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

Dynamic Bayesian network (DBN)

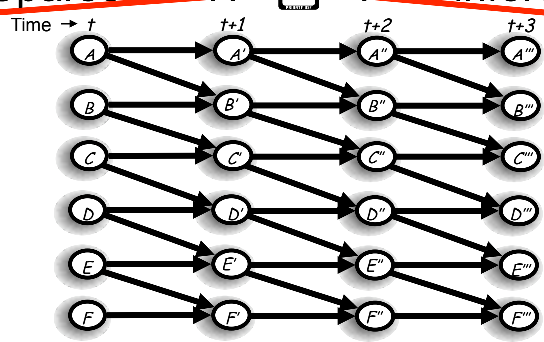
- HMM defined by
  - Transition model $P(X^{(t+1)}|X^{(t)})$
  - Observation model $P(O^{(t)}|X^{(t)})$
  - Starting state distribution $P(X^{(0)})$
- DBN – Use Bayes net to represent each of these compactly
  - Starting state distribution $P(X^{(0)})$ is a BN
  - (silly) e.g, performance in grad. school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

Unrolled DBN

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

"Sparse" DBN and fast inference

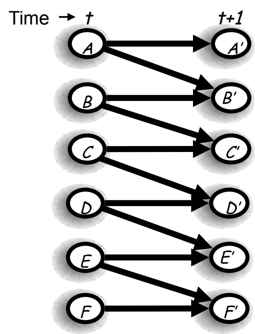
~~"Sparse" DBN ☹ Fast inference~~



4

Even after one time step!!

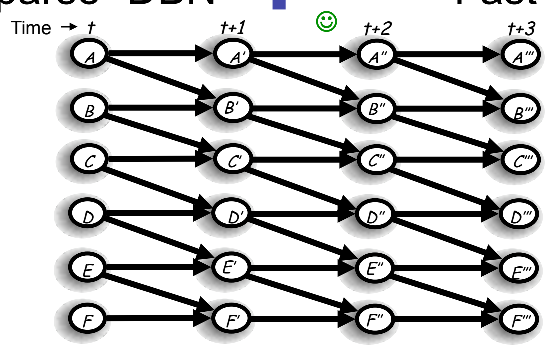
What happens when we marginalize out time t?



"Sparse" DBN and fast inference 2

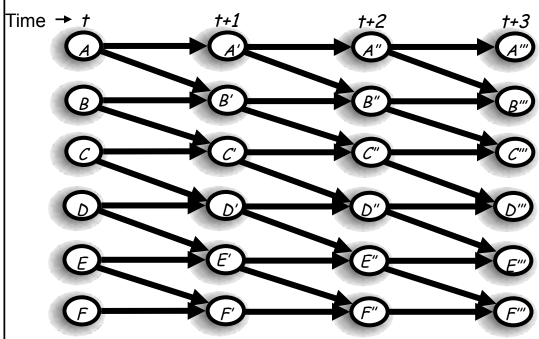
Structured representation of belief often yields good approximate

"Sparse" DBN ? Almost! ☺ Fast inference



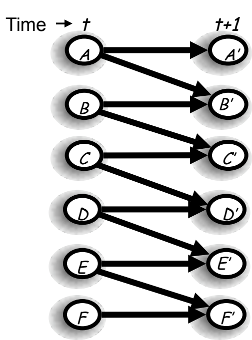
BK Algorithm for approximate DBN inference
[Boyen, Koller '98]

- 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



A simple example of BK: Fully-Factorized Distribution

- Assumed density:
  - Fully factorized



True $P(X^{(t+1)})$:

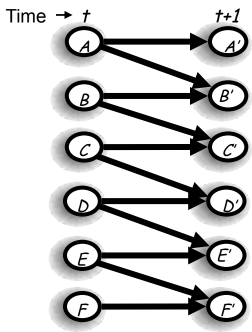
Assumed Density for $\hat{P}(X^{(t+1)})$:

Computing Fully-Factorized Distribution at time t+1

- Assumed density:
  - Fully factorized

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

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



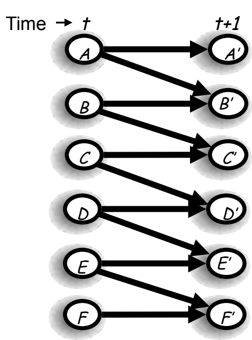
9

General case for BK: Junction Tree Represents Distribution

- Assumed density:
  - Fully factorized

True $P(X^{(t+1)})$:

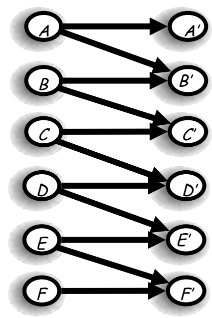
Assumed Density for $\hat{P}(X^{(t+1)})$:



10

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



11

Error accumulation

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

12

Contraction in Markov process

13

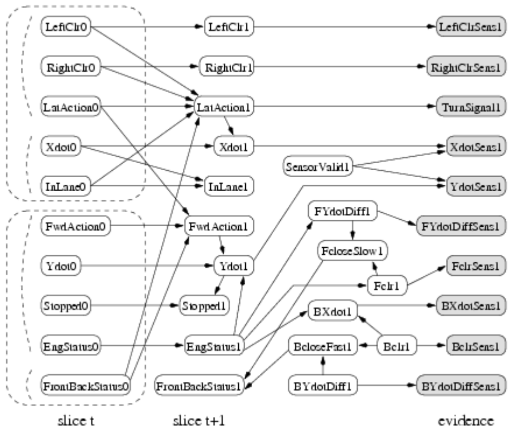
BK Theorem

- 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.

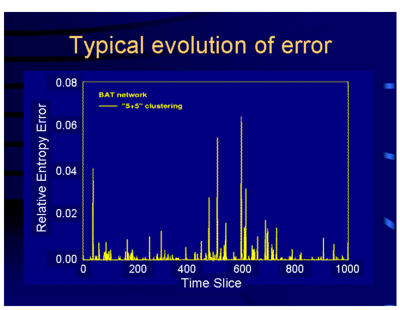
14

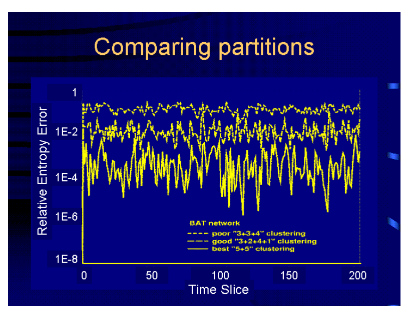
Example – BAT network [Forbes et al.]



15

BK results [Boyen, Koller '98]

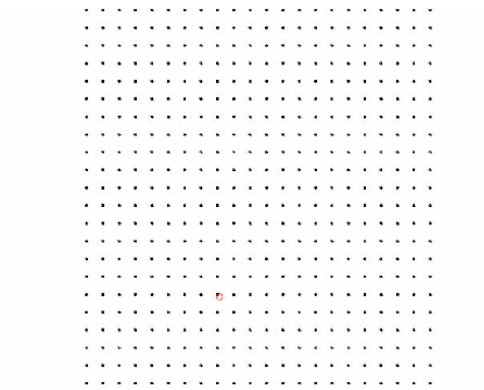




16

Thin Junction Tree Filters [Paskin '03]

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



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

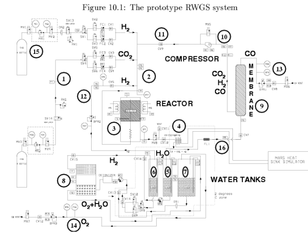


Figure 10.2: The RWGS schematic

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

Final

- Out: Later today
- Due: December 10th at NOON (STRICT DEADLINE)
- Start Early!!!

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… ☺**

21

Quick overview of some hot topics...

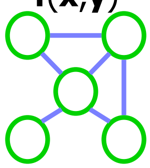
- **Maximum Margin Markov Networks**
- **Relational Probabilistic Models**
- **Influence Diagrams**

22

Max (Conditional) Likelihood

D
$x^1, t(x^1)$
…
$x^m, t(x^m)$

f(x,y)



Estimation

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

Classification

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

$$\log P_{\mathbf{w}}(\mathbf{y} \mid \mathbf{x}) = \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) - \log Z_{\mathbf{w}}(\mathbf{x})$$

Don't need to learn entire distribution!

23

OCR Example

- We want:

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

- Equivalently:

  w^T **f(**  **,"brace") > w**^{T} **f(**  **,"aaaaa")**
  w^T **f(**  **,"brace") > w**^{T} **f(**  **,"aaaab")**
  …
  w^T **f(**  **,"brace") > w**^{T} **f(**  **,"zzzzz")**

  } **a lot!**

24

Max Margin Estimation

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

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

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

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

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

$\Delta\mathbf{t}_{brace}$("craze") = 2 $\qquad$ $\Delta\mathbf{t}_{brace}$("zzzzz") = 5

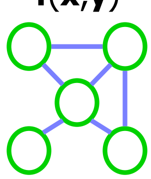
$\mathbf{w}^\top\Delta\mathbf{f}_{brace}$("craze") ≥ 2γ $\qquad$ $\mathbf{w}^\top\Delta\mathbf{f}_{brace}$("zzzzz") ≥ 5γ

25

M³Ns: Maximum Margin Markov Networks [Taskar et al. '03]

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

f(x,y)



Estimation

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

Classification

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

26

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?
- 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)

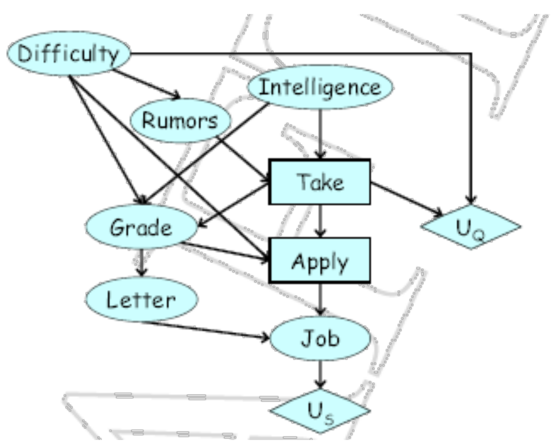
- Bayes nets defined specifically for an instance, e.g., CMU FaceBook today
  - fixed number of people
  - fixed relationships between people
  - …
- Relational and first-order probabilistic models
  - talk about objects and relations between objects
  - allow us to represent different (and unknown) numbers
  - generalize knowledge learned from one domain to other, related, but different domains

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?

29

Example of an Influence Diagram



30

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**
- **…**

31

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)
  - …

32