Readings:
K&F: 13.1, 13.2, 13.3

Dynamic Bayesian Networks

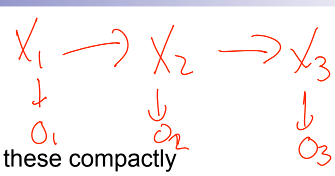
Beyond 10708

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University

December ~~1st, 2006~~ 3rd 2008

Dynamic Bayesian network (DBN)

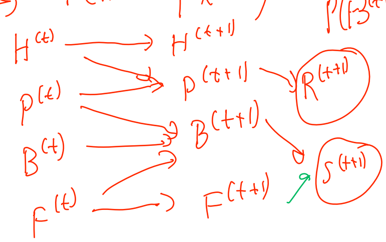
- HMM defined by
  - Transition model $P(X^{(t+1)}|X^{(t)})$
  - Observation model $P(O^{(t)}|X^{(t)})$
  - Starting state distribution $P(X^{(0)})$
- DBN – Use Bayes net to represent each of these compactly
  - Starting state distribution $P(X^{(0)})$ is a BN
  - (silly) e.g, performance in grad. school DBN
    - Vars: **H**appiness, **P**roductivity, Hira**B**lility, **F**ame
    - Observations: Pape**R**, **S**chmooze



compute, e.g. $P(B^{(5)}=t \mid R^{(1)}=t, S^{(2)}=t, R^{(3)}=t, S^{(4)}=f)$

$P(X^{(0)})$ e.g., $H^{(0)}$, $P^{(0)}$, $B^{(0)}$, $F^{(0)}$

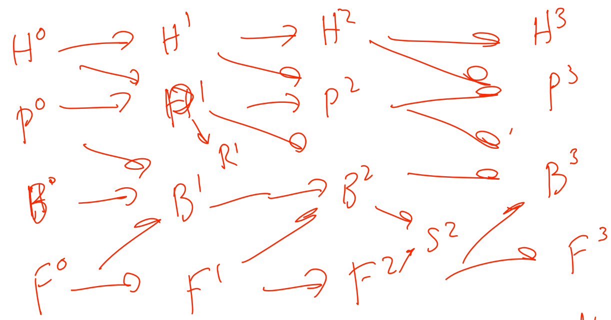
2TBN → $P(X^{(t+1)} \mid X^{(t)}) = P(H^{(t+1)}|H^{(t)})\, P(P^{(t+1)}|H^{(t)} P^{(t)})\, P(B^{(t+1)}|P^{(t)}, B^{(t)}, F^{(t)})\, P(F^{(t+1)}|F^{(t)} \ldots)$



$P(O^{(t+1)} \mid X^{(t+1)}) = P(R^{(t+1)}|P^{(t+1)})\, P(S^{(t+1)}|B^{(t+1)}, F^{(t+1)})$

Unrolled DBN

- Start with $P(X^{(0)})$
- For each time step, add vars as defined by 2-TBN



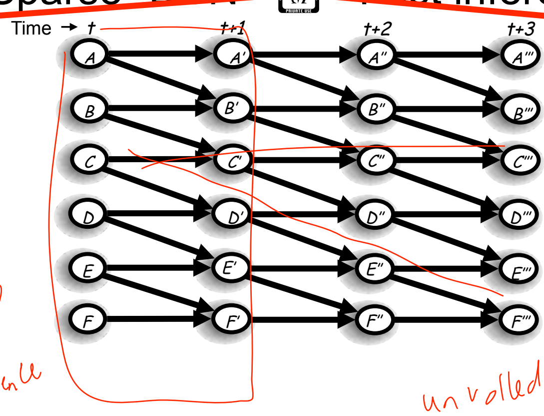
large treewidth

"Sparse" DBN and fast inference

$P(X^{(t+1)} | X^{(t)})$

~~"Sparse" DBN → Fast inference~~

learn about assumed density filtering for inference

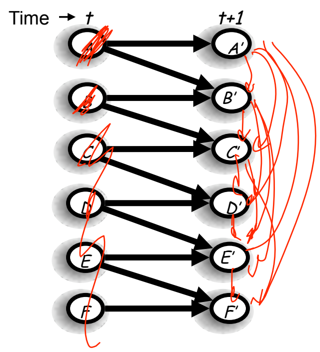


simple

unrolled complex

Even after one time step!!

What happens when we marginalize out time t?



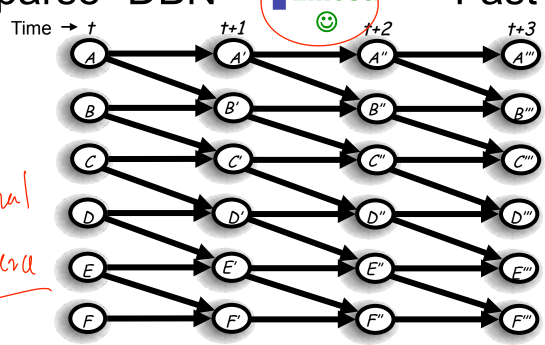
5

"Sparse" DBN and fast inference 2

Structured representation of belief often yields good approximate

"Sparse" DBN ? Almost! ☺ Fast inference

related to variational inference



Find a "simple" tractable approx. of $P(X^{(t)} | O^{(1:t)})$

6

BK Algorithm for approximate DBN inference

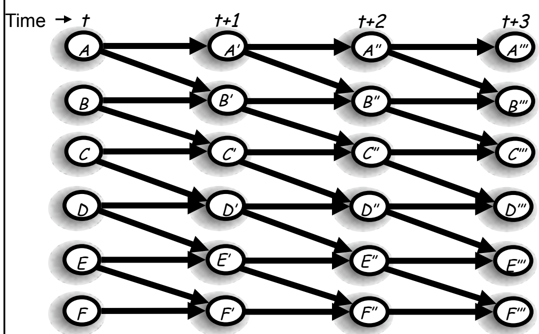
[Boyen, Koller '98]

See book for sampling

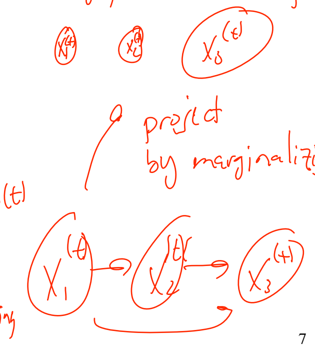
Assumed density filter

- Assumed density filtering:
  - Choose a factored representation $\hat{P}$ for the belief state
  - Every time step, belief not representable with $\hat{P}$, project into representation

$P(X^{(t)} \mid O^{(1:t)})$



$\hat{P}^{(t)}$ e.g., mean fields



project by marginalizing

$P^{(t)}$

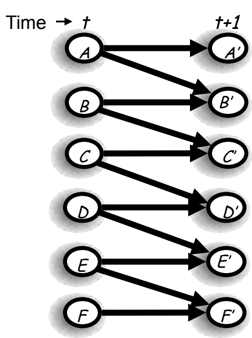
never constructing P explicitly

7

A simple example of BK: Fully-Factorized Distribution

e.g.

- Assumed density:
  - Fully factorized



True P(X^(t+1)):

A' B' C' D' ⋮

marginals of $P(X^{(t+1)})$

Assumed Density for $\hat{P}$(X^(t+1)):

A' B' C' D'

8

Computing Fully-Factorized Distribution at time t+1

$P(B'|P,H,F)$

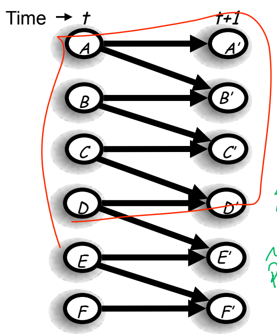
- Assumed density:
  - Fully factorized

As input given 2-TBN $P(X^{(t+1)}|X^{(t)})$

Assumed Density for $\hat{P}(X^{(t)})$:

Computing for $\hat{P}(X_i^{(t+1)})$:

$\hat{P}'(A'B'C'D') = \hat{P}(A')\hat{P}(B')\hat{P}(C')\hat{P}(D')$



$\hat{P}(X^{(t)})$ e.g., Variable elimination

$\hat{P}(A)$ A → A'

$\hat{P}(B)$ B → B'

CPT: $P(C'|BC)$

$\hat{P}(C)$ C → C'

$\hat{P}(D)$ D → D'

A'

B'

$\hat{P}(C')$ using variable elimination

C'

In this simple example

D'

$\hat{P}(c') = \sum_{b,c} P(c'|b,c)\,\hat{P}(b)\,\hat{P}(c)$

General case for BK: Junction Tree Represents Distribution

compute these

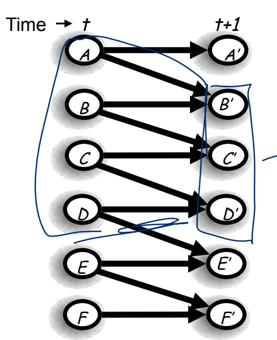
- Assumed density:
  - Fully factorized

$\hat{P}(A'B'C'D'E'F') = \hat{P}(A'B'C')\hat{P}(B'C'D')\,/\,\hat{P}(B'C')$ ... $\hat{P}(D'E'F')\,/\,\hat{P}(D')$

Assumed density

~~True~~ $\hat{P}(X^{(t)})$:

Assumed Density for $\hat{P}(X^{(t+1)})$:



ABC

BC

BCD

D

DEF

Variable elimination

← factors from JT

+

CPTs $P(B'|AB)$
$P(C'|BC)$
$P(D'|CD)$

A'B'C'

B'C'

B'C'D' ← $\hat{P}(B'C'D')$

D'

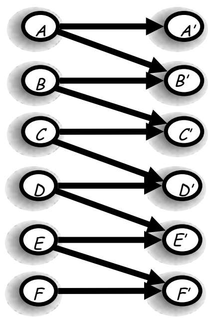
DEF'

Computing factored belief state in the next time step

- Introduce observations in current time step
  - Use J-tree to calibrate time t beliefs
- Compute $t+1$ belief, project into approximate belief state
  - marginalize into desired factors
  - corresponds to KL projection

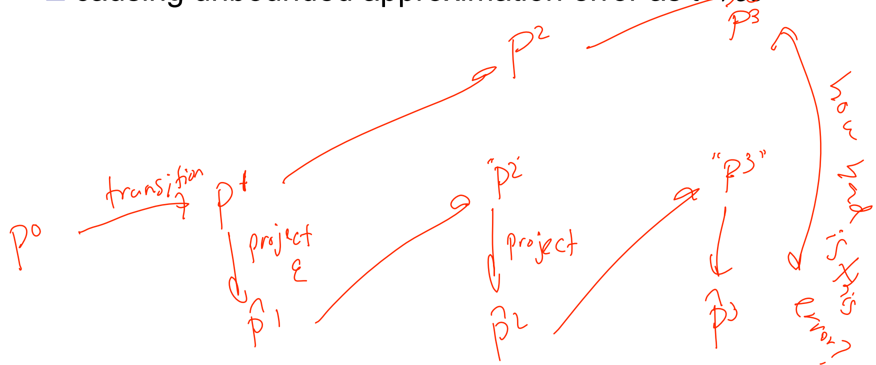
min KL

- Equivalent to computing marginals over factors directly
  - For each factor in $t+1$ step belief
    - Use variable elimination

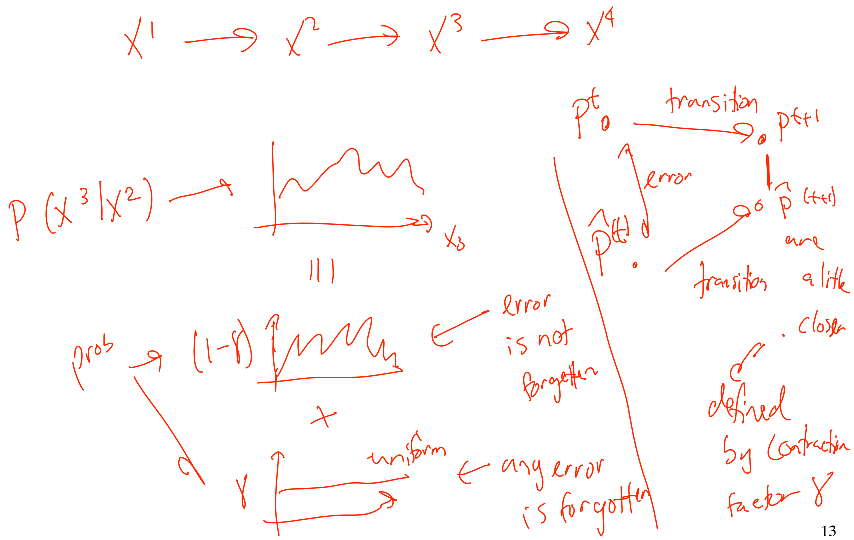


Error accumulation

- Each time step, projection introduces error
- Will error add up?
  - causing unbounded approximation error as $t \to \infty$

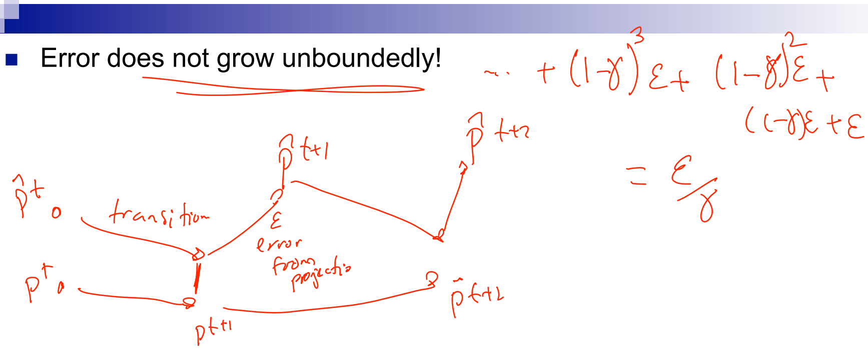


Contraction in Markov process



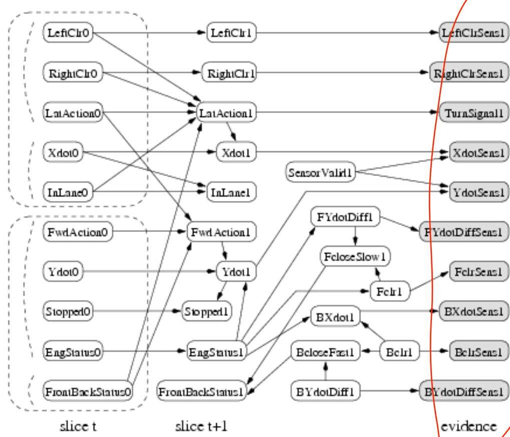
BK Theorem

- Error does not grow unboundedly!



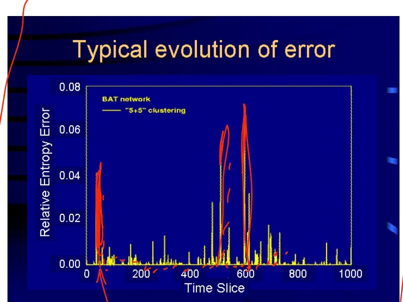
- **Theorem**: If Markov chain **contracts at a rate of** γ (usually very small), and **assumed density projection at each time step has error bounded by** ε (usually large) then the **expected error at every iteration is bounded by** ε/γ.

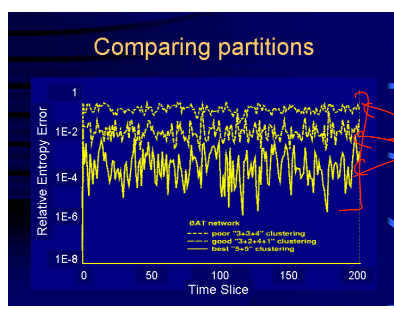
Example – BAT network [Forbes et al.]



15

BK results [Boyen, Koller '98]





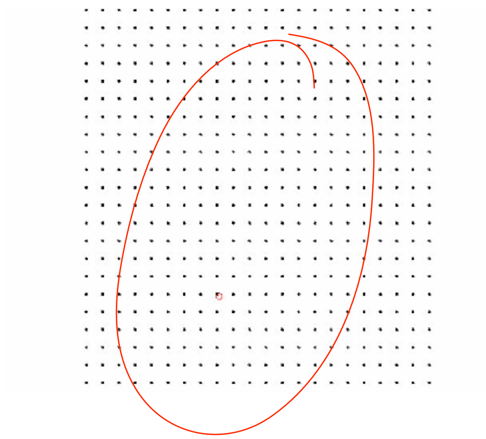
choice of assumed density

16

Thin Junction Tree Filters [Paskin '03]

- BK assumes fixed approximation clusters
- TJTF adapts clusters over time
  - attempt to minimize projection error

ε



17

Hybrid DBN (many continuous and discrete variables)

- DBN with large number of discrete and continuous variables
- # of mixture of Gaussian components blows up in one time step!
- Need many smart tricks…
  - e.g., see Lerner Thesis

Figure 10.1: The prototype RWGS system

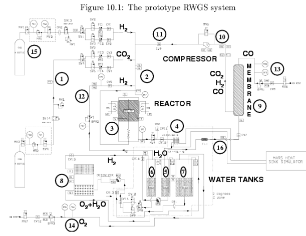


Figure 10.2: The RWGS schematic

Reverse Water Gas Shift System (RWGS) [Lerner et al. '02]

18

DBN summary

- **DBNs**
  - factored representation of HMMs/Kalman filters
  - sparse representation does not lead to efficient inference
- **Assumed density filtering**
  - BK – factored belief state representation is assumed density
  - Contraction guarantees that error does blow up (but could still be large)
  - Thin junction tree filter adapts assumed density over time
  - Extensions for hybrid DBNs
  - Sampling
  - Loopy BP (factored frontier)

19

Final

- Out: Later today Wednesday
- Due: December 10th at NOON (STRICT DEADLINE)
- Start Early!!!
  - NO LATE DAYS
  - NO COLLABORATIONS OF ANY KIND
  - NO USE GOOGLE/Similar techniques

20

And the winners are…

- Popular Vote:
  - Learning and prediction of emotion components in a conversation using dynamic bayesian networks (Ekaterina Spriggs)
- Instructors' Choice:
  - Temporal model for Enron email dataset (Leman Akoglu and Seungil Huh)
  - Learning low-treewidth CRFs via Graph cuts (Dafna Shahaf)

21

This semester…

- Bayesian networks, Markov networks, factor graphs, decomposable models, junction trees, parameter learning, structure learning, semantics, exact inference, variable elimination, context-specific independence, approximate inference, sampling, importance sampling, MCMC, Gibbs, variational inference, loopy belief propagation, generalized belief propagation, Kikuchi, Bayesian learning, missing data, EM, Chow-Liu, IPF, Gaussian and hybrid models, discrete and continuous variables, temporal and template models, Kalman filter, linearization, conditional random fields, assumed density filtering, DBNs, BK, Causality,…
- **Just the beginning… ☺**

22

Quick overview of some hot topics...

- **Maximum Margin Markov Networks**
- **Relational Probabilistic Models**
- **Influence Diagrams**

23

Max (Conditional) Likelihood

$P_w(y \mid x) = \frac{1}{Z_w(x)} e^{w^T f(x,y)}$

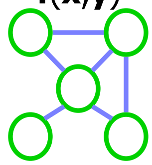
D

$\mathbf{x}^1, \mathbf{t}(\mathbf{x}^1)$

...

$\mathbf{x}^m, \mathbf{t}(\mathbf{x}^m)$

f(x,y)



Estimation CRF

maximize$_{\mathrm{w}}$ $\sum_{\mathrm{x} \in D} \log P_{\mathrm{w}}(\mathrm{t}(\mathrm{x}) \mid \mathrm{x})$

Classification

arg max$_{\mathrm{y}}$ $\mathrm{w}^\top \mathrm{f}(\mathrm{x}, \mathrm{y})$

$$\log P_{\mathrm{w}}(\mathrm{y} \mid \mathrm{x}) = \mathrm{w}^\top \mathrm{f}(\mathrm{x}, \mathrm{y}) - \underbrace{\log Z_{\mathrm{w}}(\mathrm{x})}_{\text{doesn't depend on y}}$$

Don't need to learn entire distribution!

24

OCR Example

- We want:

train' data

argmax$_{\textbf{word}}$ $\mathbf{w}^\top \mathbf{f}($$,\textbf{word}) =$ **"brace"**

- Equivalently:

max diff ⟺ max margin

$\mathbf{w}^\top \mathbf{f}($$,$**"brace"**$) > \mathbf{w}^\top \mathbf{f}($$,$**"aaaaa"**$)$

$\mathbf{w}^\top \mathbf{f}($$,$**"brace"**$) > \mathbf{w}^\top \mathbf{f}($$,$**"aaaab"**$)$

…

$\mathbf{w}^\top \mathbf{f}($$,$**"brace"**$) > \mathbf{w}^\top \mathbf{f}($$,$**"zzzzz"**$)$

generalize sums

a lot!

Max Margin Estimation

- Goal: find $\mathbf{w}$ such that

$$\mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{t}(\mathbf{x})) > \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}) \qquad \forall \mathbf{x} \in D \quad \forall \mathbf{y} \neq \mathbf{t}(\mathbf{x})$$

$$\mathbf{w}^\top \underbrace{[\mathbf{f}(\mathbf{x},\mathbf{t}(\mathbf{x})) - \mathbf{f}(\mathbf{x},\mathbf{y})]} > 0$$

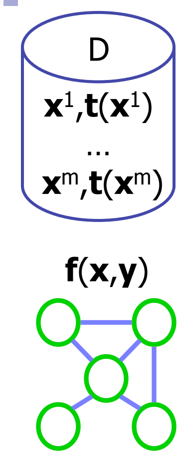
$$\boxed{\mathbf{w}^\top \Delta \mathbf{f}_{\mathbf{x}}(\mathbf{y}) \geq \gamma \Delta \mathbf{t}_{\mathbf{x}}(\mathbf{y})}$$

- Maximize margin γ
- Gain over $\mathbf{y}$ grows with # of mistakes in $\mathbf{y}$: $\Delta \mathbf{t}_{\mathbf{x}}(\mathbf{y})$

$\Delta \mathbf{t}$(**"craze"**) $= 2$ $\Delta \mathbf{t}$(**"zzzzz"**) $= 5$

$\mathbf{w}^\top \Delta \mathbf{f}$(**"craze"**) $\geq 2\gamma$ $\mathbf{w}^\top \Delta \mathbf{f}$(**"zzzzz"**) $\geq 5\gamma$

M^{3}Ns: Maximum Margin Markov Networks [Taskar et al. '03]



Estimation

$$\max_{||w||\le 1} \quad \gamma$$
$$\mathbf{w}^\top \Delta \mathbf{f}_{\mathbf{x}}(\mathbf{y}) \geq \gamma \Delta \mathbf{t}_{\mathbf{x}}(\mathbf{y})$$

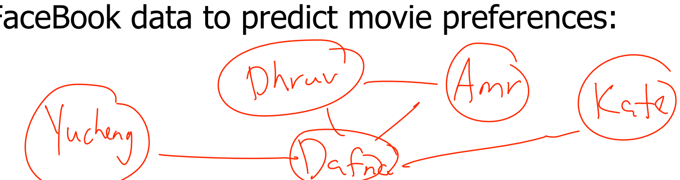
max-margin,
never compute
↳ Z

Classification

$$\arg\max_{\mathbf{y}} \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})$$

Propositional Models and Generalization

- Suppose you learn a model for social networks for CMU from FaceBook data to predict movie preferences:



- How would you apply when new people join CMU?

Yi

- Can you apply it to make predictions a some "little technical college" in Cambridge, Mass?

Generalization requires Relational Models (e.g., see tutorials by Getoor & Domingos)

- Bayes nets / Markov Nets defined specifically for an instance, e.g., CMU FaceBook today
  - fixed number of people
  - fixed relationships between people
  - …

- Relational and first-order probabilistic models
  - talk about objects and relations between objects
  - allow us to represent different (and unknown) numbers
  - generalize knowledge learned from one domain to other, related, but different domains

29

Reasoning about decisions
K&F Chapters 21 & 22

- So far, graphical models only have random variables

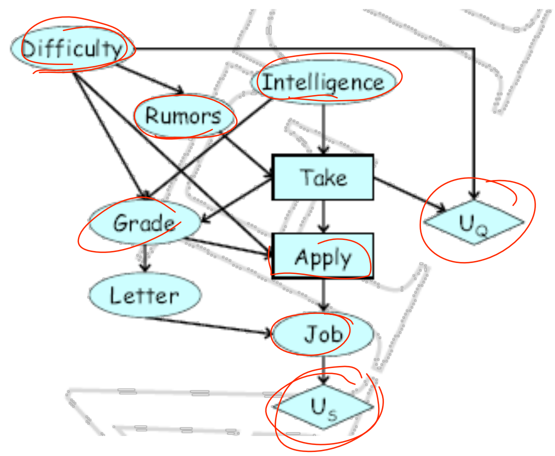
- What if we could make decisions that influence the probability of these variables?
  - e.g., steering angle for a car, buying stocks, choice of medical treatment

- How do we choose the best decision?
  - the one that maximizes the expected long-term utility

- How do we coordinate multiple decisions?

30

Example of an Influence Diagram



31

Many, many, many more topics we didn't even touch, e.g.,...

- **Graph cuts for MPE inference**
  - Exact inference in models with large treewidth, attractive/submodular potentials
- **Active learning**
  - What variables should I observe to learn?
- **Topic Models, Latent Dirichlet Allocation**
  - Unsupervised, discover topics in data
- **Non-parametric models**
  - What if you don't know the number of topics in your data?
- **Continuous time models**
  - DBNs have discrete time steps, but the world is continuous
- **Learning theory for graphical models**
  - How many samples do I need?
- **Distributed algorithms for graphical models**
  - We are moving to a parallel world… where are you?
- **Graphical models for reinforcement learning**
  - Combine DBNs with decision making to scale to huge multiagent problems
- **Applications**
- **…**

32

What next?

- Seminars at CMU:
  - Machine Learning Lunch talks: http://www.cs.cmu.edu/~learning/
  - Intelligence Seminar: http://www.cs.cmu.edu/~iseminar/
  - Machine Learning Department Seminar: http://calendar.cs.cmu.edu/ml/seminar
  - Statistics Department seminars: http://www.stat.cmu.edu/seminar
  - …
- Journal:
  - JMLR – Journal of Machine Learning Research (free, on the web)
  - JAIR – Journal of AI Research (free, on the web)
  - …
- Conferences:
  - UAI: Uncertainty in AI
  - NIPS: Neural Information Processing Systems
  - Also ICML, AAAI, IJCAI and others
- Some MLD courses:
  - 10-705 Intermediate Statistics (Fall)
  - 10-702 Statistical Foundations of Machine Learning (Spring)
  - 10-725 Optimization (Spring 2010)
  - 10-615 Art that Learns (Spring)
  - …

33