

Readings:

K&F: 18.1, 18.2, 18.3, 18.4

Dynamic Bayesian Networks

Beyond 10708

Graphical Models – 10708

Carlos Guestrin

Carnegie Mellon University

December 1st, 2006

1

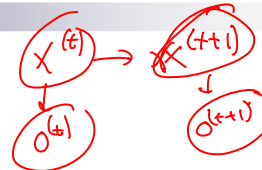
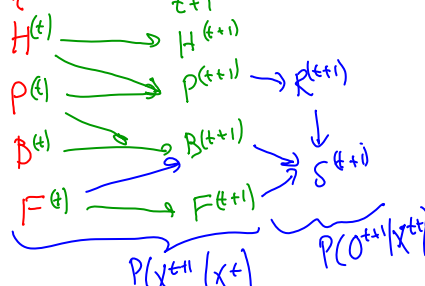
Dynamic Bayesian network (DBN)

- 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: Happiness, Productivity, HiraBility, Fame
 - Observations: Paper, Schmooze



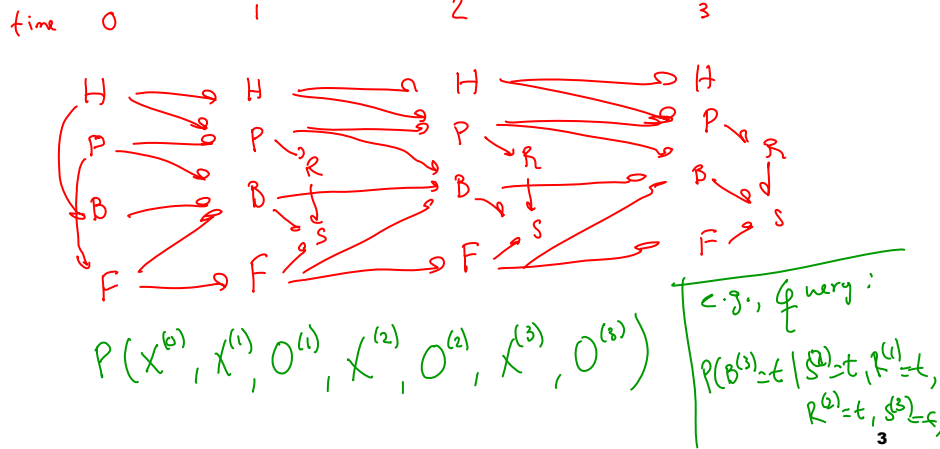
$P(X^{(t+1)} | X^{(t)})$
 how many params $(2^2 - 1)2^2$
 without DBN $2^8 - 2^2$
 with DBN
 $P(H^{(t+1)} | H^{(t)}) (2-1) \cdot 2$
 $P(P^{(t+1)} | P^{(t)}, H^{(t)}) (2-1) \cdot 2^2$
 $P(B^{(t+1)} | P^{(t)}, B^{(t)}, F^{(t)}) (2-1) \cdot 2^3$
 $P(F^{(t+1)} | F^{(t)}) (2-1) \cdot 2$

2

Unrolled DBN

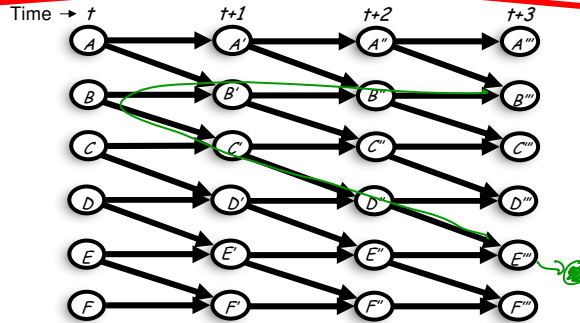
X O
 H, P, B, F R, S

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



"Sparse" DBN and fast inference

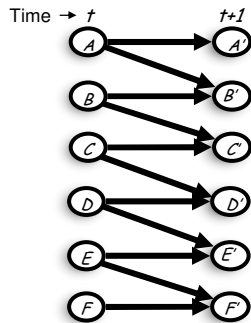
~~"Sparse" DBN → Fast inference~~



$A''' \perp E''$ true
 $B'' \perp E'''$ no!
 $A''' \perp E''$ no!!
 \vdots

Even after one time step!!

What happens when we marginalize out time t ?

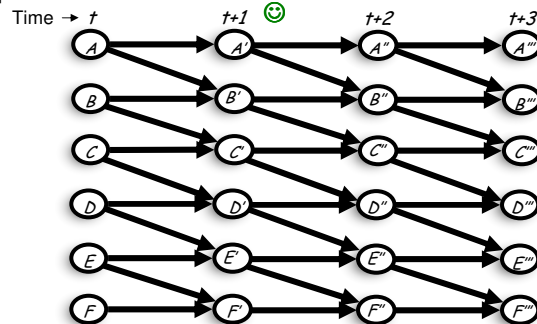


5

“Sparse” DBN and fast inference 2

Structured representation of belief often yields good approximate

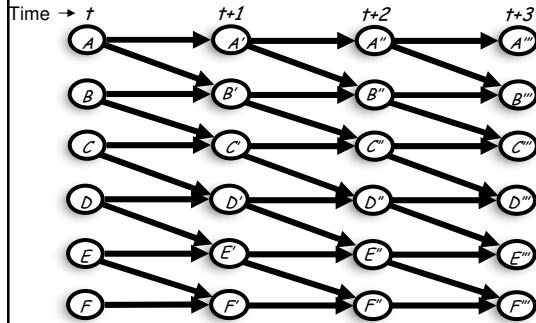
“Sparse” DBN $\xrightarrow{\text{Almost!}}$ Fast inference



6

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



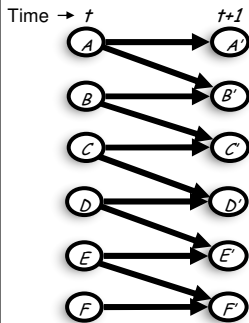
7

A simple example of BK: Fully-Factorized Distribution

- Assumed density:
 - Fully factorized

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

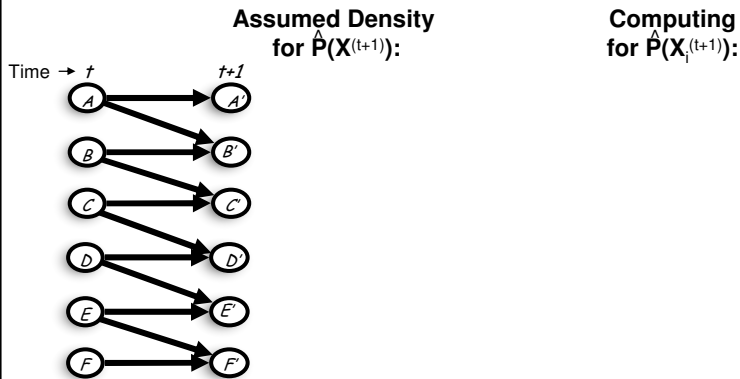
Assumed Density for $\hat{P}(X^{(t+1)})$:



8

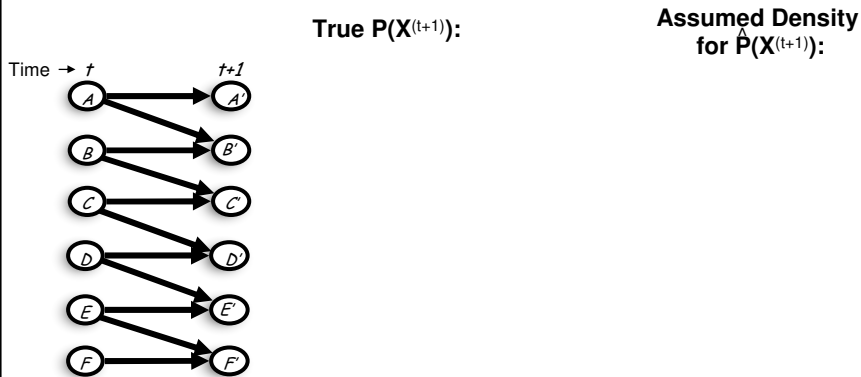
Computing Fully-Factorized Distribution at time $t+1$

- Assumed density:
 - Fully factorized



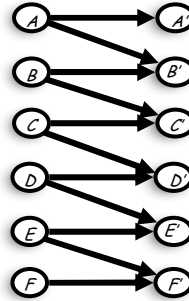
General case for BK: Junction Tree Represents Distribution

- Assumed density:
 - Fully factorized



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 $t \rightarrow \infty$

12

Contraction in Markov process

13

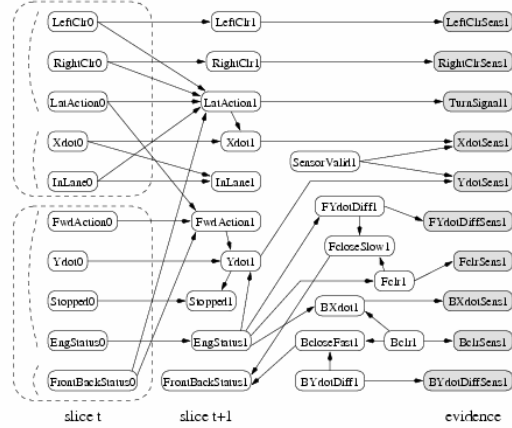
BK Theorem

- 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** .

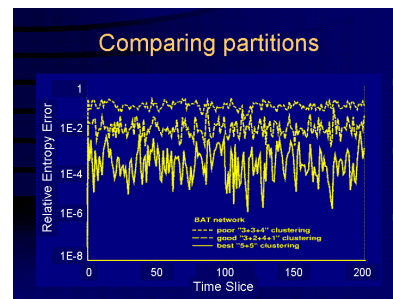
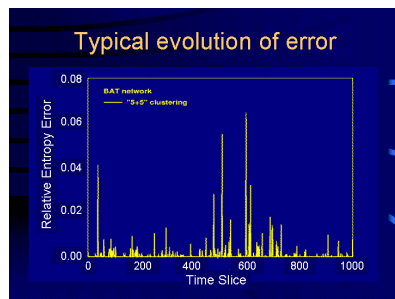
14

Example – BAT network [Forbes et al.]



15

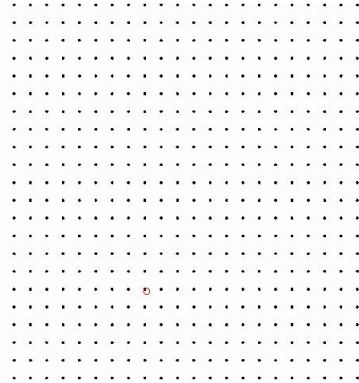
BK results [Boyen, Koller '98]



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

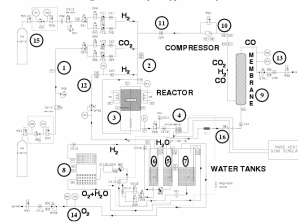


Figure 10.2: The RWGS schematic

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 not blow up (but could still be large)
- Thin junction tree filter adapts assumed density over time
- Extensions for hybrid DBNs

19

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, GIS, Gaussian and hybrid models, discrete and continuous variables, temporal and template models, Kalman filter, linearization, switching Kalman filter, assumed density filtering, DBNs, BK, Causality,...

■ **Just the beginning...** 😊

20

Quick overview of some hot topics...

- **Conditional Random Fields**
- **Maximum Margin Markov Networks**
- **Relational Probabilistic Models**
 - e.g., the parameter sharing model that you learned for a recommender system in HW1
- **Hierarchical Bayesian Models**
 - e.g., Khalid's presentation on Dirichlet Processes
- **Influence Diagrams**

21

Generative v. Discriminative models – Intuition

- **Want to Learn: $h: X \mapsto Y$**
 - X – features
 - Y – set of variables
- **Generative classifier**, e.g., Naïve Bayes, Markov networks:
 - Assume some **functional form for $P(X|Y)$, $P(Y)$**
 - Estimate parameters of **$P(X|Y)$, $P(Y)$** directly from training data
 - Use Bayes rule to calculate **$P(Y|X=x)$**
 - This is a '**generative**' model
 - **Indirect** computation of **$P(Y|X)$** through Bayes rule
 - But, **can generate a sample of the data**, **$P(X) = \sum_y P(y) P(X|y)$**
- **Discriminative classifiers**, e.g., Logistic Regression, Conditional Random Fields:
 - Assume some **functional form for $P(Y|X)$**
 - Estimate parameters of **$P(Y|X)$** directly from training data
 - This is the '**discriminative**' model
 - Directly learn **$P(Y|X)$** , can have **lower sample complexity**
 - But **cannot obtain a sample of the data**, because **$P(X)$** is not available

22

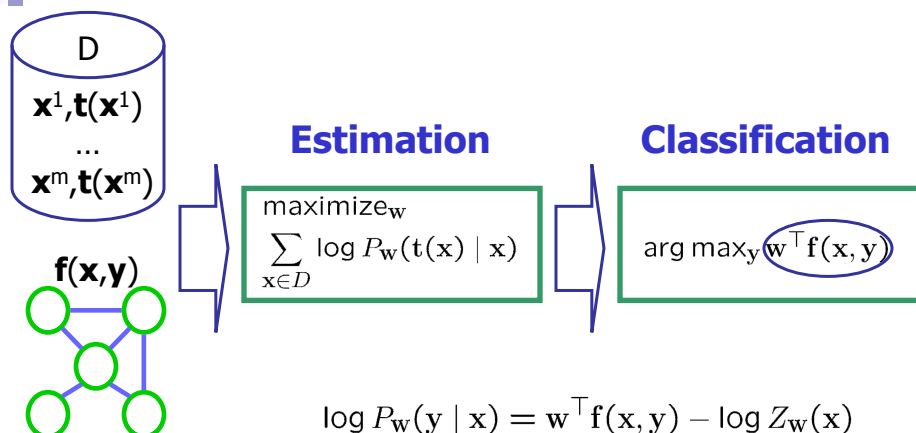
Conditional Random Fields

[Lafferty et al. '01]

- Define a Markov network using a log-linear model for $P(\mathbf{Y}|\mathbf{X})$:
- Features, e.g., for pairwise CRF:
- Learning: maximize conditional log-likelihood
 - sum of log-likelihoods you know and love...
 - learning algorithm based on gradient descent, very similar to learning MNs

23

Max (Conditional) Likelihood



Don't need to learn entire distribution!

24

OCR Example

- We want:

$$\operatorname{argmax}_{\text{word}} \mathbf{w}^T \mathbf{f}(\text{brace}, \text{word}) = \text{"brace"}$$

- Equivalently:

$$\left. \begin{aligned} \mathbf{w}^T \mathbf{f}(\text{brace}, \text{"brace"}) &> \mathbf{w}^T \mathbf{f}(\text{brace}, \text{"aaaaa"}) \\ \mathbf{w}^T \mathbf{f}(\text{brace}, \text{"brace"}) &> \mathbf{w}^T \mathbf{f}(\text{brace}, \text{"aaaab"}) \\ \dots \\ \mathbf{w}^T \mathbf{f}(\text{brace}, \text{"brace"}) &> \mathbf{w}^T \mathbf{f}(\text{brace}, \text{"zzzzz"}) \end{aligned} \right\} \text{a lot!}$$

25

Max Margin Estimation

- Goal: find \mathbf{w} such that

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

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

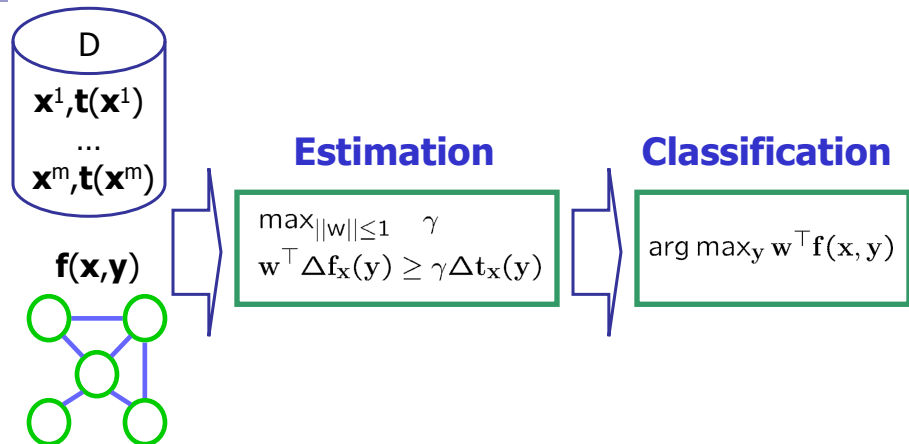
$$\mathbf{w}^T \Delta \mathbf{f}_{\mathbf{x}}(\mathbf{y}) \geq \gamma \Delta \mathbf{t}_{\mathbf{x}}(\mathbf{y})$$

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

$$\begin{aligned} \Delta \mathbf{t}_{\text{brace}}(\text{"craze"}) &= 2 & \Delta \mathbf{t}_{\text{brace}}(\text{"zzzzz"}) &= 5 \\ \mathbf{w}^T \Delta \mathbf{f}_{\text{brace}}(\text{"craze"}) &\geq 2\gamma & \mathbf{w}^T \Delta \mathbf{f}_{\text{brace}}(\text{"zzzzz"}) &\geq 5\gamma \end{aligned}$$

26

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



27

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?

28

Generalization requires Relational Models (e.g., see tutorial by Getoor)

- 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

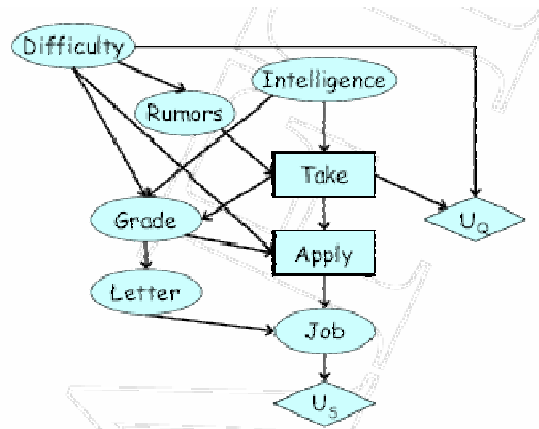
29

Reasoning about decisions K&F Chapters 20 & 21

- 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.,...

- **Active learning**
- **Non-parametric models**
- **Continuous time models**
- **Learning theory for graphical models**
- **Distributed algorithms for graphical models**
- **Graphical models for reinforcement learning**
- **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/cald/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)
- 10-801 Advanced Topics in Graphical Models: statistical foundations, approximate inference, and Bayesian methods (Spring)
- ...

33