COMS 4771 Lecture 22

1. Hidden Markov models

Hidden Markov models

Markov models

Markov model: a stochastic process $\{Y_t\}_{t\in\mathbb{N}}$ where, for each $t \in \mathbb{N}$, the conditional distribution of the next state Y_{t+1} given all previous states $\{Y_\tau : \tau \leq t\}$ only depends on the value of the current state Y_t.

Markov models

Markov model: a stochastic process $\{Y_t\}_{t \in \mathbb{N}}$ where, for each $t \in \mathbb{N}$, the conditional distribution of the next state Y_{t+1} given all previous states $\{Y_\tau : \tau \leq t\}$ only depends on the value of the current state Y_t.

Conditioned on present Y_t, past $\{Y_\tau\}_{\tau < t}$ and future $\{Y_\tau\}_{\tau > t}$ are independent.

$$\cdots \longrightarrow Y_{t-1} \longrightarrow \; Y_t \; \longrightarrow Y_{t+1} \longrightarrow \cdots$$

Markov models

Markov model: a stochastic process $\{Y_t\}_{t \in \mathbb{N}}$ where, for each $t \in \mathbb{N}$, the conditional distribution of the next state Y_{t+1} given all previous states $\{Y_\tau : \tau \leq t\}$ only depends on the value of the current state Y_t.

Conditioned on present Y_t, past $\{Y_\tau\}_{\tau < t}$ and future $\{Y_\tau\}_{\tau > t}$ are independent.

$$\cdots \longrightarrow Y_{t-1} \longrightarrow Y_t \longrightarrow Y_{t+1} \longrightarrow \cdots$$

Specifying a Markov chain (with discrete state space $[K] = \{1, 2, \ldots, K\}$):

- **Initial state distribution**: K-dimensional probability vector $\boldsymbol{\pi}$

$$\pi_i = \Pr(Y_1 = i).$$

- **Transition matrix**: $K \times K$ matrix $\boldsymbol{A}$

$$A_{i,j} = \Pr(Y_{t+1} = j \mid Y_t = i)$$

  (rows of $\boldsymbol{A}$ are probability vectors).

Hidden Markov models

Hidden Markov model (HMM): a Markov chain $\{(X_t, Y_t)\}_{t \in \mathbb{N}}$, where

Hidden Markov models

Hidden Markov model (HMM): a Markov chain $\{(X_t, Y_t)\}_{t \in \mathbb{N}}$, where

- $\{Y_t\}_{t \in \mathbb{N}}$ is also a Markov chain (with state space $[K] = \{1, 2, \ldots, K\}$) (hidden state sequence);

Hidden Markov models

Hidden Markov model (HMM): a Markov chain $\{(X_t, Y_t)\}_{t \in \mathbb{N}}$, where

- $\{Y_t\}_{t \in \mathbb{N}}$ is also a Markov chain (with state space $[K] = \{1, 2, \ldots, K\}$) (hidden state sequence);
- conditioned on Y_t, corresponding X_t is *independent of all other variables*;

Hidden Markov models

Hidden Markov model (HMM): a Markov chain $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$, where

- $\{Y_t\}_{t\in\mathbb{N}}$ is also a Markov chain (with state space $[K] = \{1, 2, \ldots, K\}$) (hidden state sequence);
- conditioned on Y_t, corresponding X_t is *independent of all other variables*;
- the Y_t are *hidden*, and the X_t are *observed*.

Hidden Markov models

Hidden Markov model (HMM): a Markov chain $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$, where

- $\{Y_t\}_{t\in\mathbb{N}}$ is also a Markov chain (with state space $[K] = \{1, 2, \ldots, K\}$) (hidden state sequence);
- conditioned on Y_t, corresponding X_t is *independent of all other variables*;
- the Y_t are *hidden*, and the X_t are *observed*.

$$\begin{array}{ccccccccc} \cdots & \longrightarrow & Y_{t-1} & \longrightarrow & Y_t & \longrightarrow & Y_{t+1} & \longrightarrow & \cdots \\ & & \downarrow & & \downarrow & & \downarrow & & \\ & & X_{t-1} & & X_t & & X_{t+1} & & \end{array}$$

Hidden Markov models

Hidden Markov model (HMM): a Markov chain $\{(X_t, Y_t)\}_{t \in \mathbb{N}}$, where

- $\{Y_t\}_{t \in \mathbb{N}}$ is also a Markov chain (with state space $[K] = \{1, 2, \ldots, K\}$) (hidden state sequence);
- conditioned on Y_t, corresponding X_t is *independent of all other variables*;
- the Y_t are *hidden*, and the X_t are *observed*.

$$
\begin{tikzcd}
\cdots \arrow[r] & Y_{t-1} \arrow[r] \arrow[d] & Y_t \arrow[r] \arrow[d] & Y_{t+1} \arrow[r] \arrow[d] & \cdots \\
& X_{t-1} & X_t & X_{t+1} &
\end{tikzcd}
$$

Time-homogeneous HMM: *conditional* distribution of X_t given Y_t does not depend on t. (We'll focus on these.)

Hidden Markov models

Hidden Markov model (HMM): a Markov chain $\{(X_t, Y_t)\}_{t \in \mathbb{N}}$, where

- $\{Y_t\}_{t \in \mathbb{N}}$ is also a Markov chain (with state space $[K] = \{1, 2, \ldots, K\}$) (hidden state sequence);
- conditioned on Y_t, corresponding X_t is *independent of all other variables*;
- the Y_t are *hidden*, and the X_t are *observed*.

$$\begin{array}{ccccccccc} \cdots & \longrightarrow & Y_{t-1} & \longrightarrow & Y_t & \longrightarrow & Y_{t+1} & \longrightarrow & \cdots \\ & & \downarrow & & \downarrow & & \downarrow & & \\ & & X_{t-1} & & X_t & & X_{t+1} & & \end{array}$$

great simplification
not depend on t!!!
only care Yt's value

Time-homogeneous HMM: *conditional* distribution of X_t given Y_t does not depend on t. (We'll focus on these.)

Useful subscript notation: $Y_{s:t} = (Y_s, Y_{s+1}, \ldots, Y_t)$ for $s \leq t$.

HMM parameters (discrete observations)

For time-homogeneous HMM where X_t takes values in $[D] = \{1, 2, \ldots, D\}$:

HMM parameters (discrete observations)

For time-homogeneous HMM where X_t takes values in $[D] = \{1, 2, \ldots, D\}$:

- **Initial state distribution**: K-dimensional probability vector $\boldsymbol{\pi}$

$$\pi_i = \Pr(Y_1 = i).$$

HMM parameters (discrete observations)

For time-homogeneous HMM where X_t takes values in $[D] = \{1, 2, \ldots, D\}$:

- **Initial state distribution**: K-dimensional probability vector $\boldsymbol{\pi}$

$$\pi_i \;=\; \Pr(Y_1 = i).$$

- **Transition matrix**: $K \times K$ matrix $\boldsymbol{A}$

$$A_{i,j} \;=\; \Pr(Y_{t+1} = j \mid Y_t = i)$$

(rows of $\boldsymbol{A}$ are probability vectors).

HMM PARAMETERS (DISCRETE OBSERVATIONS)

For time-homogeneous HMM where X_t takes values in $[D] = \{1, 2, \ldots, D\}$:

- **Initial state distribution**: K-dimensional probability vector $\boldsymbol{\pi}$

$$\pi_i \;=\; \Pr(Y_1 = i).$$

- **Transition matrix**: $K \times K$ matrix $\boldsymbol{A}$

underlying thing

$$A_{i,j} \;=\; \Pr(Y_{t+1} = j \mid Y_t = i)$$

(rows of $\boldsymbol{A}$ are probability vectors).

- **Emission matrix**: $K \times D$ matrix $\boldsymbol{B}$

emission thing

$$B_{i,j} \;=\; \Pr(X_t = j \mid Y_t = i)$$

(rows of $\boldsymbol{B}$ are probability vectors).

Mixture model

$$
\begin{array}{c}
Y \\
\downarrow \\
\boldsymbol{X}
\end{array}
$$

(Y is hidden, $\boldsymbol{X}$ is observed.)

Hidden Markov model

$$
\begin{array}{ccccccc}
Y_1 & \longrightarrow & Y_2 & \longrightarrow \cdots \longrightarrow & Y_\ell \\
\downarrow & & \downarrow & & \downarrow \\
X_1 & & X_2 & & X_\ell
\end{array}
$$

($Y_{1:\ell}$ is hidden, $X_{1:\ell}$ is observed.)

Connections to mixture models

Mixture model

$$\begin{array}{c} Y \\ \downarrow \\ \boldsymbol{X} \end{array}$$

(Y is hidden, $\boldsymbol{X}$ is observed.)

For K component mixture model, Y takes values in $[K]$.

Hidden Markov model

$$\begin{array}{ccccc} Y_1 & \longrightarrow & Y_2 & \longrightarrow \cdots \longrightarrow & Y_\ell \\ \downarrow & & \downarrow & & \downarrow \\ X_1 & & X_2 & & X_\ell \end{array}$$

($Y_{1:\ell}$ is hidden, $X_{1:\ell}$ is observed.)

Connections to mixture models

Mixture model

$$\begin{array}{c} Y \\ \downarrow \\ \boldsymbol{X} \end{array}$$

(Y is hidden, $\boldsymbol{X}$ is observed.)

For K component mixture model,
Y takes values in $[K]$.

Hidden Markov model

$$\begin{array}{ccccc} Y_1 & \longrightarrow & Y_2 & \longrightarrow \cdots \longrightarrow & Y_\ell \\ \downarrow & & \downarrow & & \downarrow \\ X_1 & & X_2 & & X_\ell \end{array}$$

($Y_{1:\ell}$ is hidden, $X_{1:\ell}$ is observed.)

For sequence of length ℓ,
$Y_{1:\ell}$ takes values in $[K]^\ell$.

Connections to mixture models

Mixture model

$$\begin{array}{c} Y \\ \downarrow \\ \boldsymbol{X} \end{array}$$

(Y is hidden, $\boldsymbol{X}$ is observed.)

For K component mixture model,
Y takes values in $[K]$.

Hidden Markov model

$$\begin{array}{ccccc} Y_1 & \longrightarrow & Y_2 & \longrightarrow \cdots \longrightarrow & Y_\ell \\ \downarrow & & \downarrow & & \downarrow \\ X_1 & & X_2 & & X_\ell \end{array}$$

($Y_{1:\ell}$ is hidden, $X_{1:\ell}$ is observed.)

For sequence of length ℓ,
$Y_{1:\ell}$ takes values in $[K]^\ell$.

Graphical diagram for HMM correctly suggests that every path—even ignoring arrow directions—is a Markov chain!

Connections to mixture models

Mixture model

$$\begin{array}{c} Y \\ \downarrow \\ \boldsymbol{X} \end{array}$$

(Y is hidden, $\boldsymbol{X}$ is observed.)

For K component mixture model, Y takes values in $[K]$.

Hidden Markov model

$$\begin{array}{ccccccc} Y_1 & \longrightarrow & Y_2 & \longrightarrow & \cdots & \longrightarrow & Y_\ell \\ \downarrow & & \downarrow & & & & \downarrow \\ X_1 & & X_2 & & & & X_\ell \end{array}$$

($Y_{1:\ell}$ is hidden, $X_{1:\ell}$ is observed.)

For sequence of length ℓ, $Y_{1:\ell}$ takes values in $[K]^\ell$.

Graphical diagram for HMM correctly suggests that every path—even ignoring arrow directions—is a Markov chain!

- $Y_1 \to Y_2 \to X_2$

Connections to mixture models

Mixture model

$$\begin{array}{c} Y \\ \downarrow \\ \boldsymbol{X} \end{array}$$

(Y is hidden, $\boldsymbol{X}$ is observed.)

For K component mixture model,
Y takes values in $[K]$.

Hidden Markov model

$$\begin{array}{ccccccc} Y_1 & \longrightarrow & Y_2 & \longrightarrow \cdots \longrightarrow & Y_\ell \\ \downarrow & & \downarrow & & \downarrow \\ X_1 & & X_2 & & X_\ell \end{array}$$

($Y_{1:\ell}$ is hidden, $X_{1:\ell}$ is observed.)

For sequence of length ℓ,
$Y_{1:\ell}$ takes values in $[K]^\ell$.

Graphical diagram for HMM correctly suggests that every path—even ignoring arrow directions—is a Markov chain!

- $Y_1 \to Y_2 \to X_2$
- $X_2 \to Y_2 \to Y_3 \to X_3$

hidden state(sub-distribution)

Mixture model

sub-distribution

$$\begin{array}{c} Y \\ \downarrow \\ \boldsymbol{X} \end{array}$$

(Y is hidden, $\boldsymbol{X}$ is observed.)

For K component mixture model,
Y takes values in $[K]$.

Hidden Markov model

$$\begin{array}{ccccccc} Y_1 & \longrightarrow & Y_2 & \longrightarrow & \cdots & \longrightarrow & Y_\ell \\ \downarrow & & \downarrow & & & & \downarrow \\ X_1 & & X_2 & & & & X_\ell \end{array}$$

($Y_{1:\ell}$ is hidden, $X_{1:\ell}$ is observed.)

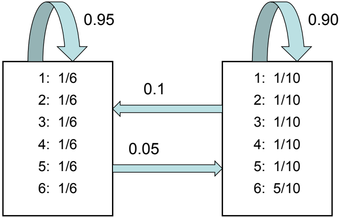
For sequence of length ℓ,
$\boxed{Y_{1:\ell} \text{ takes values in } [K]^\ell.}$

Graphical diagram for HMM correctly suggests that every path—even ignoring arrow directions—is a $\boxed{\text{Markov chain!}}$

- $Y_1 \to Y_2 \to X_2$
- $X_2 \to Y_2 \to Y_3 \to X_3$
- $X_1 \to Y_1 \to Y_{2:\ell} \to X_{2:\ell}$
- ...

note the relationship between chain!

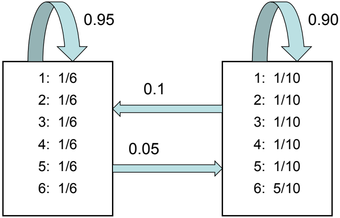
Example: dishonest casino



Casino die-rolling game:

Randomly switch between two possible dice: one is fair, the other loaded.

The dice are otherwise indistinguishable!



Casino die-rolling game:

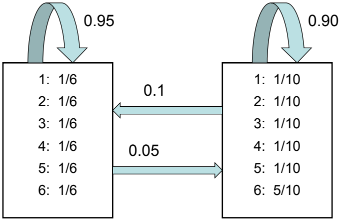
Randomly switch between two possible dice: one is fair, the other loaded.

The dice are otherwise indistinguishable!

HMM parameters:

$$\boldsymbol{A} = \begin{array}{c} \\ \text{fair die} \\ \text{loaded die} \end{array} \begin{array}{c} \begin{array}{cc} \text{fair die} & \text{loaded die} \end{array} \\ \begin{pmatrix} 0.95 & 0.05 \\ 0.10 & 0.90 \end{pmatrix} \end{array}, \quad \boldsymbol{B} = \begin{array}{c} \\ \text{fair die} \\ \text{loaded die} \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \end{pmatrix} \end{array},$$

and $\boldsymbol{\pi} = (1, 0)$ if the casino starts out with the fair die.



Casino die-rolling game:

Randomly switch between two possible dice: one is fair, the other loaded.

The dice are otherwise indistinguishable!

HMM parameters:

emission matrix

$$\boldsymbol{A} = \begin{array}{c} \\ \text{fair die} \\ \text{loaded die} \end{array} \begin{array}{cc} \text{fair die} & \text{loaded die} \\ \multicolumn{2}{c}{\begin{pmatrix} 0.95 & 0.05 \\ 0.10 & 0.90 \end{pmatrix}} \end{array}, \quad \boldsymbol{B} = \begin{array}{c} \\ \text{fair die} \\ \text{loaded die} \end{array} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \\ \multicolumn{6}{c}{\begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \end{pmatrix}} \end{array},$$

and $\boldsymbol{\pi} = (1, 0)$ if the casino starts out with the fair die.

Problem: Based on a sequence of rolls, guess which die was used at each time.

Conditional probabilities (e.g., filtering/smoothing)

- **Given**: parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, observation sequence $x_{1:\ell} \in [D]^{\ell}$.
- **Goal**: conditional distribution of $Y_{s:t}$ given $X_{1:\ell} = x_{1:\ell}$ ($1 \leq s \leq t \leq \ell$):

$$\Pr_{\boldsymbol{\theta}}\left(Y_{s:t} = y_{s:t} \,|\, X_{1:\ell} = x_{1:\ell}\right), \quad \text{for each } y_{s:t} \in [K]^{t-s+1}.$$

HMM inference/learning problems

Conditional probabilities (e.g., filtering/smoothing)

- **Given**: parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, observation sequence $x_{1:\ell} \in [D]^{\ell}$.
- **Goal**: conditional distribution of $Y_{s:t}$ given $X_{1:\ell} = x_{1:\ell}$ ($1 \leq s \leq t \leq \ell$):

$$\Pr_{\boldsymbol{\theta}}\big(Y_{s:t} = y_{s:t} \,|\, X_{1:\ell} = x_{1:\ell}\big), \quad \text{for each } y_{s:t} \in [K]^{t-s+1}.$$

Most probable state sequence (decoding)

- **Given**: parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, observation sequence $x_{1:\ell} \in [D]^{\ell}$.
- **Goal**: $\displaystyle\operatorname*{arg\,max}_{y_{1:\ell} \in [K]^{\ell}} \Pr_{\boldsymbol{\theta}}\big(Y_{1:\ell} = y_{1:\ell} \,|\, X_{1:\ell} = x_{1:\ell}\big)$.

may be the current yt, thus we could predict next state

Conditional probabilities (e.g., filtering/smoothing)

- **Given**: parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, observation sequence $x_{1:\ell} \in [D]^{\ell}$.
- **Goal**: conditional distribution of $Y_{s:t}$ given $X_{1:\ell} = x_{1:\ell}$ ($1 \leq s \leq t \leq \ell$):

$$\boxed{\Pr_{\boldsymbol{\theta}}\big(Y_{s:t} = y_{s:t} \,|\, X_{1:\ell} = x_{1:\ell}\big), \quad \text{for each } y_{s:t} \in [K]^{t-s+1}.}$$

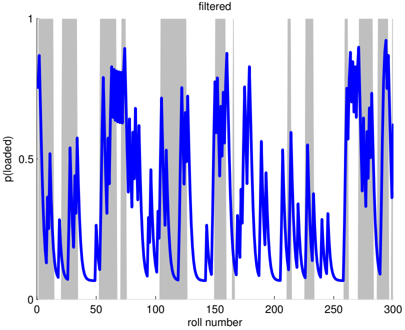
Most probable state sequence (decoding)

- **Given**: parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, observation sequence $x_{1:\ell} \in [D]^{\ell}$.
- **Goal**: $\underset{y_{1:\ell} \in [K]^{\ell}}{\arg\max} \Pr_{\boldsymbol{\theta}}\big(Y_{1:\ell} = y_{1:\ell} \,|\, X_{1:\ell} = x_{1:\ell}\big)$.

Parameter estimation

- **Given**: n observation sequences $x_{1:\ell}^{(s)}$ for $s \in [n]$.
- **Goal**: parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$.

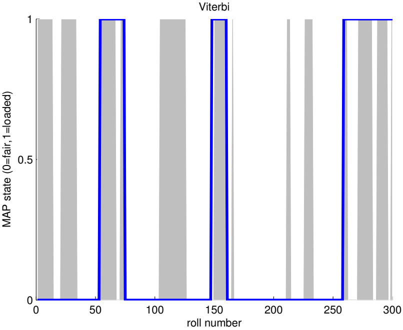
Example: dishonest casino



Conditional probability

Gray bars: Loaded dice used.

Blue: $\Pr_{\boldsymbol{\theta}}(Y_t = \text{loaded} | X_{1:\ell} = x_{1:\ell})$



Decoding

Gray bars: Loaded dice used.

Blue: Most probable state Z_t.

Some applications

- **Bioinformatics**

  Observations: amino acids in a protein

  Hidden states: indicators of evolutionary conservation

Some applications

- **Bioinformatics**

  Observations: amino acids in a protein

  Hidden states: indicators of evolutionary conservation

- **Natural language processing**

  Observations: words in a sentence

  Hidden states: words' part-of-speech or other word-type semantics

Some applications

- **Bioinformatics**

  Observations: amino acids in a protein

  Hidden states: indicators of evolutionary conservation

- **Natural language processing**

  Observations: words in a sentence

  Hidden states: words' part-of-speech or other word-type semantics

- **Speech recognition**

  Observations: recorded speech at various (discrete) times

  Hidden states: phonemes that the speaker intended to vocalize

HMM probability computations

Probability computations

HMM can be used for sequences of **arbitrary length**.

Probability computations

HMM can be used for sequences of **arbitrary length**.

Subtle difficulty: The most straightforward formulae for

- observation sequence probabilities (e.g., $\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})$) and
- conditional state sequence probabilities (e.g., $\Pr_{\boldsymbol{\theta}}(Y_\ell = y_\ell \mid X_{1:\ell} = x_{1:\ell})$)

do not suggest efficient algorithms for computation.

Probability computations

HMM can be used for sequences of **arbitrary length**.

Subtle difficulty: The most straightforward formulae for

- observation sequence probabilities (e.g., $\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})$) and
- conditional state sequence probabilities (e.g., $\Pr_{\boldsymbol{\theta}}(Y_\ell = y_\ell \mid X_{1:\ell} = x_{1:\ell})$)

do not suggest efficient algorithms for computation.

Need to exploit special structure of HMMs to get efficient algorithms.

Sequence probability computation

Probability of observation sequence $\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})$.

Sequence probability computation

Probability of observation sequence $\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})$.

We'll use the fact that conditioned on $Y_{1:\ell}$, the $\{X_t\}_{t \in [\ell]}$ are independent.

Sequence probability computation

Probability of observation sequence $\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})$.

We'll use the fact that conditioned on $Y_{1:\ell}$, the $\{X_t\}_{t \in [\ell]}$ are independent.

$$\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell}) \quad = \quad \sum_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \wedge Y_{1:\ell} = y_{1:\ell})$$

Sequence probability computation

Probability of observation sequence $\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})$.

We'll use the fact that conditioned on $Y_{1:\ell}$, the $\{X_t\}_{t \in [\ell]}$ are independent.

$$\begin{aligned}
\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell}) &= \sum_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \wedge Y_{1:\ell} = y_{1:\ell}) \\
&= \sum_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}(Y_{1:\ell} = y_{1:\ell}) \cdot \Pr_{\boldsymbol{\theta}}\left(X_{1:\ell} = x_{1:\ell} \,\middle|\, Y_{1:\ell} = y_{1:\ell}\right)
\end{aligned}$$

Sequence probability computation

Probability of observation sequence $\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})$.

We'll use the fact that conditioned on $Y_{1:\ell}$, the $\{X_t\}_{t \in [\ell]}$ are independent.

$$
\begin{aligned}
\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell}) &= \sum_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \wedge Y_{1:\ell} = y_{1:\ell}) \\
&= \sum_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}(Y_{1:\ell} = y_{1:\ell}) \cdot \Pr_{\boldsymbol{\theta}}\big(X_{1:\ell} = x_{1:\ell} \,|\, Y_{1:\ell} = y_{1:\ell}\big) \\
&= \sum_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}(Y_{1:\ell} = y_{1:\ell}) \cdot \prod_{t=1}^{\ell} \Pr_{\boldsymbol{\theta}}\big(X_t = x_t \,|\, Y_t = y_t\big)
\end{aligned}
$$

Given Y1:l, {Xt} are independent.

Probability of observation sequence $\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})$.

We'll use the fact that conditioned on $Y_{1:\ell}$, the $\{X_t\}_{t \in [\ell]}$ are independent.

$$\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell}) = \sum_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \wedge Y_{1:\ell} = y_{1:\ell})$$

$$= \sum_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}(Y_{1:\ell} = y_{1:\ell}) \cdot \Pr_{\boldsymbol{\theta}}\left(X_{1:\ell} = x_{1:\ell} \,\middle|\, Y_{1:\ell} = y_{1:\ell}\right)$$

this step, assume yt is know for current sequence

$$= \sum_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}(Y_{1:\ell} = y_{1:\ell}) \cdot \prod_{t=1}^{\ell} \Pr_{\boldsymbol{\theta}}\left(X_t = x_t \,\middle|\, Y_t = y_t\right)$$

But summation is over K^ℓ terms—seems intractable for large ℓ.

Sequence probability computation

Probability of observation sequence $\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})$.

We'll use the fact that conditioned on $Y_{1:\ell}$, the $\{X_t\}_{t\in[\ell]}$ are independent.

$$\begin{aligned}
\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell}) &= \sum_{\boxed{y_{1:\ell}\in[K]^\ell}} \Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \wedge Y_{1:\ell} = y_{1:\ell}) \\
&= \sum_{y_{1:\ell}\in[K]^\ell} \Pr_{\boldsymbol{\theta}}(Y_{1:\ell} = y_{1:\ell}) \cdot \Pr_{\boldsymbol{\theta}}\big(X_{1:\ell} = x_{1:\ell} \,|\, Y_{1:\ell} = y_{1:\ell}\big) \\
&= \sum_{y_{1:\ell}\in[K]^\ell} \Pr_{\boldsymbol{\theta}}(Y_{1:\ell} = y_{1:\ell}) \cdot \prod_{t=1}^{\ell} \Pr_{\boldsymbol{\theta}}\big(X_t = x_t \,|\, Y_t = y_t\big)
\end{aligned}$$

marigina

But summation is over K^ℓ terms—seems intractable for large ℓ.

Fortunately, the summation can be computed iteratively in time linear in ℓ.

Sequence probability computation

HMM parameters:
$\pi y = \Pr(Y_1 = y)$; $Ay, z = \Pr(Y_{t+1} = z \mid Y_t = y)$; $By, x = \Pr(X_t = x \mid Y_t = y)$.

Example: probability of observation triplet $x_{1:3} \in [D]^3$

$$\Pr_{\boldsymbol{\theta}}(X_{1:3} = x_{1:3})$$

Sequence probability computation

HMM parameters:

$\pi_y = \Pr(Y_1 = y)$; $A_{y,z} = \Pr(Y_{t+1} = z \mid Y_t = y)$; $B_{y,x} = \Pr(X_t = x \mid Y_t = y)$.

Example: probability of observation triplet $x_{1:3} \in [D]^3$

$$\Pr_{\boldsymbol{\theta}}(X_{1:3} = x_{1:3})$$

$$= \boxed{\sum_{y_{1:3} \in [K]^3} \Pr_{\boldsymbol{\theta}}(Y_{1:3} = y_{1:3}) \cdot \prod_{t=1}^{3} \Pr_{\boldsymbol{\theta}}\left(X_t = x_t \mid Y_t = y_t\right)}$$

Sequence probability computation

HMM parameters:
$\pi_y = \Pr(Y_1 = y)$; $A_{y,z} = \Pr(Y_{t+1} = z \mid Y_t = y)$; $B_{y,x} = \Pr(X_t = x \mid Y_t = y)$.

Example: probability of observation triplet $x_{1:3} \in [D]^3$

$$
\begin{aligned}
&\Pr_{\boldsymbol{\theta}}(X_{1:3} = x_{1:3}) \\
&= \sum_{y_{1:3} \in [K]^3} \Pr_{\boldsymbol{\theta}}(Y_{1:3} = y_{1:3}) \cdot \prod_{t=1}^{3} \Pr_{\boldsymbol{\theta}}\bigl(X_t = x_t \mid Y_t = y_t\bigr) \\
&= \sum_{y_{1:3} \in [K]^3} \underbrace{\bigl(\pi_{y_1} \cdot A_{y_1,y_2} \cdot A_{y_2,y_3}\bigr)}_{\text{Markov chain probabilities}} \cdot \underbrace{\bigl(B_{y_1,x_1} \cdot B_{y_2,x_2} \cdot B_{y_3,x_3}\bigr)}_{\text{emission probabilities}}
\end{aligned}
$$

Sequence probability computation

HMM parameters:
$\pi_y = \Pr(Y_1 = y)$; $A_{y,z} = \Pr(Y_{t+1} = z \mid Y_t = y)$; $B_{y,x} = \Pr(X_t = x \mid Y_t = y)$.

Example: probability of observation triplet $x_{1:3} \in [D]^3$

$$
\begin{aligned}
&\Pr_{\boldsymbol{\theta}}(X_{1:3} = x_{1:3}) \\
&= \sum_{y_{1:3} \in [K]^3} \Pr_{\boldsymbol{\theta}}(Y_{1:3} = y_{1:3}) \cdot \prod_{t=1}^{3} \Pr_{\boldsymbol{\theta}}\left(X_t = x_t \mid Y_t = y_t\right) \\
&= \sum_{y_{1:3} \in [K]^3} \underbrace{\left(\pi_{y_1} \cdot A_{y_1,y_2} \cdot A_{y_2,y_3}\right)}_{\text{Markov chain probabilities}} \cdot \underbrace{\left(B_{y_1,x_1} \cdot B_{y_2,x_2} \cdot B_{y_3,x_3}\right)}_{\text{emission probabilities}} \\
&= \underbrace{\sum_{y_1 \in [K]} \pi_{y_1} \cdot B_{y_1,x_1}}_{O(K)\text{ time}} \underbrace{\sum_{y_2 \in [K]} A_{y_1,y_2} \cdot B_{y_2,x_2}}_{O(K)\text{ time for each } y_1 \in [K]} \underbrace{\sum_{y_3 \in [K]} A_{y_2,y_3} \cdot B_{y_3,x_3}}_{O(K)\text{ time for each } y_2 \in [K]}
\end{aligned}
$$

Sequence probability computation

HMM parameters:

$\pi_y = \Pr(Y_1 = y)$; $A_{y,z} = \Pr(Y_{t+1} = z \mid Y_t = y)$; $B_{y,x} = \Pr(X_t = x \mid Y_t = y)$.

Example: probability of observation triplet $x_{1:3} \in [D]^3$

the beautiful part
there is a
transformation from
any i to any j
thus we could reduce
through following form

$$\Pr_{\boldsymbol{\theta}}(X_{1:3} = x_{1:3})$$

$$= \sum_{y_{1:3} \in [K]^3} \Pr_{\boldsymbol{\theta}}(Y_{1:3} = y_{1:3}) \cdot \prod_{t=1}^{3} \Pr_{\boldsymbol{\theta}}(X_t = x_t \mid Y_t = y_t)$$

note the x
is fixed!!!

$$= \boxed{\sum_{y_{1:3} \in [K]^3} \underbrace{(\pi_{y_1} \cdot A_{y_1,y_2} \cdot A_{y_2,y_3})}_{\text{Markov chain probabilities}} \cdot \underbrace{(B_{y_1,x_1} \cdot B_{y_2,x_2} \cdot B_{y_3,x_3})}_{\text{emission probabilities}}}$$

$$= \underbrace{\sum_{y_1 \in [K]} \pi_{y_1} \cdot B_{y_1,x_1}}_{O(K)\text{ time}} \underbrace{\sum_{y_2 \in [K]} A_{y_1,y_2} \cdot B_{y_2,x_2}}_{O(K)\text{ time for each } y_1 \in [K]} \underbrace{\sum_{y_3 \in [K]} A_{y_2,y_3} \cdot B_{y_3,x_3}}_{O(K)\text{ time for each } y_2 \in [K]}$$

Computing sums from right-to-left: $\boxed{\text{total time is } O(K^2 \ell) \text{ for length } \ell.}$

note the complexity: constrain to its own component now!!!

Conditional probabilities

A simple case: $\Pr_{\boldsymbol{\theta}}\big(Y_\ell = y_\ell \,|\, X_{1:\ell} = x_{1:\ell}\big) \;=\; \dfrac{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \,\wedge\, Y_\ell = y_\ell)}{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})}.$

Conditional probabilities

A simple case: $\Pr_{\boldsymbol{\theta}}\big(Y_\ell = y_\ell \,|\, X_{1:\ell} = x_{1:\ell}\big) \;=\; \dfrac{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \,\wedge\, Y_\ell = y_\ell)}{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})}$.

$$\begin{aligned}
&\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell}) \\
&\quad= \sum_{y_1 \in [K]} \pi_{y_1} \cdot B_{y_1, x_1} \sum_{y_2 \in [K]} A_{y_1, y_2} \cdot B_{y_2, x_2} \cdots \\
&\qquad\qquad \cdots \sum_{y_{\ell-1} \in [K]} A_{y_{\ell-2}, y_{\ell-1}} \cdot B_{y_{\ell-1}, x_{\ell-1}} \sum_{y_\ell \in [K]} A_{y_{\ell-1}, y_\ell} \cdot B_{y_\ell, x_\ell}.
\end{aligned}$$

Conditional probabilities

A simple case: $\Pr_{\boldsymbol{\theta}}\big(Y_\ell = y_\ell \,|\, X_{1:\ell} = x_{1:\ell}\big) \;=\; \dfrac{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \,\wedge\, Y_\ell = y_\ell)}{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})}$.

$$
\begin{aligned}
&\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \,\wedge\, Y_\ell = y_\ell) \\
&\quad = \sum_{y_1 \in [K]} \pi_{y_1} \cdot B_{y_1, x_1} \sum_{y_2 \in [K]} A_{y_1, y_2} \cdot B_{y_2, x_2} \cdots \\
&\qquad\quad \cdots \sum_{y_{\ell-1} \in [K]} A_{y_{\ell-2}, y_{\ell-1}} \cdot B_{y_{\ell-1}, x_{\ell-1}} \qquad A_{y_{\ell-1}, y_\ell} \cdot B_{y_\ell, x_\ell}.
\end{aligned}
$$

Conditional probabilities

A simple case: $\Pr_{\boldsymbol{\theta}}\big(Y_\ell = y_\ell \mid X_{1:\ell} = x_{1:\ell}\big) \;=\; \dfrac{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \,\wedge\, Y_\ell = y_\ell)}{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})}$.

$$\begin{aligned}
&\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell} \,\wedge\, Y_\ell = y_\ell) \\
&\quad= \sum_{y_1 \in [K]} \pi_{y_1} \cdot B_{y_1, x_1} \sum_{y_2 \in [K]} A_{y_1, y_2} \cdot B_{y_2, x_2} \cdots \\
&\qquad\quad \cdots \sum_{y_{\ell-1} \in [K]} A_{y_{\ell-2}, y_{\ell-1}} \cdot B_{y_{\ell-1}, x_{\ell-1}} \qquad A_{y_{\ell-1}, y_\ell} \cdot B_{y_\ell, x_\ell}.
\end{aligned}$$

Forward inductive computation:
Keep track of $\alpha_t(y_t) := \Pr_{\boldsymbol{\theta}}(X_{1:t} = x_{1:t} \,\wedge\, Y_t = y_t)$ for each $y_t \in [K]$.

Conditional probabilities

A simple case: $\Pr_{\boldsymbol\theta}\big(Y_\ell = y_\ell \,|\, X_{1:\ell} = x_{1:\ell}\big) \;=\; \dfrac{\Pr_{\boldsymbol\theta}(X_{1:\ell} = x_{1:\ell} \,\wedge\, Y_\ell = y_\ell)}{\Pr_{\boldsymbol\theta}(X_{1:\ell} = x_{1:\ell})}.$

$$\begin{aligned}
&\Pr_{\boldsymbol\theta}(X_{1:\ell} = x_{1:\ell} \,\wedge\, Y_\ell = y_\ell) \\
&\quad = \sum_{y_1 \in [K]} \pi_{y_1} \cdot B_{y_1,\, x_1} \sum_{y_2 \in [K]} A_{y_1,\, y_2} \cdot B_{y_2,\, x_2} \cdots \\
&\qquad \cdots \sum_{y_{\ell-1} \in [K]} A_{y_{\ell-2},\, y_{\ell-1}} \cdot B_{y_{\ell-1},\, x_{\ell-1}} \quad \boxed{A_{y_{\ell-1},\, y_\ell} \cdot B_{y_\ell,\, x_\ell}}.
\end{aligned}$$

note here, yl is fixed!
to get xl

Forward inductive computation:
Keep track of $\alpha_t(y_t) := \Pr_{\boldsymbol\theta}(X_{1:t} = x_{1:t} \,\wedge\, Y_t = y_t)$ for each $y_t \in [K]$.

Compute α_{t+1} using α_t in $\boxed{O(K^2) \text{ time:}}$

something amazing is
happening!

$$\alpha_{t+1}(y_{t+1}) \;=\; \left(\sum_{y_t \in [K]} \alpha_t(y_t) \cdot A_{y_t,\, y_{t+1}}\right) \boxed{\cdot\, B_{y_{t+1},\, x_{t+1}}} \quad \text{for each } y_{t+1} \in [K].$$

powerful! each pair from yt to
yt+1

actually, need to compute for each y(t
+1), thus the cost is O(k^2)

Conditional probabilities for $t < \ell$

For any $1 \leq t < \ell$,

$$\begin{aligned}
&\Pr_{\boldsymbol{\theta}}\big(Y_t = y_t \,|\, X_{1:\ell} = x_{1:\ell}\big) \\
&= \quad \frac{\Pr_{\boldsymbol{\theta}}(X_{1:t} = x_{1:t} \wedge Y_t = y_t) \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_t = y_t\big)}{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})}
\end{aligned}$$

(since $X_{t+1:\ell}$ is conditionally independent of $X_{1:t}$ given Y_t)

Conditional probabilities for $t < \ell$

For any $1 \le t < \ell$,

$$\begin{aligned}
&\Pr_{\boldsymbol{\theta}}\big(Y_t = y_t \,\big|\, X_{1:\ell} = x_{1:\ell}\big) \\
&\quad = \frac{\Pr_{\boldsymbol{\theta}}(X_{1:t} = x_{1:t} \,\wedge\, Y_t = y_t) \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,\big|\, Y_t = y_t\big)}{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})}
\end{aligned}$$

(since $X_{t+1:\ell}$ is conditionally independent of $X_{1:t}$ given Y_t)

$$= \frac{\alpha_t(y_t) \cdot \beta_t(y_t)}{\text{normalization term}}$$

where

$$\begin{aligned}
\alpha_t(y_t) &:= \Pr_{\boldsymbol{\theta}}(X_{1:t} = x_{1:t} \,\wedge\, Y_t = y_t), \\
\beta_t(y_t) &:= \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,\big|\, Y_t = y_t\big).
\end{aligned}$$

Conditional probabilities for $t < \ell$

For any $1 \le t < \ell$,

$$\begin{aligned}
&\Pr_{\boldsymbol{\theta}}\big(Y_t = y_t \mid X_{1:\ell} = x_{1:\ell}\big) \\
&= \quad \frac{\Pr_{\boldsymbol{\theta}}(X_{1:t} = x_{1:t} \wedge Y_t = y_t) \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \mid Y_t = y_t\big)}{\Pr_{\boldsymbol{\theta}}(X_{1:\ell} = x_{1:\ell})}
\end{aligned}$$

(since $X_{t+1:\ell}$ is conditionally independent of $X_{1:t}$ given Y_t)

$$= \quad \frac{\alpha_t(y_t) \cdot \beta_t(y_t)}{\text{normalization term}}$$

where

$$\begin{aligned}
\alpha_t(y_t) &:= \Pr_{\boldsymbol{\theta}}(X_{1:t} = x_{1:t} \wedge Y_t = y_t), \\
\beta_t(y_t) &:= \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \mid Y_t = y_t\big).
\end{aligned}$$

We already saw how to compute $\alpha_t(y_t)$ for each $y_t \in [K]$.

Conditional probabilities for $t < \ell$

$$\beta_t(y_t) \quad = \quad \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_t = y_t\big)$$

Conditional probabilities for $t < \ell$

$$
\begin{aligned}
\beta_t(y_t) &= \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_t = y_t\big) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,\wedge\, Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\big)
\end{aligned}
$$

Conditional probabilities for $t < \ell$

$$
\begin{aligned}
\beta_t(y_t) &= \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_t = y_t\big) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \wedge Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\big) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\big(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\big) \\
&\qquad\qquad \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_{t+1} = y_{t+1}\big)
\end{aligned}
$$

Conditional probabilities for $t < \ell$

$$
\begin{aligned}
\beta_t(y_t) &= \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_t = y_t\big) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \wedge Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\big) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\big(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\big) \\
&\qquad\qquad \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_{t+1} = y_{t+1}\big) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\big(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\big) \\
&\quad \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+1} = x_{t+1} \,|\, Y_{t+1} = y_{t+1}\big) \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+2:\ell} = x_{t+2:\ell} \,|\, Y_{t+1} = y_{t+1}\big)
\end{aligned}
$$

Conditional probabilities for $t < \ell$

$$
\begin{aligned}
\beta_t(y_t) &= \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_t = y_t\big) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \wedge Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\big) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\big(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\big) \\
&\qquad\qquad \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_{t+1} = y_{t+1}\big) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\big(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\big) \\
&\qquad \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+1} = x_{t+1} \,|\, Y_{t+1} = y_{t+1}\big) \cdot \Pr_{\boldsymbol{\theta}}\big(X_{t+2:\ell} = x_{t+2:\ell} \,|\, Y_{t+1} = y_{t+1}\big) \\
&= \sum_{y_{t+1} \in [K]} A_{y_t, y_{t+1}} \cdot B_{y_{t+1}, x_{t+1}} \cdot \beta_{t+1}(y_{t+1}).
\end{aligned}
$$

Conditional probabilities for $t < \ell$

$$
\begin{aligned}
\beta_t(y_t) &= \Pr_{\boldsymbol{\theta}}\left(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_t = y_t\right) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\left(X_{t+1:\ell} = x_{t+1:\ell} \wedge Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\right) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\left(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\right) \\
&\qquad \cdot \Pr_{\boldsymbol{\theta}}\left(X_{t+1:\ell} = x_{t+1:\ell} \,|\, Y_{t+1} = y_{t+1}\right) \\
&= \sum_{y_{t+1} \in [K]} \Pr_{\boldsymbol{\theta}}\left(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t\right) \\
&\qquad \cdot \Pr_{\boldsymbol{\theta}}\left(X_{t+1} = x_{t+1} \,|\, Y_{t+1} = y_{t+1}\right) \cdot \Pr_{\boldsymbol{\theta}}\left(X_{t+2:\ell} = x_{t+2:\ell} \,|\, Y_{t+1} = y_{t+1}\right) \\
&= \sum_{y_{t+1} \in [K]} A_{y_t, y_{t+1}} \cdot B_{y_{t+1}, x_{t+1}} \cdot \beta_{t+1}(y_{t+1}).
\end{aligned}
$$

Keep in mind, xt+1 is known! it could be emissioned by any yt+1

Backward inductive computation: Compute β_t using β_{t+1} in $O(K^2)$ time.

Forward-backward algorithm

Given parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ and sequence $x_{1:\ell} \in [D]^{\ell}$:

Forward-backward algorithm

Given parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ and sequence $x_{1:\ell} \in [D]^{\ell}$:

- **(Forward pass)**
  Starting with $\alpha_1(y_1) = \pi_{y_1} \cdot B_{y_1, x_1}$ for each $y_1 \in [K]$,

$$\alpha_{t+1}(y_{t+1}) \;=\; \left(\sum_{y_t \in [K]} \alpha_t(y_t) \cdot A_{y_t, y_{t+1}} \right) \cdot B_{y_{t+1}, x_{t+1}} \quad \text{for each } y_{t+1} \in [K].$$

Forward-backward algorithm

Given parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ and sequence $x_{1:\ell} \in [D]^{\ell}$:

- **(Forward pass)**
  Starting with $\alpha_1(y_1) = \pi_{y_1} \cdot B_{y_1, x_1}$ for each $y_1 \in [K]$,

  $$\alpha_{t+1}(y_{t+1}) \;=\; \left(\sum_{y_t \in [K]} \alpha_t(y_t) \cdot A_{y_t, y_{t+1}} \right) \cdot B_{y_{t+1}, x_{t+1}} \quad \text{for each } y_{t+1} \in [K].$$

- **(Backward pass)**
  Starting with $\beta_\ell(y_\ell) = 1$ for each $y_\ell \in [K]$,

  $$\beta_{t-1}(y_{t-1}) \;=\; \sum_{y_t \in [K]} A_{y_{t-1}, y_t} \cdot B_{y_t, x_t} \cdot \beta_t(y_t) \quad \text{for each } y_{t-1} \in [K].$$

Forward-backward algorithm

Given parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ and sequence $x_{1:\ell} \in [D]^{\ell}$:

- **(Forward pass)**
  Starting with $\alpha_1(y_1) = \pi_{y_1} \cdot B_{y_1, x_1}$ for each $y_1 \in [K]$,

$$\alpha_{t+1}(y_{t+1}) = \left(\sum_{y_t \in [K]} \alpha_t(y_t) \cdot A_{y_t, y_{t+1}} \right) \cdot B_{y_{t+1}, x_{t+1}} \quad \text{for each } y_{t+1} \in [K].$$

- **(Backward pass)**
  Starting with $\boxed{\beta_\ell(y_\ell) = 1 \text{ for each } y_\ell \in [K],}$

$$\beta_{t-1}(y_{t-1}) = \sum_{y_t \in [K]} A_{y_{t-1}, y_t} \cdot B_{y_t, x_t} \cdot \beta_t(y_t) \quad \text{for each } y_{t-1} \in [K].$$

- **(Also in backward pass)**
  Compute conditional probabilities:

$$\Pr_{\boldsymbol{\theta}}\left(Y_t = y_t \mid X_{1:\ell} = x_{1:\ell}\right) = \frac{\alpha_t(y_t) \cdot \beta_t(y_t)}{\text{normalization term}} \quad \text{for each } y_t \in [K].$$

Even more complicated forward-backward

Can also compute

$$\Pr_{\boldsymbol{\theta}}\big(Y_{t:t+1} = y_{t:t+1} \,|\, X_{1:\ell} = x_{1:\ell}\big) \quad \text{for each } y_{t:t+1} \in [K]^2$$

using forward-backward.

Even more complicated forward-backward

Can also compute

$$\Pr_{\boldsymbol{\theta}}\big(Y_{t:t+1} = y_{t:t+1} \,|\, X_{1:\ell} = x_{1:\ell}\big) \quad \text{for each } y_{t:t+1} \in [K]^2$$

using forward-backward.

Using Markov property, can string together these probabilities to get

$$\Pr_{\boldsymbol{\theta}}\big(Y_{s:t} = y_{s:t} \,|\, X_{1:\ell} = x_{1:\ell}\big) \quad \text{for each } y_{s:t} \in [K]^{t-s+1}.$$

Even more complicated forward-backward

Can also compute

$$\Pr_{\boldsymbol{\theta}}\big(Y_{t:t+1} = y_{t:t+1} \,|\, X_{1:\ell} = x_{1:\ell}\big) \quad \text{for each } y_{t:t+1} \in [K]^2$$

using forward-backward.

Using Markov property, can string together these probabilities to get

$$\Pr_{\boldsymbol{\theta}}\big(Y_{s:t} = y_{s:t} \,|\, X_{1:\ell} = x_{1:\ell}\big) \quad \text{for each } y_{s:t} \in [K]^{t-s+1}.$$

Similar procedure for computing most likely state sequence:

$$\operatorname*{arg\,max}_{y_{1:\ell} \in [K]^\ell} \Pr_{\boldsymbol{\theta}}\big(Y_{1:\ell} = y_{1:\ell} \,|\, X_{1:\ell} = x_{1:\ell}\big)$$

(Viterbi algorithm).

Even more complicated forward-backward

Can also compute

$$\Pr_{\boldsymbol{\theta}}\bigl(Y_{t:t+1} = y_{t:t+1} \mid X_{1:\ell} = x_{1:\ell}\bigr) \quad \text{for each } y_{t:t+1} \in [K]^2$$

using forward-backward.

Using Markov property, can string together these probabilities to get

$$\Pr_{\boldsymbol{\theta}}\bigl(Y_{s:t} = y_{s:t} \mid X_{1:\ell} = x_{1:\ell}\bigr) \quad \text{for each } y_{s:t} \in [K]^{t-s+1}.$$

Similar procedure for computing most likely state sequence:

$$\arg\max_{y_{1:\ell} \in [K]^{\ell}} \Pr_{\boldsymbol{\theta}}\bigl(Y_{1:\ell} = y_{1:\ell} \mid X_{1:\ell} = x_{1:\ell}\bigr)$$

(Viterbi algorithm).

See Rabiner's tutorial for details.

HMM parameter estimation

Parameter estimation

Parameter estimation problem:

- **Given**: n observation sequences $x_{1:\ell}^{(s)}$ for $s \in [n]$.
- **Goal**: parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$.

Parameter estimation

Parameter estimation problem:

- **Given**: n observation sequences $x_{1:\ell}^{(s)}$ for $s \in [n]$.
- **Goal**: parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$.

As is the case for mixture models, MLE for HMMs is generally intractable.

Parameter estimation

Parameter estimation problem:

- **Given**: n observation sequences $x_{1:\ell}^{(s)}$ for $s \in [n]$.
- **Goal**: parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$.

As is the case for mixture models, MLE for HMMs is generally intractable.

Nevertheless, we can use **Expectation-Maximization** to find a local maximizer of the likelihood function. (Called the **Baum-Welch** algorithm in this context.)

Expectation-Maximization for HMMs

Suppose we have current guess for parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$.

Expectation-Maximization for HMMs

Suppose we have current guess for parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$.

Compute, for each training sequence $x_{1:\ell}^{(s)}$,

$$\begin{aligned} \gamma_t^{(s)}(y_t) &:= \Pr_{\hat{\boldsymbol{\theta}}}\Big(Y_t = y_t \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\Big) \quad \text{for all } y_t \in [K] \\ \xi_t^{(s)}(y_{t-1}, y_t) &:= \Pr_{\hat{\boldsymbol{\theta}}}\Big(Y_{t-1:t} = y_{t-1:t} \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\Big) \quad \text{for all } y_{t-1:t} \in [K]^2 \end{aligned}$$

using Forward-Backward (see Rabiner tutorial for full details).

Expectation-Maximization for HMMs

Suppose we have current guess for parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$.

Compute, for each training sequence $x_{1:\ell}^{(s)}$,

$$\begin{aligned}
\gamma_t^{(s)}(y_t) &:= \Pr_{\hat{\boldsymbol{\theta}}}\Big(Y_t = y_t \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\Big) \quad \text{for all } y_t \in [K] \\
\xi_t^{(s)}(y_{t-1}, y_t) &:= \Pr_{\hat{\boldsymbol{\theta}}}\Big(Y_{t-1:t} = y_{t-1:t} \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\Big) \quad \text{for all } y_{t-1:t} \in [K]^2
\end{aligned}$$

using Forward-Backward (see Rabiner tutorial for full details).

Expected complete log likelihood of $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$:

$$\sum_{s=1}^{n} \Bigg\{ \sum_{y_1 \in [K]} \gamma_1^{(s)}(y_1) \ln \pi_{y_1} + \sum_{t=2}^{\ell} \sum_{y_{t-1} \in [K]} \sum_{y_t \in [K]} \xi_t^{(s)}(y_{t-1}, y_t) \ln A_{y_{t-1}, y_t} + \sum_{t=1}^{\ell} \sum_{y_t \in [K]} \gamma_t^{(s)}(y_t) \sum_{j=1}^{D} \mathbb{1}\{x_t^{(s)} = j\} \ln B_{y_t, j} \Bigg\}.$$

Expectation-Maximization for HMMs

Suppose we have current guess for parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$.

Compute, for each training sequence $x_{1:\ell}^{(s)}$,

$$\begin{aligned}
\gamma_t^{(s)}(y_t) &:= \Pr_{\hat{\boldsymbol{\theta}}}\left(Y_t = y_t \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\right) \quad \text{for all } y_t \in [K] \\
\xi_t^{(s)}(y_{t-1}, y_t) &:= \Pr_{\hat{\boldsymbol{\theta}}}\left(Y_{t-1:t} = y_{t-1:t} \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\right) \quad \text{for all } \boxed{y_{t-1:t} \in [K]^2}
\end{aligned}$$

using Forward-Backward (see Rabiner tutorial for full details).

Expected complete log likelihood of $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$:

$$\sum_{s=1}^{n} \left\{ \sum_{y_1 \in [K]} \gamma_1^{(s)}(y_1) \ln \pi_{y_1} + \sum_{t=2}^{\ell} \sum_{y_{t-1} \in [K]} \sum_{y_t \in [K]} \xi_t^{(s)}(y_{t-1}, y_t) \ln A_{y_{t-1}, y_t} \right.$$
$$\left. + \sum_{t=1}^{\ell} \sum_{y_t \in [K]} \gamma_t^{(s)}(y_t) \sum_{j=1}^{D} \mathbb{1}\{x_t^{(s)} = j\} \ln B_{y_t, j} \right\}.$$

Can easily find maximizing parameters $\boldsymbol{\theta}$ (subject to constraints that $\boldsymbol{\pi}$ and rows of $\boldsymbol{A}$ and $\boldsymbol{B}$ are probability distributions).

Expectation-Maximization for HMMs

Input: n observation sequences $x_{1:\ell}^{(s)}$ for $s \in [n]$.

Initialize $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ somehow.

Expectation-Maximization for HMMs

Input: n observation sequences $x_{1:\ell}^{(s)}$ for $s \in [n]$.

Initialize $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ somehow. Then repeat:

- **E step**: For each $s \in [n]$, compute

$$
\begin{aligned}
\gamma_t^{(s)}(y_t) &:= \Pr_{\boldsymbol{\theta}}\left(Y_t = y_t \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\right) \\
\xi_t^{(s)}(y_{t-1}, y_t) &:= \Pr_{\boldsymbol{\theta}}\left(Y_{t-1:t} = y_{t-1:t} \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\right).
\end{aligned}
$$

  using Forward-Backward.

Expectation-Maximization for HMMs

Input: n observation sequences $x_{1:\ell}^{(s)}$ for $s \in [n]$.

Initialize $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ somehow. Then repeat:

- **E step**: For each $s \in [n]$, compute

$$\begin{aligned} \gamma_t^{(s)}(y_t) &:= \Pr_{\boldsymbol{\theta}}\Big(Y_t = y_t \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\Big) \\ \xi_t^{(s)}(y_{t-1}, y_t) &:= \Pr_{\boldsymbol{\theta}}\Big(Y_{t-1:t} = y_{t-1:t} \,|\, X_{1:\ell} = x_{1:\ell}^{(s)}\Big). \end{aligned}$$

  using Forward-Backward.

- **M step**: Update parameters

$$\begin{aligned} \pi_i &:= \frac{\sum_{s=1}^n \gamma_1^{(s)}(i)}{\sum_{s=1}^n \sum_{j\in[K]} \gamma_1^{(s)}(j)} \\ A_{i,j} &:= \frac{\sum_{s=1}^n \sum_{t=2}^\ell \xi_t^{(s)}(i,j)}{\sum_{s=1}^n \sum_{t=2}^\ell \sum_{k\in[K]} \xi_t^{(s)}(i,k)} \\ B_{i,j} &:= \frac{\sum_{s=1}^n \sum_{t=1}^\ell \gamma_t^{(s)}(i) \cdot \mathbb{1}\{x_t^{(s)} = j\}}{\sum_{s=1}^n \sum_{t=1}^\ell \gamma_t^{(s)}(i)}. \end{aligned}$$

Generalization to continuous observations

Can have HMMs with continuous observations $\boldsymbol{X}_t$, say, taking values in $\mathbb{R}^d$.

Generalization to continuous observations

Can have HMMs with continuous observations $\boldsymbol{X}_t$, say, taking values in $\mathbb{R}^d$.

Specify conditional densities p_i of $\boldsymbol{X}_t$ given $Y_t = i$ for each $i \in [K]$
(e.g., Gaussians $\mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$).

Generalization to continuous observations

Can have HMMs with continuous observations $\boldsymbol{X}_t$, say, taking values in $\mathbb{R}^d$.

Specify conditional densities p_i of $\boldsymbol{X}_t$ given $Y_t = i$ for each $i \in [K]$ (e.g., Gaussians $\mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$).

- Forward-Backward remains the same, except with By_t, x_t replaced by density value $p_{y_t}(\boldsymbol{x}_t)$.

Generalization to continuous observations

Can have HMMs with continuous observations $\boldsymbol{X}_t$, say, taking values in $\mathbb{R}^d$.

Specify conditional densities p_i of $\boldsymbol{X}_t$ given $Y_t = i$ for each $i \in [K]$ (e.g., Gaussians $\mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$).

- Forward-Backward remains the same, except with B_{y_t, x_t} replaced by density value $p_{y_t}(\boldsymbol{x}_t)$.
- "M step" in E-M maximizes expected complete log likelihood of conditional density parameters (e.g., $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ for Gaussian densities).

$$\sum_{s=1}^{n} \Bigg\{ \sum_{y_1 \in [K]} \gamma_1^{(s)}(y_1) \ln \pi_{y_1} + \sum_{t=2}^{\ell} \sum_{y_{t-1} \in [K]} \sum_{y_t \in [K]} \xi_t^{(s)}(y_{t-1}, y_t) \ln A_{y_{t-1}, y_t} + \sum_{t=1}^{\ell} \sum_{y_t \in [K]} \gamma_t^{(s)}(y_t) \ln p_{y_t}(\boldsymbol{x}_t^{(s)}) \Bigg\}.$$

Recap

- HMM = Markov chain $\{(X_t, Y_t)\}_{t \in \mathbb{N}}$ where hidden state sequence $\{Y_t\}_{t \in \mathbb{N}}$ is a discrete Markov chain; and conditioned on Y_t, observation X_t is independent of everything else.
- Computing sequence probabilities and hidden state conditional probabilities avoids exponential computation due to Markov chain structure.
- Key algorithms: Forward-Backward algorithm (computing conditional probabilities), Viterbi (for most probably hidden state sequence), Baum-Welch (same as E-M for HMMs).
- Many applications: heavily used in speech recognition, bioinformatics, natural language processing, etc.