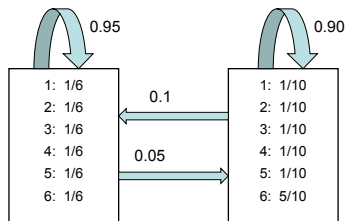


HIDDEN MARKOV MODELS

EXAMPLE: DISHONEST CASINO



Casino die-rolling game:

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

The dice are otherwise indistinguishable!

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** $[K] = \{1, 2, \dots, K\}$):

- ▶ **Initial state distribution:** K -dimensional probability vector π

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

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

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

(rows of $\mathbf{A}$ 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 $[K] = \{1, 2, \dots, K\}$)
(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 $[K] = \{1, 2, \dots, K\}$) (hidden state sequence);
- ▶ conditioned on Y_t , corresponding X_t is *independent of all other variables*;

HIDDEN MARKOV MODELS

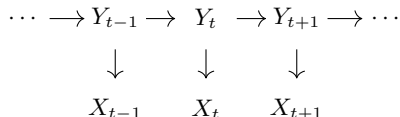
Hidden Markov model (HMM): a Markov chain $\{(X_t, Y_t)\}_{t \in \mathbb{N}}$, where

- ▶ $\{Y_t\}_{t \in \mathbb{N}}$ is also a Markov chain (with state space $[K] = \{1, 2, \dots, K\}$) (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 $[K] = \{1, 2, \dots, K\}$) (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 $[K] = \{1, 2, \dots, K\}$) (**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*.

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

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 $[K] = \{1, 2, \dots, K\}$) (**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*.

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

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

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

HMM PARAMETERS (DISCRETE OBSERVATIONS)

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

HMM PARAMETERS (DISCRETE OBSERVATIONS)

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

- ▶ **Initial state distribution:** K -dimensional probability vector π

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

HMM PARAMETERS (DISCRETE OBSERVATIONS)

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

- **Initial state distribution:** K -dimensional probability vector π

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

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

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

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

HMM PARAMETERS (DISCRETE OBSERVATIONS)

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

- ▶ **Initial state distribution:** K -dimensional probability vector π

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

- ▶ **Transition matrix:** $K \times K$ matrix A

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

(rows of A are probability vectors).

- ▶ **Emission matrix:** $K \times D$ matrix B

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

(rows of B are probability vectors).

CONNECTIONS TO MIXTURE MODELS

Mixture model

 Y $\downarrow$ X

(Y is hidden, X is observed.)

Hidden Markov model

 $Y_1 \rightarrow Y_2 \rightarrow \cdots \rightarrow Y_\ell$ $\downarrow$ $\downarrow$ $\downarrow$ X_1 X_2 X_ℓ

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

CONNECTIONS TO MIXTURE MODELS

Mixture model

 Y $\downarrow$ $\mathbf{X}$

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

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

Hidden Markov model

 $Y_1 \rightarrow Y_2 \rightarrow \cdots \rightarrow Y_\ell$ $\downarrow$ $\downarrow$ $\downarrow$ X_1 X_2 X_ℓ

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

CONNECTIONS TO MIXTURE MODELS

Mixture model

 Y $\downarrow$ X

(Y is hidden, X is observed.)

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

Hidden Markov model

 $Y_1 \rightarrow Y_2 \rightarrow \cdots \rightarrow Y_\ell$ $\downarrow$ $\downarrow$ $\downarrow$ X_1 X_2 X_ℓ

($Y_{1:\ell}$ 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

 Y $\downarrow$ X

(Y is hidden, X is observed.)

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

Hidden Markov model

 $Y_1 \rightarrow Y_2 \rightarrow \cdots \rightarrow Y_\ell$ $\downarrow$ $\downarrow$ $\downarrow$ X_1 X_2 X_ℓ

($Y_{1:\ell}$ 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

 Y $\downarrow$ X

(Y is hidden, X is observed.)

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

Hidden Markov model

 $Y_1 \rightarrow Y_2 \rightarrow \cdots \rightarrow Y_\ell$ $\downarrow$ $\downarrow$ $\downarrow$ X_1 X_2 X_ℓ

($Y_{1:\ell}$ 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!

► $Y_1 \rightarrow Y_2 \rightarrow X_2$

CONNECTIONS TO MIXTURE MODELS

Mixture model

 Y $\downarrow$ X

(Y is hidden, X is observed.)

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

Hidden Markov model

 $Y_1 \rightarrow Y_2 \rightarrow \cdots \rightarrow Y_\ell$ $\downarrow$ $\downarrow$ $\downarrow$ X_1 X_2 X_ℓ

($Y_{1:\ell}$ 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!

- ▶ $Y_1 \rightarrow Y_2 \rightarrow X_2$
- ▶ $X_2 \rightarrow Y_2 \rightarrow Y_3 \rightarrow X_3$

CONNECTIONS TO MIXTURE MODELS

Mixture model

Y

$\downarrow$

X

(Y is hidden, X is observed.)

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

Hidden Markov model

$Y_1 \rightarrow Y_2 \rightarrow \cdots \rightarrow Y_\ell$

$\downarrow$

$\downarrow$

$\downarrow$

X_1

X_2

X_ℓ

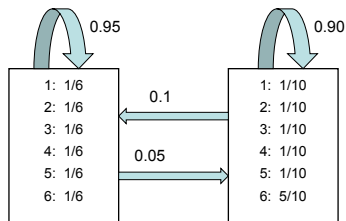
($Y_{1:\ell}$ is hidden, $X_{1:\ell}$ is observed.)

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!

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

EXAMPLE: DISHONEST CASINO

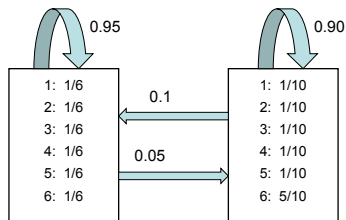


Casino die-rolling game:

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

The dice are otherwise indistinguishable!

EXAMPLE: DISHONEST CASINO



Casino die-rolling game:

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

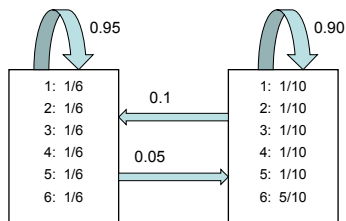
The dice are otherwise indistinguishable!

HMM parameters:

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

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

EXAMPLE: DISHONEST CASINO



Casino die-rolling game:

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

The dice are otherwise indistinguishable!

HMM parameters:

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

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

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

HMM INFERENCE/LEARNING PROBLEMS

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

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

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

HMM INFERENCE/LEARNING PROBLEMS

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

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

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

Most probable state sequence (decoding)

- ▶ **Given:** parameters $\theta = (\pi, \mathbf{A}, \mathbf{B})$, observation sequence $x_{1:\ell} \in [D]^\ell$.
- ▶ **Goal:** $\arg \max_{y_{1:\ell} \in [K]^\ell} \Pr_{\theta}(Y_{1:\ell} = y_{1:\ell} \mid X_{1:\ell} = x_{1:\ell})$.

HMM INFERENCE/LEARNING PROBLEMS

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

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

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

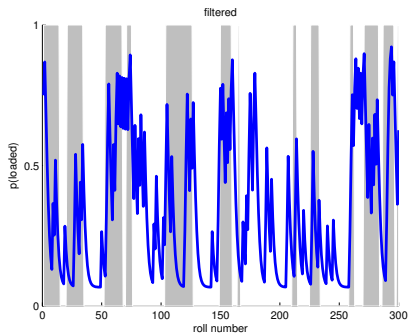
Most probable state sequence (decoding)

- ▶ **Given:** parameters $\theta = (\pi, \mathbf{A}, \mathbf{B})$, observation sequence $x_{1:\ell} \in [D]^\ell$.
- ▶ **Goal:** $\arg \max_{y_{1:\ell} \in [K]^\ell} \Pr_{\theta}(Y_{1:\ell} = y_{1:\ell} \mid X_{1:\ell} = x_{1:\ell})$.

Parameter estimation

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

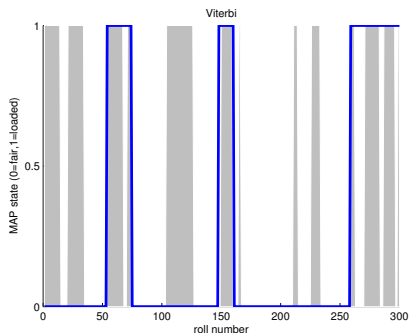
EXAMPLE: DISHONEST CASINO



Conditional probability

Gray bars: Loaded dice used.

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



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

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

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 $\Pr_{\theta}(X_{1:\ell} = x_{1:\ell})$.

SEQUENCE PROBABILITY COMPUTATION

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

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 $\Pr_{\theta}(X_{1:\ell} = x_{1:\ell})$.

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

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

SEQUENCE PROBABILITY COMPUTATION

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

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

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

SEQUENCE PROBABILITY COMPUTATION

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

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

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

SEQUENCE PROBABILITY COMPUTATION

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

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

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

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

SEQUENCE PROBABILITY COMPUTATION

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

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

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

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:

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

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

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

SEQUENCE PROBABILITY COMPUTATION

HMM parameters:

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

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

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

SEQUENCE PROBABILITY COMPUTATION

HMM parameters:

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

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

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

SEQUENCE PROBABILITY COMPUTATION

HMM parameters:

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

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

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

SEQUENCE PROBABILITY COMPUTATION

HMM parameters:

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

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

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

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

CONDITIONAL PROBABILITIES

A simple case: $\Pr_{\theta}(Y_{\ell} = y_{\ell} \mid X_{1:\ell} = x_{1:\ell}) = \frac{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell} \wedge Y_{\ell} = y_{\ell})}{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell})}.$

CONDITIONAL PROBABILITIES

A simple case: $\Pr_{\theta}(Y_{\ell} = y_{\ell} \mid X_{1:\ell} = x_{1:\ell}) = \frac{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell} \wedge Y_{\ell} = y_{\ell})}{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell})}.$

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

CONDITIONAL PROBABILITIES

A simple case: $\Pr_{\theta}(Y_{\ell} = y_{\ell} \mid X_{1:\ell} = x_{1:\ell}) = \frac{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell} \wedge Y_{\ell} = y_{\ell})}{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell})}.$

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

CONDITIONAL PROBABILITIES

A simple case: $\Pr_{\theta}(Y_{\ell} = y_{\ell} \mid X_{1:\ell} = x_{1:\ell}) = \frac{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell} \wedge Y_{\ell} = y_{\ell})}{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell})}.$

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

Forward inductive computation:

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

CONDITIONAL PROBABILITIES

A simple case: $\Pr_{\theta}(Y_{\ell} = y_{\ell} \mid X_{1:\ell} = x_{1:\ell}) = \frac{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell} \wedge Y_{\ell} = y_{\ell})}{\Pr_{\theta}(X_{1:\ell} = x_{1:\ell})}.$

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

Forward inductive computation:

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

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

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

CONDITIONAL PROBABILITIES FOR $t < \ell$

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

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

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

CONDITIONAL PROBABILITIES FOR $t < \ell$

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

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

(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

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

CONDITIONAL PROBABILITIES FOR $t < \ell$

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

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

(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

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

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

CONDITIONAL PROBABILITIES FOR $t < \ell$

$$\beta_t(y_t) = \Pr_{\theta}(X_{t+1:\ell} = x_{t+1:\ell} \mid Y_t = y_t)$$

CONDITIONAL PROBABILITIES FOR $t < \ell$

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

CONDITIONAL PROBABILITIES FOR $t < \ell$

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

CONDITIONAL PROBABILITIES FOR $t < \ell$

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

CONDITIONAL PROBABILITIES FOR $t < \ell$

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

CONDITIONAL PROBABILITIES FOR $t < \ell$

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

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

FORWARD-BACKWARD ALGORITHM

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

FORWARD-BACKWARD ALGORITHM

Given parameters $\theta = (\pi, \mathbf{A}, \mathbf{B})$ 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 $\theta = (\pi, \mathbf{A}, \mathbf{B})$ 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 $\theta = (\pi, \mathbf{A}, \mathbf{B})$ 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 $y_\ell \in [K]$,

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

► **(Also in backward pass)**

Compute conditional probabilities:

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

EVEN MORE COMPLICATED FORWARD-BACKWARD

Can also compute

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

using forward-backward.

EVEN MORE COMPLICATED FORWARD-BACKWARD

Can also compute

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

using forward-backward.

Using Markov property, can string together these probabilities to get

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

EVEN MORE COMPLICATED FORWARD-BACKWARD

Can also compute

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

using forward-backward.

Using Markov property, can string together these probabilities to get

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

Similar procedure for computing most likely state sequence:

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

(Viterbi algorithm).

HMM PARAMETER ESTIMATION

PARAMETER ESTIMATION

Parameter estimation problem:

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

PARAMETER ESTIMATION

Parameter estimation problem:

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

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 $\hat{\theta} = (\hat{\pi}, \hat{A}, \hat{B})$.

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.)

PARAMETER ESTIMATION

Parameter estimation problem:

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

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.)

Recall: basic structure of E-M algorithm

- ▶ Start with some initial guess of parameters θ .
- ▶ (E step) Using current parameters, compute conditional expectations of hidden variables given the observations, as they appear in the *complete log-likelihood*.
- ▶ (M step) Choose new parameters θ to maximize the expected complete log-likelihood function.
- ▶ Repeat E and M steps until convergence.

EXPECTATION-MAXIMIZATION FOR HMMs

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

EXPECTATION-MAXIMIZATION FOR HMMs

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

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

$$\gamma_t^{(s)}(y_t) \quad := \quad \Pr_{\hat{\theta}}\left(Y_t = y_t \mid X_{1:\ell} = x_{1:\ell}^{(s)}\right) \quad \text{for all } y_t \in [K]$$

$$\xi_t^{(s)}(y_{t-1}, y_t) \quad := \quad \Pr_{\hat{\theta}}\left(Y_{t-1:t} = y_{t-1:t} \mid X_{1:\ell} = x_{1:\ell}^{(s)}\right) \quad \text{for all } y_{t-1:t} \in [K]^2$$

using Forward-Backward

EXPECTATION-MAXIMIZATION FOR HMMs

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

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

$$\gamma_t^{(s)}(y_t) := \Pr_{\hat{\theta}}(Y_t = y_t \mid X_{1:\ell} = x_{1:\ell}^{(s)}) \quad \text{for all } y_t \in [K]$$

$$\xi_t^{(s)}(y_{t-1}, y_t) := \Pr_{\hat{\theta}}(Y_{t-1:t} = y_{t-1:t} \mid X_{1:\ell} = x_{1:\ell}^{(s)}) \quad \text{for all } y_{t-1:t} \in [K]^2$$

using Forward-Backward

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

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

EXPECTATION-MAXIMIZATION FOR HMMs

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

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

$$\gamma_t^{(s)}(y_t) := \Pr_{\hat{\theta}}(Y_t = y_t \mid X_{1:\ell} = x_{1:\ell}^{(s)}) \quad \text{for all } y_t \in [K]$$

$$\xi_t^{(s)}(y_{t-1}, y_t) := \Pr_{\hat{\theta}}(Y_{t-1:t} = y_{t-1:t} \mid X_{1:\ell} = x_{1:\ell}^{(s)}) \quad \text{for all } y_{t-1:t} \in [K]^2$$

using Forward-Backward

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

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

Can easily find maximizing parameters θ (subject to constraints that π and rows of A and B are probability distributions).

EXPECTATION-MAXIMIZATION FOR HMMs

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

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

EXPECTATION-MAXIMIZATION FOR HMMs

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

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

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

$$\gamma_t^{(s)}(y_t) \quad := \quad \Pr_{\hat{\theta}} \left(Y_t = y_t \mid X_{1:\ell} = x_{1:\ell}^{(s)} \right)$$

$$\xi_t^{(s)}(y_{t-1}, y_t) \quad := \quad \Pr_{\hat{\theta}} \left(Y_{t-1:t} = y_{t-1:t} \mid X_{1:\ell} = x_{1:\ell}^{(s)} \right).$$

using Forward-Backward.

EXPECTATION-MAXIMIZATION FOR HMMs

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

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

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

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

using Forward-Backward.

- **M step:** Update parameters

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

GENERALIZATION TO CONTINUOUS OBSERVATIONS

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

GENERALIZATION TO CONTINUOUS OBSERVATIONS

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

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

GENERALIZATION TO CONTINUOUS OBSERVATIONS

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

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

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

GENERALIZATION TO CONTINUOUS OBSERVATIONS

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

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

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

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

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.