香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 16

Hidden Markov Models

Junxian He
Nov 5, 2024

Announcements

- We have a makeup lecture this Thursday on Nov 7, 7pm-820pm, at Room 2303 after we finish HMM. Attendance is not required, zoom recording will be released

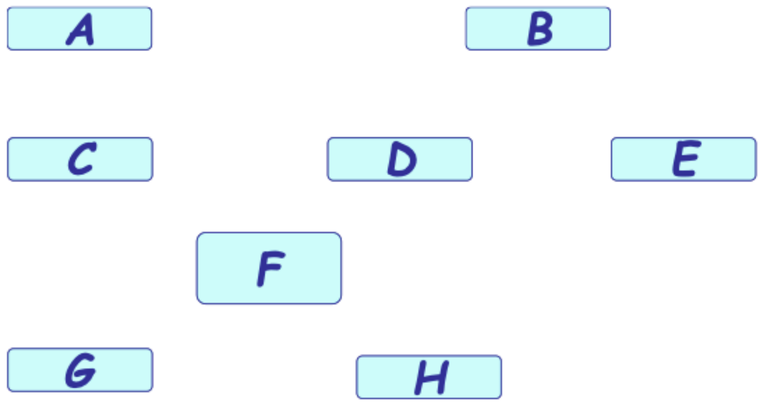
Recap: Probabilistic Graphical Models

Recap: Probabilistic Graphical Models

It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***

Recap: Probabilistic Graphical Models

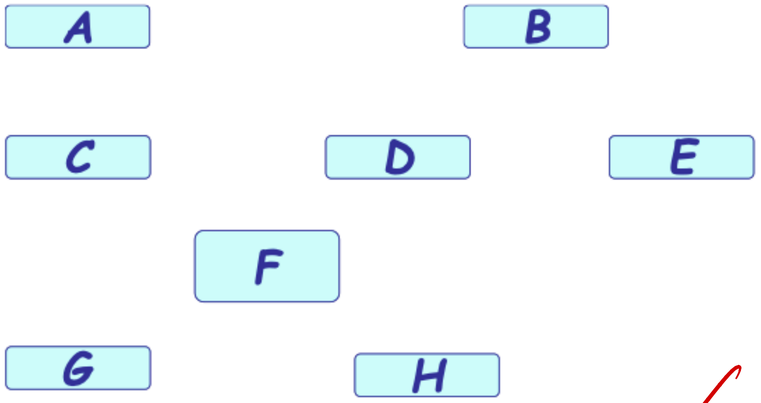
It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***

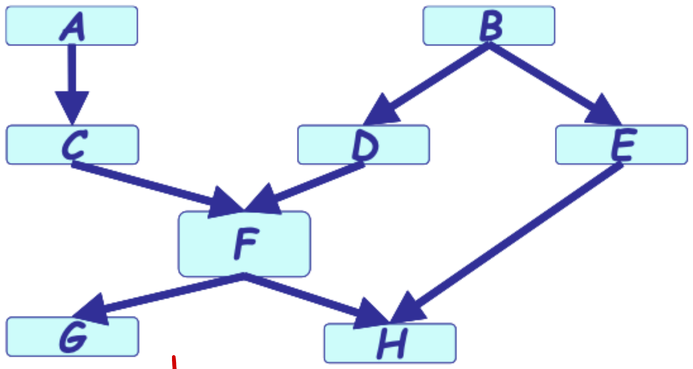


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

Recap: Probabilistic Graphical Models

It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***





$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

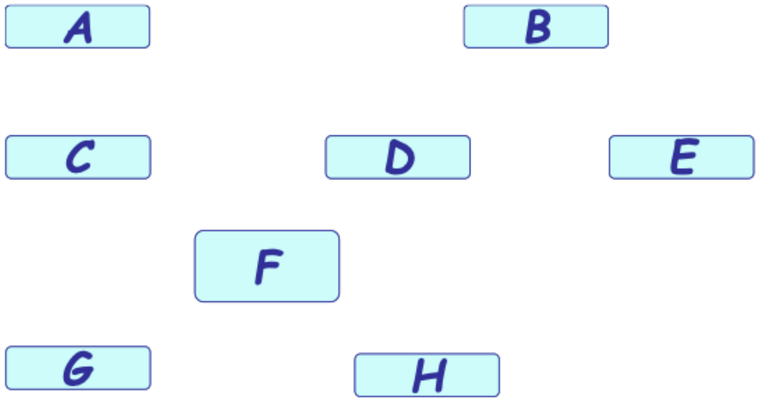
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

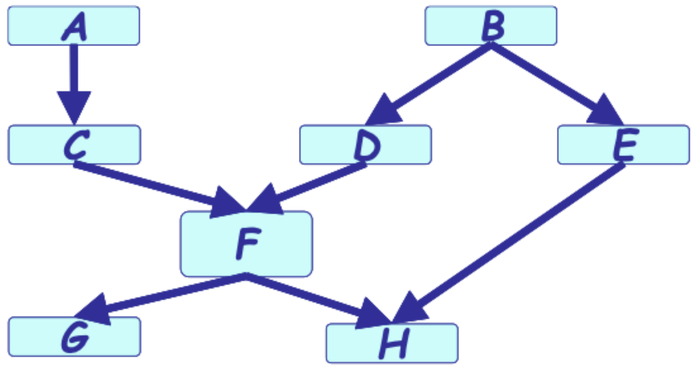
$O(K^8)$

$O(K^3)$

Recap: Probabilistic Graphical Models

It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***





$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

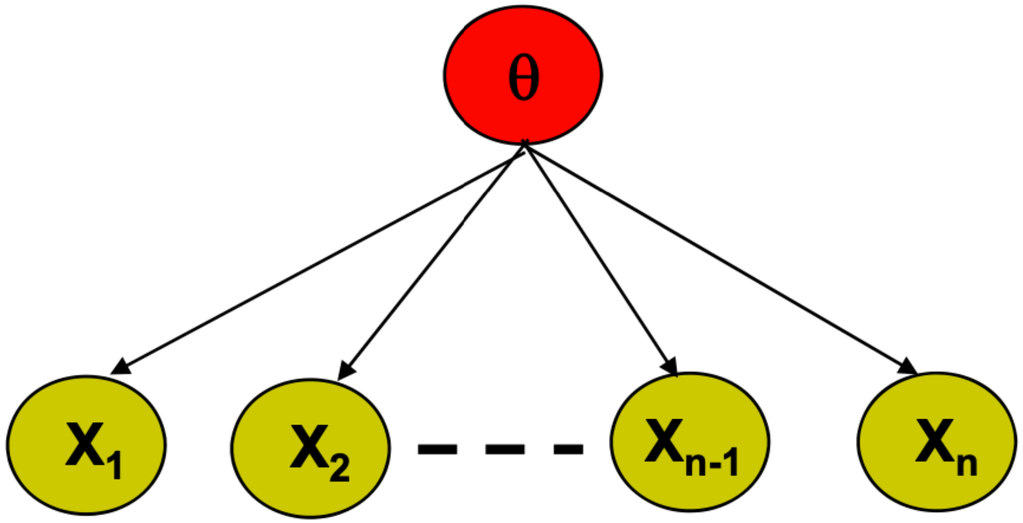
More formal definition:

It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables

Probabilistic Graphical Model is a graphical language to express conditional independence

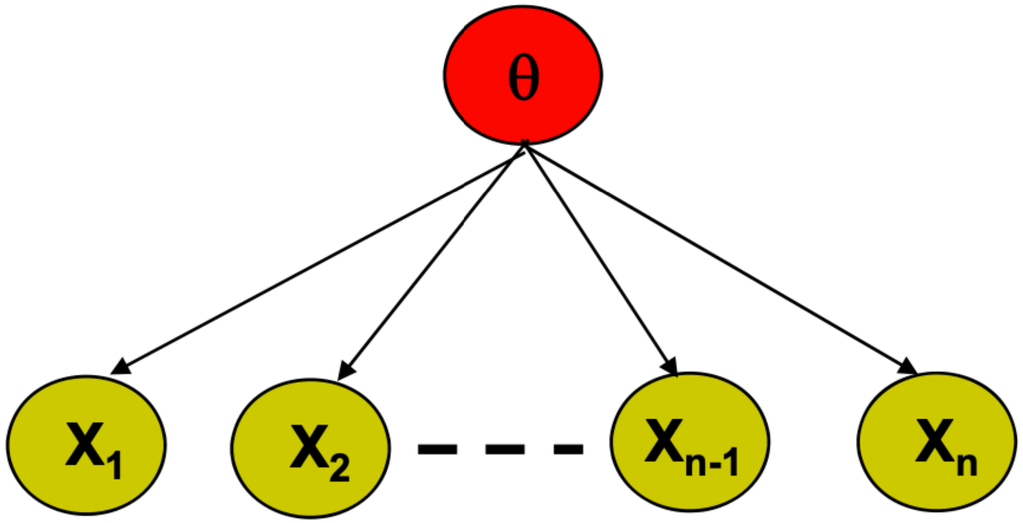
Conditionally Independent Observations

Conditionally Independent Observations



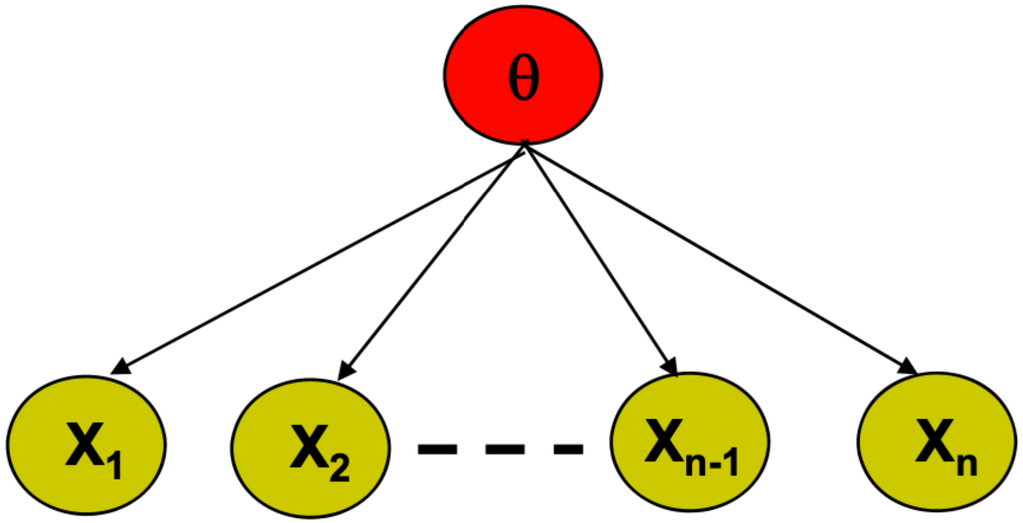
i.i.d

Conditionally Independent Observations



Model parameters

Conditionally Independent Observations

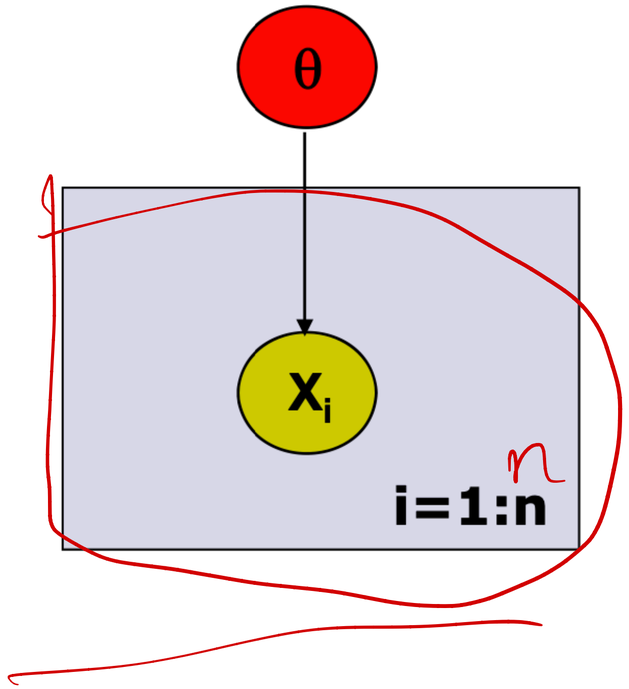


Model parameters

Data {X1, X2 …. Xn}

“Plate” Notation

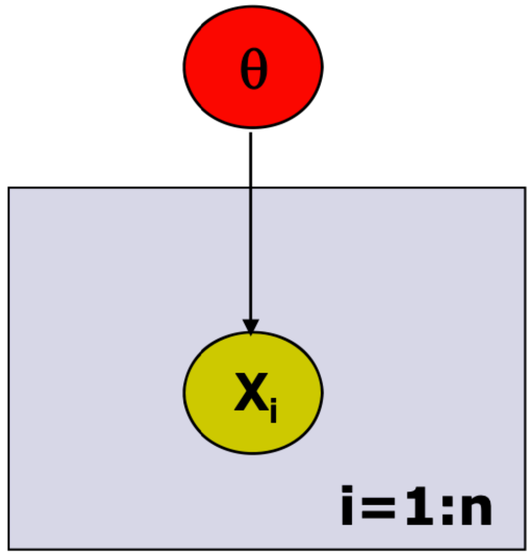
“Plate” Notation



Model parameters

Data = $\{x_1,\ldots x_n\}$

“Plate” Notation



Model parameters

Data = $\{x_1,\ldots x_n\}$

variables within a plate are replicated
in a conditionally independent manner

Example: Gaussian Model

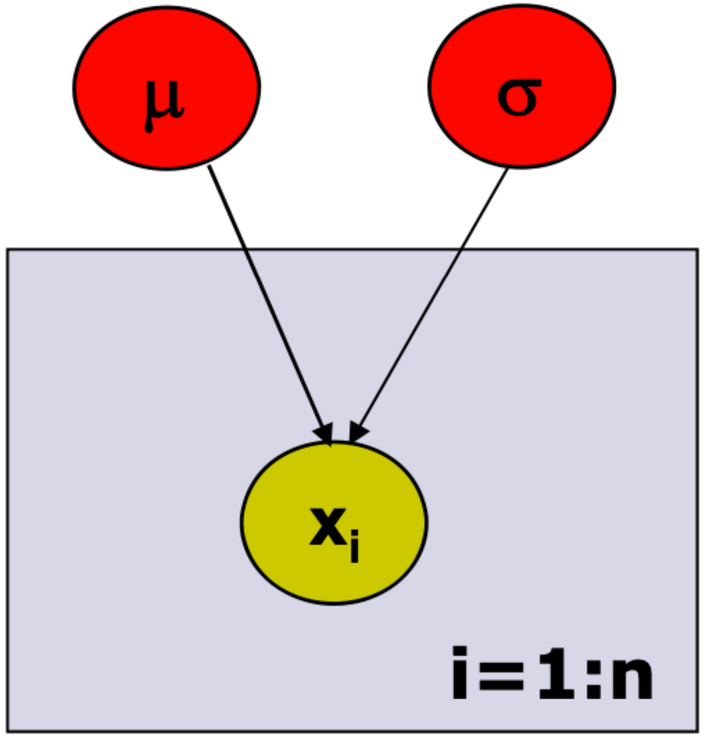
Generative model:

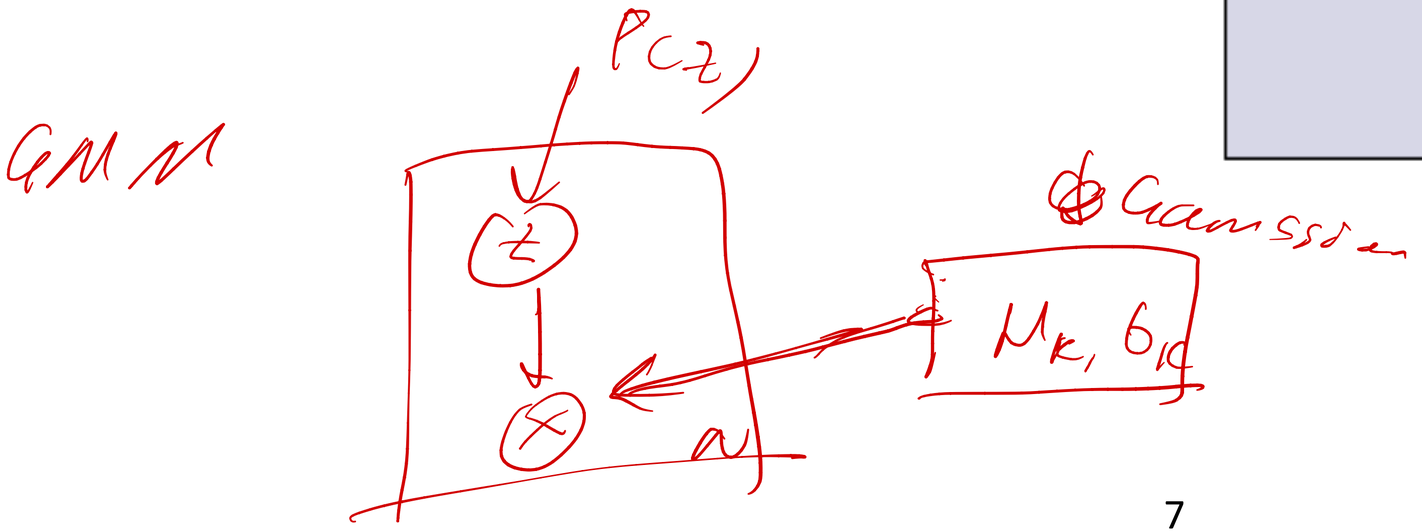
$p(x_1,\ldots x_n \mid \mu, \sigma) = \prod_i p(x_i \mid \mu, \sigma)$

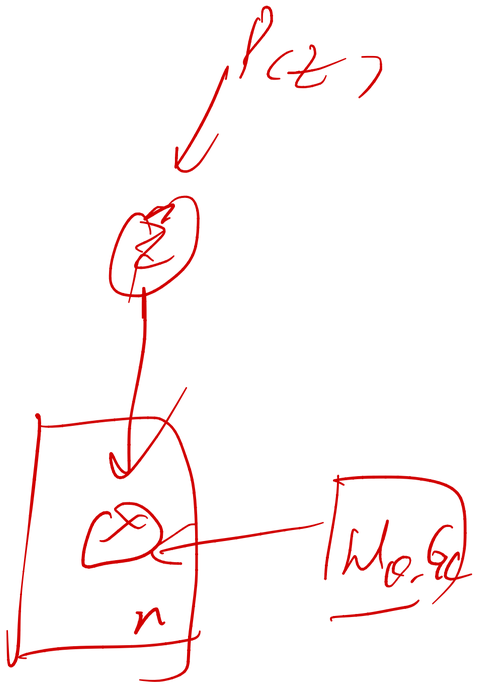
$= p(\text{data} \mid \text{parameters})$

$= p(D \mid \theta)$

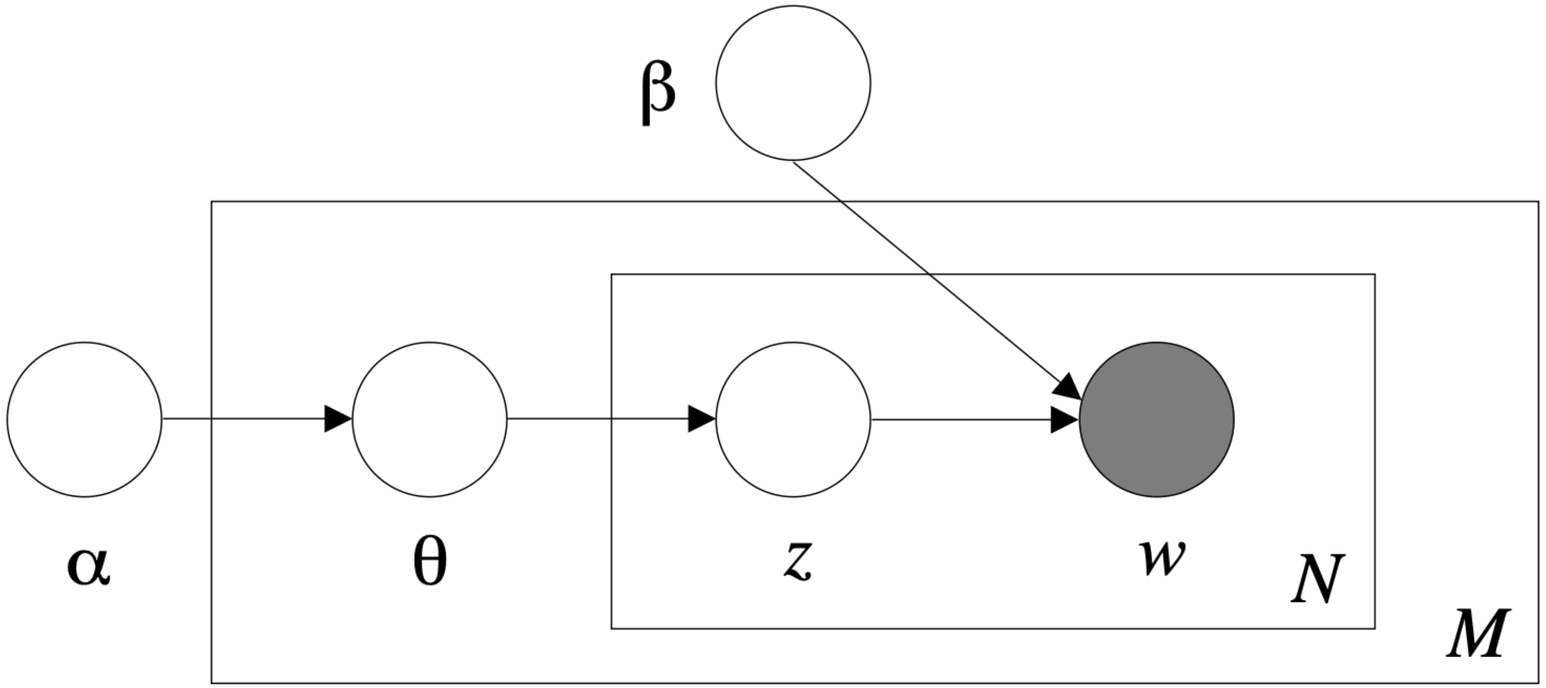
where $\theta = \{\mu, \sigma\}$



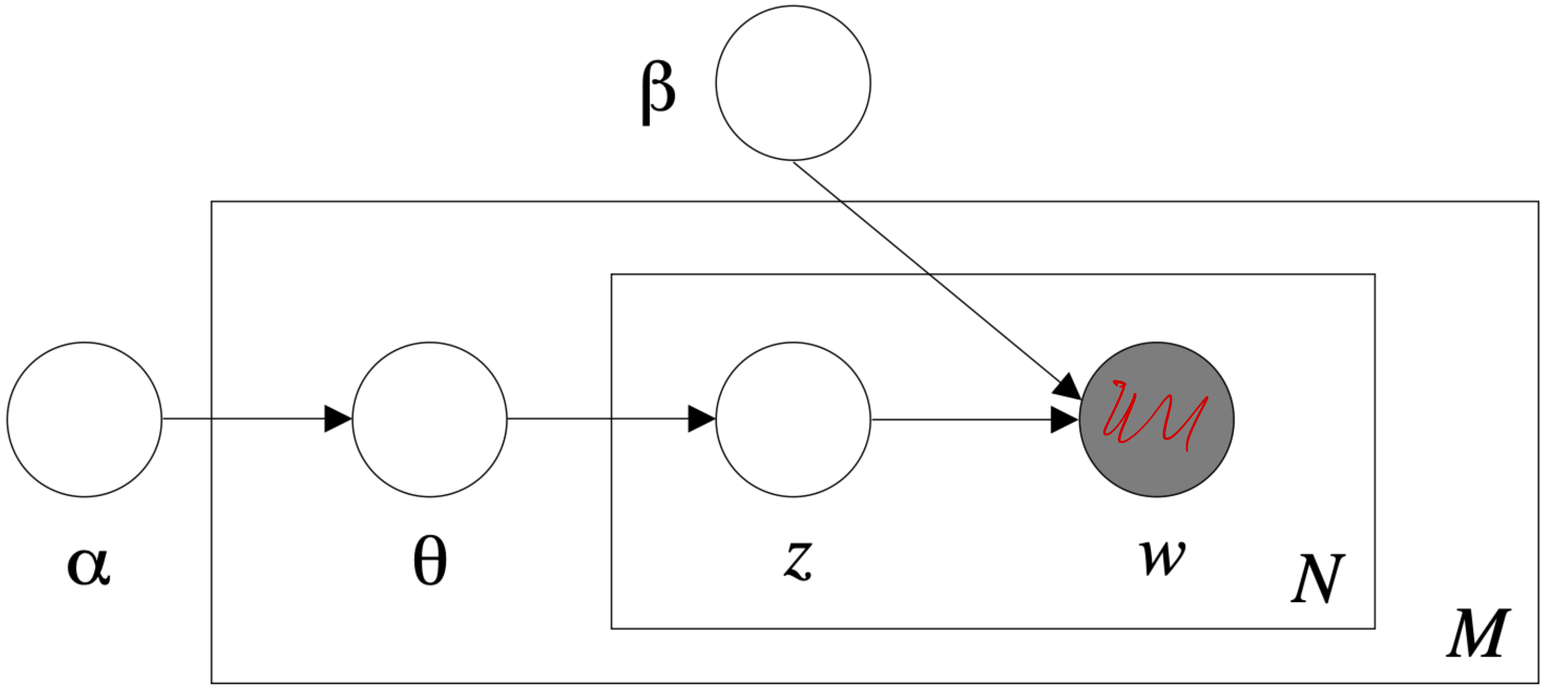




Observed Variable and Latent Variable Notations

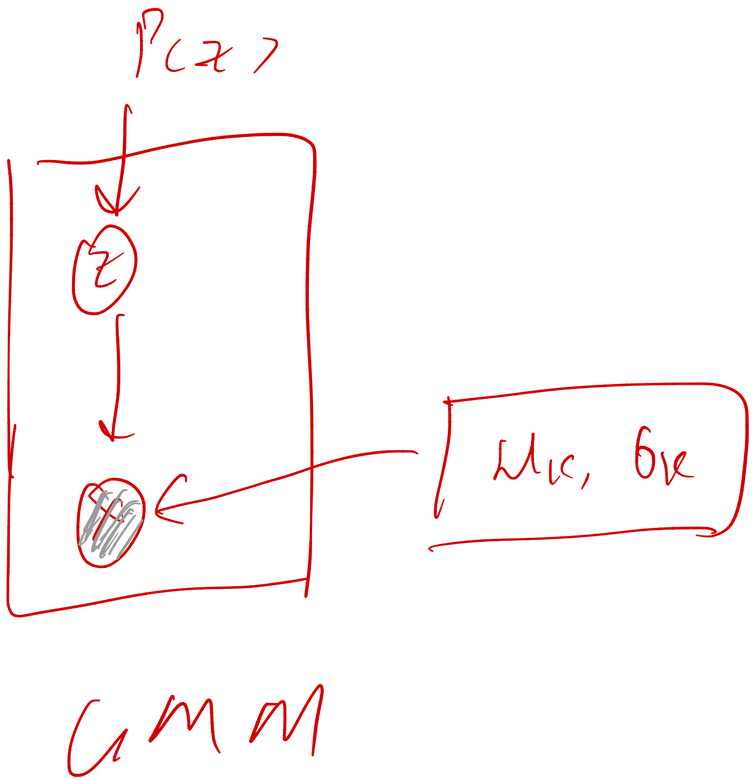


Observed Variable and Latent Variable Notations

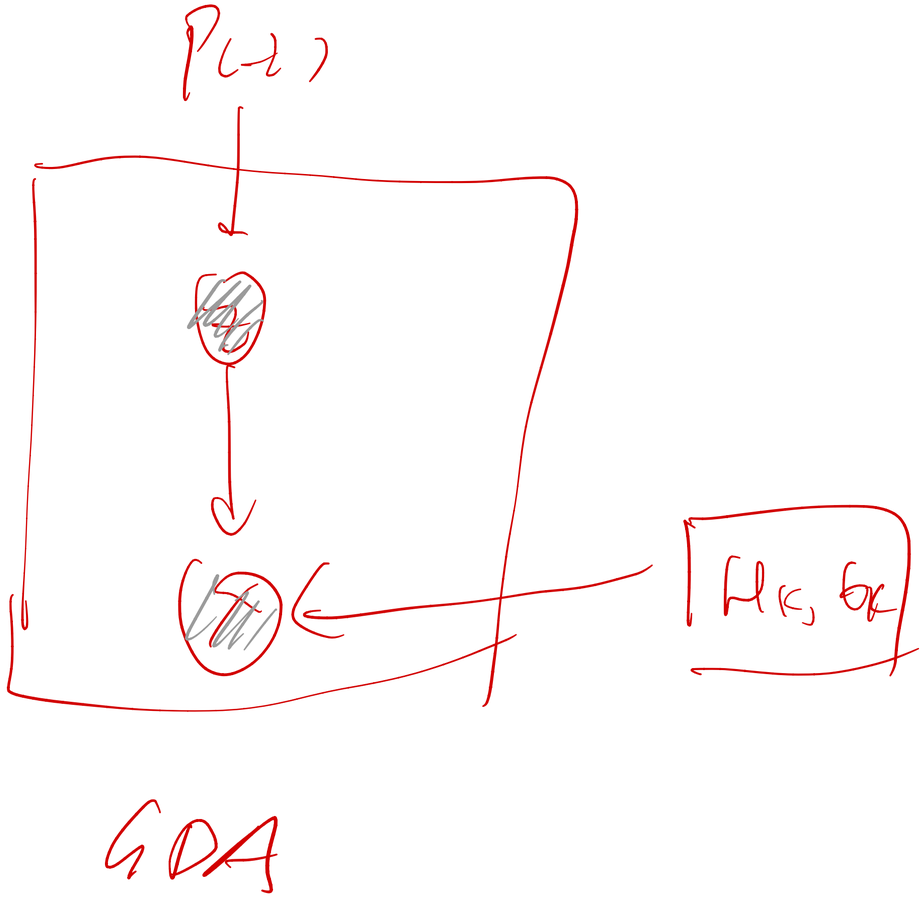


We typically use gray variables to denote observed variables

Gaussian Mixture Model / Gaussian Discriminative Analysis in PGMs



GMM



GDA

Inference and Learning

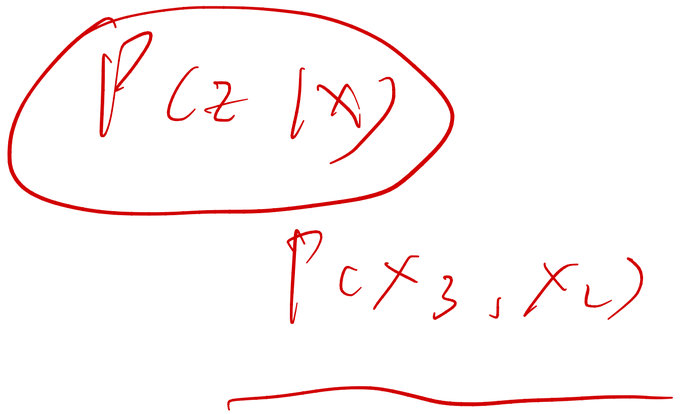
Inference and Learning

- Task 1: How do we answer **queries** about P?
  - We use **inference** as a name for the process of computing answers to such queries

Inference and Learning

Query a node (random variable) in the graph

- Task 1: How do we answer **queries** about P?
  - We use **inference** as a name for the process of computing answers to such queries



Inference and Learning

Query a node (random variable) in the graph

- Task 1: How do we answer **queries** about P?
  - We use **inference** as a name for the process of computing answers to such queries

- Task 2: How do we estimate a **plausible model** M from data D?

i. We use **learning** as a name for the process of obtaining point estimate of M.

Examples

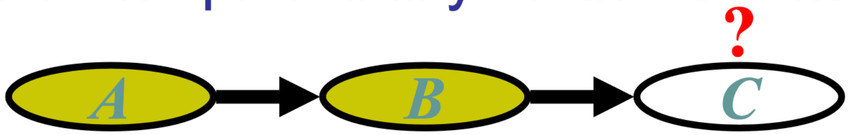
Examples

- **Prediction**: what is the probability of an outcome given the starting condition

  - the query node is a descendent of the evidence

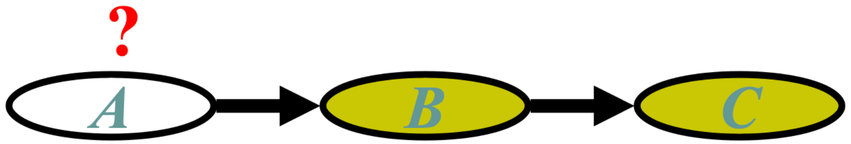
Examples

- **Prediction**: what is the probability of an outcome given the starting condition



  - the query node is a descendent of the evidence

- **Diagnosis**: what is the probability of disease/fault given symptoms

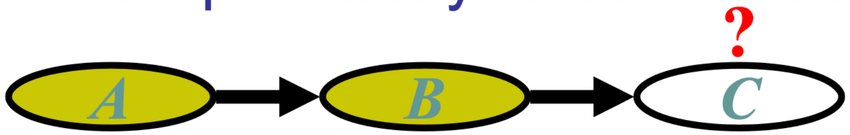


  - the query node an ancestor of the evidence

inference

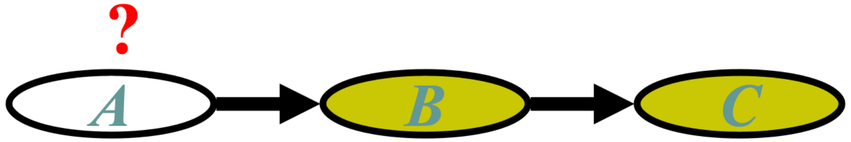
Examples

- **Prediction**: what is the probability of an outcome given the starting condition



  - the query node is a descendent of the evidence

- **Diagnosis**: what is the probability of disease/fault given symptoms



  - the query node an ancestor of the evidence

In practice, the observed variable is often the data that is on the leaf nodes

How to Learn the Parameters

How to Learn the Parameters

1. When θ is the parameter and does not have prior —> MLE

$$p(x, z; \theta)$$

How to Learn the Parameters

1. When θ is the parameter and does not have prior —> MLE

$$p(x, z; \theta)$$

2. When we add the prior over θ —> MAP (Bayesian)

$$p(x, z, \theta)$$

$P(\theta)$

$\arg\max_{\theta} P(\theta \mid x)$

$= \arg\max_{\theta} P(x \mid \theta) P(\theta)$

How to do MLE on Latent Variable Models?

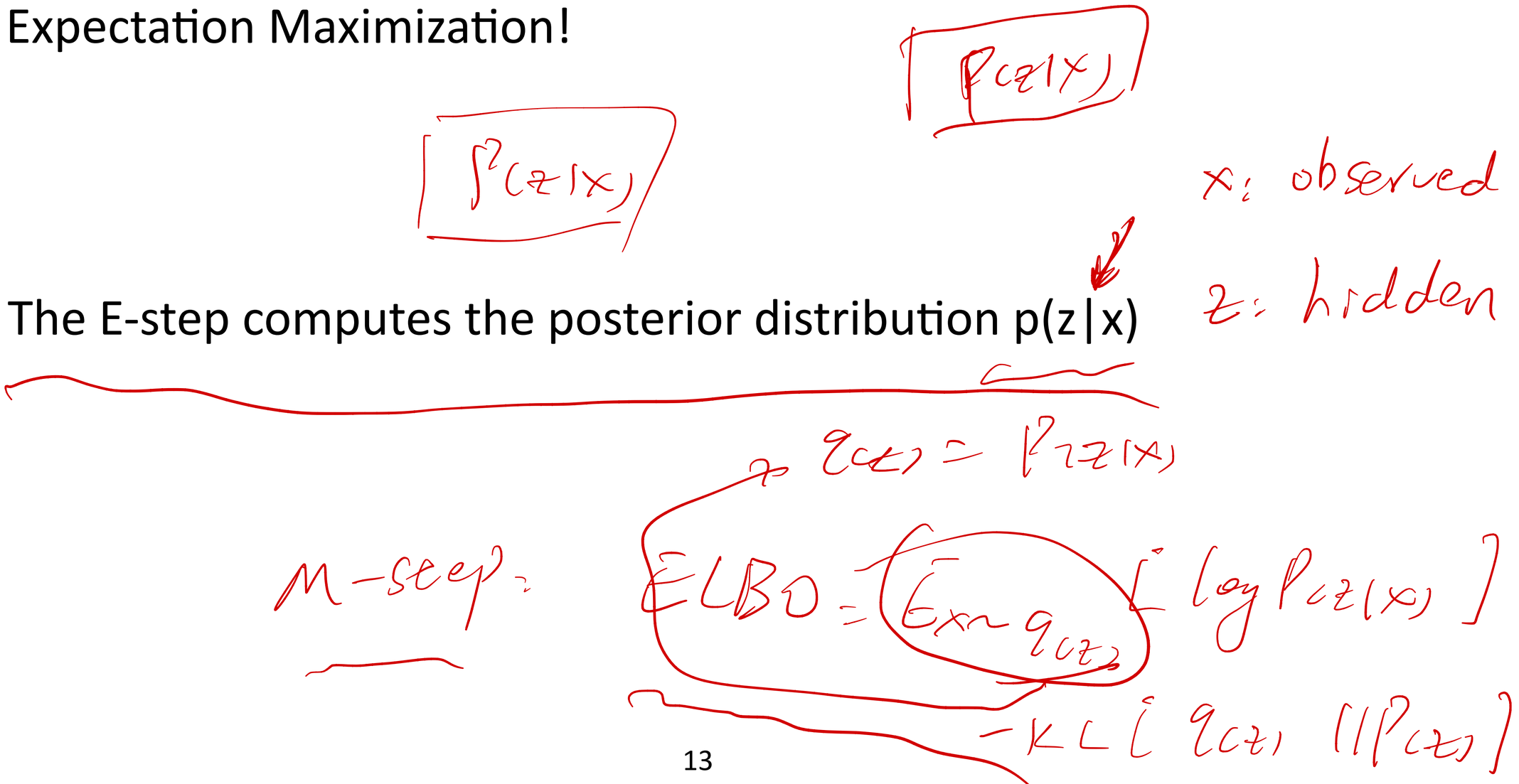
How to do MLE on Latent Variable Models?

Expectation Maximization!

How to do MLE on Latent Variable Models?

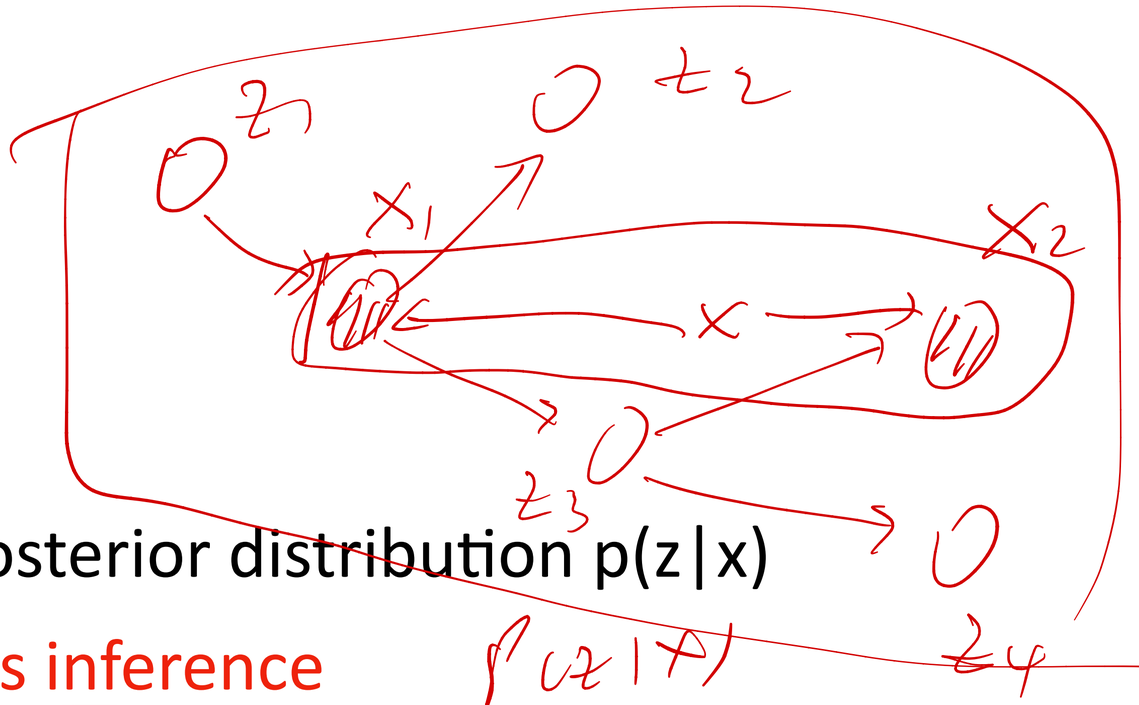
Expectation Maximization!

The E-step computes the posterior distribution p(z|x)



How to do MLE on Latent Variable Models?

Expectation Maximization!



The E-step computes the posterior distribution p(z|x)

This process is referred to as inference

$p(z|x)$

x : data

$p(z_1, z_2, z_3, z_4 | x_1, x_2)$

Approaches to Inference

Approaches to Inference

- Exact inference algorithms

  $P(Z|X)$

  - The elimination algorithm
  - Belief propagation
  - The junction tree algorithms (but will not cover in detail here)

Approaches to Inference

- Exact inference algorithms
  - The elimination algorithm
  - Belief propagation
  - The junction tree algorithms (but will not cover in detail here)
- Approximate inference techniques
  - Variational algorithms
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods

Approaches to Inference

- Exact inference algorithms
  - The elimination algorithm
  - Belief propagation
  - The junction tree algorithms (but will not cover in detail here)

- Approximate inference techniques
  - Variational algorithms
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods

Variational Autoencoders

Elimination Algorithm/ Marginalization

$$P(h) = \sum_{g}\sum_{f}\sum_{e}\sum_{d}\sum_{c}\sum_{b}\sum_{a} P(a,b,c,d,e,f,g,h)$$

a naïve summation needs to enumerate over an exponential number of terms

$P(h)$

$P(z|x)$

a, b, c, d, x

a, b, c, d : latent

x : data

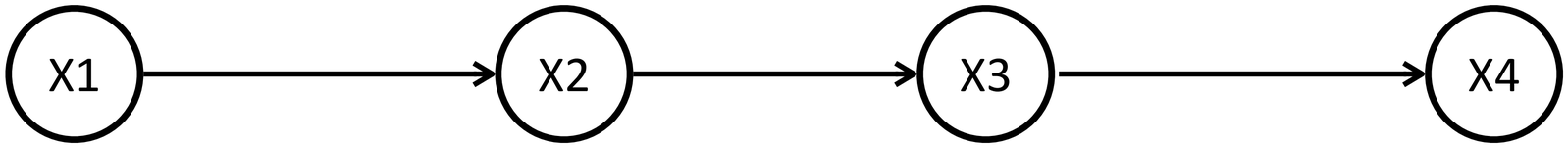
P

$$P(a|x) = \sum_{b}\sum_{c}\sum_{d} P(a,b,c,d|x)$$

Elimination Algorithm/ Marginalization

$$P(h) = \sum_{g} \sum_{f} \sum_{e} \sum_{d} \sum_{c} \sum_{b} \sum_{a} P(a,b,c,d,e,f,g,h)$$

a naïve summation needs to enumerate over an exponential number of terms



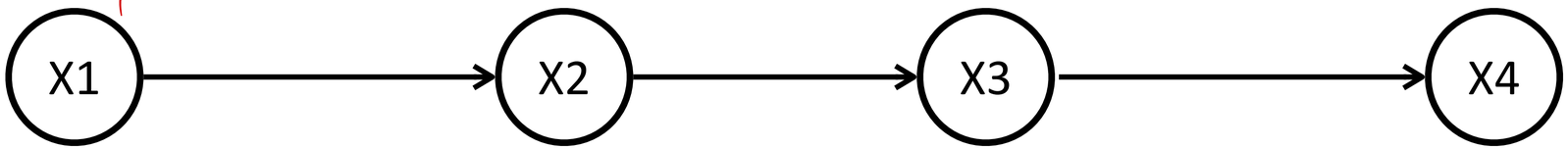
Elimination Algorithm/ Marginalization

$$P(h) = \sum_{g}\sum_{f}\sum_{e}\sum_{d}\sum_{c}\sum_{b}\sum_{a} P(a,b,c,d,e,f,g,h)$$

$O(K^3)$

a naïve summation needs to enumerate over an exponential number of terms

$\sum_{x_2}\sum_{x_3}\sum_{x_4} f(x_1, x_2, x_3, x_4) = f(x_1)$

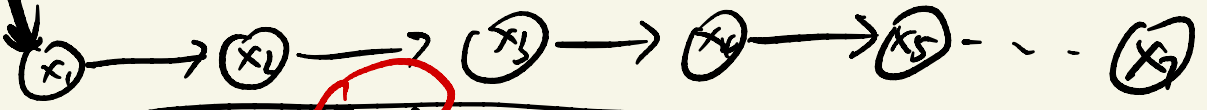


What if the random variables follow this chain structure?

$$\sum_{x_1,\dots x_7} P(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = P(x_1)$$

$O(k^7)$

$P(x_1)$



$$P(x_1, \dots x_7) = P(x_1) P(x_2|x_1) P(x_3|x_2) P(x_4|x_3) P(x_5|x_4) \dots P(x_6|x_5) P(x_7|x_6)$$

$$P(x_7) = \sum_{x_1} \cdots \sum_{x_6} f(x_5, x_4) f(x_2) P(x_3|x_2) P(x_4|x_3) \cdots P(x_7|x_6)$$

$$= \sum_{x_2} \cdots \sum_{x_6} P(x_3|x_2) P(x_4|x_3) \cdots P(x_7|x_6) \sum_{x_1} P(x_1) P(x_2|x_1)$$

$$= \sum_{x_2} \sum_{x_3} \cdots \sum_{x_6} P(x_3|x_1) P(x_4|x_3) \cdots f(x_2)$$

$$= \sum_{x_3} \cdots \sum_{x_6} P(x_4|x_3) P(x_5|x_4) \sum_{x_2} f(x_2) P(x_3|x_2)$$

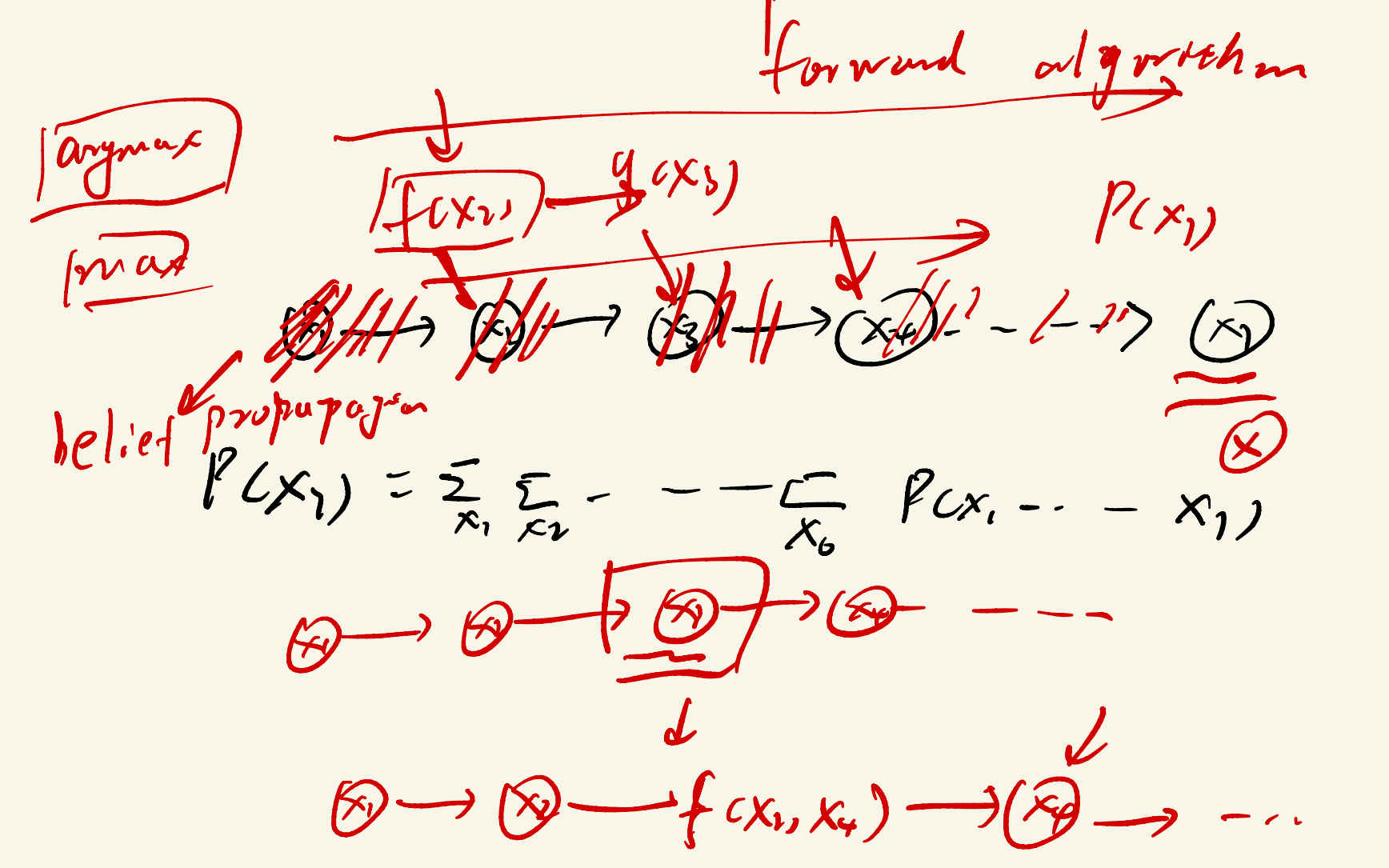
$g(x_3)$ — for each val of x_3, $O(K)$

Elimination

$O(K^2)$ $\times N$

$O(N \cdot K^2)$

$O(K^N)$

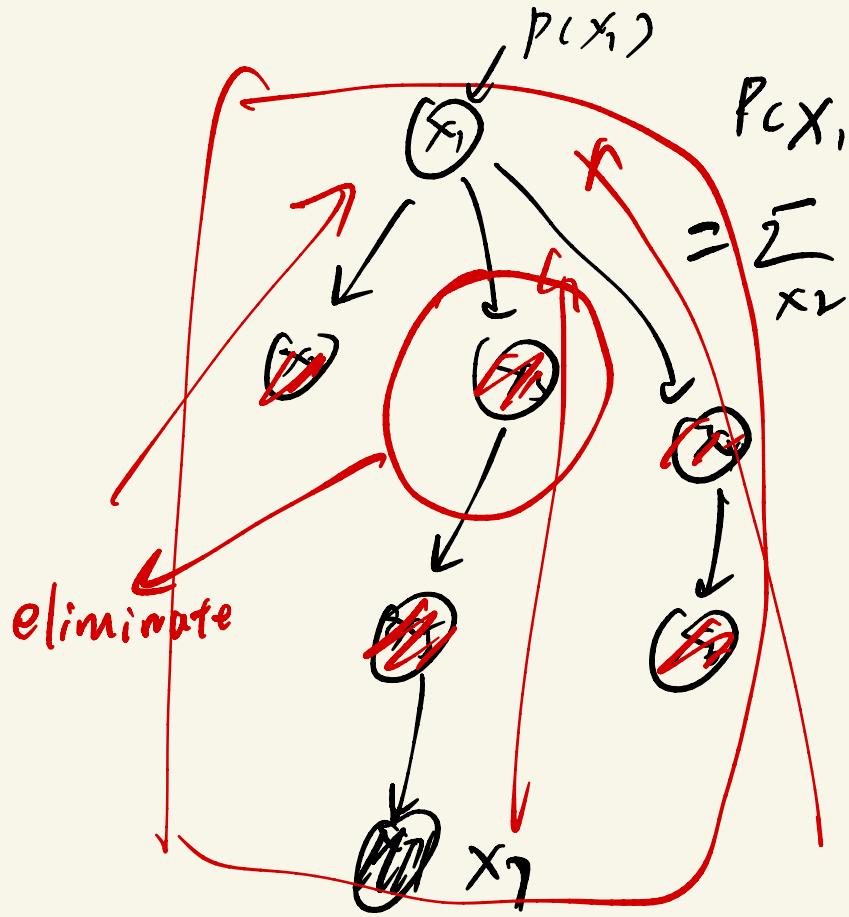

forward algorithm
argmax
max
$f(x_2)$
$g(x_3)$
$p(x_1)$
belief propagation
$p(x_1) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_6} p(x_1 \cdots x_6)$
$f(x_1, x_4)$

$\underset{x_1}{\arg\max}\ \underset{x_2}{\arg\max}\ \underset{x_3}{\arg\max} \cdots P(x_1, x_2 \cdots x_7)$

$P(x_1)\,P(x_2|x_1) \cdots$

$\underset{x_2}{\arg\max}\ \underset{x_3}{\arg\max} \cdots P(x_3|x_2) \cdots \left[\underset{x_1}{\arg\max}\ P(x_1)\,P(x_2|x_1)\right]$

$f(x_2)$



$$P(x_1 | x_7) = \sum_{x_2} \sum_{x_3} \cdots \sum_{x_6} P(x_1, x_2, \ldots, x_6 | x_7)$$

$P(x_5|x_1)\ P(x_6|x_3)$

inside algorithm

Hidden Markov Models

i.i.d to sequential data

i.i.d to sequential data

- So far we assumed independent, identically distributed data

$$\{X_i\}_{i=1}^{n} \overset{iid}{\sim} p(\mathbf{X})$$

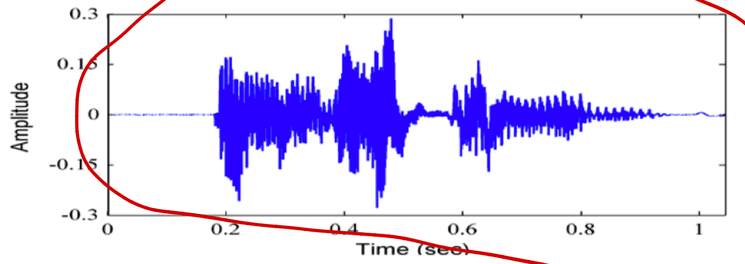
i.i.d to sequential data

- So far we assumed independent, identically distributed data

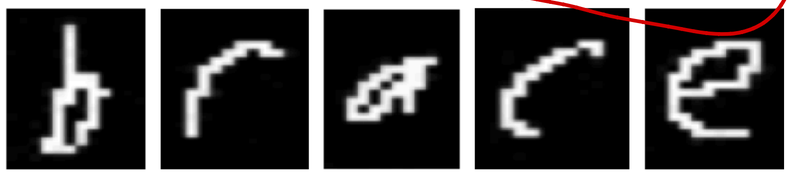
$$\{X_i\}_{i=1}^{n} \overset{iid}{\sim} p(\mathbf{X})$$

- Sequential (non i.i.d.) data
  - Time-series data

    E.g. Speech



  - Characters in a sentence



  - Base pairs along a DNA strand

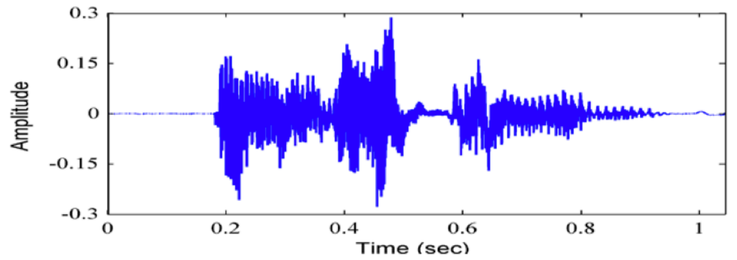
i.i.d to sequential data

- So far we assumed independent, identically distributed data

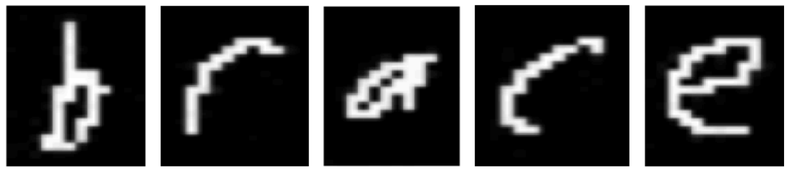
$$\{X_i\}_{i=1}^{n} \overset{iid}{\sim} p(\mathbf{X})$$

- Sequential (non i.i.d.) data
  - Time-series data

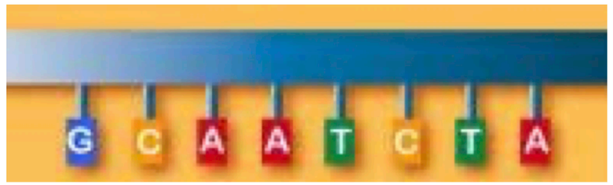
    E.g. Speech



  - Characters in a sentence



  - Base pairs along a DNA strand



(Sequential data is still i.i.d on the sequence level)

Markov Models

Markov Models

- Joint distribution of *n* arbitrary random variables

$$
\begin{aligned}
p(\mathbf{X}) &= p(X_1, X_2, \ldots, X_n) \\
&= p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)\ldots p(X_n|X_{n-1}, ..., X_1) \\
&= \prod_{i=1}^{n} p(X_n|X_{n-1}, \ldots, X_1) \qquad \text{Chain rule}
\end{aligned}
$$

naive Bayes

Markov Models

- Joint distribution of *n* arbitrary random variables

$$
\begin{aligned}
p(\mathbf{X}) &= p(X_1, X_2, \ldots, X_n) \\
&= p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)\ldots p(X_n|X_{n-1}, ..., X_1) \\
&= \prod_{i=1}^{n} p(X_n|X_{n-1}, \ldots, X_1) \qquad \text{Chain rule}
\end{aligned}
$$

- Markov Assumption (m^{th} order)

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

m

Current observation only depends on past m observations

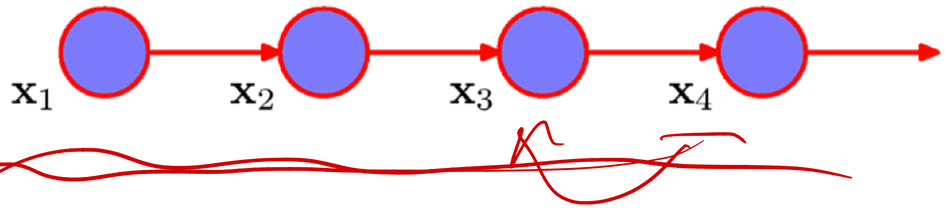
Markov Models

Markov Models

- Markov Assumption

1st order

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1})$$

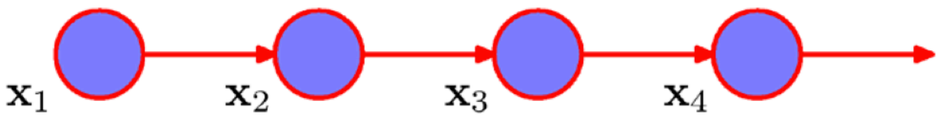


Markov Models

- Markov Assumption

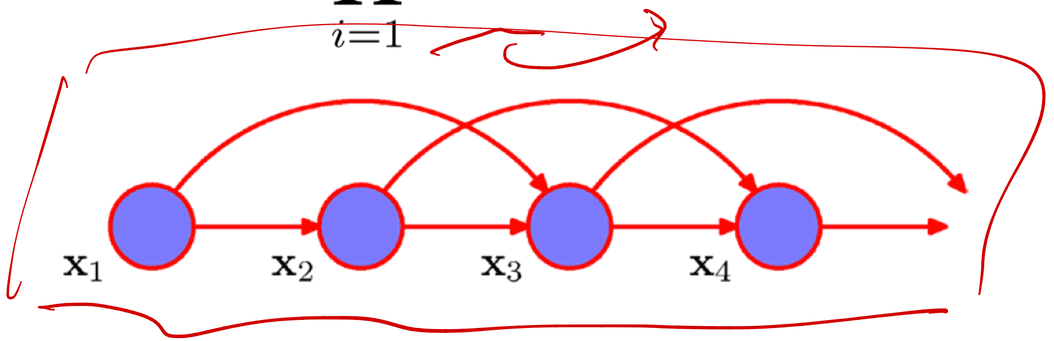
1st order

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1})$$



2nd order

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



$P(X_2 = 2 \mid X_1 = 1)$

$\|$

$P(X_4 = 2 \mid X_3 = 1)$

Markov Models

Markov Models

Homogeneous/stationary Markov model (probabilities don't depend on n)

Markov Models

Homogeneous/stationary Markov model (probabilities don't depend on n)

- Markov Assumption

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

parameters in stationary model
K-ary variables

$O(K^2)$

$P(X_n|X_{n-1}) = \begin{bmatrix} & \\ & \end{bmatrix}_{K \times K}$

Markov Models

Homogeneous/stationary Markov model (probabilities don't depend on n)

parameters in stationary model
K-ary variables

- Markov Assumption

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

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

Markov Models

Homogeneous/stationary Markov model (probabilities don't depend on n)

parameters in stationary model K-ary variables

- Markov Assumption

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

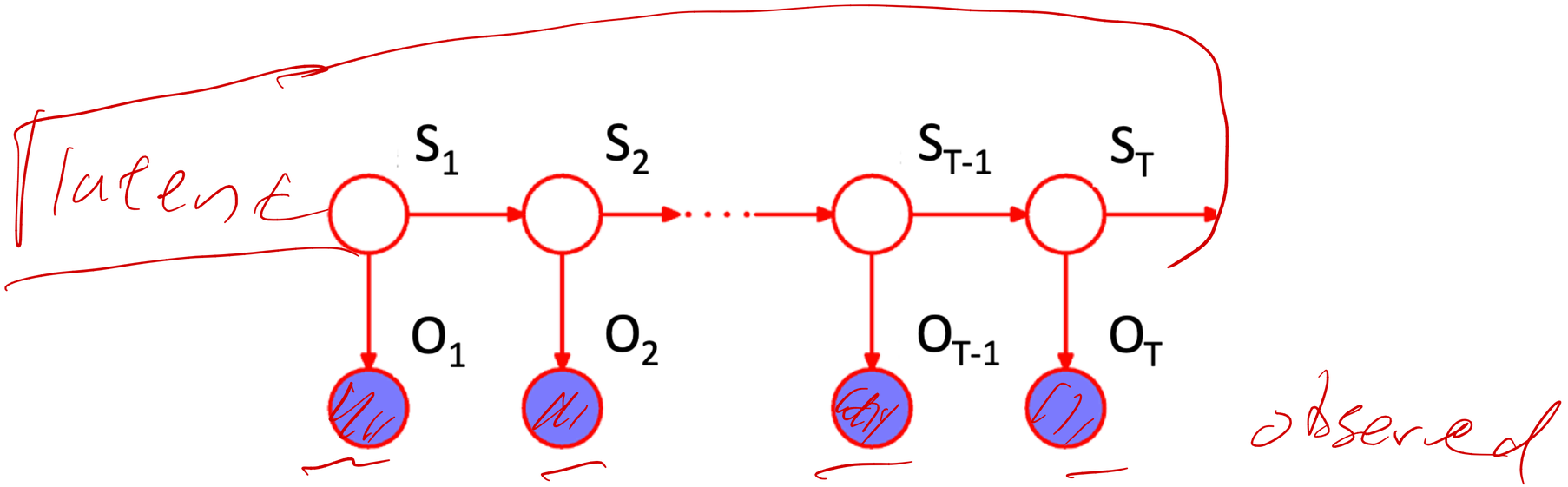
mth order $p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1},\ldots,X_{n-m})$ $O(K^{m+1})$

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

$O(cK^n)$

≡ no assumptions – complete (but directed) graph

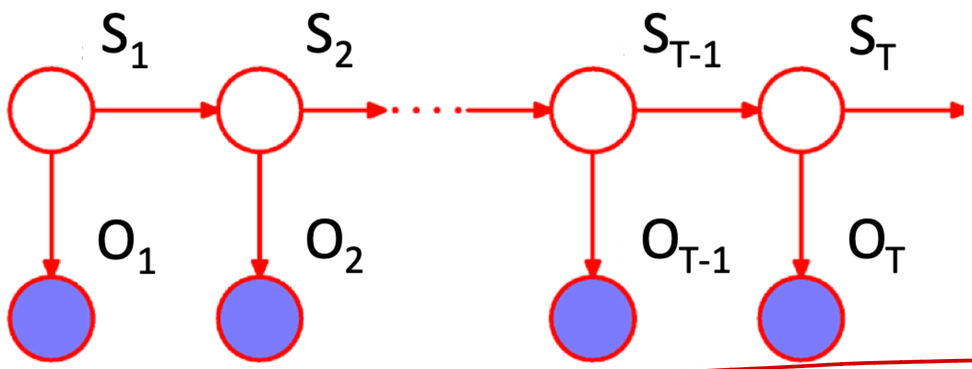
Hidden Markov Models



Observation space	$O_t \in \{y_1, y_2, \ldots, y_K\}$
Hidden states	$S_t \in \{1, \ldots, I\}$

Hidden Markov Models

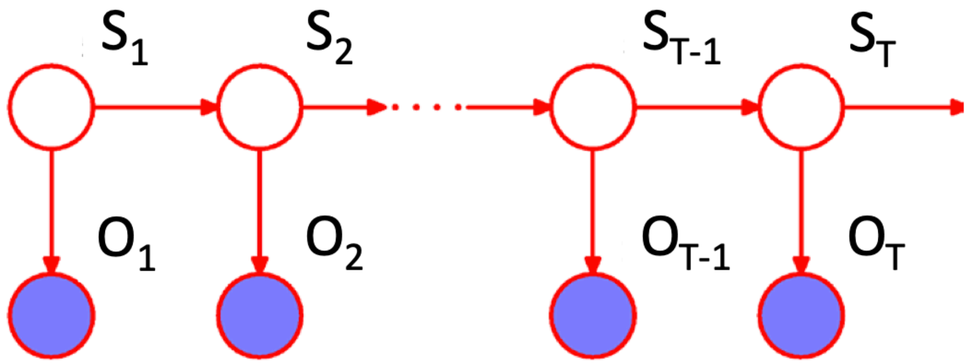
Hidden Markov Models



$$p(S_1, \ldots, S_T, O_1, \ldots, O_T) = \prod_{t=1}^{T} p(O_t|S_t) \prod_{t=1}^{T} p(S_t|S_{t-1})$$

Markov model

Hidden Markov Models

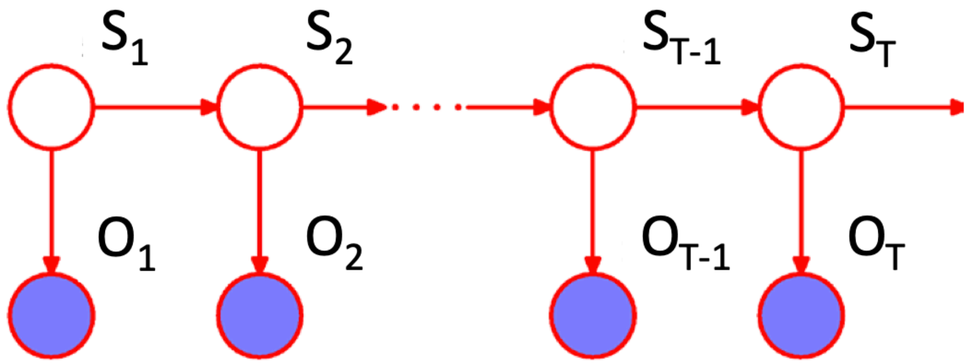


$$p(S_1, \ldots, S_T, O_1, \ldots, O_T) = \prod_{t=1}^{T} p(O_t|S_t) \prod_{t=1}^{T} p(S_t|S_{t-1})$$

1st order Markov assumption on hidden states $\{S_t\}$ t = 1, …, T

(can be extended to higher order).

Hidden Markov Models



$$p(S_1, \ldots, S_T, O_1, \ldots, O_T) = \prod_{t=1}^{T} p(O_t|S_t) \prod_{t=1}^{T} p(S_t|S_{t-1})$$

1st order Markov assumption on hidden states $\{S_t\}$ t = 1, …, T

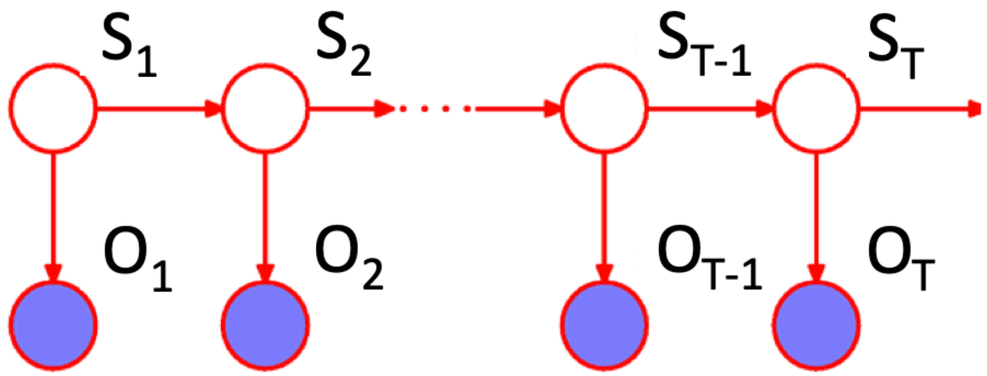
(can be extended to higher order).

Bayes Ball

Is O_T and O_2 independent?

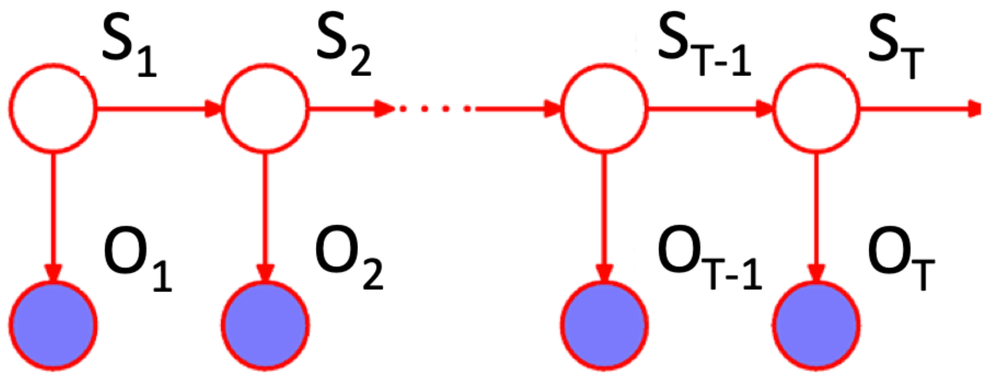
$O_T \perp O_2$ given S_{T-1}

Hidden Markov Models



Hidden Markov Models

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

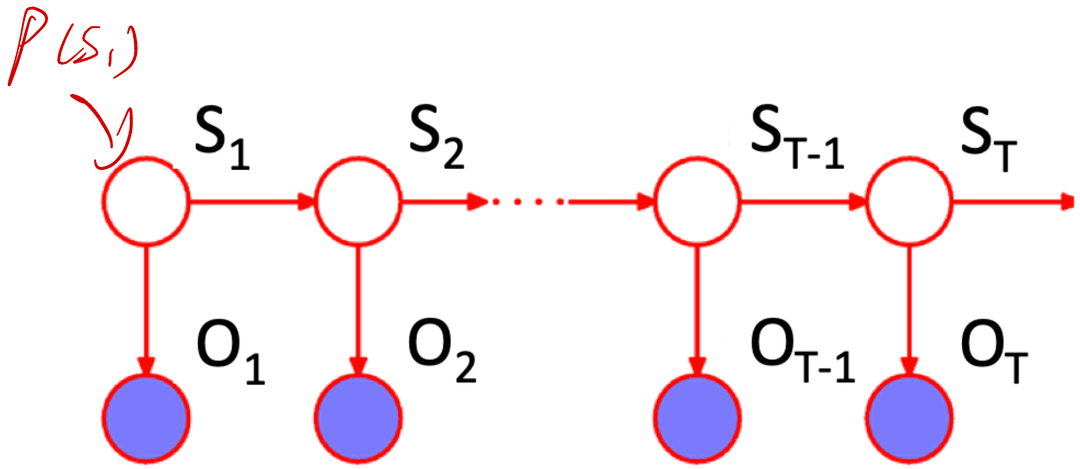


Hidden Markov Models

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

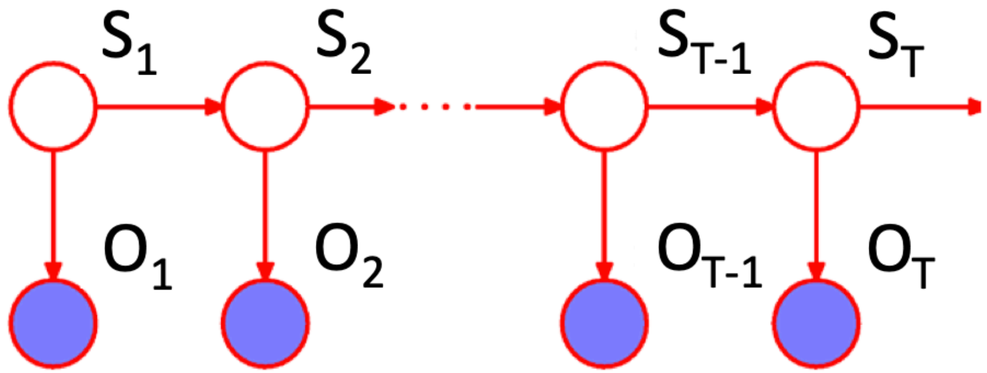
Initial probabilities

$$p(S_1 = i) = \pi_i$$



Hidden Markov Models

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



Initial probabilities

$$p(S_1 = i) = \pi_i$$

Transition probabilities

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

Hidden Markov Models

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

Initial probabilities

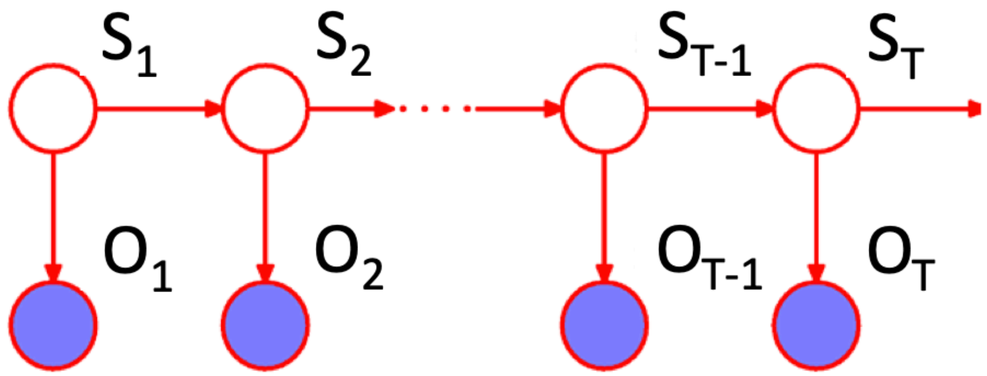
$$p(S_1 = i) = \pi_i$$

Transition probabilities

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

Emission probabilities

$$p(O_t = y \mid S_t = i) = q_i^y$$



Hidden Markov Models

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

Initial probabilities

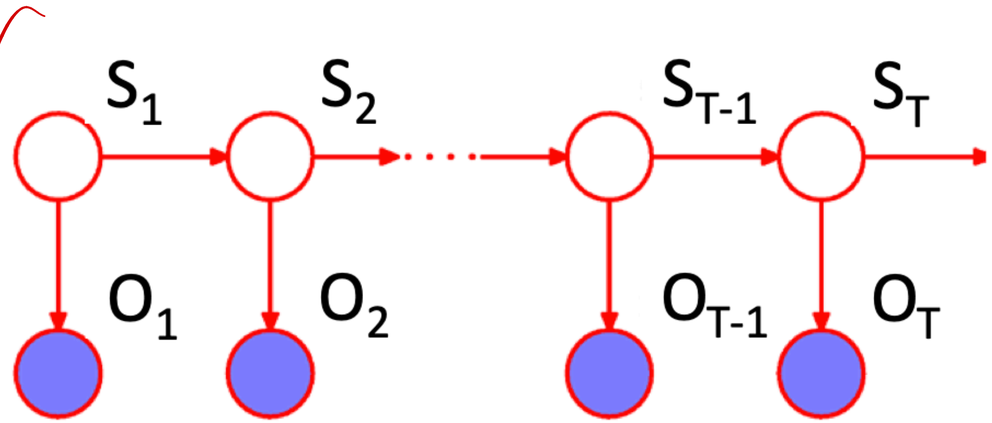
$p(S_1 = i) = \pi_i$

Transition probabilities

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

Emission probabilities

$p(O_t = y \mid S_t = i) = q_i^y$



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

HMM 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

emission

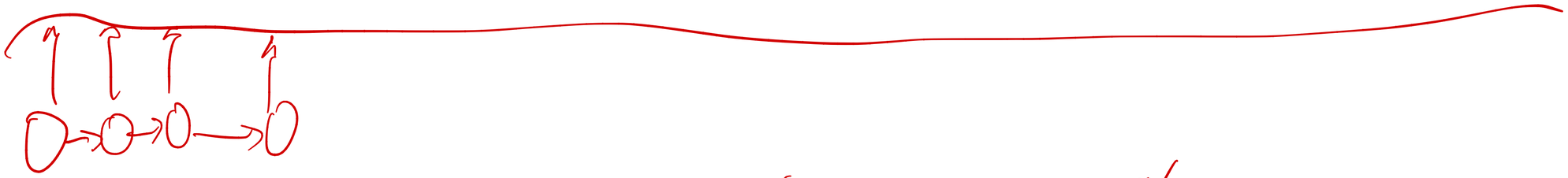
transition

HMM Example

HMM Example

GIVEN: A sequence of rolls by the casino player

12455264621461461361366616646616366163661636165156151151461235623 44



$P(S_t = \text{fair} \mid S_{t-1} = \text{loaded}) = 5\%$

HMM Example

GIVEN: A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

Question

1. How likely is the sequence given our model?
   This is the evaluation problem in HMMs

HMM Example

GIVEN: A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

Question

1. How likely is the sequence given our model?
   This is the evaluation problem in HMMs

2. 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

HMM Example

GIVEN: A sequence of rolls by the casino player

12455264621461461361366616646616366163661636165156151151461235623 44

Question

1. How likely is the sequence given our model?
   This is the evaluation problem in HMMs

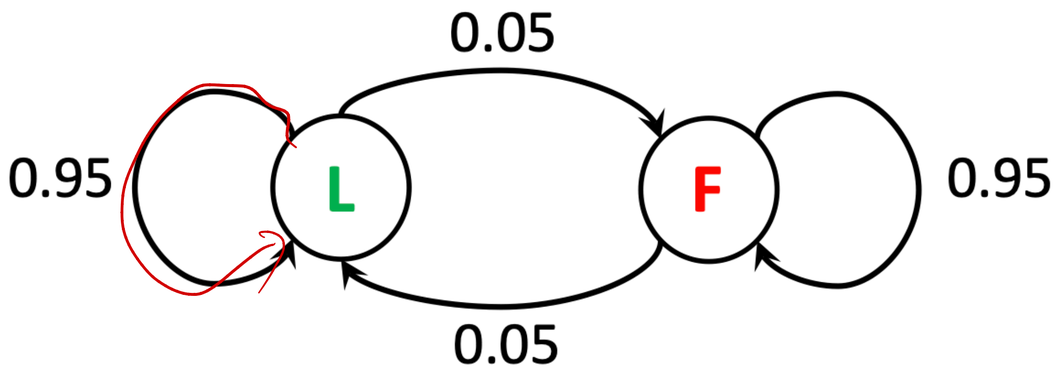
2. What portion of the sequence was generated with the fair die, and what portion with the loaded die
   This is the decoding question in HMMs

3.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

State Space Representation

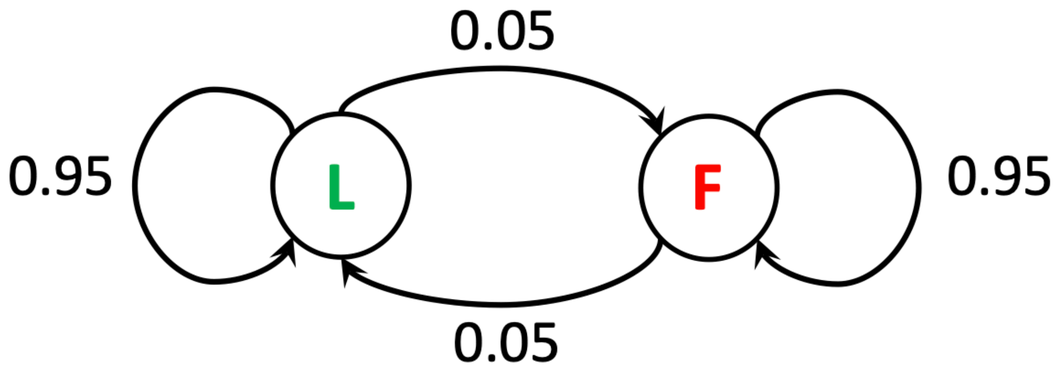
Fair Loaded

- Switch between **F** and **L** with 5% probability



State Space Representation

- 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	$P(O_t = y \mid S_t = F) = 1/6$ $\quad y = 1,2,3,4,5,6$
	$P(O_t = y \mid S_t = L) = 1/10$ $\quad y = 1,2,3,4,5$
	$= 1/2$ $\quad y = 6$

learn

Three Main Problems in HMMs

Three Main Problems in HMMs

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

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

Three Main Problems in HMMs

- 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 $\arg\max_{s_1,\ldots,s_T} p(\{S_t\}_{t=1}^T|\{O_t\}_{t=1}^T,\theta)$ most probable

  sequence of hidden states

Three Main Problems in HMMs

- **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 $\arg\max_{s_1,\ldots,s_T} p(\{S_t\}_{t=1}^T|\{O_t\}_{t=1}^T,\theta)$ 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

HMM Algorithms

HMM Algorithms

- Evaluation – What is the probability of the observed sequence? Forward Algorithm

HMM Algorithms

- 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

HMM Algorithms

- 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

HMM Algorithms

- 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)

Evaluation Problem

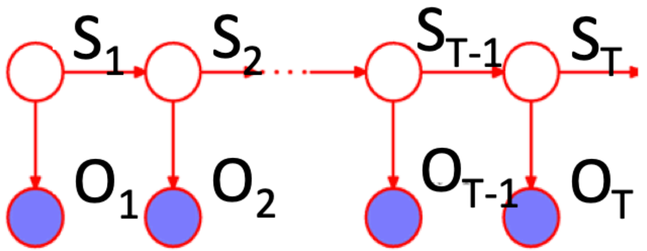
Evaluation Problem

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

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

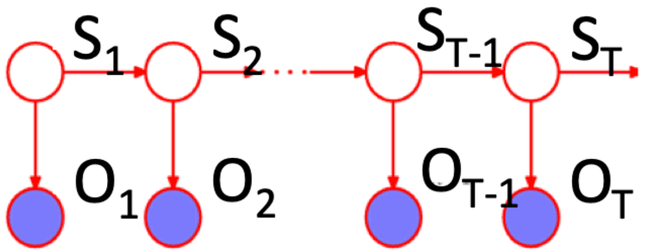


$$p(\{O_t\}_{t=1}^T) = \sum_{S_1,\ldots,S_T} p(\{O_t\}_{t=1}^T, \{S_t\}_{t=1}^T)$$

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

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!

Forward Probability

Forward Probability

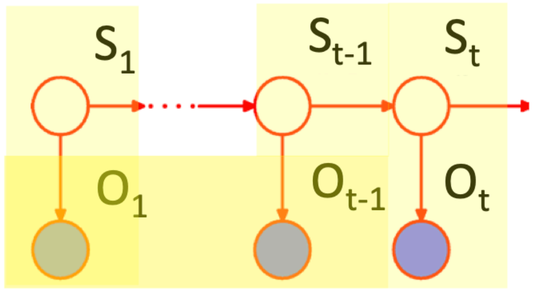
$$p(\{O_t\}_{t=1}^T) = \sum_k p(\{O_t\}_{t=1}^T, S_T = k) = \sum_k \alpha_T^k$$

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

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



Introduce S_{t-1}

.
.
.

Chain rule

Markov assumption

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

Forward Algorithm

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

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

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

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

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

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

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

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

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

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

Can we do in the backward direction?

Decoding Problem 1

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

Decoding Problem 1

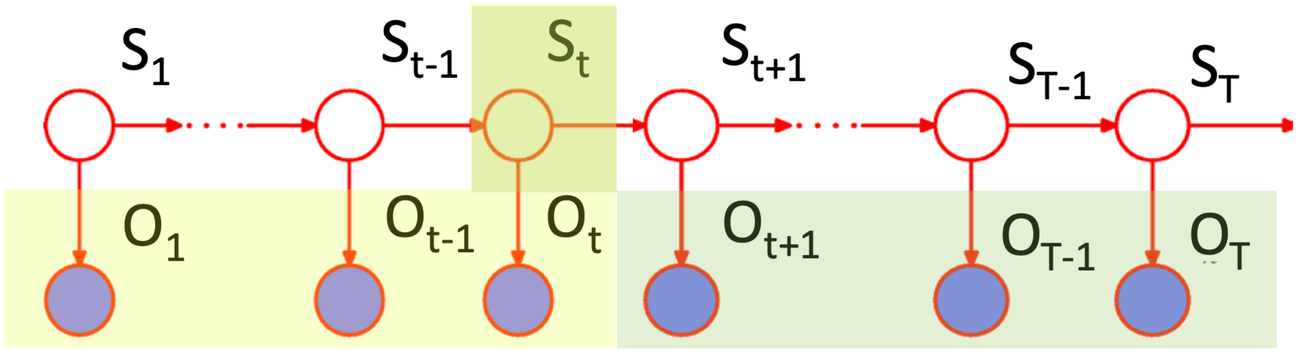
- Given HMM parameters $p(S_1), p(S_t|S_{t-1}), p(O_t|S_t)$ & observation sequence $\{O_t\}_{t=1}^T$

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

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

Compute recursively

α_t^k β_t^k



Forward-Backward Algorithm

Forward-Backward Algorithm

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

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

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

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

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

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

- 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 vs. Most Likely Sequence

Most Likely State vs. 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 State vs. 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

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

Most Likely State vs. 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

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

Are the solutions the same?

Decoding Problem 2

Decoding Problem 2

- 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 most likely assignment of state sequence

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

Decoding Problem 2

- 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 most likely assignment of state sequence

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

$$= \arg \max_k \underbrace{\max_{\{S_t\}_{t=1}^{T-1}} p(S_T = k, \{S_t\}_{t=1}^{T-1}, \{O_t\}_{t=1}^T)}_{V_T^k}$$

Compute recursively

Decoding Problem 2

- 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 most likely assignment of state sequence

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

$$= \arg \max_k \max_{\{S_t\}_{t=1}^{T-1}} p(S_T = k, \{S_t\}_{t=1}^{T-1}, \{O_t\}_{t=1}^T)$$

V_T^k

Compute recursively

V_T^k - probability of most likely sequence of states ending at state $S_T = k$

Viterbi Decoding

Viterbi Decoding

$$\max_{\{S_t\}_{t=1}^T} p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T) = \max_k V_T^k$$

Viterbi Decoding

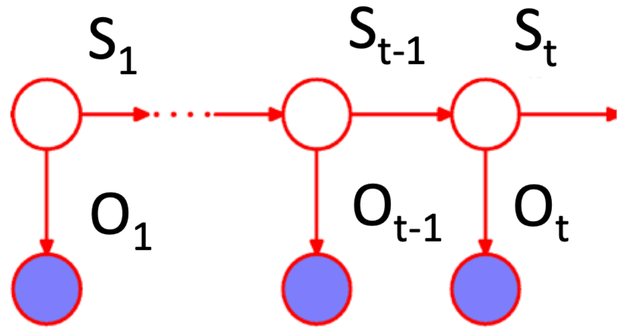
$$\max_{\{S_t\}_{t=1}^T} p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T) = \max_k V_T^k$$

Compute probability V_t^k recursively over t

$$V_t^k := \max_{S_1,\ldots,S_{t-1}} p(S_t = k, S_1, \ldots, S_{t-1}, O_1, \ldots, O_t)$$

. Bayes rule

. Markov assumption



$$= p(O_t|S_t = k) \max_i p(S_t = k|S_{t-1} = i) V_{t-1}^i$$

Viterbi Algorithm

Viterbi Algorithm

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

- Initialize: $V_1^k = p(O_1|S_1=k)p(S_1 = k)$ for all k
- Iterate: for t = 2, …, T

$$V_t^k = p(O_t|S_t = k) \max_i p(S_t = k|S_{t-1} = i)V_{t-1}^i \quad \text{for all k}$$

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

Viterbi Algorithm

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

- Initialize: $V_1^k = p(O_1|S_1=k)p(S_1 = k)$ for all k

- Iterate: for t = 2, …, T

$$V_t^k = p(O_t|S_t = k) \max_i p(S_t = k|S_{t-1} = i)V_{t-1}^i \quad \text{for all k}$$

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

Traceback:

$$S_T^* = \arg\max_k V_T^k$$

$$S_{t-1}^* = \arg\max_i p(S_t^*|S_{t-1} = i)V_{t-1}^i$$

Viterbi Algorithm

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

- Initialize: $V_1^k = p(O_1|S_1=k)p(S_1 = k)$ for all k

- Iterate: for t = 2, …, T

$$V_t^k = p(O_t|S_t = k) \max_i p(S_t = k|S_{t-1} = i)V_{t-1}^i \quad \text{for all k}$$

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

Traceback:

$$S_T^* = \arg\max_k V_T^k$$

$$S_{t-1}^* = \arg\max_i p(S_t^*|S_{t-1} = i)V_{t-1}^i$$

Can we do in the backward direction?

Computational Complexity

Computational Complexity

- What is the running time for Forward, Backward, Viterbi?

$$\alpha_t^k = q_k^{O_t} \sum_i \alpha_{t-1}^i \; p_{i,k}$$

$$\beta_t^k = \sum_i p_{k,i} \; q_i^{O_{t+1}} \; \beta_{t+1}^i$$

$$V_t^k = q_k^{O_t} \; \max_i p_{i,k} \; V_{t-1}^i$$

$O(K^2T)$ linear in T instead of $O(K^T)$ exponential in T!

Learning with EM

- Start with random initialization of parameters
- **E-step** – Fix parameters, find expected state assignments

$$\gamma_i(t) = p(S_t = i|O,\theta) = \frac{\alpha_t^i \beta_t^i}{\sum_j \alpha_t^j \beta_t^j} \qquad \mathbf{O} = \{O_t\}_{t=1}^T$$

Forward-Backward algorithm

$$\xi_{ij}(t) = p(S_{t-1}= i, S_t = j|O,\theta)$$

$$= \frac{p(S_{t-1} = i|O,\theta)p(S_t = j, O_t, \ldots, O_T|S_{t-1} = i,\theta)}{p(O_t, \ldots, O_T|S_{t-1} = i,\theta)}$$

$$= \frac{\gamma_i(t-1)\ p_{ij}\ q_j^{O_t}\ \beta_t^j}{\beta_{t-1}^i}$$

Learning with EM

- Start with random initialization of parameters
- **E-step** – Fix parameters, find expected state assignments

$$\gamma_i(t) = p(S_t = i|O,\theta) = \frac{\alpha_t^i \beta_t^i}{\sum_j \alpha_t^j \beta_t^j} \qquad \mathbf{O} = \{O_t\}_{t=1}^T$$

Forward-Backward algorithm

$$\xi_{ij}(t) = p(S_{t-1}= i, S_t = j|O,\theta)$$

$$= \frac{p(S_{t-1} = i|O,\theta)p(S_t = j, O_t, \ldots, O_T|S_{t-1} = i,\theta)}{p(O_t, \ldots, O_T|S_{t-1} = i,\theta)}$$

$$= \frac{\gamma_i(t-1)\ p_{ij}\ q_j^{O_t}\ \beta_t^j}{\beta_{t-1}^i}$$

You will derive the EM in your HW

Thank You!
Q & A