Probability and Time: Markov Models

CPSC 322 Lecture 30

Lecture Overview

- **Recap**
- **Temporal Probabilistic Models**
- Start Markov Models
  - Markov Chain
  - Markov Chains in Natural Language Processing

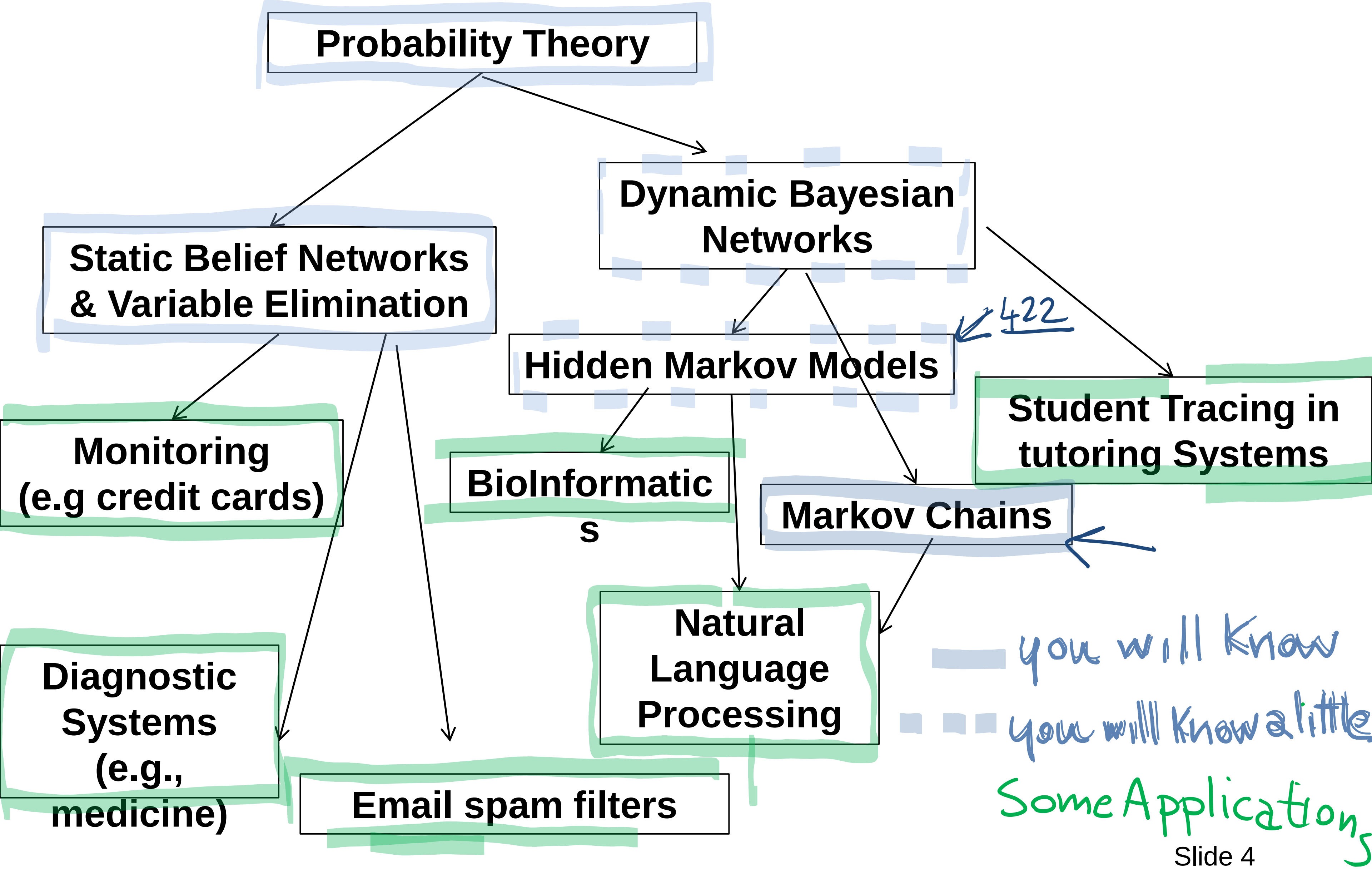
R&R systems we'll cover in this course

		Environment	
Problem		**Deterministic**	**Stochastic**
Static	Constraint Satisfaction	*Variables + Constraints* Search Arc Consistency Local Search	
	Query	*Logics* Search	*Bayesian (Belief) Networks* Variable Elimination
Sequential	Planning	*STRIPS* Search	*Decision Networks* Variable Elimination

Representation

Reasoning Technique

Answering Query under Uncertainty



Learning Goals for today's class

You can:

- Specify a Markov Chain and compute the probability of a sequence of states
- Justify and apply Markov Chains to compute the probability of a Natural Language sentence

Lecture Overview

- Recap
- **Temporal Probabilistic Models**
- Start Markov Models
  - Markov Chain
  - Markov Chains in Natural Language Processing

Modelling static Environments

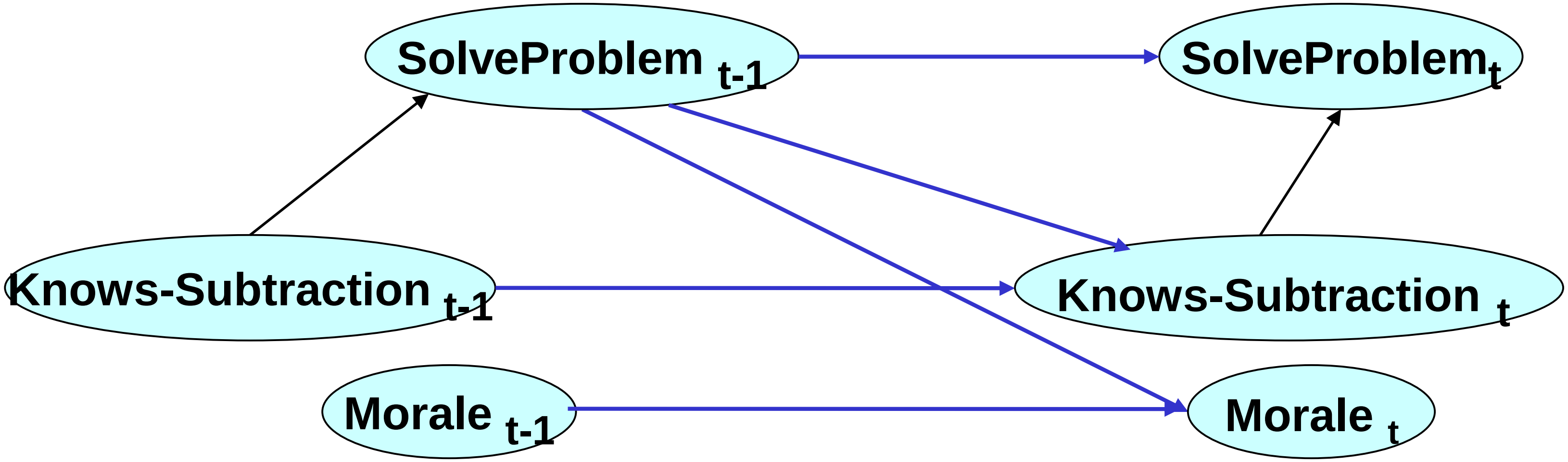
So far we have used Bnets to perform inference in **static environments**

- For instance, the system keeps collecting evidence to diagnose the cause of a fault in a system (e.g., a car).
- The environment (values of the evidence, the true causes of the fault) does not change as I gather new evidence
- What does change?

The system's beliefs over possible causes

Modeling Evolving Environments

- Often we need to make inferences about evolving environments.
- Represent the state of the world at each specific point in time via a series of snapshots, or ***time slices***,



Tutoring system tracing student *knowledge* and *morale*

Lecture Overview

- **Recap**
- **Temporal Probabilistic Models**
- Start Markov Models
    - **Markov Chain**
    - Markov Chains in Natural Language Processing

Simplest Possible DBN

- **One random variable** for each time slice: let's assume S_t represents the state at time t. with domain $\{v_1 \ldots v_n\}$

- **We assume that each random variable depends only on the previous one**

- Thus $P(S_{t+1} \mid S_0 \cdots S_t) = P(\underline{S_{t+1}} \mid \underline{S_t})$
- Intuitively S_t conveys all of the information about the history that can affect the future states.
- "The future is independent of the past given the present."

Simplest Possible DBN (cont')

- How many CPTs do we need to specify?

iclicker.

A. *1* **B.** *4* **C .** *2* **D.** *3* **E.** *42*

- *Stationary process assumption:* the mechanism that regulates how state variables change overtime is **stationary**, that is it can be described by a single transition model

- **$P(S_t|S_{t-1})$** is the same for all t

Stationary Markov Chain (SMC)

A stationary Markov Chain : for all t >0

- $P\ (S_{t+1}|\ S_0,\ldots,S_t) = P\ (S_{t+1}|S_t)$ Markov assumption
- $P\ (S_{t\ +1}|S_t)$ is the same stationary assumption

So now we only need to specify…

A. $P\ (S_{t\ +1}\ |S_t)$ and $P\ (S_0)$

B. $P\ (S_0)$

C . $P\ (S_{t\ +1}|S_t)$

D. $P\ (S_t\ |S_{t+1}\ )$

Stationary Markov Chain (SMC)

A stationary Markov Chain : for all t >0

- $P(S_{t+1}| S_0,\ldots,S_t) = P(S_{t+1}|S_t)$ Markov assumption
- $P(S_{t+1}|S_t)$ is the same stationary assumption

We only need to specify $P(S_0)$ and $P(S_{t+1}|S_t)$

- Simple Model, easy to specify
- Often the natural model
- The network can extend indefinitely
- **Variations of SMC are at the core of many Natural Language Processing (NLP) applications!** PageRank

Stationary Markov-Chain: Example

Domain of variable S_i is {t , q, p, a, h, e}

Probability of initial state $P(S_0)$

t	.6
q	.4
p	
a	
h	
e	

Stochastic Transition Matrix $P(S_{t+1}|S_t)$

Which of these two is a possible STM?

S_{t+1}

S_t

	t	q	p	a	h	e
t	0	.3	0	.3	.4	0
q	.4	0	.6	0	0	0
p	0	0	1	0	0	0
a	0	0	.4	.6	0	0
h	0	0	0	0	0	1
e	1	0	0	0	0	0

S_{t+1}

S_t

	t	q	p	a	h	e
t	1	0	0	0	0	0
q	0	1	0	0	0	0
p	.3	0	1	0	0	0
a	0	0	0	1	0	0
h	0	0	0	0	0	1
e	0	0	0	.2	0	1

A. Left only

B. Right only

C . Both

D. Neither

E. 42?

Stationary Markov-Chain: Example

Domain of variable S_i is {t , q, p, a, h, e}

We only need to specify…

$P(S_0)$

Probability of initial state

t	.6
q	.4
p	
a	
h	
e	

Stochastic Transition Matrix

$P(S_{t+1}|S_t)$

S_{t+1}

S_t	t	q	p	a	h	e
t	0	.3	0	.3	.4	0
q	.4	0	.6	0	0	0
p	0	0	1	0	0	0
a	0	0	.4	.6	0	0
h	0	0	0	0	0	1
e	1	0	0	0	0	0

Markov-Chain: Inference

Probability of a sequence of states $S_0 \ldots S_T$

$$P(S_0,\ldots,S_T) = P(S_0)\,P(S_1 \mid S_0)\,P(S_2 \mid S_1) - - - - - - -$$

$P(S_0)$

t	.6
q	.4
p	0
a	0
h	0
e	0

$P(S_{t+1}|S_t)$

	t	q	p	a	h	e
t	0	.3	0	.3	.4	0
q	.4	0	.6	0	0	0
p	0	0	1	0	0	0
a	0	0	.4	.6	0	0
h	0	0	0	0	0	1
e	1	0	0	0	0	0

Example:

$$P(t,q,p) =$$

A. 0.42 B. 0 C. 0.24 D. 0.054 E. 0.108

Lecture Overview

- Recap
- Temporal Probabilistic Models
- **Markov Models**
  - **Markov Chain**
  - Markov Chains in Natural Language Processing

Key problems in NLP

“Book me a room near UBC”

w_1 w_2 w_3 w_4 w_5 w_6

$$\boxed{P(w_1,..,w_n)?}$$

Assign a probability to a sentence (a sequence of words)

- Part-of-speech tagging
- Word-sense disambiguation,
- Probabilistic Parsing

Summarization, Machine Translation……..

Predict the next word

- Speech recognition
- Hand-writing recognition
- Augmentative communication for the disabled

$$P(w_n \mid w_1 \ldots\ldots w_{n-1}) =$$

$$= P(w_1 \ldots w_n) / P(w_1 \ldots w_{n-1})$$

$$\boxed{P(w_1,..,w_n)?}$$

Impossible to estimate ☹

Key problems in NLP

$\boxed{P(w_1,..,w_n)?}$

Impossible to estimate!

Assuming 10^5 words and average sentence contains 10 words …….

$(10^5)^{10} = 10^{50}$

would contain ↑ probabilities

Google language repository (22 Sept. 2006) contained “only” 95,119,665,584 sentences

Most sentences will not appear or appear only once ☹

What can we do?

Make a strong simplifying assumption!

Assume sentences are generated by a Markov Chain

$$P(w_1,..,w_n) = P(w_1 |< S >) \prod_{k=2}^{n} P(w_k \mid w_{k-1})$$

P(The big red dog barks)=

P(The|<S>) * ______________

Estimates for Bigrams $P(w_i | w_{i-1})$

Silly language repository with **only two sentences**:

“<S>*The big red dog barks at the big pink dog*”

“<S>*The big pink dog is much smaller*”

Count How many times in your documents you have "big red" and "big"

$$P(red \mid big) = \frac{P(big, red)}{P(big)} = \frac{\dfrac{C(big, red)}{N_{pairs}}}{\dfrac{C(big)}{N_{words}}} = \frac{C(big, red)}{C(big)} =$$

$P(w_i | w_{i-1})$

$10^5 * 10^5$ matrix

$$\boxed{P(w_i \mid w_{i-1}, w_{i-2})}$$

some models use two preceeding words

Bigrams in practice…

If you have 10^5 words in your dictionary

$P(w_i \mid w_{i-1})$ will contain how many numbers.. ??

A. $2 * 10^5$

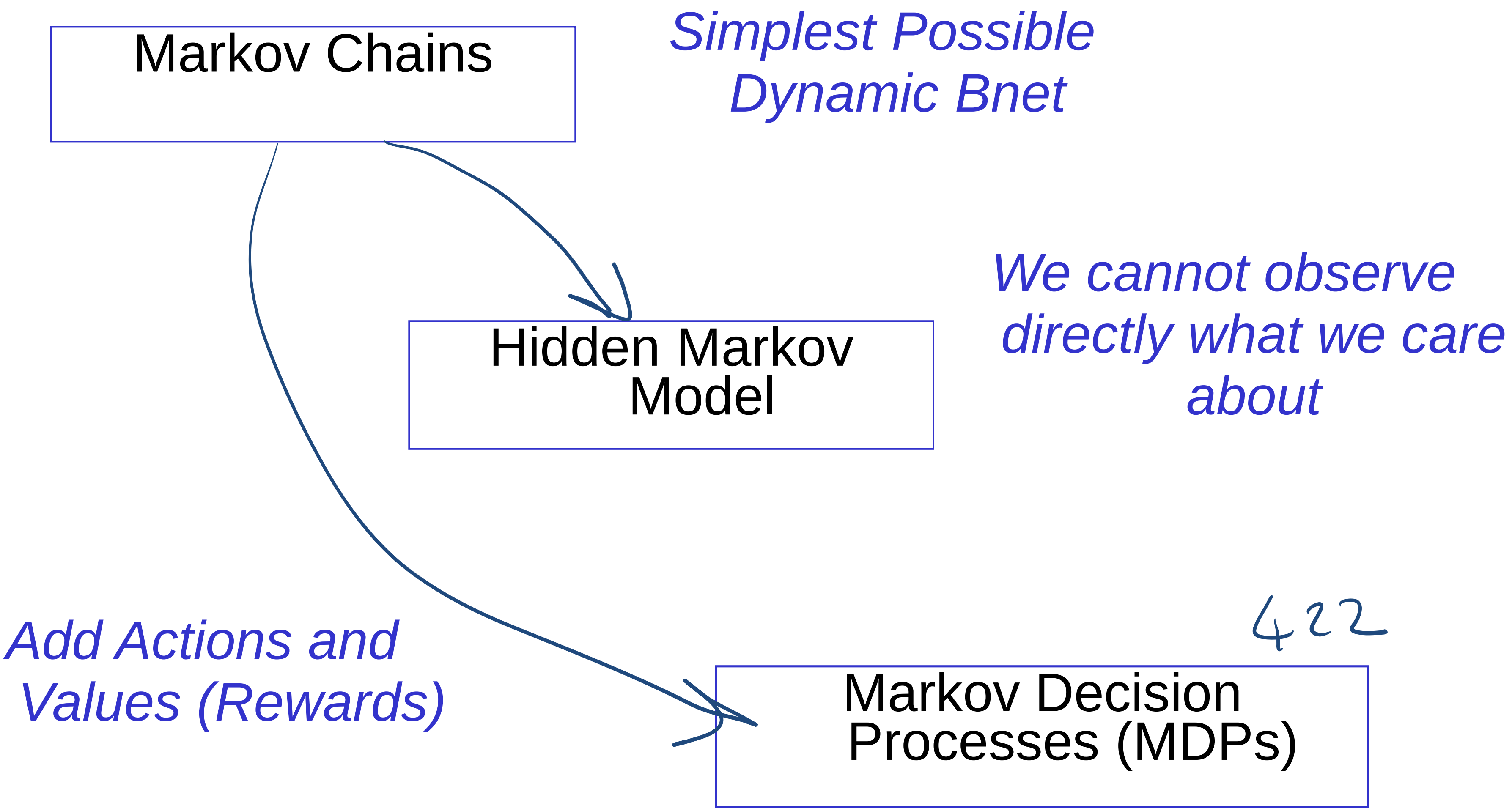
B. 10^{10}

C. $5 * 10^5$

D. $2 * 10^{10}$

E. 42

Markov Models



Markov Chains

Simplest Possible Dynamic Bnet

Hidden Markov Model

We cannot observe directly what we care about

422

Add Actions and Values (Rewards)

Markov Decision Processes (MDPs)

Next Class

- **Finish Probability and Time:** Hidden Markov Models (HMM) *(TextBook 8.5.2)*
- **Start Decision networks** *(TextBook chpt 9)*