Hidden Markov Model

Anil Sharma

Slides are prepared from various online sources and I don't claim authorship for the content.

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Example: Assistive technology

- *Assume you have a little robot that is trying to **estimate** the **posterior** probability that you are **happy** or **sad**, given that the robot has observed whether you are **watching Game of Thrones (w)**, **sleeping (s)**, **crying (c)** or **face booking (f)**.*

- *Let the **unknown state** be **X=h** if you're happy and **X=s** if you're sad.*

- *Let **Y** denote the **observation**, which can be **w**, **s**, **c** or **f** .*

- *We want to answer queries, such as:*
  P(X=h|Y=f) ?
  P(X=s|Y=c) ?

Assistive technology contd..

- *Assume that an expert has compiled the following **prior** and **likelihood** models:*

  P[X=h] = 0.2
  P[Y=w|X=h] = 0.4, P[Y=s|X=h] = 0.1, P[Y=c|X=h] = 0.1,
  P[Y=w|X=s] = 0.2, P[Y=s|X=s] = 0.4, P[Y=c|X=s] = 0.1,

- ***P(X=h|Y=f) =?***

- Please do the computation on your notebook

Assistive technology contd..

- *But what if instead of an absolute prior, what we have instead is a temporal (**transition prior**). That is, we assume a **dynamical system***

$$P[X_t=s|X_{t-1}=s] = 0.90$$
$$P[X_t=h|X_{t-1}=h] = 0.95$$

- *Given a history of observations, say **Y_1=w, Y_2=f, Y_3=c**, we want to compute the posterior distribution that you are happy at step 3. That is, we want to estimate:*

$$\boldsymbol{P(X3=h|\ Y1=w, Y2=f,\ Y3=c)}$$

- Formulation on board

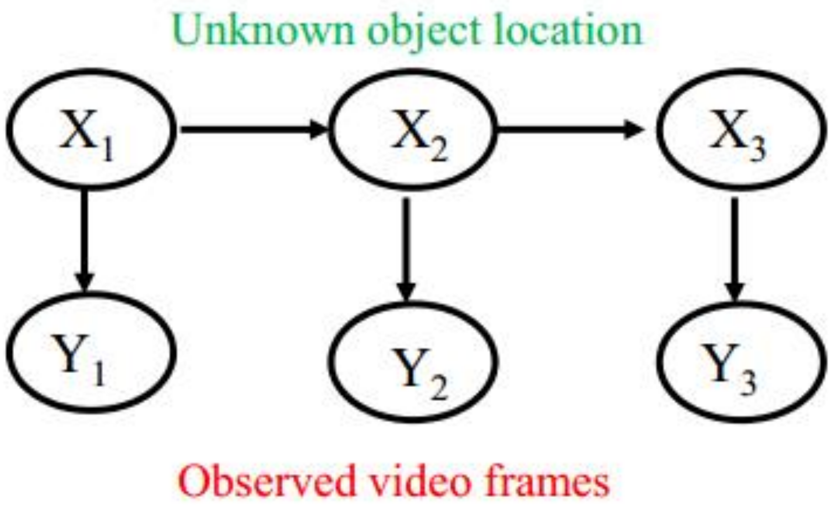
Formulation

- *In general, we assume we have an* ***initial*** *distribution* $\boldsymbol{P(X_0)}$*, a* ***transition*** *model* $\boldsymbol{P(X_t | X_{t-1})}$*, and an* ***observation*** *model* $\boldsymbol{P(Y_t | X_t)}$*.*

- **Filtering:** $\boldsymbol{P(X_t | Y_{1:t}) = P(X_t | Y_1, Y_2, \ldots, Y_t)}$
- **Prediction:** $\boldsymbol{P(X_t | Y_{1:t-1})}$

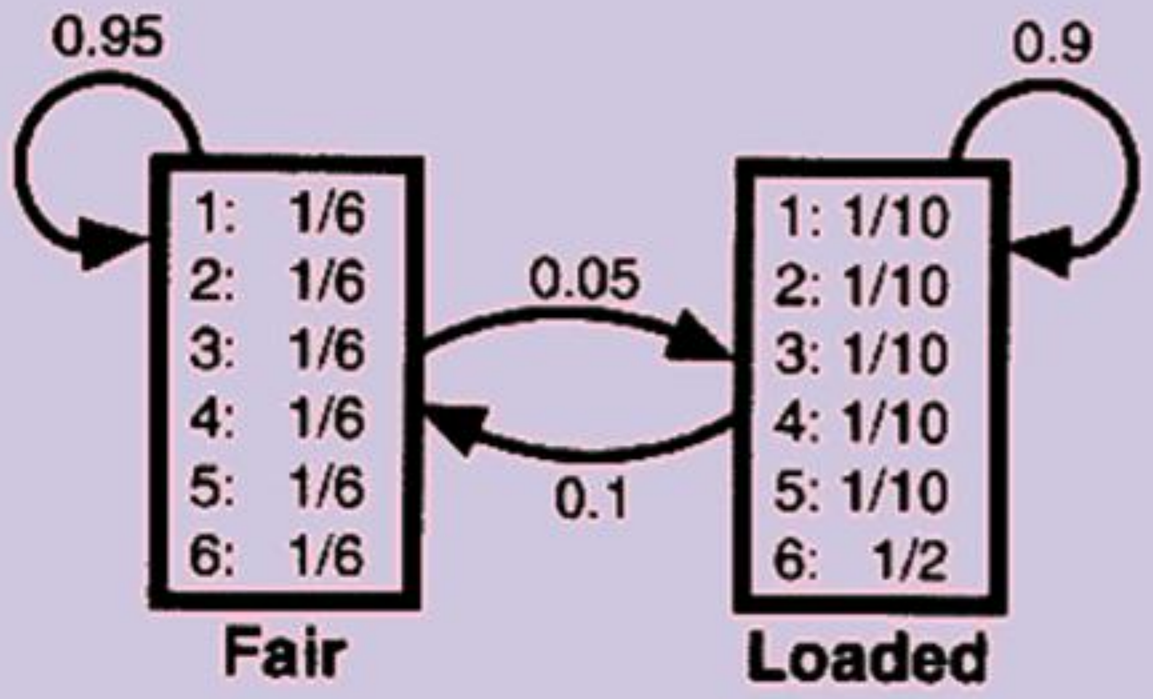
- Computation on board

Example 1: Image tracking



Task: Estimate motion of targets in 3D world from indirect, potentially noisy measurements

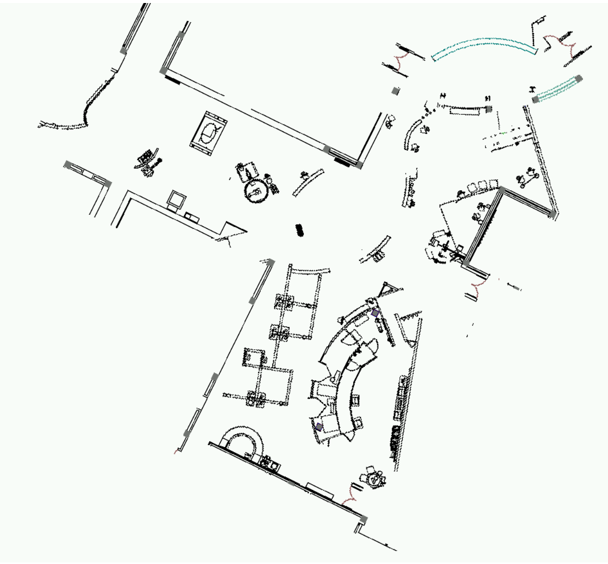
Example 2: Dishonest casino



Actually, what is hidden in this model?

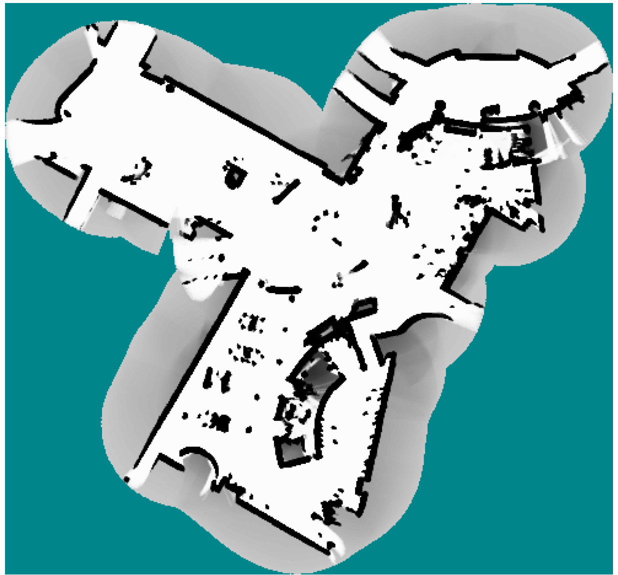
Example 3: Robot Navigation- *SLAM*

Simultaneous Localization and Mapping



CAD Map

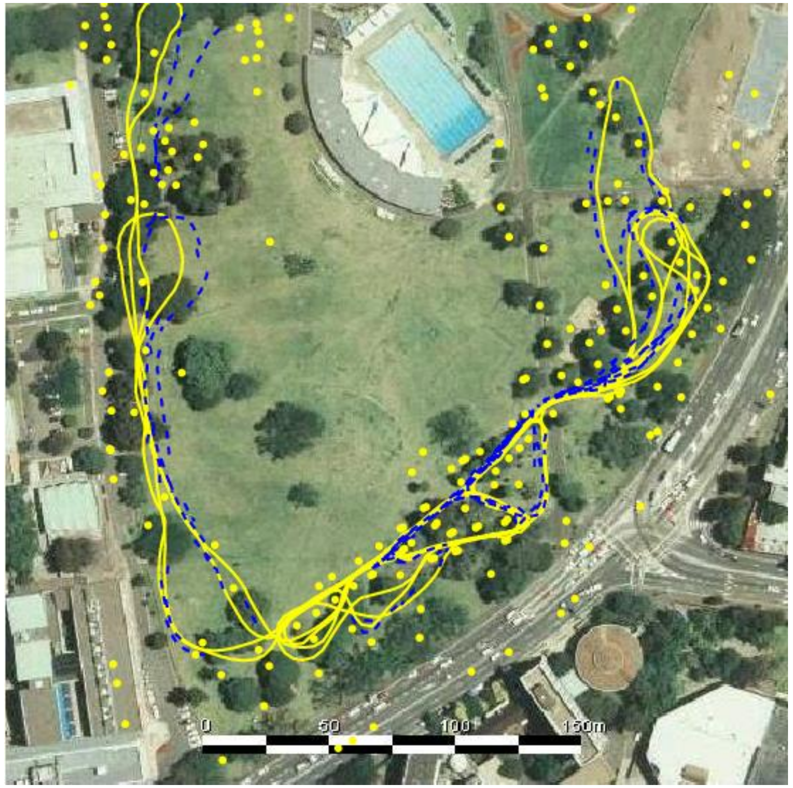
(S. Thrun, San Jose Tech Museum)



Estimated Map

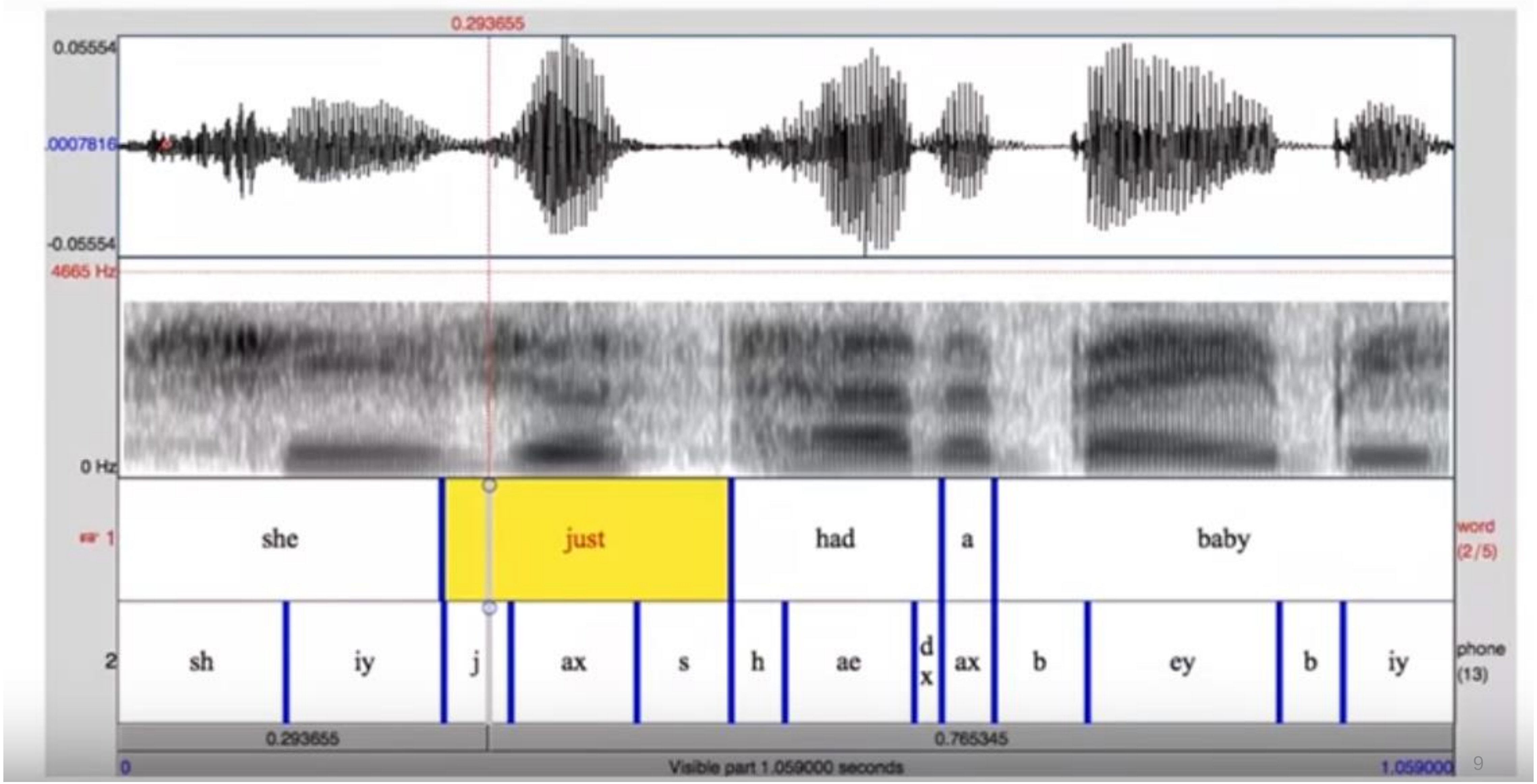
Landmark SLAM

(E. Nebot, Victoria Park)



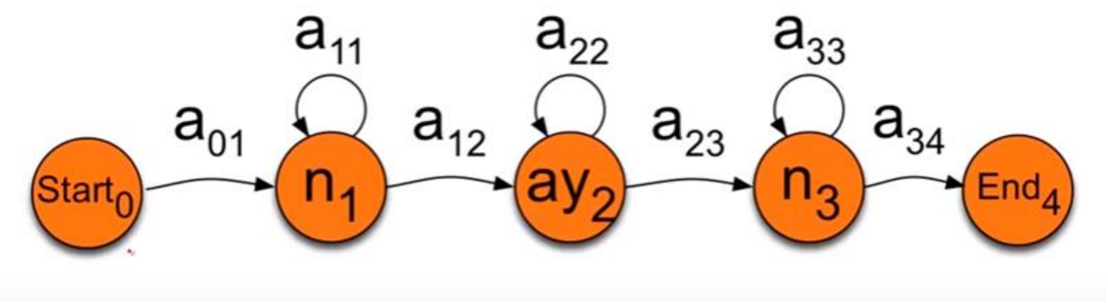
- As robot moves, estimate its pose & world geometry

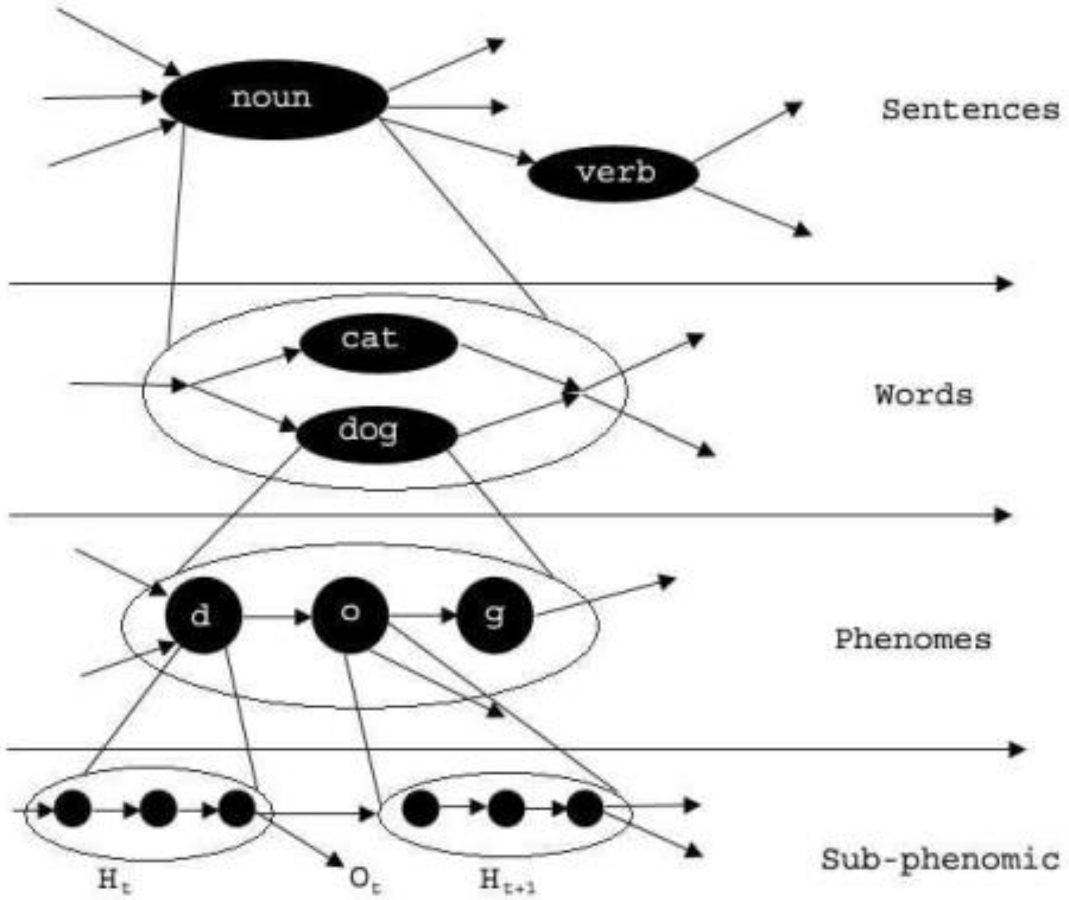
Example 4: Segmentation of acoustic signal



Contd..

- Make a word HMM





More examples..

- DNA sequence alignment
- Human genome project
- Next word prediction
- Text annotation
- Human identification using Gait

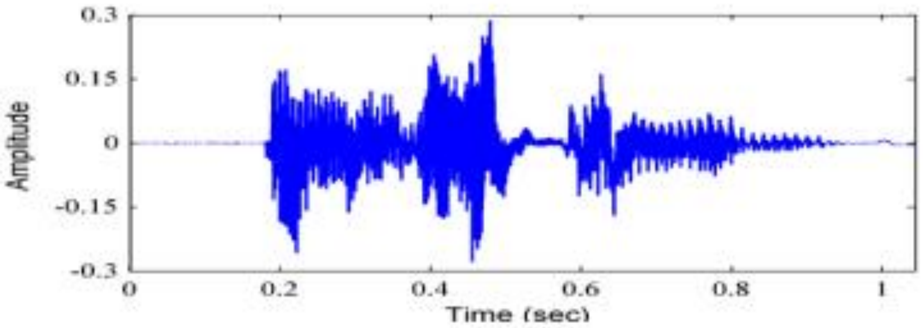
HMM contd..

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

i.i.d. data

- So far we assumed independent, identically distributed data $\{X_i\}_{i=1}^{n} \overset{iid}{\sim} p(\mathbf{X})$
- Sequential (non i.i.d.) data
  - Time-series data

    E.g. Speech



  - Characters in a sentence

  - Base pairs along a DNA strand

Markov Models

- Markov Assumption

parameters in stationary model K-ary variables

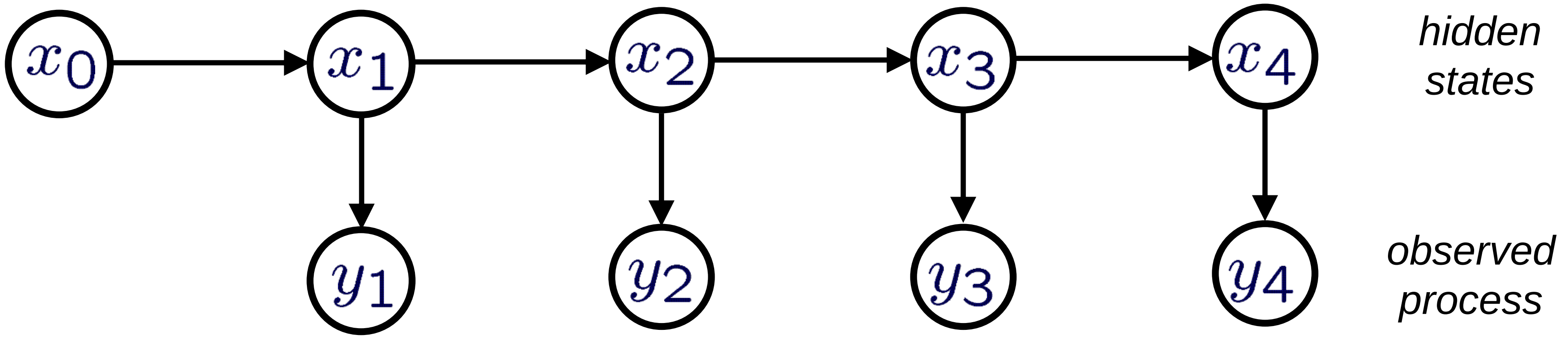
1st order $p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1})$ $O(K^2)$

mth order $p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1}, \ldots, X_{n-m})$ $O(K^{m+1})$

n-1th order $p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1}, \ldots, X_1)$ $O(K^n)$

≡ no assumptions – complete (but directed) graph

Hidden states



- Given x_t, earlier observations provide no *additional information* about the future:

$$p(y_t, y_{t+1}, \dots \mid x_t, y_{t-1}, y_{t-2}, \dots) = p(y_t, y_{t+1}, \dots \mid x_t)$$

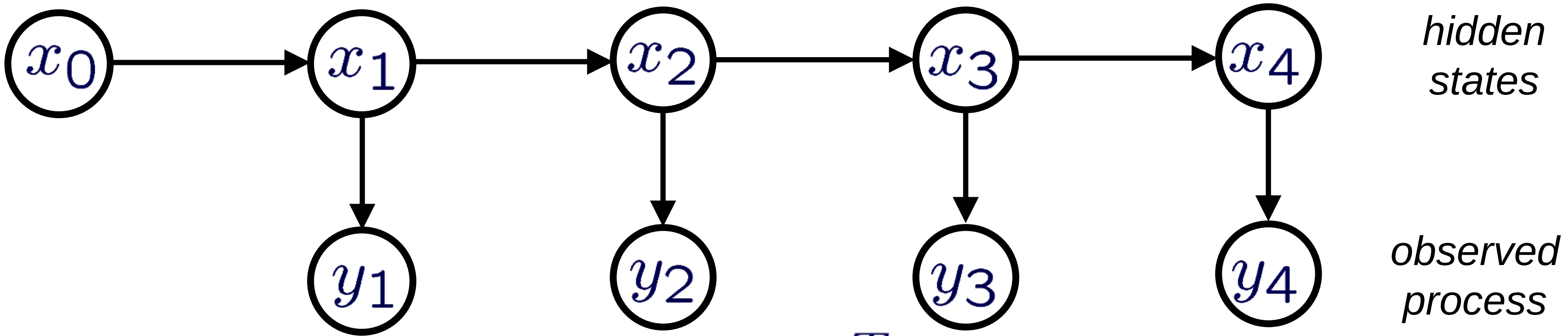
- Transformation of process under which dynamics take a *simple*, first-order form

Hidden Markov Models

- Few realistic time series directly satisfy the assumptions of Markov processes:

"Conditioned on the present,
the past & future are independent"

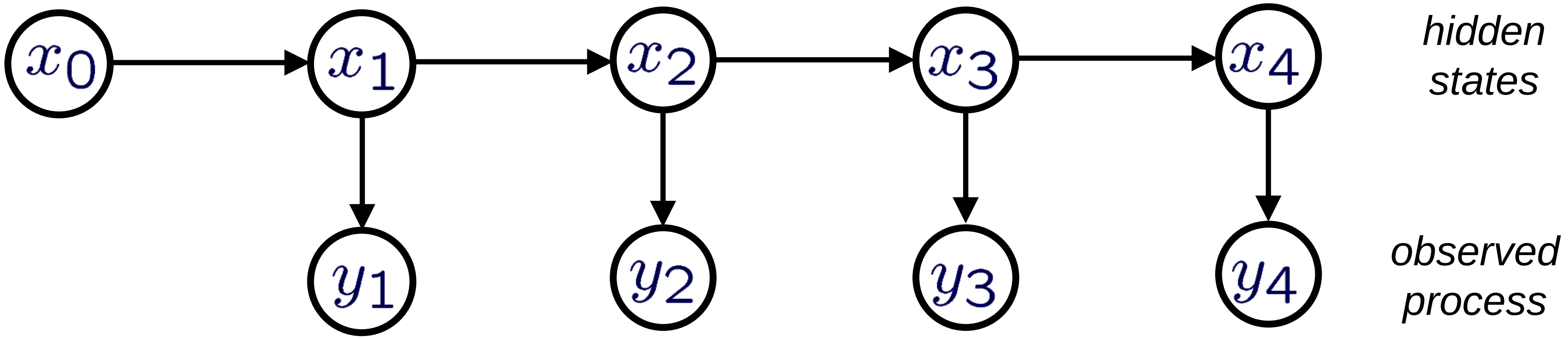
- Motivates *hidden Markov models (HMM):*



$$p(x_0, x_1, \ldots, x_T) = p(x_0) \prod_{t=1}^{T} p(x_t \mid x_{t-1}) p(y_t \mid x_t)$$

Discrete State HMMs

$$x_t \in \{1, 2, \ldots, N\}$$

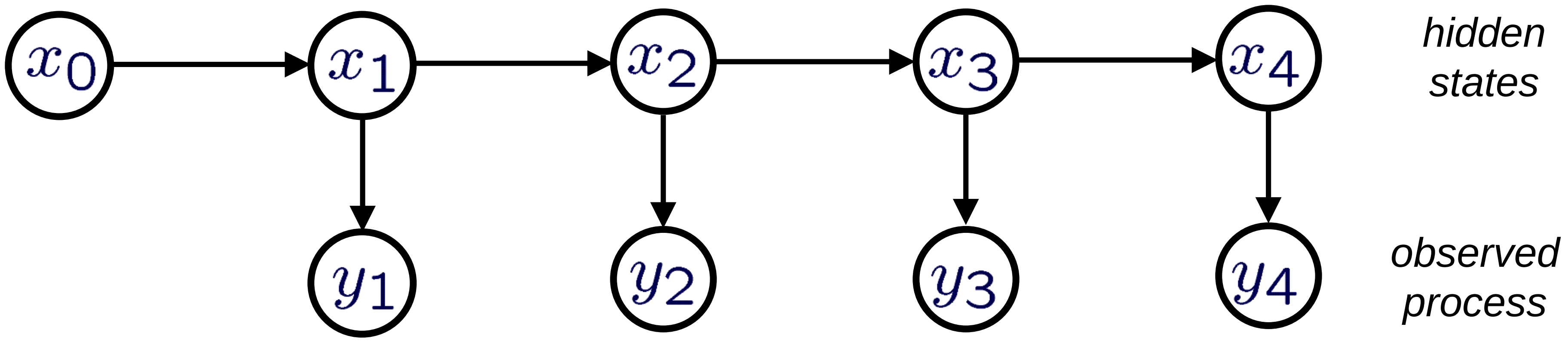


- Associate each of the *N* hidden states with a different observation distribution:

$$p(y_t \mid x_t = 1) \qquad p(y_t \mid x_t = 2) \quad \cdots$$

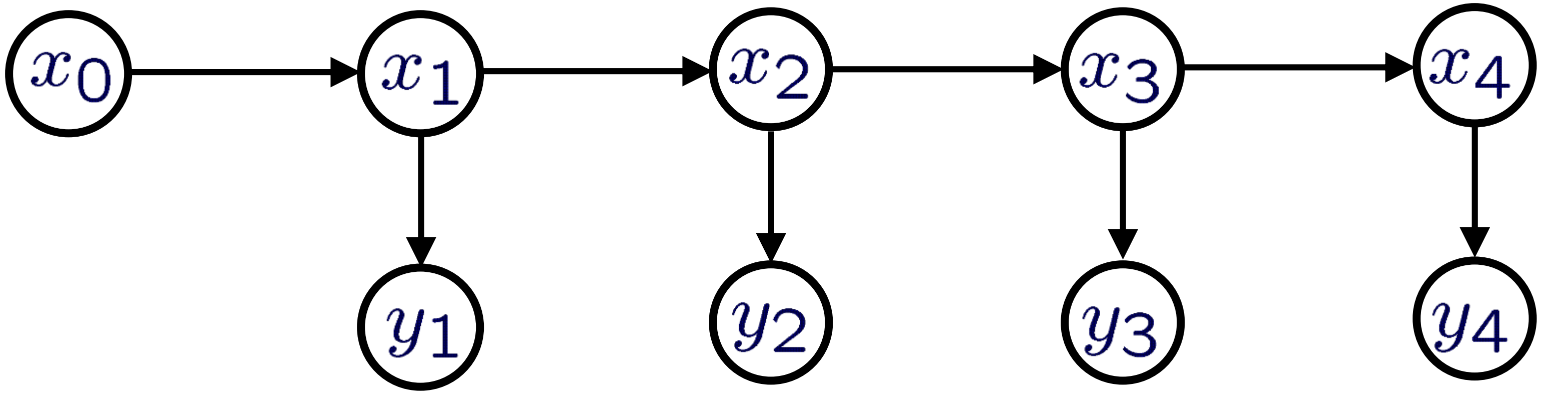
- Observation densities are typically chosen to encode domain knowledge

Where do states come from?



- Analysis of a *physical phenomenon*:
  - Dynamical models of an aircraft or robot
  - Geophysical models of climate evolution
- Discovered from *training data*:
  - Recorded examples of spoken English
  - Historic behavior of stock prices

Specifying an HMM



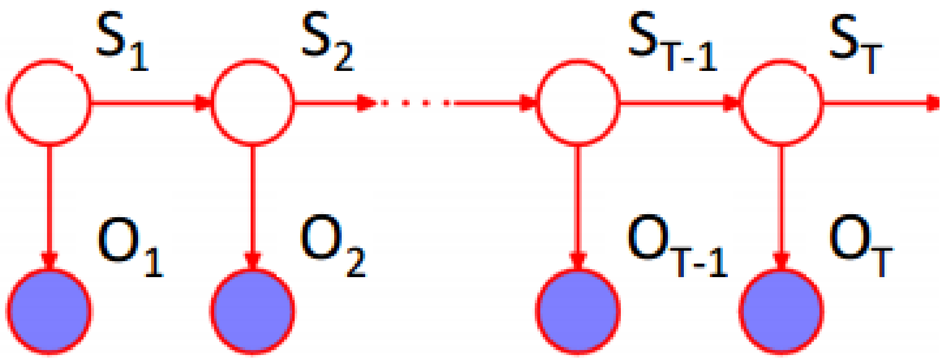
- Observation/emission probabilities: $P(y_i|x_i)$
- Transition probabilities: $P(x_i|x_{i-1})$
- Initial state distribution: $P(x_0)$

Hidden Markov Model (changed notations)

- Parameters – stationary/homogeneous markov model (independent of time t)

Initial probabilities

$p(S_1 = i) = \pi_i$



Transition probabilities

$p(S_t = j | S_{t-1} = i) = p_{ij}$

Emission probabilities

$p(O_t = y | S_t = i) = q_i^y$

$$p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T) = p(S_1) \prod_{t=2}^{T} p(S_t|S_{t-1}) \prod_{t=1}^{T} p(O_t|S_t)$$

Consider an Example

- The Dishonest Casino

A casino has two dices:

Fair dice

P(1) = P(2) = P(3) = P(5) = P(6) = 1/6

Loaded dice

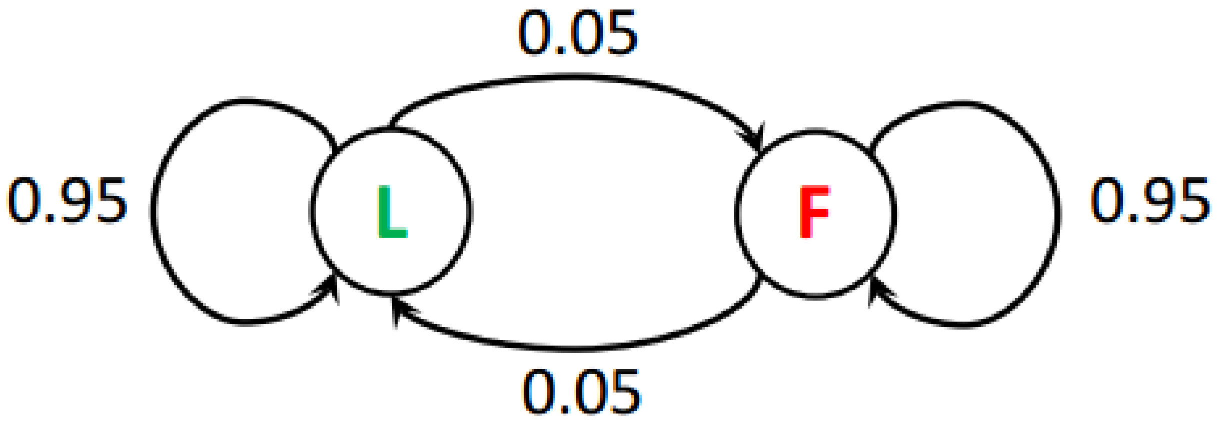
P(1) = P(2) = P(3) = P(5) = 1/10

P(6) = ½

- Casino player switches back-&-forth between fair and loaded die with 5% probability

Example contd..

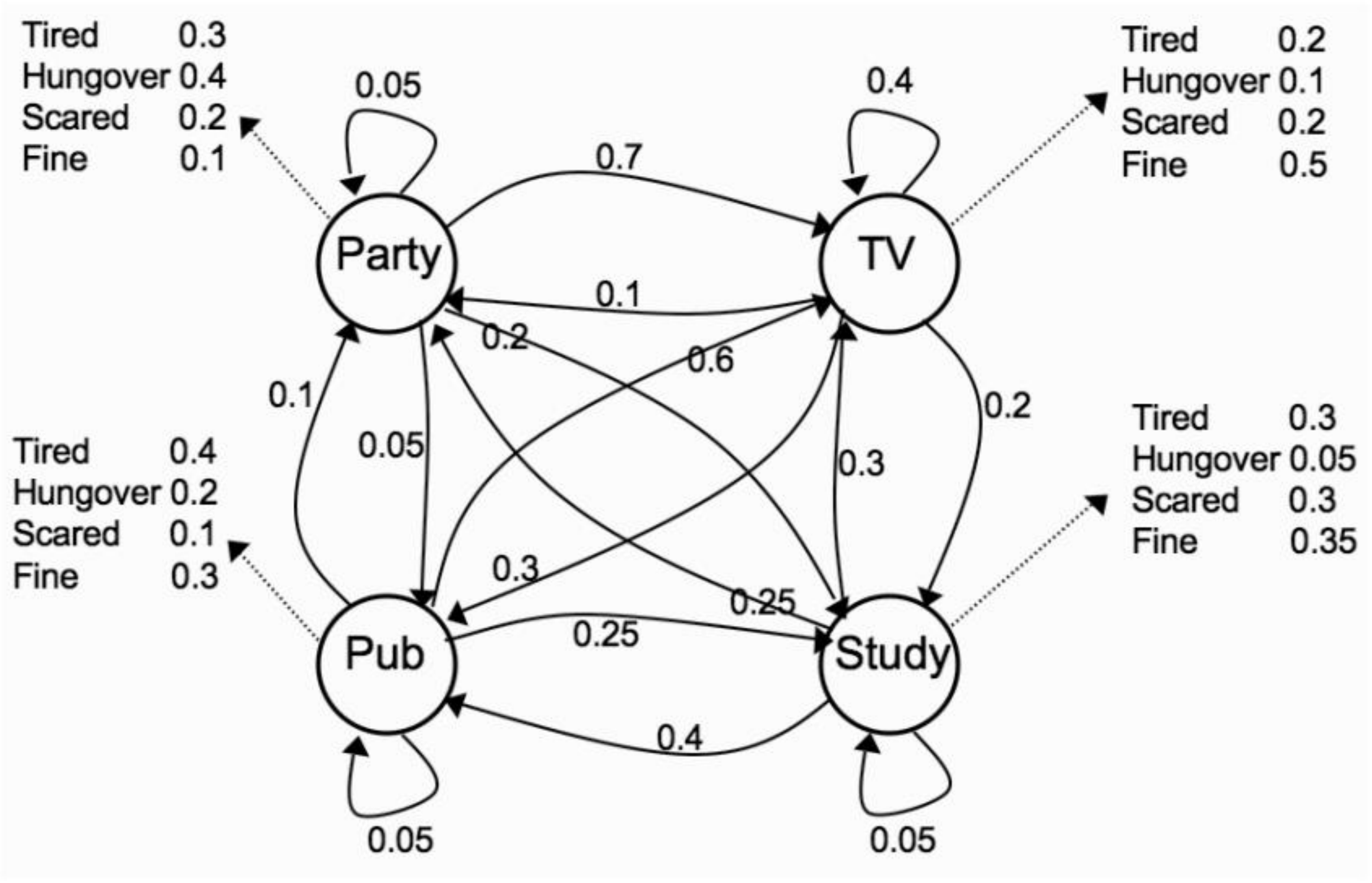
- Switch between F and L with 5% probability



- **HMM Parameters**

Initial probs	$P(S_1 = L) = 0.5 = P(S_1 = F)$
Transition probs	$P(S_t = L/F \mid S_{t-1} = L/F) = 0.95$
	$P(S_t = F/L \mid S_{t-1} = L/F) = 0.05$
Emission probabilities	$P(O_t = y \mid S_t = F) = 1/6 \quad y = 1,2,3,4,5,6$
	$P(O_t = y \mid S_t = L) = 1/10 \quad y = 1,2,3,4,5$
	$= 1/2 \quad y = 6$

Another Example



Contd..

	Previous night			
	TV	Pub	Party	Study
TV	0.4	0.6	0.7	0.3
Pub	0.3	0.05	0.05	0.4
Party	0.1	0.1	0.05	0.25
Study	0.2	0.25	0.2	0.05

Transition probabilities

Observation/ emission probabilities

	TV	Pub	Party	Study
Tired	0.2	0.4	0.3	0.3
Hungover	0.1	0.2	0.4	0.05
Scared	0.2	0.1	0.2	0.3
Fine	0.5	0.3	0.1	0.35

Three Basic Problems in HMM

- Evaluation – What is the probability of the observed sequence? - **Forward Algorithm**
- Decoding – What is the probability that the third roll was loaded given the observed sequence? - **Forward-Backward Algorithm**

– What is the most likely die sequence given the observed sequence? - **Viterbi Algorithm**

- Learning – Under what parameterization is the observed sequence most probable? **Baum-Welch Algorithm (EM)**

Three main problems in HMM

- **Evaluation** – Given HMM parameters & observation seqn $\{O_t\}_{t=1}^T$

  find $p(\{O_t\}_{t=1}^T|\theta)$ prob of observed sequence

- **Decoding** – Given HMM parameters & observation seqn $\{O_t\}_{t=1}^T$

  find $\arg\max_{s_1,\dots,s_T} p(\{S_t\}_{t=1}^T|\{O_t\}_{t=1}^T,\theta)$ most probable sequence of hidden states

- **Learning** – Given HMM with unknown parameters and $\{O_t\}_{t=1}^T$ observation sequence

  find $\arg\max_{\theta} p(\{O_t\}_{t=1}^T|\theta)$ parameters that maximize likelihood of observed data

Queries

GIVEN: A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

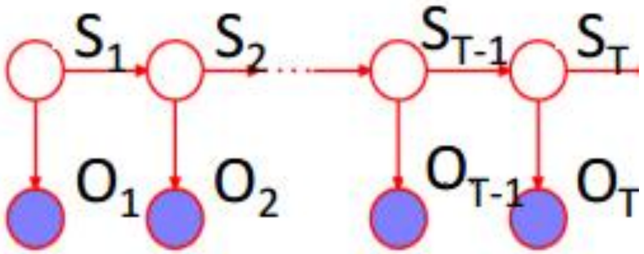
QUESTION

- How likely is this sequence, given our model of how the casino works?
  - This is the **EVALUATION** problem in HMMs
- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
  - This is the **DECODING** question in HMMs
- How “loaded” is the loaded die? How “fair” is the fair die? How often does the casino player change from fair to loaded, and back?
  - This is the **LEARNING** question in HMMs

1. Evaluation Problem

- Given HMM parameters $p(S_1), p(S_t|S_{t-1}), p(O_t|S_t)$ & observation sequence $\{O_t\}_{t=1}^T$

find probability of observed sequence



$$p(\{O_t\}_{t=1}^T) = \sum_{S_1,\ldots,S_T} p(\{O_t\}_{t=1}^T, \{S_t\}_{t=1}^T)$$

$$= \sum_{S_1,\ldots,S_T} p(S_1) \prod_{t=2}^T p(S_t|S_{t-1}) \prod_{t=1}^T p(O_t|S_t)$$

requires summing over all possible hidden state values at all times – K^T exponential # terms!

Instead: $$p(\{O_t\}_{t=1}^T) = \sum_k \underbrace{p(\{O_t\}_{t=1}^T, S_T = k)}_{\alpha_T^k}$$

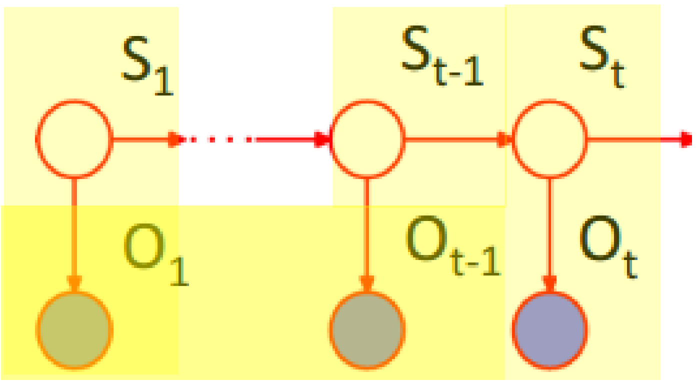
Compute recursively

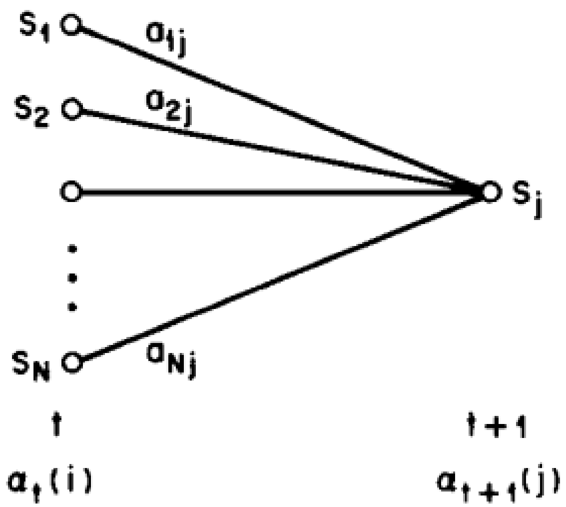
Forward Probability

$$p(\{O_t\}_{t=1}^T) = \sum_k p(\{O_t\}_{t=1}^T, S_T = k) = \sum_k \alpha_T^k$$

Compute forward probability α_t^k recursively over t

$$\alpha_t^k \quad := \quad p(O_1, \ldots, O_t, S_t = k)$$





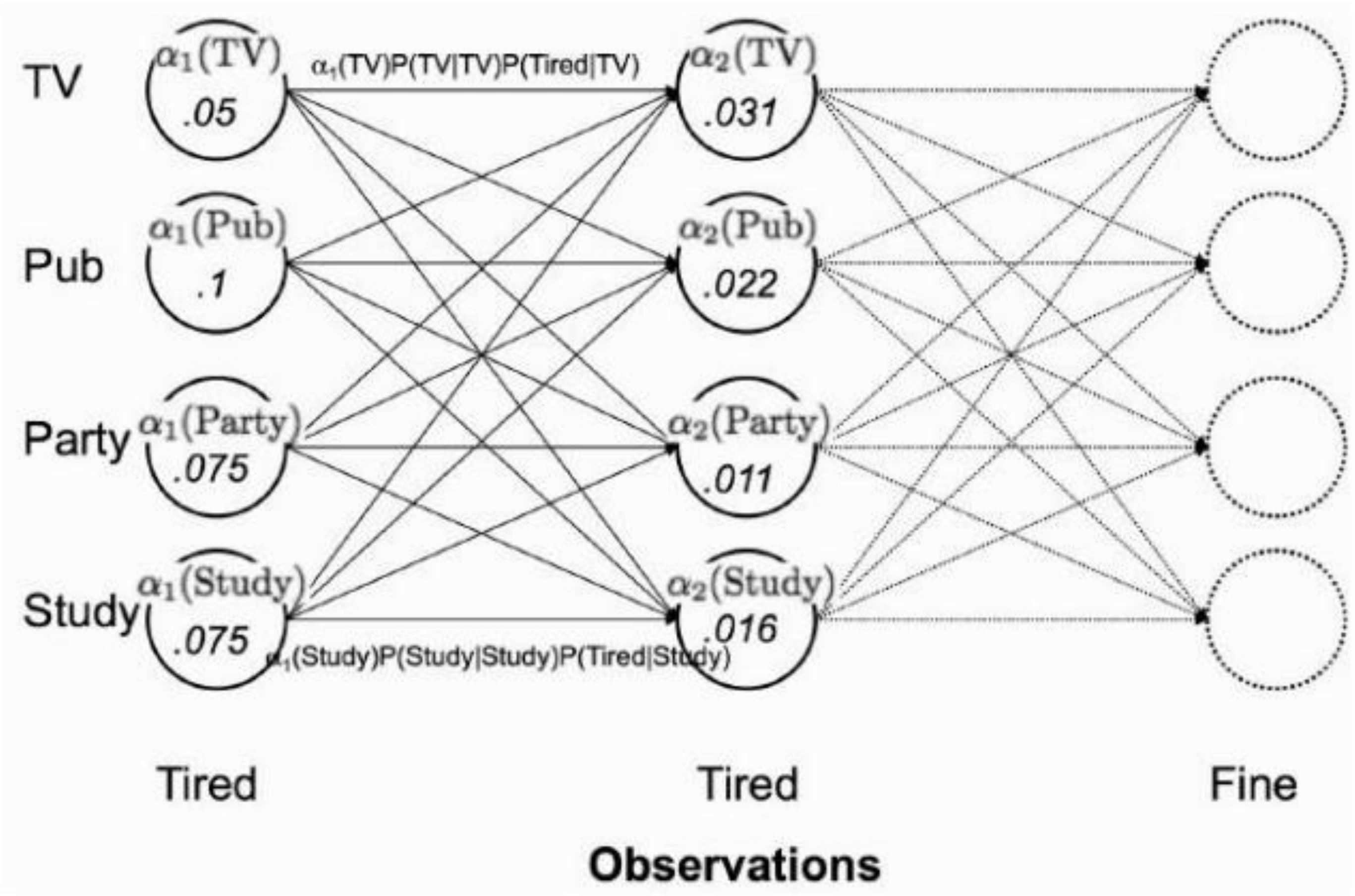
Introduce S_{t-1}

Chain rule

Markov assumption

$$= \quad p(O_t|S_t = k) \sum_i \alpha_{t-1}^i p(S_t = k|S_{t-1} = i)$$

Forward variables (Graphically)



Forward Algorithm

Can compute α_t^k for all k, t using dynamic programming:

- Initialize: $\alpha_1^k = p(O_1 | S_1 = k)\, p(S_1 = k)$ for all k

- Iterate: for t = 2, ..., T

  $\alpha_t^k = p(O_t | S_t = k) \sum_i \alpha_{t-1}^i\, p(S_t = k | S_{t-1} = i)$ for all k

- Termination: $p(\{O_t\}_{t=1}^T) = \sum_k \alpha_T^k$

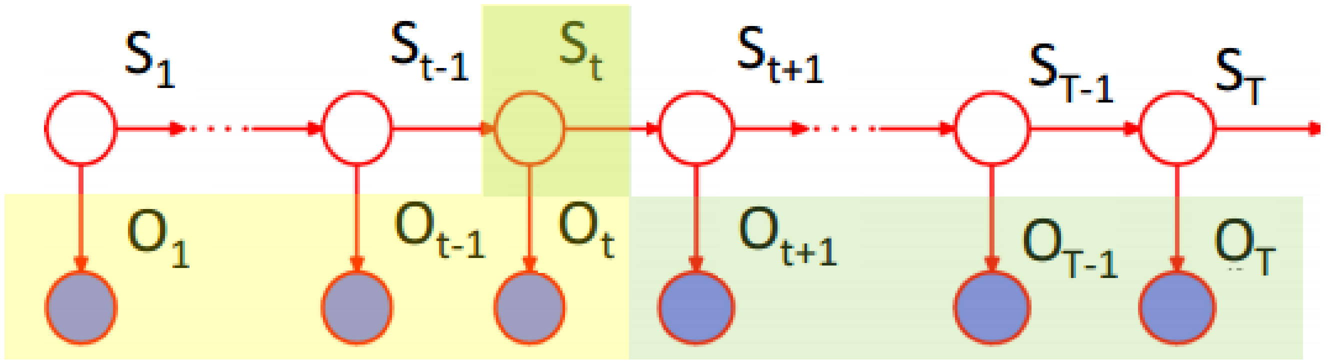
2. Decoding Problem -1

- Given HMM parameters $p(S_1), p(S_t|S_{t-1}), p(O_t|S_t)$ & observation sequence $\{O_t\}_{t=1}^T$

  find probability that hidden state at time t was k $p(S_t = k|\{O_t\}_{t=1}^T)$

$$
\begin{aligned}
p(S_t = k, \{O_t\}_{t=1}^T) &= p(O_1, \dots, O_t, S_t = k, O_{t+1}, \dots, O_T) \\
&= \underbrace{p(O_1, \dots, O_t, S_t = k)}_{\alpha_t^k} \underbrace{p(O_{t+1}, \dots, O_T | S_t = k)}_{\beta_t^k}
\end{aligned}
$$

Compute recursively

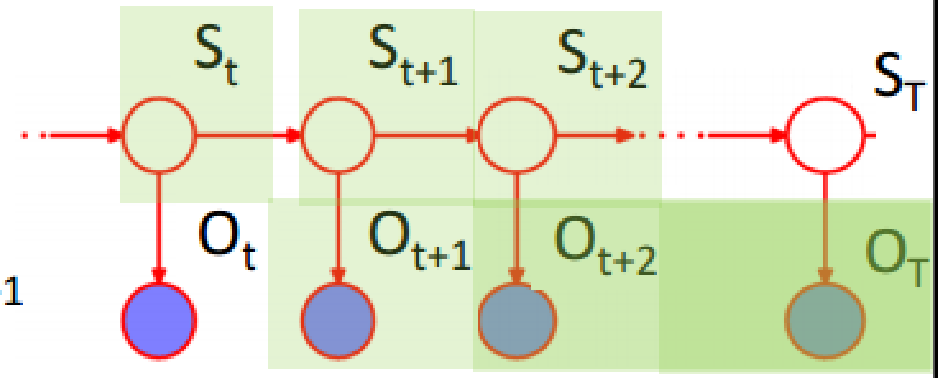


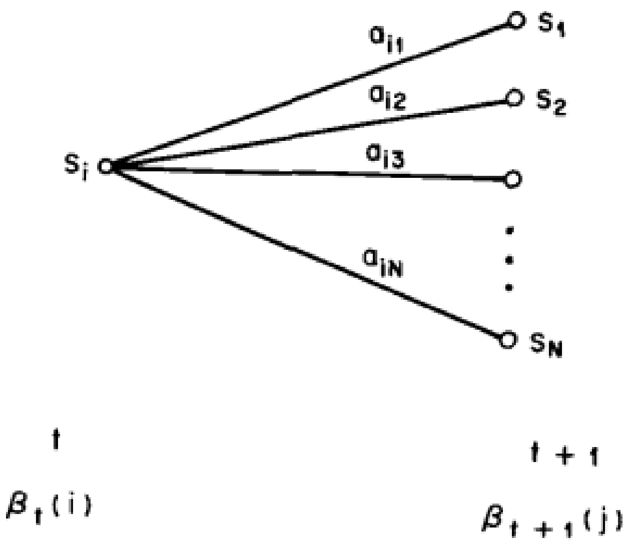
Backward Probability

$$p(S_t = k, \{O_t\}_{t=1}^{T}) = p(O_1, \ldots, O_t, S_t = k)p(O_{t+1}, \ldots, O_T | S_t = k) = \alpha_t^k \beta_t^k$$

Compute forward probability β_t^k recursively over t

$$\beta_t^k := p(O_{t+1}, \ldots, O_T | S_t = k)$$





Introduce S_{t+1}

Chain rule

Markov assumption

$$= \sum_i p(S_{t+1} = i | S_t = k) p(O_{t+1} | S_{t+1} = i) \beta_{t+1}^i$$

Forward-Backward Algorithm

Can compute β_t^k for all k, t using dynamic programming:

- Initialize: $\beta_T^k = 1$ for all k

- Iterate: for t = T-1, …, 1

$$\beta_t^k = \sum_i p(S_{t+1} = i | S_t = k) p(O_{t+1} | S_{t+1} = i) \beta_{t+1}^i$$ for all k

- Termination: $p(S_t = k, \{O_t\}_{t=1}^T) = \alpha_t^k \beta_t^k$

$$p(S_t = k | \{O_t\}_{t=1}^T) = \frac{p(S_t = k, \{O_t\}_{t=1}^T)}{p(\{O_t\}_{t=1}^T)} = \frac{\alpha_t^k \beta_t^k}{\sum_i \alpha_t^i \beta_t^i}$$

Most likely state & most likely sequence

- Most likely state assignment at time t

$$\arg\max_k p(S_t = k | \{O_t\}_{t=1}^T) = \arg\max_k \alpha_t^k \beta_t^k$$

E.g. Which die was most likely used by the casino in the third roll given the observed sequence?

- Most likely assignment of state sequence

$$\arg \max_{\{S_t\}_{t=1}^T} p(\{S_t\}_{t=1}^T | \{O_t\}_{t=1}^T)$$

E.g. What was the most likely sequence of die rolls used by the casino given the observed sequence?

Not the same solution !

MLA of x?
MLA of (x,y)?

x	y	P(x,y)
0	0	0.35
0	1	0.05
1	0	0.3
1	1	0.3

Next Lecture

- Decoding problem -2 (find most likely state sequence) – Viterbi algorithm
- One example problem on the board.

References

- Rabiner, Lawrence R. "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE* 77.2 (1989): 257-286.
- Machine Learning: An algorithm perspective by Stephen Marsland
- Slides reference: http://www.cs.cmu.edu/~aarti/Class/10701_Spring14/slides/HMM.pdf