment up to $t-1$

$$\psi_t(i) := \arg\max_{x_{t-1}} \max_{x_{0:t-2}} p(x_{0:t-1}, x_t = i, z_{0:t})$$

State from the previous time that leads to the maximum for the current state at time t

- **Initialize**: $$\delta_0(i) = p(z_0 \mid x_0 = i)p(x_0 = i) = B(z_0, i)\pi(i)$$ $$\psi_0(i) = 0$$
- **Forward Pass** for $t = 1, \ldots, T$

$$\delta_t(i) = \max_j p(z_t \mid x_t = i) p_f(x_t = i \mid x_{t-1} = j)\delta_{t-1}(j) = \max_j B(z_t, i)A(i,j)\delta_{t-1}(j)$$

$$\psi_t(i) = \arg\max_j p(z_t \mid x_t = i) p_f(x_t = i \mid x_{t-1} = j)\delta_{t-1}(j) = \arg\max_j B(z_t, i)A(i,j)\delta_{t-1}(j)$$

$$\boxed{p(x^*_{0:T}, z_{0:T}) = \max_i \delta_T(i)} \qquad \boxed{x^*_T = \arg\max_i \delta_T(i)}$$

- **Backward Pass** for $t = T-1, \ldots, 0$: $\boxed{x^*_t = \psi_{t+1}(x^*_{t+1})}$

HMM Parameter Estimation

- Given labeled data $D := \left\{ \left(z_{0:T}^{(k)}, x_{0:T}^{(k)} \right) \right\}_{k=1}^{K}$, estimate the model parameters $\theta := (\pi, A, B)$
- For a model with N hidden states and M observations there are: $N - 1 + N(N-1) + N(M-1)$ parameters
- **Maximum Likelihood Estimation**:

$$\max_{\pi, T, B} \sum_{k=1}^{K} \log p\left(z_{0:T}^{(k)}, x_{0:T}^{(k)}\right) \xlongequal[\text{Assumption}]{\text{Markov}} \sum_{k=1}^{K} \log \left[\pi\left(x_0^{(k)}\right) \prod_{t=1}^{T} B\left(z_t^{(k)}, x_t^{(k)}\right) A\left(x_t^{(k)}, x_{t-1}^{(k)}\right) \right]$$

$$= \sum_{k=1}^{K} \log \pi\left(x_0^{(k)}\right) + \sum_{k=1}^{K} \sum_{t=1}^{T} \log B\left(z_t^{(k)}, x_t^{(k)}\right) + \sum_{k=1}^{K} \sum_{t=1}^{T} \log A\left(x_t^{(k)}, x_{t-1}^{(k)}\right)$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}\left\{x_0^{(k)} = i\right\} \log \pi(i) + \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}\left\{z_t^{(k)} = i, x_t^{(k)} = j\right\} \log B(i,j)$$

$$+ \sum_{j=1}^{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}\left\{x_t^{(k)} = j, x_{t-1}^{(k)} = i\right\} \log A(j,i)$$

- The parameters can be estimated separately even for each state and state-observation pair

HMM Parameter Estimation

- Given labeled data $D := \left\{ \left(z_{0:T}^k, x_{0:T}^k \right) \right\}_{k=1}^K$, estimate $\theta := (\pi, A, B)$
- **Maximum Likelihood Estimation**:

$$\pi(j) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\left\{ x_0^{(k)} = j \right\}$$

$$A(i,j) = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}\left\{ x_t^{(k)} = i, x_{t-1}^{(k)} = j \right\}}{\sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}\left\{ x_{t-1}^{(k)} = j \right\}} \qquad B(i,j) = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}\left\{ z_t^{(k)} = i, x_t^{(k)} = j \right\}}{\sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}\left\{ x_t^{(k)} = j \right\}}$$

- **Continuous Observations**: $p_h(z_t \mid x_t = j) = \phi(z_t; \mu_j, \Sigma_j)$, where:

$$\mu_j = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T} z_t^{(k)} \mathbb{1}\left\{ x_t^{(k)} = j \right\}}{\sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}\left\{ x_t^{(k)} = j \right\}} \qquad \Sigma_j = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T} \left(\mu_j - z_t^{(k)} \right) \left(\mu_j - z_t^{(k)} \right)^T \mathbb{1}\left\{ x_t^{(k)} = j \right\}}{\sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}\left\{ x_t^{(k)} = j \right\}}$$

Baum-Welch Algorithm (EM for HMMs)

- Given unlabeled data $D := \left\{z_{0:T}^{(k)}\right\}_{k=1}^{K}$, jointly estimate the parameters $\theta := (\pi, A, B)$ and the hidden variables $x_{0:T}$
- **Baum-Welch procedure**: use **Jensen's inequality** to obtain a lower bound:

$$\max_{\theta:=(\pi,A,B)} \log p_\theta(z_{0:T}) = \max_\theta \log \sum_{x_{0:T}} p_\theta(z_{0:T}, x_{0:T}) \frac{q(x_{0:T})}{q(x_{0:T})}$$
$$\geq \max_\theta \sum_{x_{0:T}} q(x_{0:T}) \log \frac{p_\theta(z_{0:T}, x_{0:T})}{q(x_{0:T})}$$

- **E-step**: we saw that the choice for q that makes the lower bound touch the objective function at the current parameters $\theta^{(0)}$ is the pdf of the hidden variables conditioned on the data: $\boxed{q^*(x_{0:T}) := p_{\theta^{(0)}}(x_{0:T} \mid z_{0:T})}$
- **Example**: used for word alignment in machine translation: given pairs of sentences in 2 languages, the model translates word by word. The word alignment variables are hidden and the parameters are learned using EM

Baum-Welch Algorithm

- **Initialization**: $\theta^{(0)} = \left(\pi^{(0)}, A^{(0)}, B^{(0)}\right)$
- **E-step**: Given the current parameters $\theta^{(l)}$, compute the hidden state pdf $p_{\theta^{(l)}}(x_{0:T} \mid z_{0:T})$ via the **Forward-Backward procedure**: $\alpha_t^{(k)}(i)$, $\beta_t^{(k)}(j)$, $\gamma_t^{(k)}(i)$, $\xi_t^{(k)}(i,j)$
- **M-step**: compute $\theta^{(l+1)}$ based on the inferred hidden states (**Weighted MLE**):

$$\pi^{(l+1)}(j) = \frac{1}{K}\sum_{k=1}^{K}\gamma_0^{(k)}(j)$$

$$A^{(l+1)}(i,j) = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T}\xi_{t-1}^{(k)}(i,j)}{\sum_{k=1}^{K}\sum_{t=1}^{T}\gamma_{t-1}^{(k)}(j)} \qquad B^{(l+1)}(i,j) = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{1}\left\{z_t^{(k)} = i\right\}\gamma_t^{(k)}(j)}{\sum_{k=1}^{K}\sum_{t=1}^{T}\gamma_t^{(k)}(j)}$$

Scaling

- The recursive definition of $\alpha_t(i)$ consist of the sum of a large number of terms of the form:

$$\alpha_t(i) = B(z_t, i) \sum_{j=1}^{N} A(i,j)\alpha_{t-1}(j) = B(z_t, i) \sum_{x_{1:t-1}} \prod_{s=0}^{t-1} A(x_{s+1}, x_s) B(z_s, x_s) \pi(x_0)$$

- Since each $A(i,j)$ and $B(z,i)$ is less than 1, it can be seen that as t increases, the terms required to compute $\alpha_t(i)$ start to head exponentially to zero. Thus, a **scaling procedure** is required.
- **Scaling**:
  1. Define $\hat{\alpha}_t(i) := C_t \alpha_t(i) := \frac{1}{\sum_{j=1}^{N} \alpha_t(j)} \alpha_t(i)$
  2. Given $\hat{\alpha}_t(i)$ compute $c_{t+1} := \sum_{i=1}^{N} B(z_{t+1}, i) \sum_{j=1}^{N} A(i,j)\hat{\alpha}_t(j)$ and set $\hat{\alpha}_{t+1}(i) = c_{t+1} B(z_{t+1}, i) \sum_{j=1}^{N} A(i,j)\hat{\alpha}_t(j)$
  3. Thus, $C_t = \prod_{s=0}^{t} c_s$
  4. Use the same scales during the backward procedure, i.e., $\hat{\beta}_T(i) = c_T$ and $\hat{\beta}_t(i) = c_t \sum_{j=1}^{N} B(z_{t+1}, j) A(j,i) \hat{\beta}_{t+1}(j) = \left[\prod_{s=t}^{T} c_s\right] \beta_t(i) := D_t \beta_t(j)$

HMM Inference with Scaled $\alpha_t(i)$ and $\beta_t(i)$

- **Observation sequence likelihood**:
  $\log p_\theta(z_{0:T}) = \log \sum_{i=1}^{N} \alpha_T(i) = \log \sum_{i=1}^{N} \frac{1}{C_T} \hat{\alpha}_T(i) = -\sum_{t=0}^{T} \log c_t$
- **Viterbi**: no scaling required – we can just take log:

$$\log \delta_t(i) = \max_j \left\{ \log B(Z_t, i) + \log A(i,j) + \log \delta_{t-1}(j) \right\}$$

- **Baum-Welch**: the terms $\gamma_t(i)$, $\xi_t(i,j)$ are unchanged when α_t and β_t are replaced by $\hat{\alpha}_t$ and $\hat{\beta}_t$:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)} = \frac{\cancel{C_t D_t}}{\cancel{C_t D_t}} \frac{\hat{\alpha}_t(i)\hat{\beta}_t(i)}{\sum_j \hat{\alpha}_t(j)\hat{\beta}_t(j)}$$

$$\xi_t(i,j) = \frac{\alpha_t(j)A(i,j)B(z_{t+1},i)\beta_{t+1}(i)}{\sum_{i',j'} \alpha_t(j')A(i',j')B(z_{t+1},i')\beta_{t+1}(i')}$$

$$= \frac{\cancel{C_t D_{t+1}}}{\cancel{C_t D_{t+1}}} \frac{\hat{\alpha}_t(j)A(i,j)B(z_{t+1},i)\hat{\beta}_{t+1}(i)}{\sum_{i',j'} \hat{\alpha}_t(j')A(i',j')B(z_{t+1},i')\hat{\beta}_{t+1}(i')}$$