HIDDEN MARKOV MODEL

- NIPUN BATRA

SOME APPLICATIONS

1. ACTIVITY MONITORING

SOME APPLICATIONS

1 ACTIVITY MONITORING

Accelerometer

Activity WALKING RUNNING SITTING

SOME APPLICATIONS

① ACTIVITY MONITORING

Accelerometer 

Activity WALKING RUNNING SITTING

② SPECH RECOGNITION

RAW SPECH 

SOME APPLICATIONS

1. ACTIVITY MONITORING

Accelerometer

Activity WALKING RUNNING SITTING

2. SPECH RECOGNITION

RAW SPECH

FEATURES F_1, F_2 F_1, F_2 . . — — . . .

WORDS I LIKE MACHINE LEARNING...

SOME APPLICATIONS

1. ACTIVITY MONITORING
2. SPECH RECOGNITION
3. PART OF SPEECH TAGGING
4. GENE FINDING

⋮

SEQUENTIAL DATA

1) Amount of rainfall daily

2) Price of stock

3) Words in a sentence

4) Energy consumption of a home

WEATHER EXAMPLE

Let $x_t \in \begin{cases} R: \text{Rainy} \\ S: \text{Sunny} \\ C: \text{Cloudy} \end{cases}$

be observed outlook at 't^{th}' day

Sample observations: R R R C C R R S S S C S S

Modelling Sequential Data as i.i.d

* Easiest way to treat sequential data ⟶ ignore sequential aspect and treat observations as i.i.d.

DOWNSIDE?

Modelling Sequential Data as i.i.d

* Easiest way to treat sequential data → ignore sequential aspect and treat observations as i.i.d.

DOWNSIDE?

* Fails to exploit sequential patterns
  → Nearby observations "correlated"

Modelling Sequential Data as i.i.d

* Easiest way to treat sequential data $\longrightarrow$ ignore sequential aspect and treat observations as i.i.d.

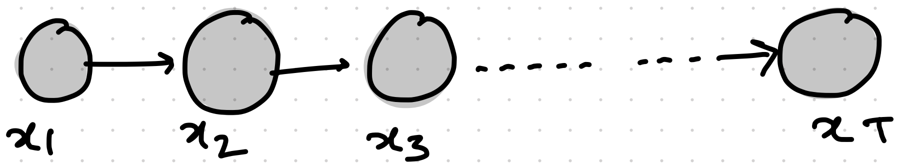
DOWNSIDE?

* Fails to exploit sequential patterns
  $\longrightarrow$ Nearby observations "correlated"

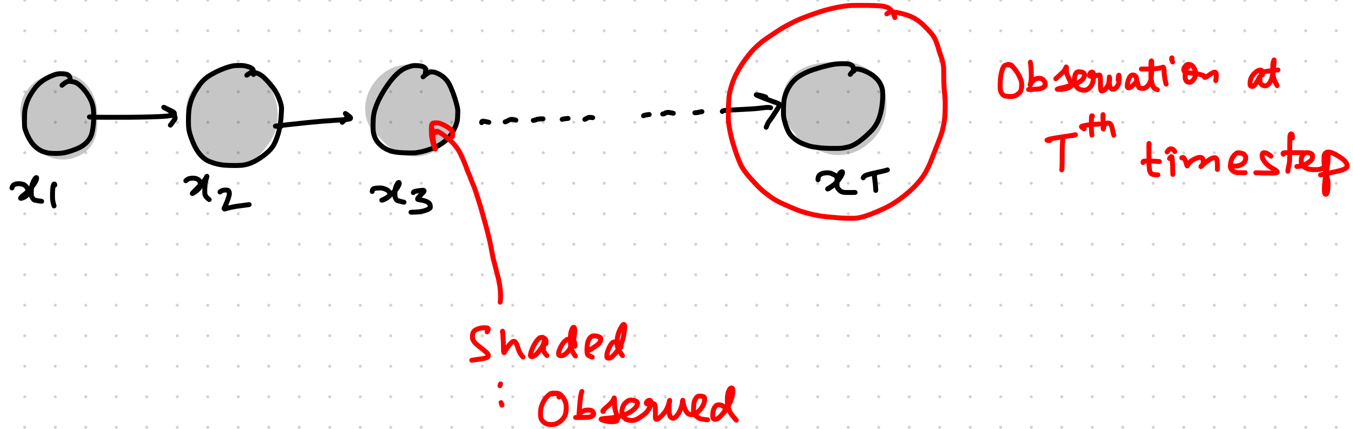
* Q: Predict if it will rain today or not?

  $\because$ Rain today $\xrightarrow[\text{w/ high probability}]{}$ Rain tomorrow

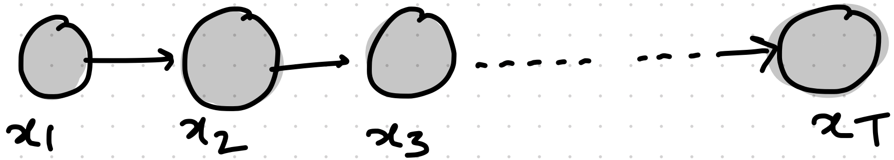
MARKOV MODEL



MARKOV MODEL



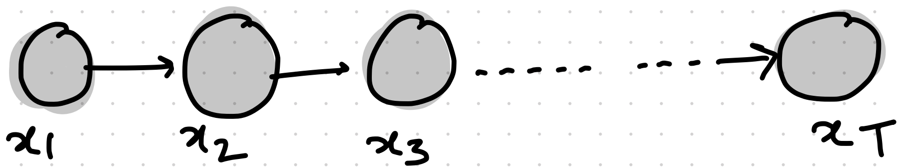
MARKOV MODEL



FIRST ORDER MARKOV CHAIN

Future prediction independent of past given present

MARKOV MODEL

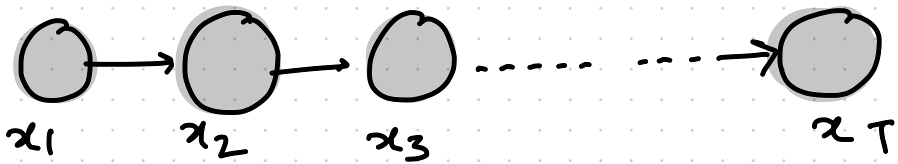


FIRST ORDER MARKOV CHAIN

Future prediction independent of past given present

$$\Rightarrow \quad p(x_{t+1} \mid x_1, x_2 \dots x_t) = p(x_{t+1} \mid x_t)$$

MARKOV MODEL



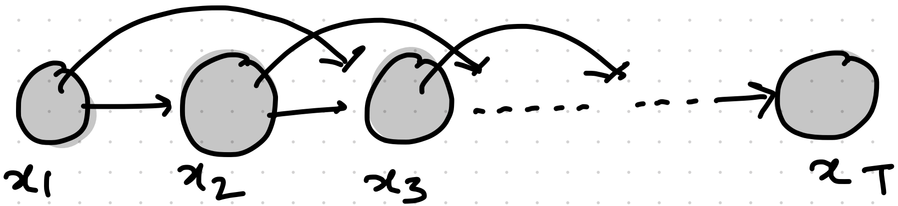
FIRST ORDER MARKOV CHAIN

Future prediction independent of past given present

$$\Rightarrow \quad p(x_{t+1} \mid x_1, x_2 \ldots x_t) = p(x_{t+1} \mid x_t)$$

JOINT PROB $= p(x_1, \ldots, x_T) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2) \cdots p(x_T \mid x_{T-1})$

MARKOV MODEL

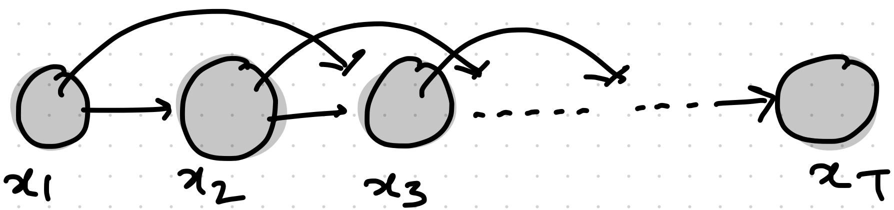


SECOND ORDER MARKOV CHAIN

$$p(x_{t+1} \mid x_1, x_2 \ldots x_t) = p(x_{t+1} \mid x_t, x_{t-1})$$

JOINT PROB $= p(x_1, \ldots, x_T) =$

MARKOV MODEL

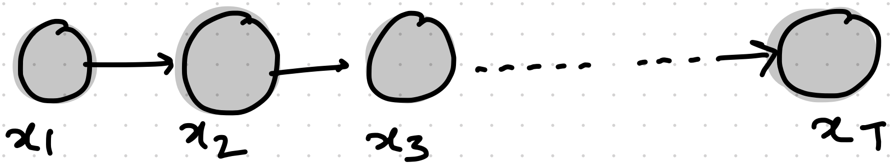


SECOND ORDER MARKOV CHAIN

$$p(x_{t+1} \mid x_1, x_2 \ldots x_t) = p(x_{t+1} \mid x_t, x_{t-1})$$

JOINT PROB $= p(x_1, \ldots, x_T) = p(x_1) \cdot p(x_2 \mid x_1) \cdot \prod_{n=3}^{T} p(x_n \mid x_{n-1}, x_{n-2})$

PARAMETERS OF FIRST ORDER MARKOV MODEL



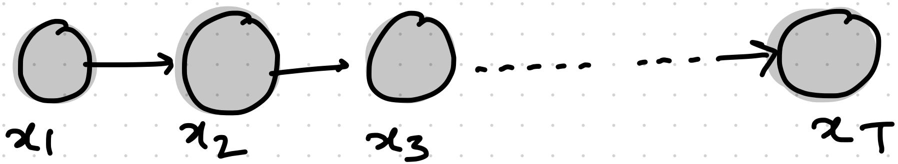
FIRST ORDER MARKOV CHAIN

Future prediction independent of past given present

$$\Rightarrow \; p(x_{t+1} \mid x_1, x_2 \dots x_t) = p(x_{t+1} \mid x_t)$$

JOINT PROB $= p(x_1, \dots, x_T) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2) \dots p(x_T \mid x_{T-1})$

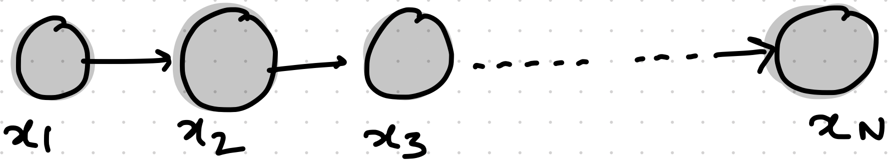
PARAMETERS OF FIRST ORDER MARKOV MODEL



① MAIN ASSUMPTION : STATIONARITY

Data evolves over time, but distribution from which data is generated is fixed.

PARAMETERS OF FIRST ORDER MARKOV MODEL

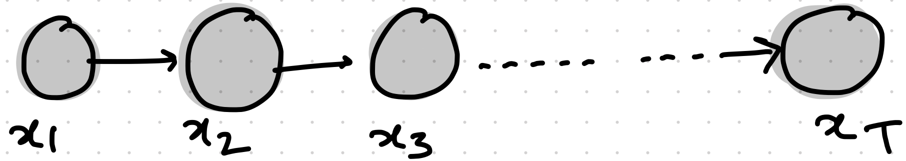


① MAIN ASSUMPTION : STATIONARITY

Data evolves over time, but distribution from which data is generated is fixed.

② We would like to have parameter sharing (parameters independent of time)

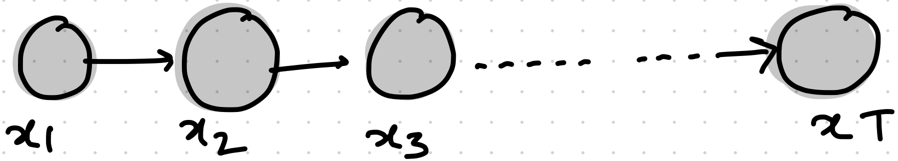
PARAMETERS OF FIRST ORDER MARKOV MODEL



we would like to have parameter sharing

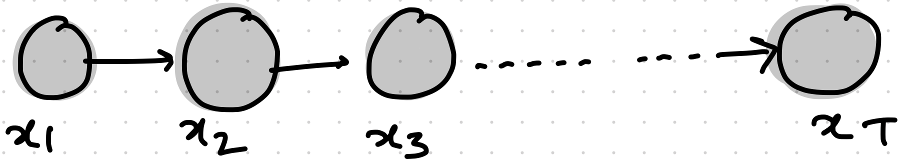
$$p(x_t \mid x_{t-1}) \text{ is same } \forall\, t$$

PARAMETERS OF FIRST ORDER MARKOV MODEL



$$p(x_1, x_2 \ldots x_T) = p(x_1) \prod_{t=2}^{T} p(x_t | x_{t-1})$$

PARAMETERS OF FIRST ORDER MARKOV MODEL



$$p(x_1, x_2 \dots x_T) = p(x_1) \prod_{t=2}^{T} p(x_t | x_{t-1})$$

$$\text{Parameters} = \theta = \{ \pi , A \}$$

Prior probability

Transition matrix

PARAMETERS OF FIRST ORDER MARKOV MODEL

TRANSITION MATRIX (A)

ASSUME x_t can take 1 of K states

$$A_{jk} \equiv p(x_t = k \mid x_{t-1} = j)$$

'j' and 'k' $\in \{1, \ldots K\}$

PARAMETERS OF FIRST ORDER MARKOV MODEL

ASSUME x_t can take 1 of K states

$$A_{jk} \equiv p(x_t = k \mid x_{t-1} = j)$$

'j' and 'k' $\in \{1, \dots K\}$

$j \longrightarrow k$

PARAMETERS OF FIRST ORDER MARKOV MODEL

Representing $A_{jk} \equiv p(x_t = k \mid x_{t-1} = j)$

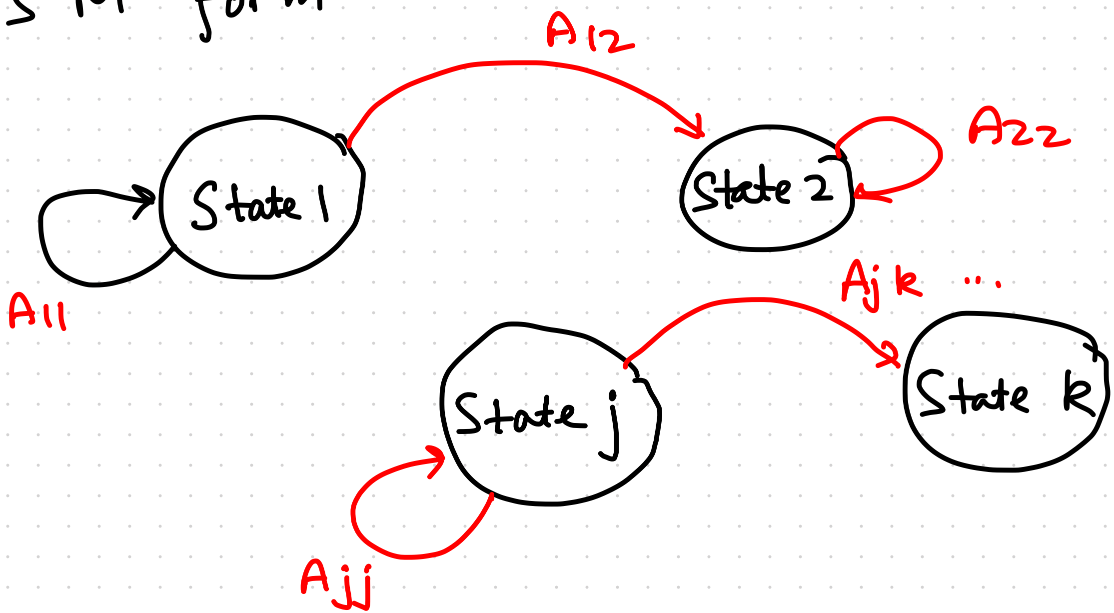
① Tabular form

x_{t-1}	x_t	$p(x_t \mid x_{t-1})$
1	1	A_{11}
2	1	A_{21}
$\vdots$		$\cdot$
j	k	A_{jk}
K	K	A_{KK}

PARAMETERS OF FIRST ORDER MARKOV MODEL

Representing $A_{jk} \equiv p(x_t = k \mid x_{t-1} = j)$

② FSM form



PARAMETERS OF FIRST ORDER MARKOV MODEL

Representing $A_{jk} \equiv p(x_t = k \mid x_{t-1} = j)$

③ Adjacency Matrix

$$
\begin{array}{c|ccccc}
x_{t-1} \backslash x_t & 1 & 2 & \cdots & k & K \\
\hline
1 & & & & & \\
2 & & & & & \\
\vdots & & & & & \\
j & & & & A_{jk} & \\
\end{array}
$$

PARAMETERS OF FIRST ORDER MARKOV MODEL

Prior Probability (π)

$$\pi_k = p(z_1 = k)$$

Probability of 1st observation

Q) Given A & π as:

$$A = \begin{array}{c} \\ R \\ C \\ S \end{array} \begin{array}{c} \begin{array}{ccc} R & C & S \end{array} \\ \begin{bmatrix} .4 & .3 & .3 \\ .2 & .6 & .2 \\ .1 & .1 & .8 \end{bmatrix} \end{array}$$

$$\pi = \begin{array}{c} \begin{bmatrix} .2 & .4 & .4 \end{bmatrix} \\ \begin{array}{ccc} R & C & S \end{array} \end{array}$$

WHAT IS PROBALITY OF SEQUENCE

$$x = \{R\ R\ S\ S\ C\}$$

Q) Given A & π as:

$$A = \begin{array}{c|ccc} & R & C & S \\ \hline R & .4 & .3 & .3 \\ C & .2 & .6 & .2 \\ S & .1 & .1 & .8 \end{array}$$

$$\pi = \begin{array}{ccc} .2 & .4 & .4 \\ R & C & S \end{array}$$

WHAT IS PROBALITY OF SEQUENCE

$$x = \{R\ R\ S\ S\ C\}$$

$$P(x \mid \theta = \{A, \pi\}) = P(x_1 = R) \cdot P(x_2 = R \mid x_1 = R) \cdot P(x_3 = S \mid x_2 = R) \cdot P(x_4 = S \mid x_3 = S) \cdot P(x_5 = C \mid x_4 = S)$$

$$= \pi_R \cdot A_{RR} \cdot A_{RS} \cdot A_{SS} \cdot A_{SC}$$

Q) Given A & π as:

$$A = \begin{array}{c} \\ R \\ C \\ S \end{array} \begin{array}{c} \begin{array}{ccc} R & C & S \end{array} \\ \begin{bmatrix} .4 & .3 & .3 \\ .2 & .6 & .2 \\ .1 & .1 & .8 \end{bmatrix} \end{array}$$

$$\pi = \begin{array}{c} \begin{bmatrix} .2 & .4 & .4 \end{bmatrix} \\ \begin{array}{ccc} R & C & S \end{array} \end{array}$$

$p(x_1 = R) = 0.2$. what is $p(x_2 = R)$?

Q) Given A & π as:

$$A = \begin{array}{c|ccc} & R & C & S \\ \hline R & .4 & .3 & .3 \\ C & .2 & .6 & .2 \\ S & .1 & .1 & .8 \end{array} \qquad \pi = \begin{array}{ccc} R & C & S \\ \hline .2 & .4 & .4 \end{array}$$

$p(x_1 = R) = 0.2$. what is $p(x_2 = R)$?

$$p(x_2 = R) = p(x_1 = R) . p(x_2 = R \mid x_1 = R) + p(x_1 = S) . p(x_2 = R \mid x_1 = S) + p(x_1 = C) . p(x_2 = R \mid x_1 = C)$$

Q) Given A & π as:

$$A = \begin{array}{c c} & \begin{array}{ccc} R & C & S \end{array} \\ \begin{array}{c} R \\ C \\ S \end{array} & \begin{bmatrix} .4 & .3 & .3 \\ .2 & .6 & .2 \\ .1 & .1 & .8 \end{bmatrix} \end{array} \qquad \pi = \begin{array}{c} \begin{bmatrix} .2 & .4 & .4 \end{bmatrix} \\ \begin{array}{ccc} R & C & S \end{array} \end{array}$$

$p(x_1 = R) = 0.2$. what is $p(x_2 = R)$?

$$p(x_2 = R) = p(x_1 = R) . p(x_2 = R \mid x_1 = R) +$$
$$p(x_1 = S) . p(x_2 = R \mid x_1 = S) +$$
$$p(x_1 = C) . p(x_2 = R \mid x_1 = C)$$

$$= (.2) * (.4) + (.4) * (0.1) + (.4) * (0.2) = 0.2 = p(x_1 = R)$$

Hidden Markov Model

what is the hidden component

Hidden Markov Model

what is the hidden component

Example: Coin Toss Model

Assume 2 coins
- Biased ($p(H) = 0.7$)
- Unbiased / Fair ($p(H) = 0.5$)

Hidden Markov Model

What is the hidden component

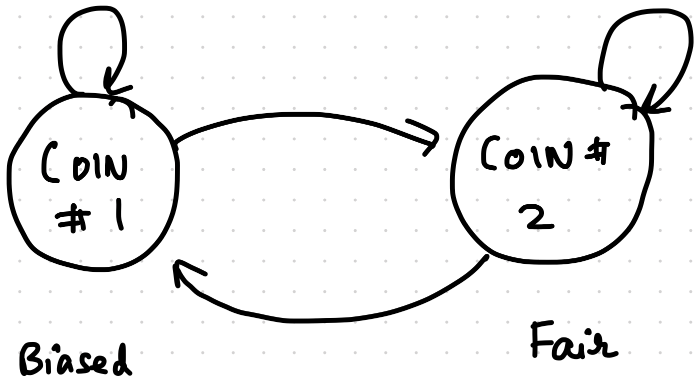
Example: Coin Toss Model

Assume 2 coins → Biased $(p(H) = 0.7)$
→ Unbiased / Fair $(p(H) = 0.5)$

We see only sequence of observations $x = \{H, T \ldots\}$

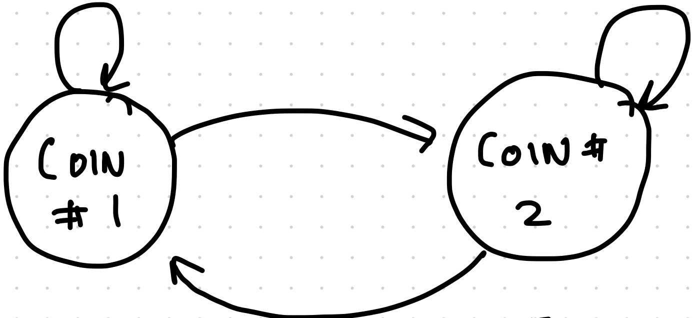
Hidden: which coin was tossed!

HMM (example)



Markov Model

HMM (example)



Markov Model

Biased

Fair

Emission

$p(H) = 0.7$
$p(T) = 0.3$

$p(H) = 0.5$
$p(T) = 0.5$

HMM (example)

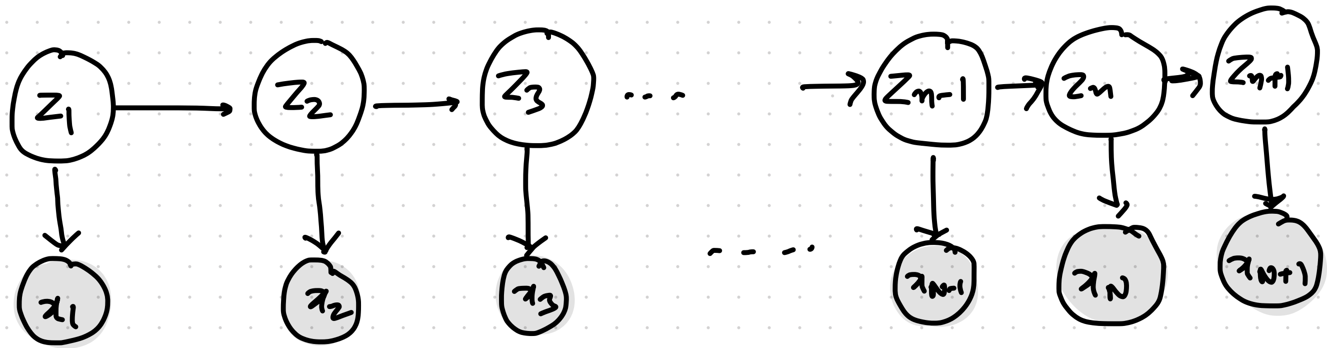
$z_i = \{\text{coin } 1, \text{coin } 2\}$

$$z_1 \to z_2 \to z_3 \cdots \to z_{n-1} \to z_n \to z_{n+1}$$

HMM (example)

$Z_i = \{\text{coin 1}, \text{coin 2}\}$

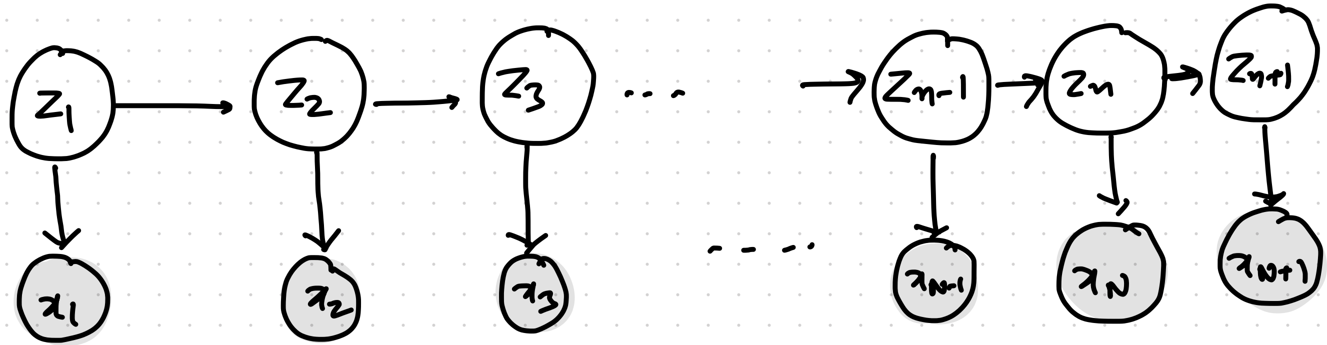
$x_i = \{H, T\}$



HMM (example)

$z_i = \{\text{coin 1}, \text{coin 2}\}$

$x_i = \{H, T\}$



x_i : Observed

z_i : Latent state

HMM Parameters

1. Transition matrix A:

$$A_{jk} = p(z_n = k \mid z_{n-1} = j)$$

HMM Parameters

① Transition matrix A:

$$A_{jk} = p(z_n = k \mid z_{n-1} = j)$$

② Prior prob. π:

$$\pi_k = p(z_1 = k)$$

HMM Parameters

1. Transition matrix A:

$$A_{jk} = p(z_n = k \mid z_{n-1} = j)$$

2. Prior prob. π:

$$\pi_k = p(z_1 = k)$$

3. Emission Probability : ϕ define $p(x_n \mid z_n, \phi)$

HMM Parameters

(3) Emission Probability : ϕ define $p(x_n | z_n, \phi)$

Case I : ϕ is discrete.

e.g. $z_i = \{\text{Fair}, \text{Biased}\}$

ϕ defined in terms of conditional probability

Fair	Biased
$P(H) = 0.5$	$P(H) = 0.7$
$P(T) = 0.5$	$P(T) = 0.3$

HMM Parameters

③ Emission Probability : ϕ define $p(x_n | z_n, \phi)$

<u>Case II</u> ϕ is continuous

State: Appliance is ON / OFF

ϕ : ON : Power $\sim N(100, 10)$

OFF : Power $\sim N(0, 5)$

SOME APPLICATIONS

(1) ACTIVITY MONITORING

Accelerometer

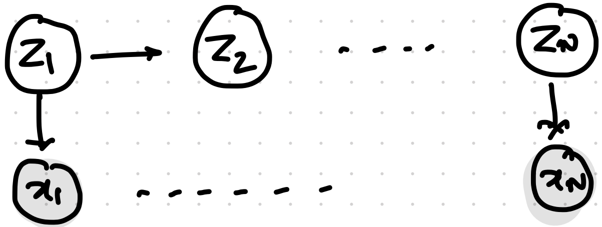
Activity

WALKING z_1 | RUNNING z_2 | SITTING z_3

$z_i \in \{$ WALKING, RUNNING, SITTING $\}$

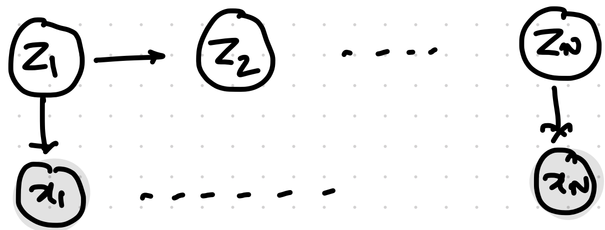
$x_i \in$ (SOME TRANSFORMATION OF RAW DATA)

PI: HMM sampling



Given π, A, ϕ generate $X = \{x_1 \dots x_N\}$ & $Z = \{z_1 \dots z_N\}$

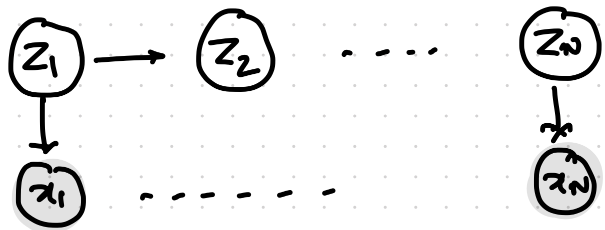
PI: HMM sampling



Given π, A, ϕ generate $X = \{x_1 \ldots x_N)$ & $Z = \{z_1 \ldots z_N\}$

① CHOOSE z_1 as per π

PI: HMM sampling

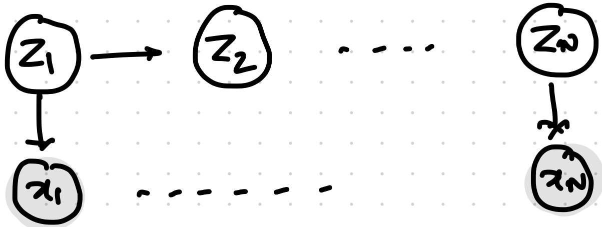


Given π, A, ϕ generate $X = \{x_1 \ldots x_N\}$ & $Z = \{z_1 \ldots z_N\}$

① CHOOSE z_1 as per π

② Sample x_1 using ϕ and z_1

PI: HMM sampling



Given π, A, ϕ generate $X = \{x_1 \ldots x_N)$ & $Z = \{z_1 \ldots z_N\}$

① CHOOSE z_1 as per π

② Sample x_1 using ϕ and z_1

③ For $n = 2 : N$

(3.1) Sample z_n from z_{n-1} using A and z_{n-1}

(3.2) Sample x_n from z_n using ϕ and z_n

PI: HMM sampling

Notebooks: Discrete HMM Simulatⁿ (Unfair Casino)

Continuous HMM Simulatⁿ (Power Appliance)

P II HMM Evidence likelihood

$X = \{x_1 \dots x_T\}$

Given X, $\underbrace{\pi, A, \phi}_{\theta}$ what is $L(X|\theta)$?

PII: HMM Evidence likelihood

$X = \{x_1 \dots x_T\}$

Given X, $\underbrace{\pi, A, \phi}_{\theta}$ what is $L(X|\theta)$?

Likelihood $= p(X|\theta) = \sum_{z} p(X, z|\theta)$

(Marginalisat^n)

FOR $X = \{H, H\}$ what is $L(X|\theta)$ where $\theta = ?$

$$\pi = \begin{array}{cc} [\,0.6 & 0.4\,] \\ \text{Bias} & \text{Fair} \end{array}$$

$$A = \begin{array}{ccc} & B & F \\ \text{Bias} & 0.9 & 0.1 \\ \text{Fair} & 0.1 & 0.9 \end{array}$$

$\phi =$

Bias

$p(H) = 0.7$

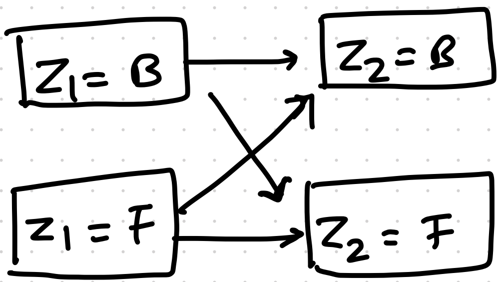
$p(T) = 0.3$

Fair

$p(H) = 0.5$

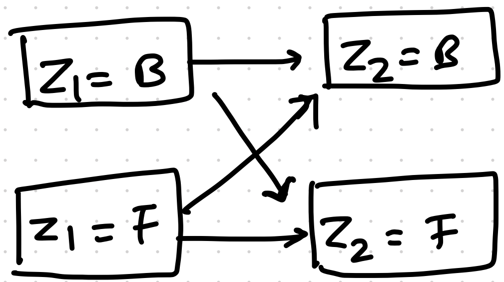
$p(T) = 0.5$

FOR $X = \{H, H\}$ what is $L(X|\theta)$ where $\theta = ?$



Trellis Diagram
for 2 states
and 2 timesteps

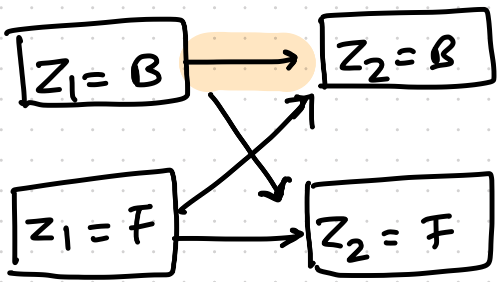
FOR $X = \{H, H\}$ what is $L(X|\theta)$ where $\theta =$



Trellis Diagram for 2 states and 2 timesteps

$$L(X|\theta) = p(z_1 = B) \cdot p(x_1 = H | z = B) \cdot p(z_2 = B | z_1 = B) \cdot p(x_2 = H | z = B)$$

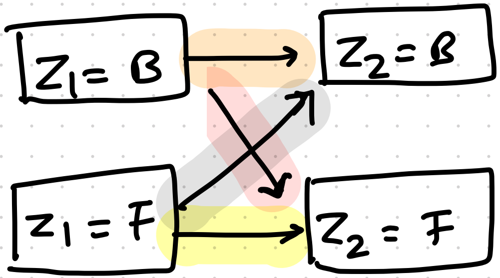
FOR $X = \{H, H\}$ what is $L(X|\theta)$ where $\theta =$



Trellis Diagram
for 2 states
and 2 timesteps

$$L(X|\theta) = p(z_1 = B) \cdot p(x_1 = H \mid z = B) \cdot p(z_2 = B \mid z_1 = B) \cdot p(x_2 = H \mid z = B)$$

FOR $X = \{H, H\}$ what is $L(X|\theta)$ where $\theta = ?$



Trellis Diagram for 2 states and 2 timesteps

$$L(X|\theta) = p(z_1 = B) \cdot p(x_1 = H | z = B) \cdot p(z_2 = B | z_1 = B) \cdot p(x_2 = H | z = B)$$

$$+ \; p(z_1 = F) \cdot p(x_1 = H | z = F) \cdot p(z_2 = B | z_1 = F) \cdot p(x_2 = H | z = B)$$

$$+ \; p(z_1 = B) \cdot p(x_1 = H | z = B) \cdot p(z_2 = F | z_1 = B) \cdot p(x_2 = H | z = F)$$

$$+ \; p(z_1 = F) \cdot p(x_1 = H | z = F) \cdot p(z_2 = F | z_1 = F) \cdot p(x_2 = H | z = F)$$

FOR $X = \{H, H\}$ what is $L(X|\theta)$ where $\theta = 2$

$$L(X|\theta) = p(z_1 = B) \cdot p(x_1 = H | z = B) \cdot p(z_2 = B | z_1 = B) \cdot p(x_2 = H | z = B)$$

$$+ \; p(z_1 = F) \cdot p(x_1 = H | z = F) \cdot p(z_2 = B | z_1 = F) \cdot p(x_2 = H | z = B)$$

$$+ \; p(z_1 = B) \cdot p(x_1 = H | z = B) \cdot p(z_2 = F | z_1 = B) \cdot p(x_2 = H | z = F)$$

$$+ \; p(z_1 = F) \cdot p(x_1 = H | z = F) \cdot p(z_2 = F | z_1 = F) \cdot p(x_2 = H | z = F)$$

$$= (0.6)(0.7)(0.9)(0.7) \qquad = 0.2646 \qquad = 0.3896$$

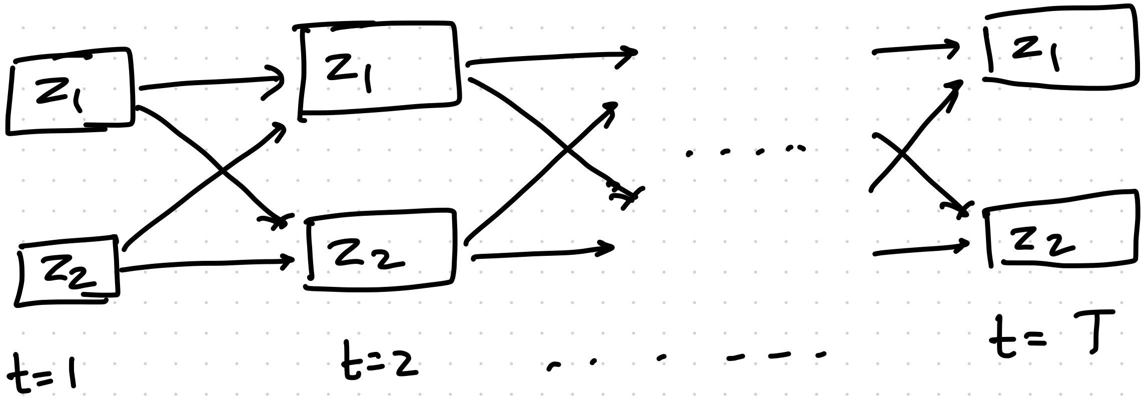
$$+ (0.4)(0.5)(0.1)(0.7) \qquad + 0.0140$$

$$+ (0.6)(0.7)(0.1)(0.5) \qquad + 0.0210$$

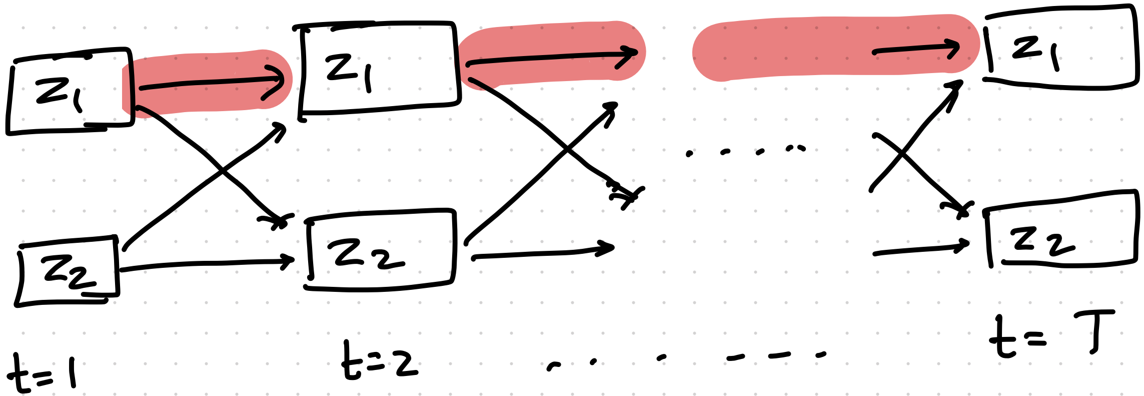
$$+ (0.4)(0.5)(0.9)(0.5) \qquad + 0.0900$$

FOR $X = \{\underbrace{H, H, \ldots}_{T \text{ timesteps.}}\}$ what is $L(X|\theta)$ where $\theta = 1$

FOR $X = \{\underbrace{H, H, \ldots}_{T \text{ timesteps.}}\}$ what is $L(X|\theta)$ where $\theta =$?



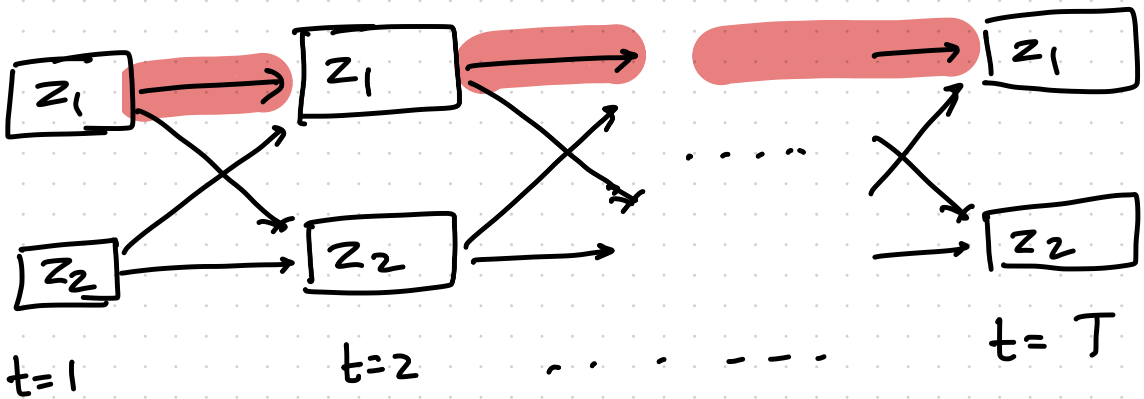
FOR $X = \{\underbrace{H, H, \ldots}_{T \text{ timesteps}}\}$ what is $L(X|\theta)$ where $\theta = ?$



SNO	1	2	TIME	T
1	z_1	z_1		z_1
⋮				
K^T	...	...		..

FOR $X = \{H, H, \ldots\}$ what is $L(X|\theta)$ where $\theta = ?$

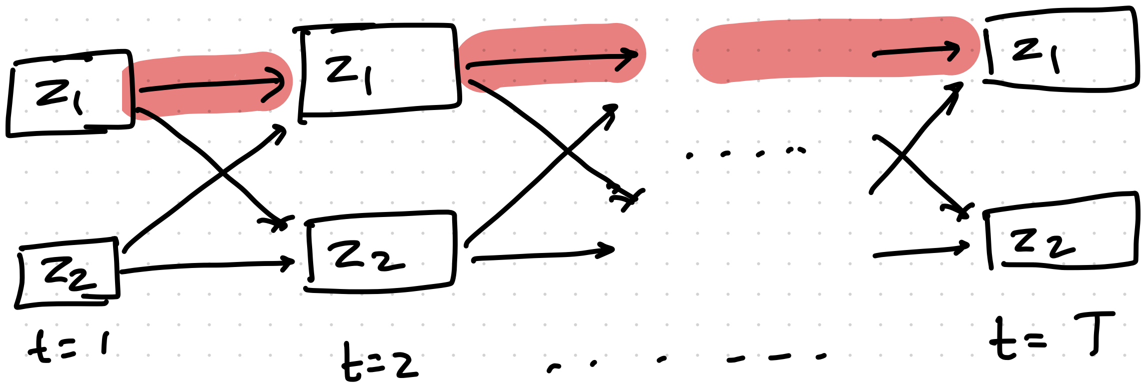
$\underline{\hphantom{H, H, \ldots}}$ T timesteps.



SNO	TIME 1	2	...	T
1	z_1	z_1		z_1
⋮				
K^T	...	...		..

Total K^T paths required for $L(X|\theta)$

FOR $X = \underbrace{\{H, H, \ldots\}}_{T \text{ timesteps.}}$ what is $L(X|\theta)$ where $\theta = ;$



Length of each path $= T$

$\Rightarrow O(T)$ multiplications

Total K^T paths required for $L(X|\theta)$

$P(X|\theta)$ computat^n is $O(TK^T)$

Can we do better? Hint: Dynamic Programming

FORWARD PROCEDURE

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$

FORWARD PROCEDURE

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$

i.e. $\alpha_t(i)$ = Probability of being in state 'i' at time t and observatⁿ $x_{1:t}$

FORWARD PROCEDURE (Base Step $t=1$)

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$

For our example for $t=1$

$$\alpha_1(B) = P(x_1, z_1 = \text{Biased}) = \pi_B \cdot P(x_1 = H \mid z_1 = B)$$

$$= 0.6 * 0.7 = 0.42$$

FORWARD PROCEDURE (Base Step $t=1$)

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$

For our example for $t=1$

$$\alpha_1(\text{Bias}) = P(x_1, z_1 = \text{Biased}) = \pi_B \cdot P(x_1 = H \mid z_1 = B)$$

$$= 0.6 * 0.7 = 0.42$$

$$\alpha_1(\text{Fair}) = P(x_1 = H, z_1 = F) = \pi_F \cdot P(x_1 = H \mid z_1 = F)$$

$$= 0.4 * 0.5 = 0.20$$

FORWARD PROCEDURE (INDUCTION STEP)

Define $\alpha_t(i) = P(x_{1:t}, Z_t = i)$

$\alpha_1 = 0.42$

$\boxed{Z_B}$

$\boxed{Z_F}$

$\alpha_1 = 0.2$

FORWARD PROCEDURE (INDUCTION STEP)

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$

$\alpha_1 = 0.42$ $\qquad\qquad$ $\alpha_2(B) = ?$

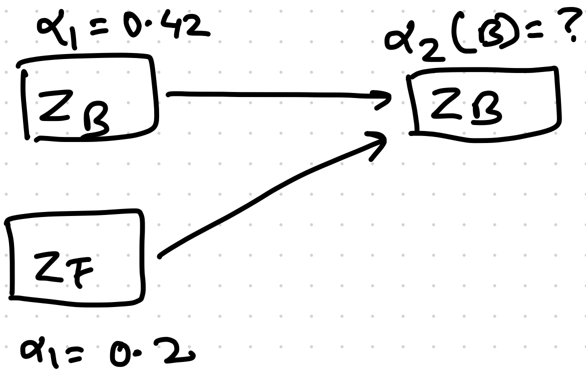
$\boxed{Z_B}$ $\qquad\qquad$ $\boxed{Z_B}$

$\boxed{Z_F}$

$\alpha_1 = 0.2$

FORWARD PROCEDURE (INDUCTION STEP)

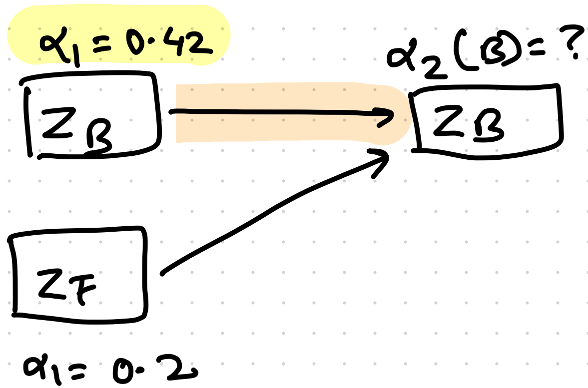
Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$



2 paths (K) to Bias state at time step 2

FORWARD PROCEDURE (INDUCTION STEP)

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$

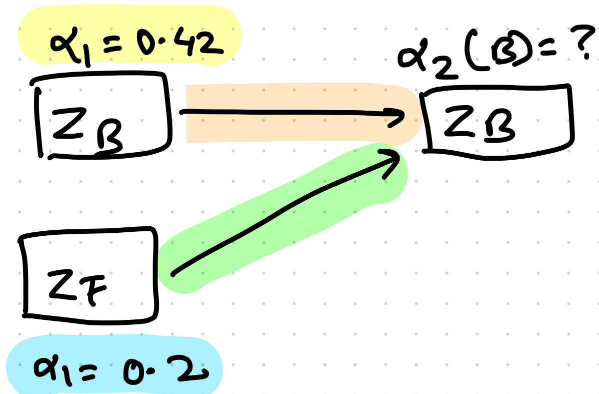


2 paths (K) to Bias state at time step 2

$$P(x_1 = H, x_2 = H, z_2 = B) = \left[P(x_1 = H, z_1 = B) \cdot A_{BB}\right] P(H \mid Bias)$$

FORWARD PROCEDURE (INDUCTION STEP)

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$



2 paths (K) to Bias state at time step 2

$$P(x_1 = H, x_2 = H, z_2 = B) = [P(x_1 = H, z_1 = B) \cdot A_{BB}] \, P(H \mid Bias)$$

$$+ [P(x_1 = H, z_1 = F) \cdot A_{FB}] \, P(H \mid B)$$

FORWARD PROCEDURE (INDUCTION STEP)

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$

$$P(x_1 = H, x_2 = H, z_2 = B) = [P(x_1 = H, z_1 = B) . A_{BB}] P(H|Bias)$$
$$+ [P(x_1 = H, z_1 = F) . A_{FB}] P(H|B)$$

$$\Rightarrow \alpha_2(B) = (\alpha_1(B) A_{BB}) \phi_B(x=H) + (\alpha_1(F) A_{FB}) \phi_B(x=H)$$

$$= (0.42) * (0.9) * (0.7) + (0.2) * (0.1) * (0.7)$$

$$= 0.2646 + 0.014$$

$$= 0.2786$$

FORWARD PROCEDURE (INDUCTION STEP)

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$

$$\alpha_{t+1}(j) = \left\{ \sum_i \alpha_t(i) * A_{ij} \right\} \phi_j(x_{t+1})$$

FORWARD PROCEDURE (INDUCTION STEP)

Define $\alpha_t(i) = P(x_{1:t}, z_t = i)$

$\alpha_1 = 0.42$ $\qquad$ $\alpha_2(B) = 0.2786$

$\boxed{Z_B}$ $\qquad$ $\boxed{Z_B}$

$\boxed{Z_F}$ $\qquad$ $\boxed{Z_F}$

$\alpha_1 = 0.2$ $\qquad$ $\alpha_2(F) = 0.1110$

FORWARD PROCEDURE (Termination)

$$\alpha_T(B) = P(x_{1:T}, Z_T = B)$$

$$\alpha_T(F) = P(x_{1:T}, Z_T = F)$$

$$\therefore \; P(x_{1:T}) = \alpha_T(B) + \alpha_T(F)$$

FORWARD PROCEDURE (Termination)

$$\alpha_T(B) = P(x_{1:T}, Z_T = B)$$

$$\alpha_T(F) = P(x_{1:T}, Z_T = F)$$

$$\therefore \; P(x_{1:T}) = \alpha_T(B) + \alpha_T(F)$$

In general,

$$P(x_{1:T} \mid \theta) = \sum_i \alpha_T(i)$$

or

$$L(X_{1:T} \mid \theta) = \sum_i \alpha_T(i)$$

P III HMM Filtering

Compute belief state

$p(z_t \mid x_{1:t})$ online (or recursively)

P III HMM Filtering

Compute belief state

$p(z_t \mid x_{1:t})$ online (or recursively)

$$= \frac{p(z_t, x_{1:t})}{p(x_{1:t})} = \frac{\alpha_t(i)}{\sum_i \alpha_t(i)}$$

P III HMM Filtering.

$$P(Z_1 = B \mid X_1 = H) = \frac{\alpha_1(B)}{\alpha_1(B) + \alpha_1(F)} = \frac{0.42}{0.62} \approx \frac{2}{3}$$

$$P(Z_1 = F \mid X_1 = H) = \frac{\alpha_1(F)}{\alpha_1(B) + \alpha_1(F)} = \frac{0.2}{0.62} \approx \frac{1}{3}$$

P III HMM Filtering.

$$P(Z_2 = B \mid X = \{HH\}) = \frac{\alpha_2(B)}{\alpha_2(B) + \alpha_2(F)} = 0.715$$

$$P(Z_2 = F \mid X = \{HH\}) = \frac{\alpha_2(F)}{\alpha_2(B) + \alpha_2(F)} = 0.289$$

P III HMM Filtering.

$$P(Z_2 = B \mid X = \{HH\}) = \frac{\alpha_2(B)}{\alpha_2(B) + \alpha_2(F)} = 0.715$$

$$P(Z_2 = F \mid X = \{HH\}) = \frac{\alpha_2(F)}{\alpha_2(B) + \alpha_2(F)} = 0.289$$

Note

$$P(Z_2 = B \mid \{H\,H\}) > P(Z_1 = B \mid \{H\})$$

$\because$ Biased coin $\Rightarrow$ more heads!

PIV : $P(X_{t+1:T} \mid Z_t = i, \theta)$?

Backwards algorithm

$$\beta_t(i) = P(X_{t+1:T} \mid Z_t = i, \theta)$$

= Probability of sequence $X_{t+1:T}$ given state is 'i' at t^{th} time instance

$\underline{P\overline{IV}}$: $P(X_{t+1:T} \mid Z_t = i, \theta)$?

$$\beta_t(i) = P(X_{t+1:T} \mid Z_t = i, \theta)$$

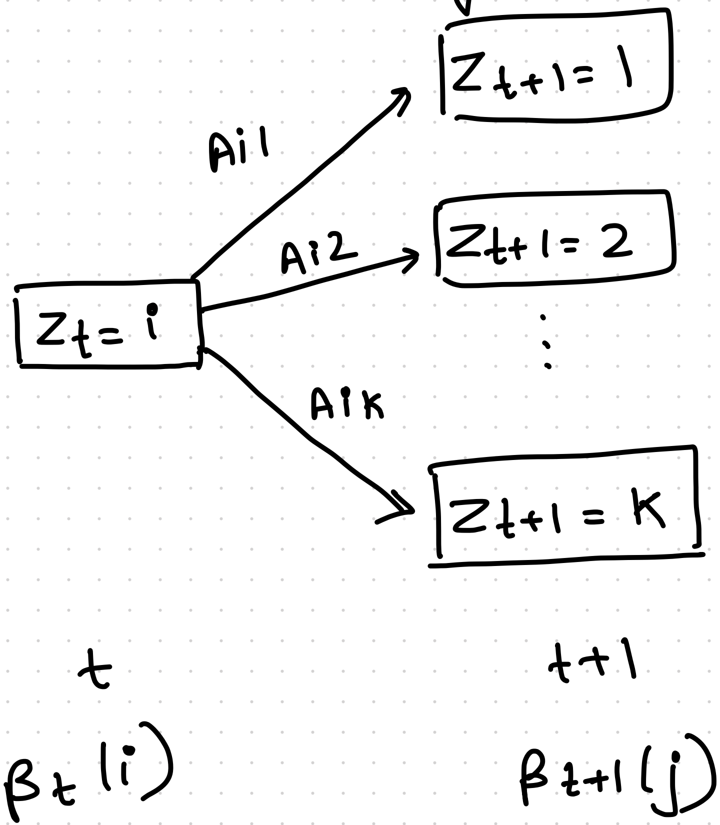
e.g. $X = \{H, H, H\}$

$T = 3$

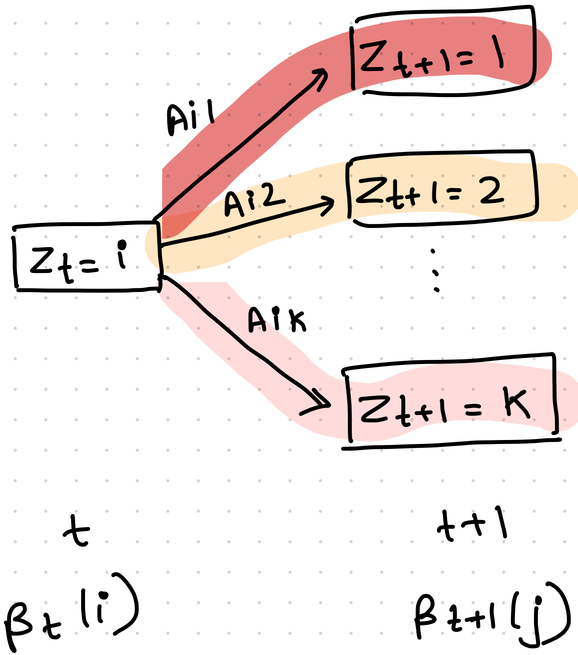
$t = 1$

$Z_1 =$ Biased

Backwards Algorithm



Backwards Algorithm



$$\beta_t(i) = P(x_{t+1:T} \mid z_t = i)$$

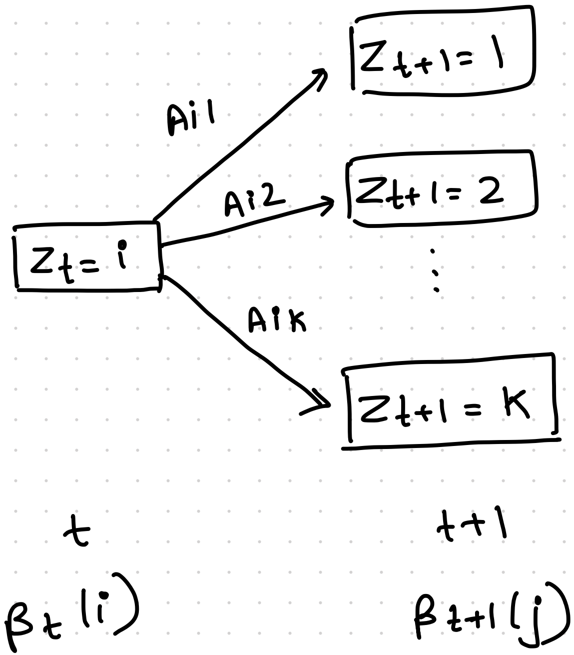
$$= P(x_{t+2:T} \mid z_{t+1} = 1) \cdot A_{i1} \cdot \phi_1(x_{t+1})$$

$$+ P(x_{t+2:T} \mid z_{t+1} = 2) \cdot A_{i2} \cdot \phi_2(x_{t+1})$$

$$\vdots$$

$$+ P(x_{t+2:T} \mid z_{t+1} = K) \cdot A_{iK} \cdot \phi_K(x_{t+1})$$

Backwards Algorithm

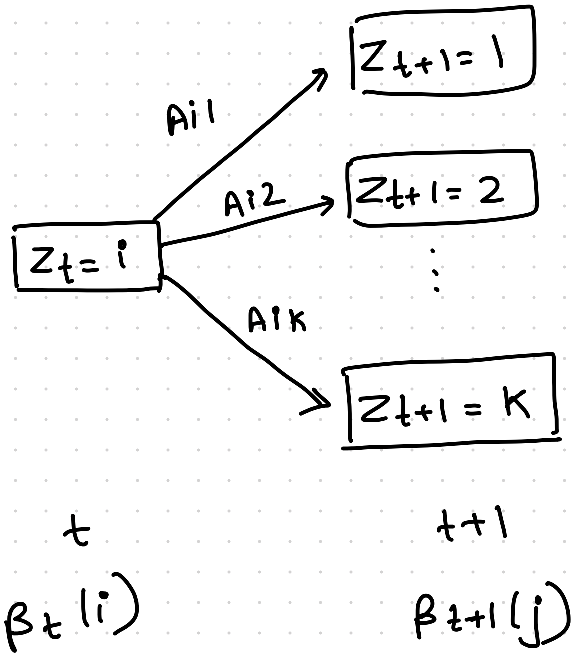


INDUCTION STEP

$$\beta_t(i) = P(x_{t+1:T} \mid z_t = i)$$

$$= \sum_{j=1}^{K} \beta_{t+1}(j) \cdot A_{ij} \cdot \phi_j(x_{t+1})$$

Backwards Algorithm



INDUCTION STEP

$$\beta_t(i) = P(x_{t+1:T} \mid z_t = i)$$

$$= \sum_{j=1}^{K} \beta_{t+1}(j) \cdot A_{ij} \cdot \phi_j(x_{t+1})$$

$t = T-1, \ldots 1$

INIT. STEP

$$\beta_T(i) = 1 \quad \forall i \in \{1, \ldots K\}$$

(Arbitrarily defined)

Backwards Algorithm (example)

$X = \{H, H, H\}$

Find $\beta_1, \beta_2, \beta_3$

INIT

$\beta_3(B) = 1 = \beta_3(F)$ (Arbitrary)

Backwards Algorithm (example)

$X = \{H, H, H\}$ Find $\beta_1, \beta_2, \beta_3$

INIT

$$\beta_3(B) = 1 = \beta_3(F) \quad \text{(Arbitrary)}$$

STEP

$$\beta_2(B) = \sum_{j=\{B,F\}} \beta_3(j) \cdot A_{Bj} \cdot \phi_j(H)$$

$$= \beta_3(B) . A_{BB} \cdot \phi_B(H) + \beta_3(F) . A_{BF} \cdot \phi_F(H)$$

$$= 1 \cdot (0 \cdot 9) \cdot (0 \cdot 7) + (1)(0 \cdot 1)(0 \cdot 5) = 0 \cdot 68$$

Backwards Algorithm (example)

$X = \{H, H, H\}$ Find $\beta_1, \beta_2, \beta_3$

INIT

$$\beta_3(B) = 1 = \beta_3(F) \quad \text{(Arbitrary)}$$

STEP

$$\beta_2(B) = \sum_{j=\{B,F\}} \beta_3(j) \cdot A_{Bj} \cdot \phi_j(H)$$

$$= \beta_3(B) \cdot A_{BB} \cdot \phi_B(H) + \beta_3(F) \cdot A_{BF} \cdot \phi_F(H)$$

$$= 1 \cdot (0.9) \cdot (0.7) + (1)(0.1)(0.5) = 0.68$$

$$\beta_2(F) = \beta_3(B) \cdot A_{FB}\, \phi_B(H) + \beta_3(F) \cdot A_{FF} \cdot \phi_F(H)$$

$$= (1) \cdot (0.1)(0.7) + (1)(0.9)(0.5) = 0.52$$

Backwards Algorithm (example)

$X = \{H, H, H\}$ Find $\beta_1, \beta_2, \beta_3$

INIT

$\beta_3(B) = 1 = \beta_3(F)$ (Arbitrary)

STEP

$$\beta_1(B) = \beta_2(B) . A_{BB} . \phi_B(H) + \beta_2(F) . A_{BF} . \phi_F(H)$$
$$= (0.68)(0.9)(0.7) + (0.52).(0.1)(0.5) = 0.4544$$

$$\beta_1(F) = \beta_2(B) . A_{FB} . \phi_B(H) + \beta_2(F) . A_{FF} . \phi_F(H)$$
$$= (0.68)(0.1)(0.7) + (0.52)(0.9)(0.5) = 0.2816$$

$\underline{PV}$ "OPTIMAL" sequence of states given model and observations.

Given $X = \{X_1, \dots X_T\}$

$$\Theta = \{\pi, A, \phi\}$$

How to choose $Z = \{Z_1, \dots Z_T\}$ optimally?

<u>PV</u> 'OPTIMAL' sequence of states : MPM

OPTIMALITY DEFINITION #1

Choose state z_t which is individually most likely

PV "OPTIMAL" sequence of states : MPM

OPTIMALITY DEFINITION #1

Choose state z_t which is individually most likely

$$\hat{z} = \left(\underset{z_1}{\arg\max}\ p(z_1 \mid x_{1:T}), \ldots, \underset{z_T}{\arg\max}\ p(z_T \mid x_{1:T}) \right)$$

PV "OPTIMAL" sequence of states : MPM

OPTIMALITY DEFINITION #1

Choose state z_t which is individually most likely

$$\hat{z} = \left(\underset{z_1}{\arg\max}\, p(z_1 \mid x_{1:T}), \ldots, \underset{z_T}{\arg\max}\, p(z_T \mid x_{1:T}) \right)$$

Also called Sequence of Marginally Most Probable States (MPM)

PV 'OPTIMAL' sequence of states : MPM

OPTIMALITY DEFINITION #1

Choose state z_t which is individually most likely

Let $\gamma_t(i) = p(z_t = i \mid x_{1:T})$

PV "OPTIMAL" sequence of states : MPM

OPTIMALITY DEFINITION #1

Choose state z_t which is individually most likely

Let $\gamma_t(i) = p(z_t = i \mid x_{1:T})$

$$\propto p(z_t = i \mid x_{1:t}) \cdot p(x_{t+1:T} \mid z_t = i, x_{1:t})$$

<u>PV</u> 'OPTIMAL' sequence of states : MPM

OPTIMALITY DEFINITION #1

Choose state z_t which is individually most likely

Let $\gamma_t(i) = p(z_t = i \mid x_{1:T})$

$$\propto p(z_t = i \mid x_{1:t}) \cdot p(x_{t+1:T} \mid z_t = i, \cancel{x_{1:t}})$$

Independent

$$\propto p(z_t = i \mid x_{1:t}) \cdot p(x_{t+1:T} \mid z_t = i)$$

$$\gamma_t(i) \propto \alpha_t(i) \cdot \beta_t(i)$$

<u>PV</u> "OPTIMAL" sequence of states : MPM

FORWARD BACKWARD ALGORITHM

$$\gamma_t(i) = p(z_t = i \mid x_{1:T})$$

$$\gamma_t(i) \propto \alpha_t(i) . \beta_t(i)$$

$$\gamma_t(i) = \frac{\alpha_t(i) . \beta_t(i)}{\sum_i \alpha_t(i) . \beta_t(i)}$$

$\overline{PV}$ 'OPTIMAL' sequence of states : MPM

FORWARD BACKWARD ALGORITHM

Example: For some Θ, $X = \{H, H, H\}$ Find 'smooth' set of states

$$\gamma_t(i) = \frac{\alpha_t(i) \cdot \beta_t(i)}{\sum_i \alpha_t(i) \cdot \beta_t(i)}$$

$\overline{\text{PV}}$ "OPTIMAL" sequence of states : MPM

FORWARD BACKWARD ALGORITHM

Example: For some Θ, $X = \{H, H, H\}$ Find 'smooth' set of states

$$\gamma_t(i) = \frac{\alpha_t(i) \cdot \beta_t(i)}{\sum_i \alpha_t(i) \cdot \beta_t(i)}$$

$$\gamma_1(B) = \frac{\alpha_1(B) \cdot \beta_1(B)}{\alpha_1(B) \cdot \beta_1(B) + \alpha_1(F) \cdot \beta_1(F)} = \frac{0.42 \times 0.4544}{.42 \times 0.4544 + 0.2 \times 0.2816}$$

$$= \frac{.190848}{.247168} = 0.77$$

PV "OPTIMAL" sequence of states : MPM

FORWARD BACKWARD ALGORITHM

Example: For some θ, $X = \{H, H, H\}$ Find 'smooth' set of states

$$\gamma_t(i) = \frac{\alpha_t(i) \cdot \beta_t(i)}{\sum_i \alpha_t(i) \cdot \beta_t(i)}$$

$\gamma_1(B) = 0.77$

$\Rightarrow \gamma_1(F) = 0.23$

Much more likely to have $z_1 = B$ if $\{H, H, H\}$ and θ given

$\overline{PV}$ 'OPTIMAL' sequence of states : MPM

FORWARD BACKWARD ALGORITHM

Example: For some Θ, $X = \{H, H, H\}$ Find 'smooth' set of states

$$\gamma_t(i) = \frac{\alpha_t(i) \cdot \beta_t(i)}{\sum_i \alpha_t(i) \cdot \beta_t(i)}$$

$$\hat{Z} = \left(\underset{i}{\operatorname{argmax}}\ \gamma_1(i), \ldots\right) = (\text{Biased}, \text{Biased}, \text{Biased})$$

$\overline{\text{PV}}$ 'OPTIMAL' sequence of states : MAP

OPTIMALITY DEFINITION # 2

CHOOSE MOST PROBABLE SEQUENCE OF STATES

$$z^* = \underset{z_{1:T}}{\arg\max}\; p(z_{1:T} \mid x_{1:T})$$

$\overline{PV}$ 'OPTIMAL' sequence of states : MAP v/s MPM

CONSIDER FOLLOWING 2 timesteps joint prob

	$Z_1 = 1$	$Z_1 = 2$
$Z_2 = 1$	0.04	0.3
$Z_2 = 2$	0.36	0.3

$\overline{PV}$ "OPTIMAL" sequence of states : MAP v/s MPM

CONSIDER FOLLOWING 2 timesteps joint prob

	$Z_1 = 1$	$Z_1 = 2$	
$Z_2 = 1$	0.04	0.3	0.34
$Z_2 = 2$	0.36	0.3	0.66
	.4	.6	

MPM = sequence of Marginally MOST PROB. STATES

$$= \left(Z_1 = 2 \quad , \quad Z_2 = 2 \right)$$

$\because 0.6 > 0.4$ $\qquad \because 0.66 > 0.34$

PV "OPTIMAL" sequence of states : MAP v/s MPM

CONSIDER FOLLOWING 2 timesteps joint prob

	$z_1 = 1$	$z_1 = 2$	
$z_2 = 1$	0.04	0.3	0.34
$z_2 = 2$	0.36	0.3	0.66
	.4	.6	

MAP = $(z_1 = 1, z_2 = 2)$

$\because$ .36 is highest number

<u>PV</u> 'OPTIMAL' sequence of states : MAP

OPTIMALITY DEFINITION # 2

CHOOSE MOST PROBABLE SEQUENCE OF STATES

$$z^* = \underset{z_{1:T}}{\arg\max}\; p(z_{1:T} \mid x_{1:T})$$

VITERBI ALGORITHM (Dynamic Programming)

VITERBI ALGORITHM

Define $\delta_t(i) = \max\limits_{z_1, z_2, \ldots z_{t-1}} P\left[z_1, z_2, z_3 \ldots, z_{t-1}, z_t = i, x_1, x_2 \ldots x_t \mid \theta\right]$

VITERBI ALGORITHM

Define $\delta_t(i) = \max_{z_1, z_2, \ldots z_{t-1}} P[z_1, z_2, z_3 \ldots, z_{t-1}, z_t = i, x_1, x_2 \ldots x_t \mid \theta]$

Best score (highest prob.) along a single path at time t, which accounts for first t observations and ends in $z_t = i$

Difference B/w $\alpha_t(i)$ and $\delta_t(i)$

$$\delta_t(i) = \max_{z_1, z_2, \dots z_{t-1}} P\left[z_1, z_2, z_3 \dots, z_{t-1}, z_t = i, x_1, x_2 \dots x_t \mid \theta\right]$$

$$\alpha_t(i) = P\left(x_{1:t}, z_t = i \mid \theta\right)$$

Difference B/w $\alpha_t(i)$ and $\delta_t(i)$

$$\delta_t(i) = \max_{z_1, z_2, \ldots z_{t-1}} P\left[z_1, z_2, z_3 \ldots, z_{t-1}, z_t = i, x_1, x_2 \ldots x_t \mid \theta\right]$$

$$\alpha_t(i) = P\left(x_{1:t}, z_t = i \mid \theta\right)$$

FOCUS ON MOST LIKELY SEQUENCE

FOCUS ON MOST LIKELY STATE AT 't'

VITERBI ALGORITHM

Define $\delta_t(i) = \max\limits_{z_1, z_2, \ldots z_{t-1}} P\left[z_1, z_2, z_3 \ldots, z_{t-1}, z_t = i, x_1, x_2 \ldots x_t \mid \theta\right]$

Relation b/w $\delta_t(i)$ AND $\delta_{t+1}(j)$?

VITERBI ALGORITHM

Define $$\delta_t(i) = \max_{z_1, z_2, \dots z_{t-1}} P\left[z_1, z_2, z_3 \dots, z_{t-1}, z_t = i, x_1, x_2 \dots x_t \,\middle|\, \theta\right]$$

Relation b/w $\delta_t(i)$ AND $\delta_{t+1}(j)$?

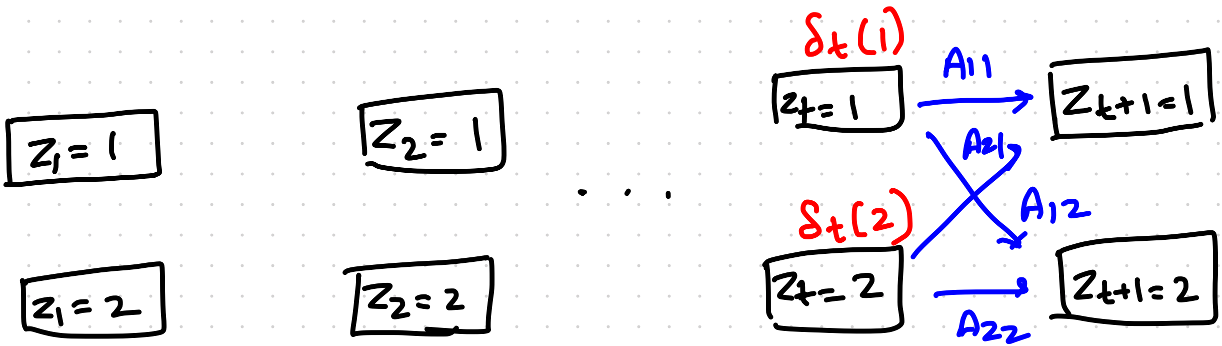
$$\boxed{z_1 = 1} \quad \boxed{z_2 = 1} \quad \dots \quad \boxed{z_t = 1} \quad \boxed{z_{t+1} = 1}$$

$$\boxed{z_1 = 2} \quad \boxed{z_2 = 2} \quad \quad \boxed{z_t = 2} \quad \boxed{z_{t+1} = 2}$$

VITERBI ALGORITHM

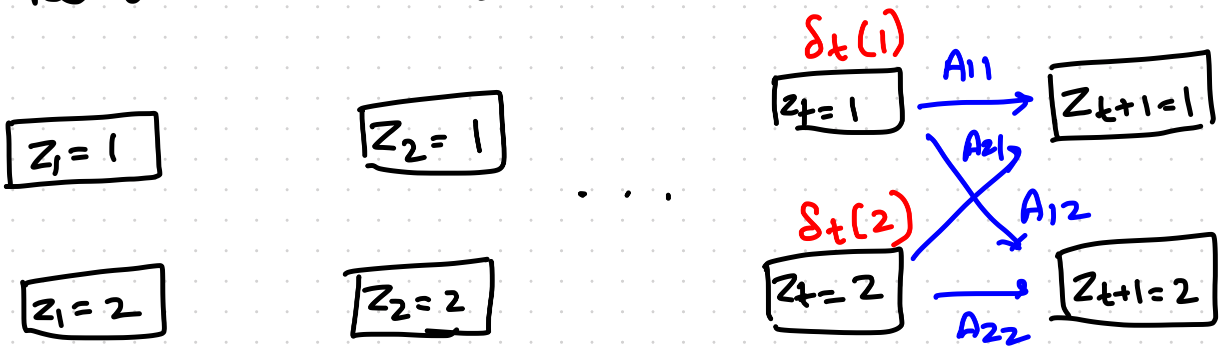
Define $$\delta_t(i) = \max_{z_1, z_2, \dots z_{t-1}} P\left[z_1, z_2, z_3 \dots, z_{t-1}, z_t = i, x_1, x_2 \dots x_t \mid \theta\right]$$

Relation b/w $\delta_t(i)$ AND $\delta_{t+1}(j)$?



VITERBI ALGORITHM

Relation b/w $\delta_t(i)$ AND $\delta_{t+1}(j)$?

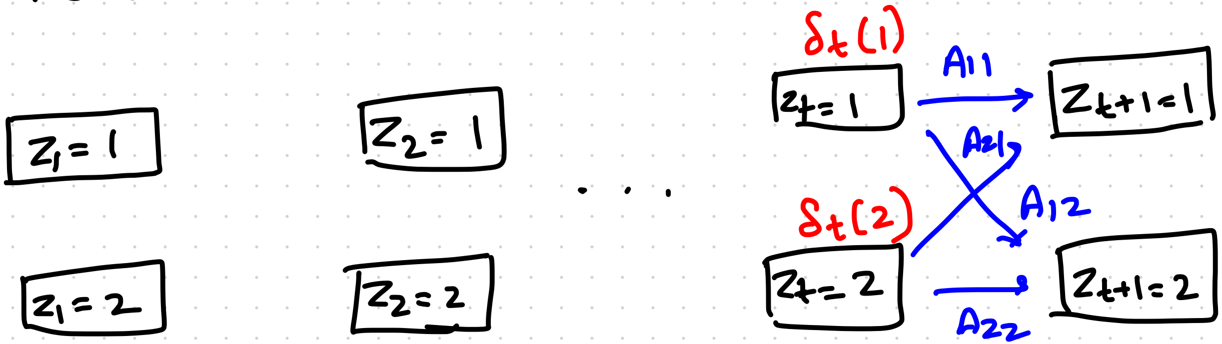


* We could reach $z_{t+1} = j$ from any $i \in \{1, .. K\}$ via a transition with prob. A_{ij}

* Once we reach $z_{t+1} = j$, prob. of observing x_{t+1} is $\phi_j(x_{t+1})$

VITERBI ALGORITHM

Relation b/w $\delta_t(i)$ AND $\delta_{t+1}(j)$?



* We could reach $z_{t+1} = j$ from any $i \in \{1, .. K\}$ via a transition with prob. A_{ij}. We take 'best path'
* Once we reach $z_{t+1} = j$, prob. of observing x_{t+1} is $\phi_j(x_{t+1})$

$$\delta_{t+1}(j) = \left[\max_i \{ \delta_t(i) \cdot A_{ij} \} * \phi_j(x_{t+1}) \right]$$

VITERBI ALGORITHM

$$\delta_{t+1}(j) = \left[\max_i \{\delta_t(i) \cdot A_{ij}\} * \phi_j(x_{t+1})\right]$$

FOR EACH t and j, we need to store argument 'i' which maximized above equation.

called $\psi_t(j)$

VITERBI ALGORITHM

(i) INITIALISATION

$$\delta_1(i) =$$

VITERBI ALGORITHM

(i) INITIALISATION

$$\delta_1(i) = \pi_i * \phi_i(x_1)$$

$$\psi_1(i) = 0 \quad \text{(ARBITRARY)}$$

VITERBI ALGORITHM

① INITIALISATION

$$\delta_1(i) = \pi_i * \phi_i(x_1)$$

$$\psi_1(i) = 0 \quad \text{(ARBITRARY)}$$

② RECURSION

FOR t in 2 ~~to~~ T

FOR j in 1 ~~to~~ K

$$\delta_t(j) = \left(\max_{i \in \{1,K\}} \left(\delta_{t-1}(i) \cdot A_{ij}\right)\right) \phi_j(x_t)$$

$$\psi_t(j) = \underset{i \in \{1,..K\}}{\operatorname{argmax}} \; \delta_{t-1}(i) \cdot A_{ij}$$

VITERBI ALGORITHM

③ Termination

$$P^* = \max_{i \in \{1,..K\}} \delta_T(i)$$

$$Z_T^* = \underset{i \in \{1,..K\}}{\operatorname{argmax}} \delta_T(i)$$

VITERBI ALGORITHM

③ Termination

$$P^* = \max_{i \in \{1,..K\}} \delta_T(i)$$

$$Z_T^* = \underset{i \in \{1,..,K\}}{\operatorname{argmax}} \delta_T(i)$$

④ Backtracking

For t in T-1, ... 1 :

$$Z_t^* = \Psi_{t+1}(Z_{t+1}^*)$$

VITERBI ALGORITHM

Example

Given $\Theta = \{\pi, A, \phi\}$ and $x_{1:T} = \{H, H, H\}$

determine MAP z^*

$$\pi = \begin{array}{c} \begin{bmatrix} 0.6 & 0.4 \end{bmatrix} \\ \text{Bias} \quad \text{Fair} \end{array}$$

$$A = \begin{array}{c|cc} & B & F \\ \hline \text{Bias} & 0.9 & 0.1 \\ \text{Fair} & 0.1 & 0.9 \end{array}$$

$\phi =$

Bias	Fair
$p(H) = 0.7$	$p(H) = 0.5$
$p(T) = 0.3$	$p(T) = 0.5$

VITERBI ALGORITHM

Example

Given $\Theta = \{\pi, A, \phi\}$ and $x_{1:T} = \{H, H, H\}$

determine MAP z^*

$$\delta_1(\text{Bias}) = \pi_{\text{Bias}} * \phi_{\text{Bias}}(H) = 0.6 * 0.7 = 0.42$$

$$\delta_1(\text{Fair}) = \pi_{\text{Fair}} * \phi_{\text{Fair}}(H) = 0.4 * 0.5 = 0.20$$

VITERBI ALGORITHM

Example

Given $\Theta = \{\pi, A, \phi\}$ and $x_{1:T} = \{H, H, H\}$

determine MAP z^*

$$\delta_1(\text{Bias}) = \pi_{\text{Bias}} * \phi_{\text{Bias}}(H) = 0.6 * 0.7 = 0.42$$

$$\delta_1(\text{Fair}) = \pi_{\text{Fair}} * \phi_{\text{Fair}}(H) = 0.4 \times 0.5 = 0.20$$

$$\psi_1(\text{Bias}) = \psi_1(\text{Fair}) = 0 \text{ (Arbitrary)}$$

VITERBI ALGORITHM

Example

$$\delta_1(\text{Bias}) = \pi_{\text{Bias}} * \phi_{\text{Bias}}(H) = 0.6 * 0.7 = 0.42$$

$$\delta_1(\text{Fair}) = \pi_{\text{Fair}} * \phi_{\text{Fair}}(H) = 0.4 * 0.5 = 0.20$$

$$\psi_1(\text{Bias}) = \psi_1(\text{Fair}) = 0 \quad (\text{Arbitrary})$$

$$\delta_2(\text{Bias}) = \max \begin{Bmatrix} \delta_1(\text{Fair}) . A_{\text{Fair, Bias}} \\ \delta_1(\text{Bias}) . A_{\text{Bias, Bias}} \end{Bmatrix} * \phi_{\text{Bias}}(H)$$

$$= \max \begin{Bmatrix} 0.2 * 0.1 \\ 0.42 * 0.9 \end{Bmatrix} * 0.7 = 0.42 * 0.9 * 0.7$$

VITERBI ALGORITHM

$$\delta_1(\text{Bias}) = \pi_{\text{Bias}} * \phi_{\text{Bias}}(H) = 0.6 * 0.7 = 0.42$$

$$\delta_1(\text{Fair}) = \pi_{\text{Fair}} * \phi_{\text{Fair}}(H) = 0.4 * 0.5 = 0.20$$

$$\psi_1(\text{Bias}) = \psi_1(\text{Fair}) = 0 \text{ (Arbitrary)}$$

$$\delta_2(\text{Bias}) = \max \left\{ \begin{array}{l} \delta_1(\text{Fair}) . A_{\text{Fair, Bias}} \\ \delta_1(\text{Bias}) . A_{\text{Bias, Bias}} \end{array} \right\} * \phi_{\text{Bias}}(H)$$

$$= \max \left\{ \begin{array}{l} 0.2 * 0.1 \\ 0.42 * 0.9 \end{array} \right\} * 0.7 = 0.42 * 0.9 * 0.7$$

$$= 0.26$$

$$\psi_2(\text{Bias}) = \text{Bias}$$

VITERBI ALGORITHM

$$\delta_1(\text{Bias}) = \pi_{\text{Bias}} * \phi_{\text{Bias}}(H) = 0.6 * 0.7 = 0.42$$

$$\delta_1(\text{Fair}) = \pi_{\text{Fair}} * \phi_{\text{Fair}}(H) = 0.4 * 0.5 = 0.20$$

$$\psi_1(\text{Bias}) = \psi_1(\text{Fair}) = 0 \quad (\text{Arbitrary})$$

$$\delta_2(\text{Fair}) = \max \begin{Bmatrix} \delta_1(\text{Fair}) . A_{\text{Fair},\text{Fair}} \\ \delta_1(\text{Bias}) . A_{\text{Bias},\text{Fair}} \end{Bmatrix} * \phi_{\text{Fair}}(H)$$

$$= \max \begin{Bmatrix} 0.2 * 0.9 \\ 0.42 * 0.1 \end{Bmatrix} * 0.5 = 0.2 * 0.9 * 0.5$$

$$= 0.09$$

$$\psi_2(\text{Fair}) = \text{Fair}$$

VITERBI ALGORITHM

$\delta_2(\text{Bias}) = 0.2646$ $\qquad$ $\psi_2(\text{Bias}) = \text{Bias}$

$\delta_2(\text{Fair}) = 0.09$ $\qquad$ $\psi_2(\text{Fair}) = \text{Fair}$

$$\delta_3(\text{Bias}) = \max \begin{Bmatrix} \delta_2(\text{Bias}) * A_{\text{Bias},\text{Bias}} \\ \delta_2(\text{Fair}) * A_{\text{Fair},\text{Bias}} \end{Bmatrix} * \phi_{\text{Bias}}(H)$$

$$\delta_3(\text{Bias}) = \max \begin{Bmatrix} .2646 * 0.9 \\ 0.09 * 0.1 \end{Bmatrix} * 0.7 = 0.166698$$

$\psi_3(\text{Bias}) = \text{Bias}$

VITERBI ALGORITHM

$\delta_2(\text{Bias}) = 0.2646$ $\quad\quad$ $\psi_2(\text{Bias}) = \text{Bias}$

$\delta_2(\text{Fair}) = 0.09$ $\quad\quad$ $\psi_2(\text{Fair}) = \text{Fair}$

$$\delta_3(\text{Fair}) = \max \begin{Bmatrix} \delta_2(\text{Fair}) \cdot A_{\text{Fair},\text{Fair}} \\ \delta_2(\text{Bias}) \cdot A_{\text{Bias},\text{Fair}} \end{Bmatrix} * \phi_{\text{Fair}}(H)$$

$$= \max \begin{Bmatrix} 0.09 * 0.9 \\ 0.2646 * 0.1 \end{Bmatrix} * 0.5 = \max \begin{Bmatrix} 0.081 \\ 0.026 \end{Bmatrix} * 0.5$$

$$= 0.0405$$

$\psi_3(\text{Fair}) = \text{Fair}$

VITERBI ALGORITHM

$\delta_1(\text{Bias}) = 0.42$

$\delta_1(\text{Fair}) = 0.2$

$\delta_2(\text{Bias}) = 0.2646$

$\delta_2(\text{Fair}) = 0.09$

$\psi_2(\text{Bias}) = \text{Bias}$

$\psi_2(\text{Fair}) = \text{Fair}$

$\delta_3(\text{Bias}) = 0.167$

$\delta_3(\text{Fair}) = 0.04$

$\psi_3(\text{Bias}) = \text{Bias}$

$\psi_3(\text{Fair}) = \text{Fair}$

VITERBI ALGORITHM

$\delta_1(\text{Bias}) = 0.42$

$\delta_1(\text{Fair}) = 0.2$

$\delta_2(\text{Bias}) = 0.2646$ $\quad\quad \psi_2(\text{Bias}) = \text{Bias}$

$\delta_2(\text{Fair}) = 0.09$ $\quad\quad \psi_2(\text{Fair}) = \text{Fair}$

$\delta_3(\text{Bias}) = 0.167$ $\quad\quad \psi_3(\text{Bias}) = \text{Bias}$

$\delta_3(\text{Fair}) = 0.04$ $\quad\quad \psi_3(\text{Fair}) = \text{Fair}$

$$Z_3^* = \operatorname{argmax}\left[\delta_3(\text{Bias}), \delta_3(\text{Fair})\right] = \text{Bias}$$

VITERBI ALGORITHM

$\delta_1(\text{Bias}) = 0.42$

$\delta_1(\text{Fair}) = 0.2$

$\delta_2(\text{Bias}) = 0.2646$

$\delta_2(\text{Fair}) = 0.09$

$\delta_3(\text{Bias}) = 0.167$

$\delta_3(\text{Fair}) = 0.04$

$\psi_2(\text{Bias}) = \text{Bias}$

$\psi_2(\text{Fair}) = \text{Fair}$

$\psi_3(\text{Bias}) = \text{Bias}$

$\psi_3(\text{Fair}) = \text{Fair}$

$$z_3^* = \operatorname{argmax}\left[\delta_3(\text{Bias}), \delta_3(\text{Fair})\right] = \text{Bias}$$

$$z_2^* = \psi_3(\text{Bias}) = \text{Bias}$$

$$z_1^* = \psi_2(z_2^*) = \psi_2(\text{Bias}) = \text{Bias} \Rightarrow z^* = \{\text{Bias}, \text{Bias}, \text{Bias}\}$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

* Fully observed: Observe all hidden state sequences

* $\theta = \{\pi, A, \phi\}$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

* Fully observed: observe all hidden state sequences

* $\theta = \{\pi, A, \phi\}$

* Given $D = \Big\{ \big((z_{11}, z_{12} \ldots z_{1T_1}), (x_{11}, x_{12} \ldots x_{1T_1})\big),$

$\vdots$

$\big((z_{N1}, \ldots z_{NT_N}), (x_{N1}, \ldots x_{NT_N})\big)$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

* Fully observed: observe all hidden state sequences

* $\theta = \{\pi, A, \phi\}$

* Given $D = \Big\{ \big((z_{11}, z_{12} \dots z_{1T_1}), (x_{11}, x_{12} \dots x_{1T_1})\big),$

$\vdots$

N sequences $\longleftarrow$

$\big((z_{N1}, \dots z_{NT_N}), (x_{N1}, \dots x_{NT_N})\big)$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

* Fully observed: observe all hidden state sequences

* $\theta = \{\pi, A, \phi\}$

* Given $D = \Big\{ \big((z_{11}, z_{12} \dots z_{1T_1}), (x_{11}, x_{12} \dots x_{1T_1})\big),$

$\vdots$

N sequences each of T_i length (may be different) ←

$\big((z_{N1}, \dots z_{NT_N}), (x_{N1}, \dots x_{NT_N})\big)$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

SOME MATHS FOR A SINGLE SEQUENCE FIRST

Likelihood of a sequence $Z_{1:T}$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

Likelihood of a sequence $z_{1:T}$

$$= \pi(z_1)\, A(z_1, z_2) \cdots A(z_{T-1}, z_T)$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

Likelihood of a sequence $z_{1:T}$

$$= \pi(z_1)\, A(z_1, z_2) \;\cdots\; A(z_{T-1}, z_T)$$

Let us write $\pi(z_1)$ in terms of $j \in \{1, .. K\}$

$\underline{\overline{\text{PVI}}}$: HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

Likelihood of a sequence $z_{1:T}$

$$= \pi(z_1)\, A(z_1, z_2) \cdots A(z_{T-1}, z_T)$$

Let us write $\pi(z_1)$ in terms of $j \in \{1, .. K\}$

$$\pi = \underset{z_1=1}{[\text{----}}\;\underset{z_1=2}{\text{-----}}\;\underset{z_1=j}{\boxed{}}\;\underset{z_1=K}{\text{----}]}$$

$$\pi(z_1) = \pi_j \quad : \text{if } z_1 = j$$

$\underline{\text{PVI}}$: HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

Likelihood of a sequence $z_{1:T}$

$$= \pi(z_1)\, A(z_1, z_2) \;\ldots\; A(z_{T-1}, z_T)$$

Let us write $\pi(z_1)$ in terms of $j \in \{1, .. K\}$

$$\pi = \underset{z_1 = 1 \quad z_1 = 2 \quad z_1 = j \quad z_1 = K}{[\text{-----------------------}]}$$

$$\pi(z_1) = \pi_j \;:\; \text{if } z_1 = j \qquad \text{or} \qquad \pi(z_1) = \prod_{j=1}^{K} (\pi_j)^{I(z_1 = j)}$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

Likelihood of a sequence $z_{1:T}$

$$= \pi(z_1)\, A(z_1, z_2) \cdots A(z_{T-1}, z_T)$$

Let us write $A(z_1, z_2)$ in terms of $j, k \in \{1, \dots K\}$

$\underline{\text{PVI}}$: HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

Likelihood of a sequence $z_{1:T}$

$$= \pi(z_1) A(z_1, z_2) \ldots A(z_{T-1}, z_T)$$

Let us write $A(z_1, z_2)$ in terms of $j, k \in \{1, \ldots K\}$

$$A(z_1, z_2) = \prod_{j=1}^{K} \prod_{k=1}^{K} (A_{jk})^{I(z_1 = j, z_2 = k)}$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

Likelihood of a sequence $z_{1:T}$

$$= \pi(z_1)\, A(z_1, z_2) \cdots A(z_{T-1}, z_T)$$

Now, $A(z_1, z_2) = \prod_{j=1}^{K} \prod_{k=1}^{K} (A_{jk})^{\mathbb{I}(z_1 = j,\, z_2 = k)}$

$$\Rightarrow A(z_1, z_2) \cdots A(z_{T-1}, z_T) = \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} (A_{jk})^{\mathbb{I}(z_t = k,\, z_{t-1} = j)}$$

$\underline{\text{PVI}}$: HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

Likelihood of a sequence $z_{1:T}$

$$= \pi(z_1)\, A(z_1, z_2) \ldots A(z_{T-1}, z_T)$$

$$LL(D \mid \theta = \{A, \pi\}) = \log \left(\prod_{j=1}^{K} (\pi_j)^{I(z_1 = j)} \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} (A_{jk})^{I(z_t = k,\, z_{t-1} = j)} \right)$$

$$= \log (F \cdot G)$$

$$= \log F + \log G$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

$$LL(D \mid \theta = \{A, \pi\}) = \log\left(\prod_{j=1}^{K} (\pi_j)^{I(z_1 = j)} \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} (A_{jk})^{I(z_t = k,\, z_{t-1} = j)}\right)$$

$$= \log(F \cdot G)$$

$$= \log F + \log G$$

Now $\log F = \log\left(\pi_1^{I(z_1=1)}\, \pi_2^{I(z_1=2)} \cdots \pi_K^{I(z_1=K)}\right)$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

$$LL(D \mid \theta = \{A, \pi\}) = \log \left(\prod_{j=1}^{K} (\pi_j)^{I(z_1 = j)} \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} (A_{jk})^{I(z_t = k,\, z_{t-1} = j)} \right)$$

$$= \log (F \cdot G)$$

$$= \log F + \log G$$

Now $\log F = \log \left(\pi_1^{I(z_1 = 1)} \pi_2^{I(z_1 = 2)} \cdots \pi_K^{I(z_1 = K)} \right)$

$$= I(z_1 = 1) \log \pi_1 + \cdots$$

$$\log F = \sum_{j=1}^{K} I(z_1 = j) \log \pi_j$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

$$LL(D \mid \theta = \{A, \pi\}) = \log \left(\prod_{j=1}^{K} (\pi_j)^{I(z_1 = j)} \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} (A_{jk})^{I(z_t = k,\, z_{t-1} = j)} \right)$$

$$= \log (F \cdot G)$$

$$\log G = \log \left(\prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{I(z_1 = j,\, z_2 = k)} * \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{I(z_2 = j,\, z_3 = k)} \cdots \right)$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

$$LL(D \mid \theta = \{A, \pi\}) = \log \left(\prod_{j=1}^{K} (\pi_j)^{I(z_1 = j)} \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} (A_{jk})^{I(x_t = k,\, x_{t-1} = j)} \right)$$

$$= \log (F \cdot G)$$

$$\log G = \log \left(\prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{I(x_1 = j,\, x_2 = k)} * \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{I(x_2 = j,\, x_3 = k)} \; \text{----} \right)$$

$$= \sum_{t=2}^{T} \log \left(\prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{I(x_t = k,\, x_{t-1} = j)} \right)$$

$\underline{\text{PVI}}$: HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

$$\log G = \log\left(\prod_{j=1}^{K}\prod_{k=1}^{K} A_{jk}^{I(x_1=j,\,x_2=k)} * \prod_{j=1}^{K}\prod_{k=1}^{K} A_{jk}^{I(x_2=j,\,x_3=k)} \cdots\right)$$

$$= \sum_{t=2}^{T} \log\left(\prod_{j=1}^{K}\prod_{k=1}^{K} A_{jk}^{I(x_t=k,\,x_{t-1}=j)}\right)$$

$$\log G = \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} I(x_t=k,\, x_{t-1}=j)\log A_{jk}$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

BACK TO 'N' sequences now.

* Given $D = \{ ((z_{11}, z_{12} \ldots z_{1T_1}), (x_{11}, x_{12} \ldots x_{1T_1})),$

$\vdots$

$((z_{N1}, \ldots z_{NT_N}), (x_{N1}, \ldots x_{NT_N}))$

N sequences each of T_i length (may be different) ←

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

STEP I

Estimate π and A given $D' = ((z_{i1} \ldots z_{iT_i}) \; \forall i \in (1, N))$

$$\log p(D' \mid \theta' = \{A, \pi\}) = \sum_{i=1}^{N} \log p(z_i \mid \theta')$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

STEP I

Estimate π and A given $D' = ((z_{i1} \ldots z_{iT_i})\ \forall\, i \in (1, N))$

$$\log p(D' \mid \theta' = \{A, \pi\}) = \sum_{i=1}^{N} \log p(z_i \mid \theta')$$

$$\boxed{\begin{aligned} N'_j &\triangleq \sum_{i=1}^{N} \mathbb{I}(z_{i1} = j) \\ N_{jk} &\triangleq \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} \mathbb{I}(x_{i,t} = j,\ x_{i,t+1} = k) \end{aligned}}$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

STEP I

Estimate π and A given $D' = ((z_{i1} \ldots\ldots z_{iT_i})\ \forall\, i \in (1, N))$

$$\log p(D' \mid \theta' = \{A, \pi\}) = \sum_{i=1}^{N} \log p(z_i \mid \theta')$$

$$N_j^{1} \triangleq \sum_{i=1}^{N} \mathbb{I}(z_{i1} = j)$$ ← COUNT OF SEQUENCES WHERE $z_1 = j$

$$N_{jk} \triangleq \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} \mathbb{I}(x_{i,t} = j,\ x_{i,t+1} = k)$$

PVI: HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

STEP I

Estimate π and A given $D' = ((z_{i1} \ldots z_{iT_i}) \; \forall i \in (1, N))$

$$\log p(D' \mid \theta' = \{A, \pi\}) = \sum_{i=1}^{N} \log p(z_i \mid \theta')$$

$$N_j^1 \triangleq \sum_i I(z_{i1} = j)$$ ← COUNT OF SEQUENCES WHERE $z_1 = j$

COUNT OF TRANSITIONS FROM 'j' to 'k' →

$$N_{jk} \triangleq \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} I(x_{i,t} = j, \; x_{i,t+1} = k)$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

FROM MLE,

$$\hat{\pi}_j = \frac{N^1_j}{\sum_{j=1}^{K} N^1_j} \qquad \hat{A}_{jk} = \frac{N_{jk}}{\sum_{k} N_{jk}}$$

$$N^1_j \triangleq \mathbb{I}(z_{i1} = j)$$

← COUNT OF SEQUENCES WHERE $z_1 = j$

COUNT OF TRANSITIONS FROM 'j' to 'k' →

$$N_{jk} \triangleq \sum_{i=1}^{N} \sum_{t=1}^{T_i-1} \mathbb{I}(x_{i,t} = j,\ x_{i,t+1} = k)$$

PVI : HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

FROM MLE,

$$\hat{\pi}_j = \frac{N_j^1}{\sum_{j=1}^{K} N_j^1} \qquad \hat{A}_{jk} = \frac{N_{jk}}{\sum_k N_{jk}}$$

HOW TO ESTIMATE ϕ:

① Assume MULTINOULLI EMISSION

$$\Rightarrow \phi_{jl} = p(x_t = l \mid z_t = j) \quad l \in \{1 \ldots L\}$$

$\overline{\text{PVI}}$: HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

FROM MLE,

$$\hat{\pi}_j = \frac{N_j^1}{\sum_{j=1}^{K} N_j^1} \qquad \hat{A}_{jk} = \frac{N_{jk}}{\sum_{k} N_{jk}}$$

HOW TO ESTIMATE ϕ:

① ASSUME MULTINOULLI EMISSION

$$\Rightarrow \phi_{jl} = p(x_t = l \mid z_t = j) \quad l \in \{1 \dots L\}$$

FROM MLE,

$$\boxed{\hat{\phi}_{jl} = \frac{N_{jl}^X}{N_j}} \; ; \quad N_{jl}^X \triangleq \sum_{i=1}^{N} \sum_{t=1}^{T_i} \mathbb{I}(z_{i,t} = j, x_{i,t} = l)$$

$\overline{\text{PVI}}$: HMM PARAMETER LEARNING GIVEN FULLY OBSERVED DATA

FROM MLE,

$$\hat{\pi}_j = \frac{N_j^1}{\sum_{j=1}^{K} N_j^1} \qquad \hat{A}_{jk} = \frac{N_{jk}}{\sum_k N_{jk}}$$

HOW TO ESTIMATE ϕ FOR NORMALLY DISTRIBUTED ϕ

$$\hat{\mu}_k = \frac{\bar{x}_k}{N_k} \qquad \hat{\Sigma}_k = \frac{\overline{(xx)}^T_k - N_k \hat{\mu}_k \hat{\mu}_k^T}{N_k}$$

$$\bar{x}_k \triangleq \sum_{i=1}^{N} \sum_{t=1}^{T_i} I(z_{i,t} = k)\, x_{i,t} \qquad \overline{(xx)}^T_k \triangleq \sum_{i=1}^{N} \sum_{t=1}^{T_i} I(z_{i,t} = k)\, x_{i,t} x_{i,t}^T$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1	Bias	Fair	Fair	Bias
	H	T	H	H
S_2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1	Bias	Fair	Fair	Bias
	H	T	H	H
S_2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

N_j^1 = # Times z_j occurs at $t=1$

$$= \sum_{i=1}^{N} I(z_{i1} = j)$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1	Bias	Fair	Fair	Bias
	H	T	H	H
S_2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

N_j^1 = # Times z_j occurs at $t=1$

$$= \sum_{i=1}^{N} I(z_{i1} = j)$$

$N_B^1 = 2$ (S1 and S_3)

$N_F^1 = 1$ (S2)

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

S_1	Bias	Fair	Fair	Bias
	H	T	H	H
S_2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

$$N'_B = 2, \quad N'_F = 1$$

$$\hat{\pi} = [\hat{\pi}_B, \hat{\pi}_F] = \left[\frac{N'_B}{N'_B + N'_F}, \frac{N'_F}{N'_F + N'_B}\right] = \left[\frac{2}{3}, \frac{1}{3}\right]$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1	Bias	Fair	fair	Bias
	H	T	H	H
S2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	h	h	H

N_{jk} = COUNT OF TRANSITIONS FROM Z_j to Z_k.

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

S_1	Bias	Fair	fair	Bias
	H	T	H	H
S_2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

N_{jk} = COUNT OF TRANSITIONS FROM z_j to z_k.

$N_{BB} = 3$ $\quad$ $N_{FF} = 4$

$N_{BF} = 1$ $\quad$ $N_{FB} = 1$

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1	Bias	Fair	fair	Bias
	H	T	H	H
S_2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

$N_{BB} = 3$ $\qquad N_{FF} = 4$

$N_{BF} = 1$ $\qquad N_{FB} = 1$

$$\hat{A}_{BB} = \frac{N_{BB}}{N_{BB} + N_{BF}} = \frac{3}{4}$$

$$\hat{A}_{BF} = \frac{N_{BF}}{N_{BB} + N_{BF}} = \frac{1}{4}$$

$$\therefore \hat{A} = \begin{bmatrix} \hat{A}_{BB} & \hat{A}_{BF} \\ \hat{A}_{FB} & \hat{A}_{FF} \end{bmatrix}$$

$$\hat{A} = \begin{bmatrix} 0.75 & 0.25 \\ .2 & .8 \end{bmatrix}$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1	B	F	F	B
	H	T	H	H
S_2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

$$\hat{\Phi} = \begin{bmatrix} \hat{\Phi}_{BH} & \hat{\Phi}_{BT} \\ \hat{\Phi}_{FH} & \hat{\Phi}_{FT} \end{bmatrix}$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1	B	F	F	B
	H	T	H	H
S_2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

$$\hat{\Phi} = \begin{bmatrix} \hat{\Phi}_{BH} & \hat{\Phi}_{BT} \\ \hat{\Phi}_{FH} & \hat{\Phi}_{FT} \end{bmatrix}$$

$$\hat{\Phi}_{BH} = \frac{\text{COUNT of (Bias \& H)}}{\text{COUNT OF BIAS}}$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1	B	F	F	B
	H	T	H	H
S_2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

$$\hat{\Phi} = \begin{bmatrix} \hat{\Phi}_{BH} & \hat{\Phi}_{BT} \\ \hat{\Phi}_{FH} & \hat{\Phi}_{FT} \end{bmatrix}$$

$$\hat{\Phi}_{BH} = \frac{\text{COUNT of (Bias \& H)}}{\text{COUNT OF BIAS}}$$

$$= \frac{\text{COUNT }(B,H)}{\text{COUNT}(B,T) + \text{COUNT}(B,H)}$$

$$= \frac{4}{1 + 4}$$

$$= 0.8$$

$\hat{\Phi}_{BT} = 0.2$

$\hat{\Phi}_{FH} = 0.4$

$\hat{\Phi}_{FT} = 0.6$

PVII: HMM PARAMETER LEARNING WITHOUT FULLY OBSERVED DATA

* WITH MISSING DATA | LATENT VARIABLES (z_t) COMPUTING MLE | MAP IS HARD

* COULD USE GRADIENT BASED METHODS

PVII: HMM PARAMETER LEARNING WITHOUT FULLY OBSERVED DATA

* WITH MISSING DATA | LATENT VARIABLES (z_t) COMPUTING MLE | MAP IS HARD

* COULD USE GRADIENT BASED METHODS BUT TRICKY TO ENFORCE CONSTRAINTS like $\sum_j \pi_j = 1$ etc

* OFTEN USE EXPECTATION MAXIMISATION (EM)
  → ITERATIVE, CLOSED FORM UPDATES

EM ALGORITHM

Let x_i be observed variable
z_i be latent / hidden

{ General case
DON'T CONFUSE
NOTATION WITH HMM }

GOAL: MAXIMIZE LOG. LIKELIHOOD

$$l(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

EM ALGORITHM

Let x_i be observed variable
z_i be latent / hidden

{ General case

DON'T CONFUSE NOTATION WITH HMM }

* GOAL: MAXIMIZE LOG. LIKELIHOOD

$$l(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta) = \sum_{i=1}^{N} \log \left[\sum_{z_i} p(x_i, z_i \mid \theta) \right]$$

HARD TO SOLVE! (LOG CAN'T BE PUSHED INSIDE)

EM ALGORITHM

* GOAL: MAXIMIZE LOG. LIKELIHOOD

$$\ell(\theta) = \sum_{i=1}^{N} \log p(x_i | \theta) = \sum_{i=1}^{N} \log \left[\sum_{z_i} p(x_i, z_i | \theta) \right]$$

* DEFINE COMPLETE DATA LOG. LIKELIHOOD

$$\ell_c(\theta) \triangleq \sum_{i=1}^{N} \log p(x_i, z_i | \theta)$$

EM ALGORITHM

* GOAL: MAXIMIZE LOG. LIKELIHOOD

$$l(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta) = \sum_{i=1}^{N} \log \left[\sum_{z_i} p(x_i, z_i \mid \theta) \right]$$

* DEFINE COMPLETE DATA LOG. LIKELIHOOD

$$l_c(\theta) \triangleq \sum_{i=1}^{N} \log p(x_i, z_i \mid \theta)$$

$l_c(\theta)$ can not be computed $\because$ z_i is unknown

EM ALGORITHM

* GOAL: MAXIMIZE LOG. LIKELIHOOD

$$\ell(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta) = \sum_{i=1}^{N} \log \left[\sum_{z_i} p(x_i, z_i \mid \theta) \right]$$

* DEFINE COMPLETE DATA LOG. LIKELIHOOD

$$\ell_c(\theta) \triangleq \sum_{i=1}^{N} \log p(x_i, z_i \mid \theta)$$

* DEFINE EXPECTED COMPLETE DATA LIKELIHOOD

$$Q(\theta, \theta^{t-1}) = E\left[\ell_c(\theta) \mid D, \theta^{t-1} \right]$$

t: CURREN ITERATION

Q: AUXILIARY FUNCTION

EM ALGORITHM

* GOAL: MAXIMIZE LOG. LIKELIHOOD

$$\ell(\theta) = \sum_{i=1}^{N} \log p(x_i|\theta) = \sum_{i=1}^{N} \log \left[\sum_{z_i} p(x_i, z_i|\theta) \right]$$

* DEFINE COMPLETE DATA LOG. LIKELIHOOD

$$\ell_c(\theta) \triangleq \sum_{i=1}^{N} \log p(x_i, z_i|\theta)$$

* DEFINE EXPECTED COMPLETE DATA LIKELIHOOD

$$Q(\theta, \theta^{t-1}) = E\left[\ell_c(\theta)|D, \theta^{t-1}\right]$$

* E-step : COMPUTE Q $\Rightarrow$ COMPUTE EXPECTED SUFFICIENT STATS (ESS)

EM ALGORITHM

* GOAL: MAXIMIZE LOG. LIKELIHOOD

$$\ell(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta) = \sum_{i=1}^{N} \log \left[\sum_{z_i} p(x_i, z_i \mid \theta) \right]$$

* DEFINE COMPLETE DATA LOG. LIKELIHOOD

$$\ell_c(\theta) \triangleq \sum_{i=1}^{N} \log p(x_i, z_i \mid \theta)$$

* DEFINE EXPECTED COMPLETE DATA LIKELIHOOD

$$Q(\theta, \theta^{t-1}) = E\left[\ell_c(\theta) \mid D, \theta^{t-1} \right]$$

* E-step : COMPUTE Q
* M-step : OPTIMIZE Q wrt θ , i.e. $\theta^{t} = \underset{\theta}{\operatorname{argmax}} \, (\theta, \theta^{t-1})$

EM ALGORITHM (WITHOUT THE DERIVATION)

TWO STEPS (TILL CONVERGENCE)

① E : INFERRING MISSING VALUES $(Z_{i,t})$ GIVEN MODEL PARAMETERS $(\theta = (\pi, A, \phi))$

EM ALGORITHM

TWO STEPS (TILL CONVERGENCE)

① E : INFERRING MISSING VALUES $(Z_{i,t})$
GIVEN MODEL PARAMETERS $(\theta = \{\pi, A, \phi\})$

② M : OPTIMIZING PARAMETERS $(\theta = \{\pi, A, \phi\})$
USING MLE AND
FILLED IN DATA $(Z_{i,t})$

COMPUTING EXPECTED SUFFICIENT STATS (ESS)
(OR TERMS FROM AUX. FUNCTION ON WHICH MLE DEPENDS ON)

COMPUTING EXPECTED SUFFICIENT STATS (ESS)
(OR TERMS FROM AUX. FUNCTION ON
WHICH MLE DEPENDS ON)

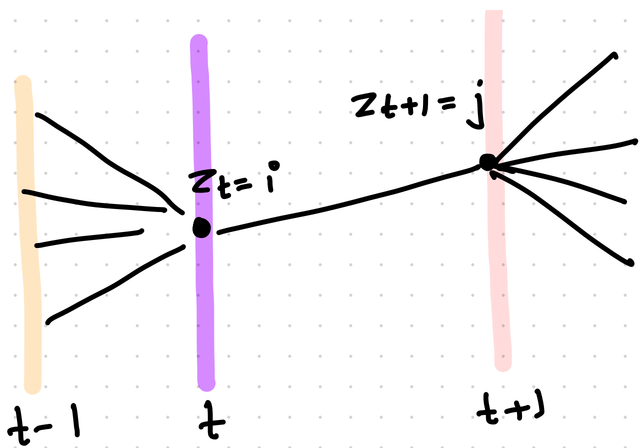
$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$

Probability of being in state 'i' at
't' and state 'j' at 't+1' given
model and observations

COMPUTING EXPECTED SUFFICIENT STATS (ESS)

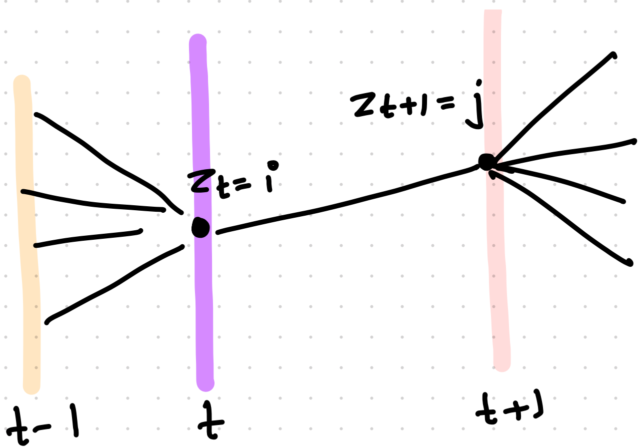
(OR TERMS FROM AUX. FUNCTION ON WHICH MLE DEPENDS ON)

$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$



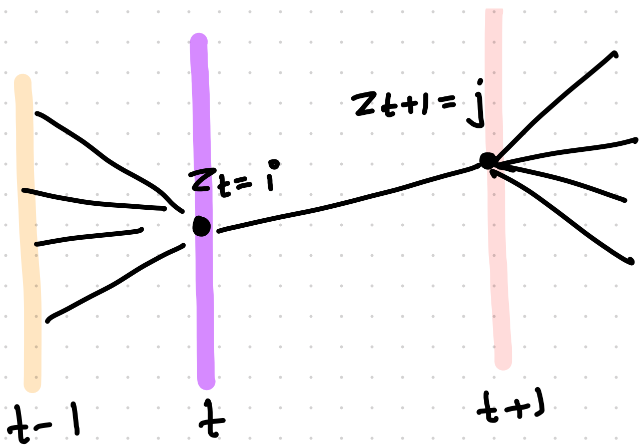
$$\xi_t(i,j) = p(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$

$$= \frac{p(z_t = i, z_{t+1} = j, x_{1:T} \mid \theta)}{p(x_{1:T} \mid \theta)}$$



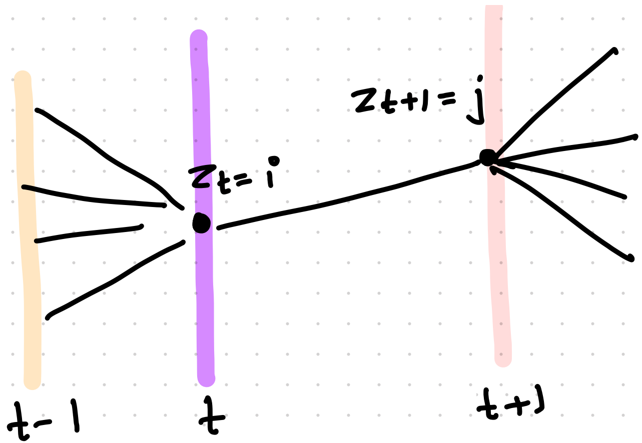
$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$

$$= \frac{P(z_t = i, z_{t+1} = j, x_{1:T} \mid \theta)}{P(x_{1:T} \mid \theta)}$$



$$\xi_t(i,j) = \frac{P(z_t = i, x_{1:t})\, A_{ij}\, P(x_{t+1} \mid z_{t+1} = j)\, P(x_{t+1:T} \mid z_{t+1} = j)}{P(x_{1:T} \mid \theta)}$$

$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$



$$\xi_t(i,j) = \frac{P(z_t = i, x_{1:t})\, A_{ij}\, P(x_{t+1} \mid z_{t+1} = j)\, P(x_{t+1:T} \mid z_{t+1} = j)}{P(x_{1:T} \mid \theta)}$$

$$= \frac{\alpha_t(i)\, \phi_j(x_{t+1})\, \beta_{t+1}(j)\, A_{ij}}{\sum_i \sum_j \alpha_t(i)\, \phi_j(x_{t+1})\, \beta_t(j)\, A_{ij}}$$

$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$

NOW; $\gamma_t(i)$ = Probability of being in state 'i' at time 't' given observations and model.

$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$

Now; $\gamma_t(i) =$ Probability of being in state 'i' at time 't' given observations and model.

$$\Rightarrow \gamma_t(i) = \sum_{j=1}^{K} \xi_t(i,j)$$

$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$

Now; $\gamma_t(i) =$ Probability of being in state 'i' at time 't' given observations and model.

$$\Rightarrow \gamma_t(i) = \sum_{j=1}^{K} \xi_t(i,j)$$

Q: What does $\sum_t \gamma_t(i)$ mean?

$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$

Now; $\gamma_t(i) =$ Probability of being in state 'i' at time 't' given observations and model.

$$\Rightarrow \gamma_t(i) = \sum_{j=1}^{K} \xi_t(i,j)$$

Q: What does $\sum_t \gamma_t(i)$ mean?

Expected # of visits to state i

$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$

Now; $\gamma_t(i) =$ Probability of being in state 'i' at time 't' given observations and model.

$$\Rightarrow \gamma_t(i) = \sum_{j=1}^{K} \xi_t(i,j)$$

Q: What does $\sum_{t=1}^{T-1} \gamma_t(i)$ mean

Expected # of visits to state i

Or, Expected # of transitions from state i (ignoring $t = T$)

$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$

Q: What does $\sum_{t=1}^{T-1} \gamma_t(i)$ mean

Expected # of visits to state i

Or, Expected # of transitions from state i (ignoring $t = T$)

Q: What does $\sum_{t=1}^{T-1} \xi_t(i,j)$ mean?

$$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid x_{1:T}, \theta)$$

Q: What does $\sum_{t=1}^{T-1} \gamma_t(i)$ mean

Expected # of visits to state i

Or, Expected # of transitions from state i (ignoring $t = T$)

Q: What does $\sum_{t=1}^{T-1} \xi_t(i,j)$ mean?

Expected # of transitions from state i to j.

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

S_1	Bias	Fair	fair	Bias
	H	T	H	H
S2 :	F	F	F	F
	H	T	T	T
S_3 :	B	B	B	B
	T	H	H	H

Unknown

Let us assume:

← 1^{st} iteration (RANDOM INIT.)

$$\pi^1 = [0.9 \quad 0.1] \qquad A^1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$$

$$\phi_B^1 = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ \begin{bmatrix} 0.8 & 0.2 \end{bmatrix} \end{array} \qquad \phi_F^1 = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ \begin{bmatrix} 0.4 & 0.6 \end{bmatrix} \end{array}$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

CONSIDER S1 ONLY FOR NOW

S_1

H T H H

$\pi^1 = [0.9 \quad 0.1]$

$A^1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$

$\phi_B^1 = \begin{array}{cc} H & T \\ \left[0.8 \right. & \left. 0.2 \right] \end{array}$ $\quad \phi_F^1 = \begin{array}{cc} H & T \\ \left[0.4 \right. & \left. 0.6 \right] \end{array}$

$\eta_1(B, B) = ?$ $\quad \eta_1(B, F) = ?$ $\quad \eta_1(F, F) = ?$ $\quad \eta_1(F, B) = ?$

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

CONSIDER S1 ONLY FOR NOW

S_1 H T H H

$$\pi^1 = [0.9 \quad 0.1] \qquad A^1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$$

$$\phi^1_B = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ [0.8 \quad 0.2] \end{array} \qquad \phi^1_F = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ [0.4 \quad 0.6] \end{array}$$

$$\xi_1(B,B) = \frac{\alpha_1(B) . A_{BB}\, \phi_B(x_2 = T) . \beta_2(B)}{\text{NORMALIZATION CONSTANT}}$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

CONSIDER S1 ONLY FOR NOW

S_1	H	T	H	H

$$\pi^1 = [0.9 \quad 0.1] \qquad A^1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$$

$$\phi^1_B = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ [0.8 \quad 0.2] \end{array} \qquad \phi^1_F = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ [0.4 \quad 0.6] \end{array}$$

$$\xi_1(B,B) = \frac{\alpha_1(B) \cdot A_{BB}\, \phi_B(x_2 = T) \cdot \beta_2(B)}{\text{NORMALIZATION CONSTANT}}$$

$$\alpha_1(B) = \pi_B\, \phi_B(x_1 = H) = (0.9)(0.8) = 0.72$$

$$\beta_2(B) = ?$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

CONSIDER S1 ONLY FOR NOW

S_1	H	T	H	H

$\pi' = [0.9 \quad 0.1]$ $A' = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$

$\phi'_B = \begin{array}{cc} H & T \\ \end{array} [0.8 \quad 0.2]$ $\phi'_F = \begin{array}{cc} H & T \\ \end{array} [0.4 \quad 0.6]$

$\beta_4(B) = 1$ & $\beta_4(F) = 1$

$\beta_3(B) = A_{BB}\,\phi_B(x_4 = H)\,\beta_4(B) + A_{BF}\,\phi_F(x_4 = H)\,\beta_4(F)$

$= (0.7)(0.8) + (0.3)(0.4) = .56 + .12 = 0.68$

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

CONSIDER S1 ONLY FOR NOW

S_1	H	T	H	H

$\pi^1 = [0.9 \quad 0.1]$ $\qquad A^1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$

$\phi_B^1 = \begin{array}{cc} H & T \\ \hline 0.8 & 0.2 \end{array}$ $\qquad \phi_F^1 = \begin{array}{cc} H & T \\ \hline 0.4 & 0.6 \end{array}$

$\beta_4(B) = 1 \quad \& \quad \beta_4(F) = 1$

$\beta_3(B) = A_{BB}\,\phi_B(x_4 = H)\,\beta_4(B) + A_{BF}\,\phi_F(x_4 = H)\,\beta_4(F)$

$= (0.7)(0.8) + (0.3)(0.4) = .56 + .12 = 0.68$

$\beta_2(B) = A_{BB}\,\phi_B(x_3 = H)\,\beta_3(B) + A_{BF}\,\phi_F(x_3 = H)\,\beta_3(F)$

$= (0.7)(0.8)(0.68) + (0.3)(0.4)(1 - .68) = 0.4192$

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

CONSIDER S1 ONLY FOR NOW

S_1	H	T	H	H

$\pi^1 = [0.9 \quad 0.1]$ $\qquad A^1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$

$\phi_B^1 = \begin{array}{cc} H & T \\ \left[0.8\right. & \left.0.2\right] \end{array}$ $\qquad \phi_F^1 = \begin{array}{cc} H & T \\ \left[0.4\right. & \left.0.6\right] \end{array}$

$$\xi_1(B,B) = \frac{\alpha_1(B) \cdot A_{BB}\, \phi_B(x_2 = T) \cdot \beta_2(B)}{\text{NORMALIZATION CONSTANT } (C)}$$

$$= \frac{0.72 * 0.7 * 0.2 * .4192}{C}$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

CONSIDER S1 ONLY FOR NOW

S_1 H T H H

$\pi^1 = [0.9 \quad 0.1]$

$A^1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$

$\phi_B^1 = \begin{array}{cc} H & T \\ \end{array}$ $[0.8 \quad 0.2]$

$\phi_F^1 = \begin{array}{cc} H & T \\ \end{array}$ $[0.4 \quad 0.6]$

We can similarly find $\eta_1(B,F)$, $\eta_1(F,F)$, $\eta_1(F,B)$ for this sequence

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

CONSIDER S1 ONLY FOR NOW

S_1	H	T	H	H

$$\pi^1 = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix} \qquad A^1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$$

$$\Phi^1_B = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ \begin{bmatrix} 0.8 & 0.2 \end{bmatrix} \end{array} \qquad \Phi^1_F = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ \begin{bmatrix} 0.4 & 0.6 \end{bmatrix} \end{array}$$

we can find.

$$\gamma_1(B) = \xi_1(B, F) + \xi_1(B, B)$$

$$\gamma_1(F) = 1 - \gamma_1(B)$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1

H T H H

CONSIDER ONLY THE STATES AT $t=1$

$$P(Z_1 = B) = \gamma_1(B)$$

$$P(Z_1 = F) = \gamma_1(F)$$

HMM LEARNING EXAMPLE (ESTIMATE π, A, Φ)

S_1

H T H H

CONSIDER ONLY THE STATES AT $t=1$

$$P(Z_1 = B) = \gamma_1(B)$$

$$P(Z_1 = F) = \gamma_1(F)$$

we don't know for sure whether Z_1 is B or F, but we know the probabilies.

KEY INSIGHT: CONSIDER $\gamma_1(j)$ = EXPECTED # OF TIMES $Z_1 = j$

HMM LEARNING EXAMPLE (ESTIMATE π, A, ϕ)

KEY INSIGHT: CONSIDER $\gamma_1(j)$ = EXPECTED # OF TIMES $z_1 = j$

NOW AS WE DID FOR FULLY OBSERVED CASE, WE CAN USE MLE to estimate

π^2, A^2, ϕ^2

EM ALGORITHM

TWO STEPS (TILL CONVERGENCE)

① E : INFERRING MISSING VALUES $(z_{i,t})$
GIVEN MODEL PARAMETERS $(\theta = \{\pi, A, \phi\})$

COMPUTE ESS

$$\gamma_{i,t}(j) \triangleq p(z_t = j \mid x_{i,1:T_i}, \theta)$$

$$\xi_{i,t}(j,k) \triangleq p(z_{t-1} = j, z_t = k \mid x_{i,1:T_i}, \theta)$$

EM ALGORITHM

TWO STEPS (TILL CONVERGENCE)

② M : OPTIMIZING PARAMETERS ($\theta = \{\pi, A, \phi\}$)

USING MLE AND

FILLED IN DATA ($z_{i,t}$) OR <u>ESS</u>

$$\hat{\pi}_k = \text{Expected Freq. of } z_{i,t} = \text{state } k$$

$$= \frac{\sum_{i=1}^{N} \gamma_{i,1}(k)}{N}$$

EM ALGORITHM

TWO STEPS (TILL CONVERGENCE)

② M : OPTIMIZING PARAMETERS ($\theta = \{\pi, A, \phi\}$)

USING MLE AND

FILLED IN DATA ($z_{i,t}$) OR ESS

$$\hat{A}_{jk} = \frac{\text{Expected \# transitions } z_t = j \text{ to } z_{t+1} = k}{\text{Expected \# transitions from } z_t = j}$$

EM ALGORITHM

TWO STEPS (TILL CONVERGENCE)

② M : OPTIMIZING PARAMETERS ($\theta = \{\pi, A, \phi\}$)

USING MLE AND

FILLED IN DATA ($z_{i,t}$) OR <u>ESS</u>

$$\hat{A}_{jk} = \frac{\text{Expected \# transitions } z_t = j \text{ to } z_{t+1} = k}{\text{Expected \# transitions from } z_t = j}$$

$$= \frac{\sum_{i=1}^{N} \sum_{t=2}^{T_i} \eta_{i,t}(j,k)}{\sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \eta_{i,t}(j,k)} \leftarrow \text{NORMALIZATION}$$

EM ALGORITHM

TWO STEPS (TILL CONVERGENCE)

② M : OPTIMIZING PARAMETERS ($\theta = \{\pi, A, \phi\}$)

USING MLE AND

FILLED IN DATA ($z_{i,t}$) OR ESS

$$\hat{\phi}_{jl} = \frac{\text{Expected \# times in state 'j' and observing 'l'}}{\text{Expected \# times in state j}}$$

$$= \frac{\sum_{i=1}^{N} \sum_{t: x_{i,t} = l} \gamma_{i,t}(j)}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j)}$$

HOW TO INITIALIZE θs?

HOW TO INITIALIZE θs?

1. RANDOMLY INITIALIZE W/ MULTIPLE RESTARTS
   - CHOOSE THE "BEST" (MAX LL) PARAMS

HOW TO INITIALIZE θs?

1. RANDOMLY INITIALIZE W/ MULTIPLE RESTARTS
   - CHOOSE THE "BEST" (MAX LL) PARAMS

2. INIT. USING "SOME" TRAINING DATA
   - KNOWN zs

HOW TO INITIALIZE θs?

1. RANDOMLY INITIALIZE W/ MULTIPLE RESTARTS
   - CHOOSE THE "BEST" (MAX LL) PARAMS

2. INIT. USING "SOME" TRAINING DATA
   - KNOWN Zs

3. INIT. USING VITERBI TRAINING