School of Computer Science
Carnegie Mellon

Spectral Probabilistic Modeling and Applications to Natural Language Processing

Doctoral Thesis Proposal

Ankur Parikh

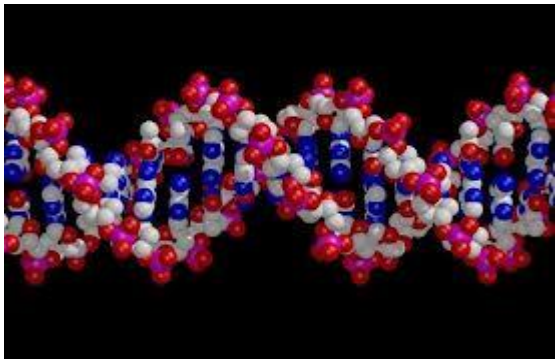
Thesis Committee:

Eric Xing, Geoff Gordon, Noah Smith, Le Song, Ben Taskar

Probabilistic Modeling

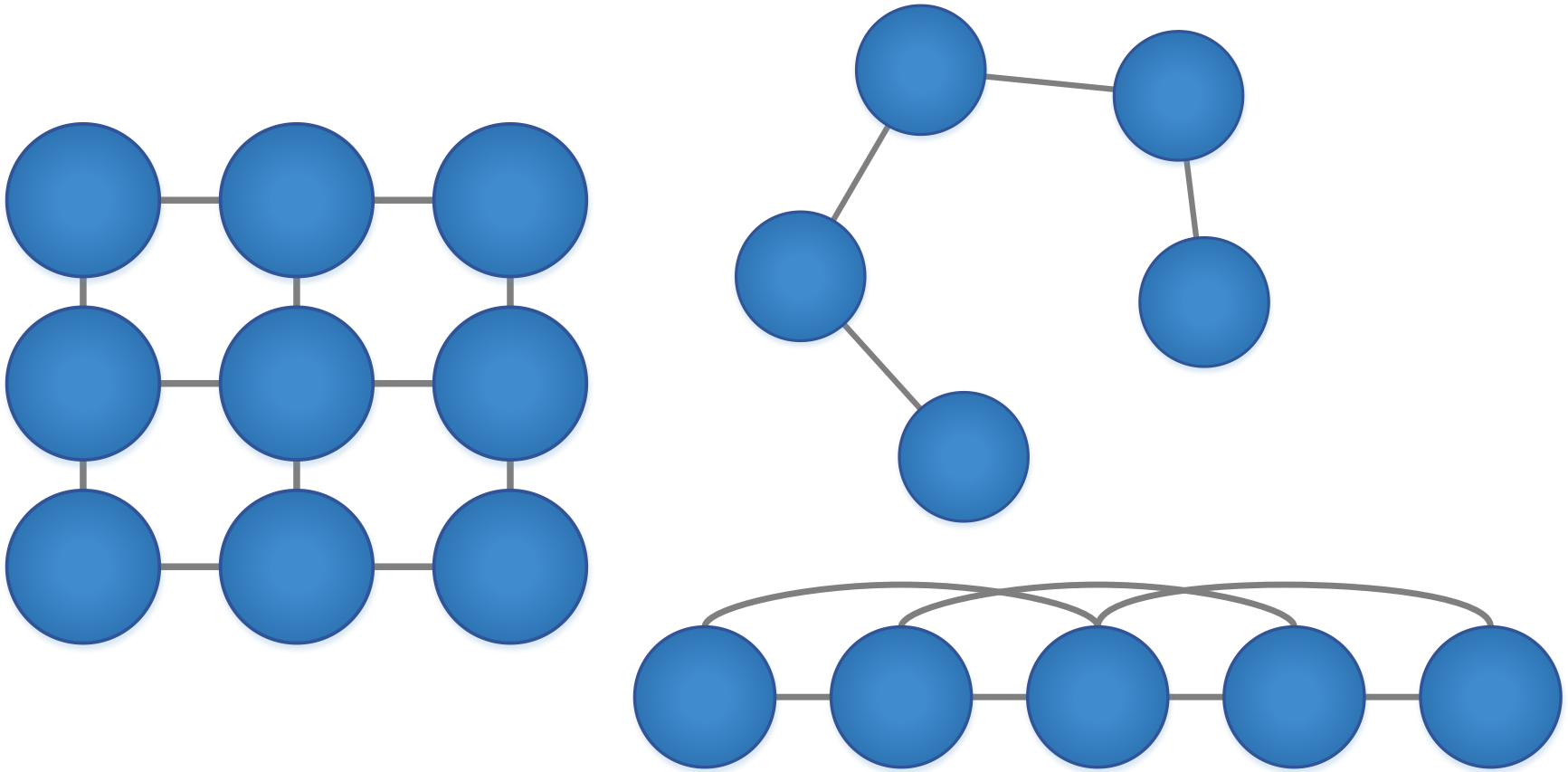
- Modeling uncertainty among a set of random variables.

- Key to advances many aspects of AI: natural language processing, bioinformatics, computer vision



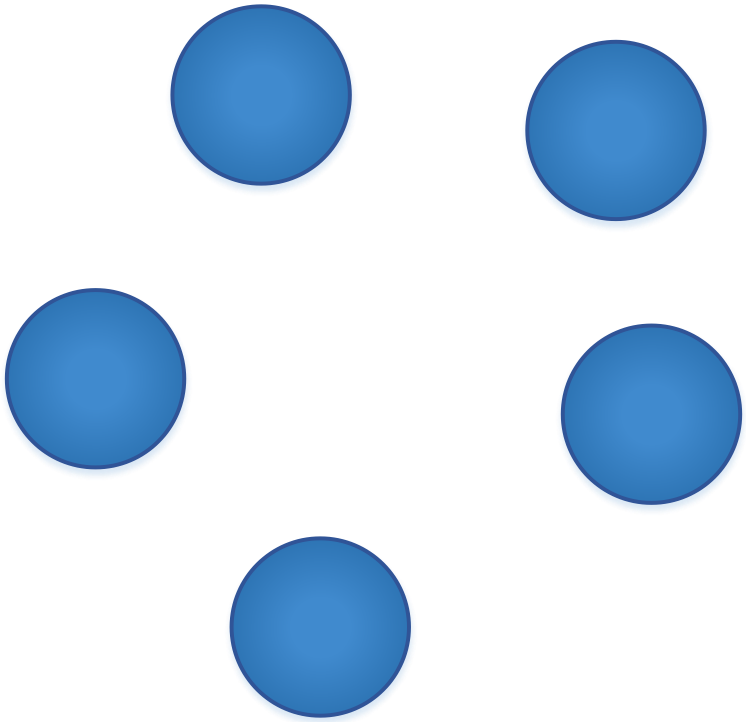
Probabilistic Graphical Models

- Sophisticated formalism for probabilistic modeling



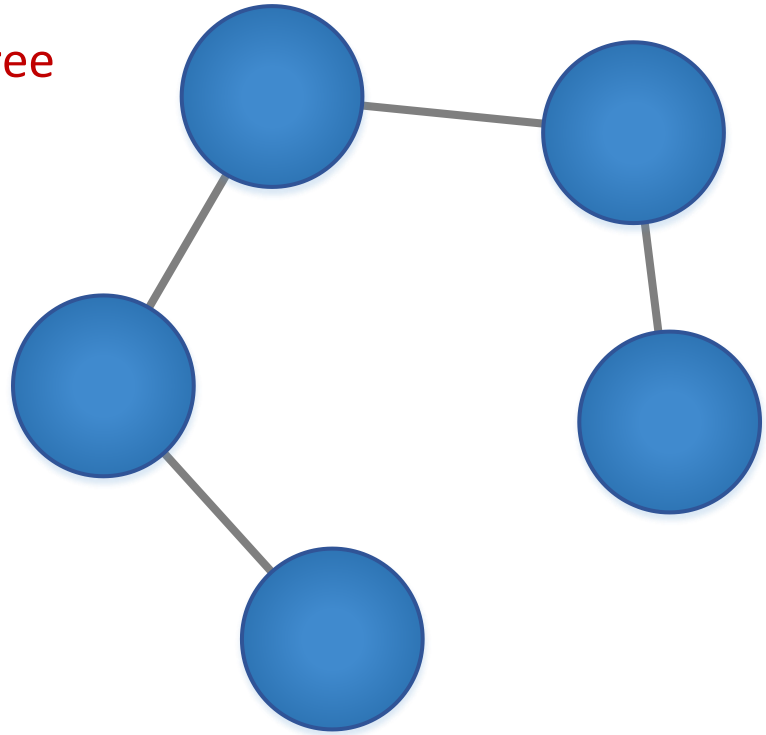
Key Aspects of Graphical Models

- Structure Learning



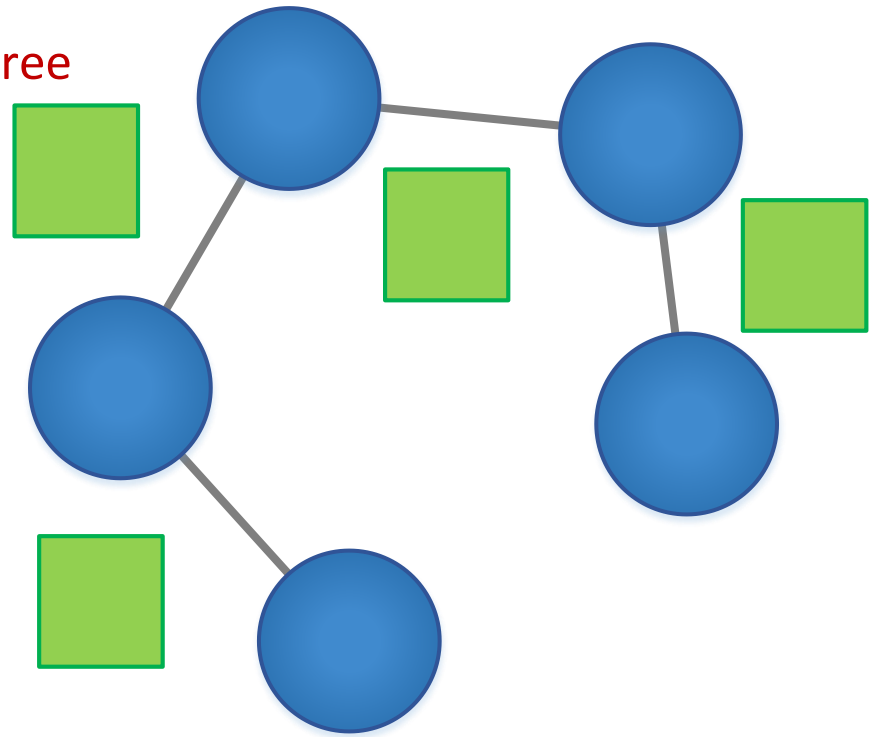
Key Aspects of Graphical Models

- Structure Learning
  - Chow Liu Algorithm can find optimal tree



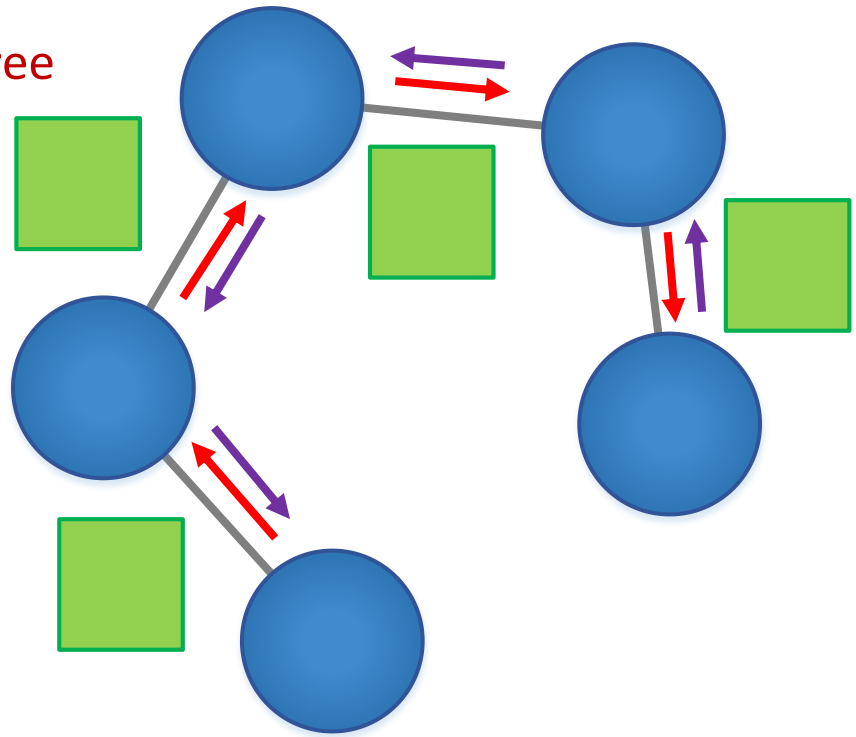
Key Aspects of Graphical Models

- Structure Learning
  - Chow Liu Algorithm can find optimal tree
- Parameter Learning
  - Max Likelihood



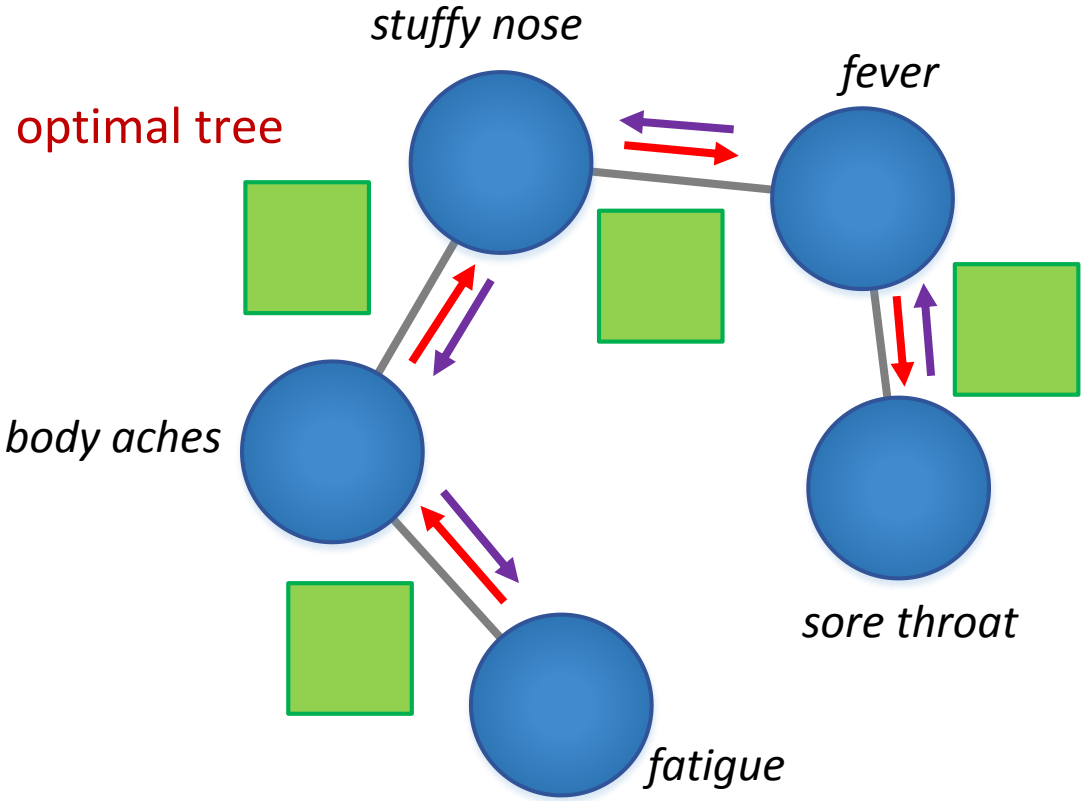
Key Aspects of Graphical Models

- Structure Learning
  - Chow Liu Algorithm can find optimal tree
- Parameter Learning
  - Max Likelihood
- Inference
  - Belief Propagation

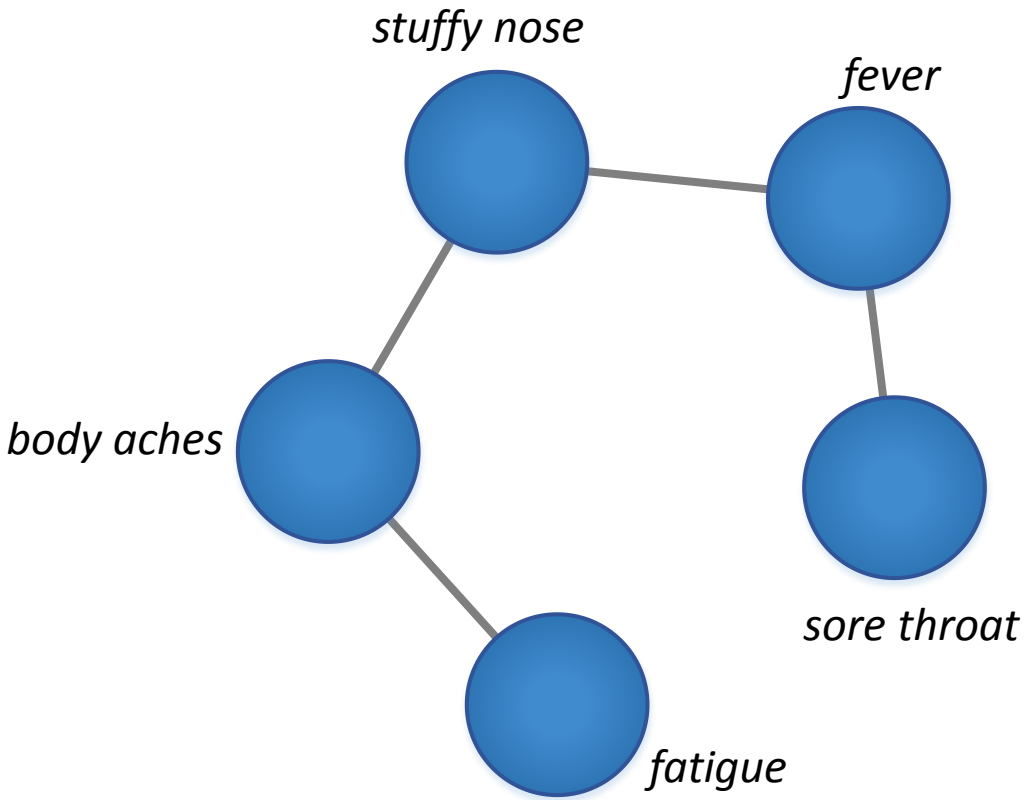


Key Aspects of Graphical Models

- Structure Learning
  - Chow Liu Algorithm can find optimal tree
- Parameter Learning
  - Max Likelihood
- Inference
  - Belief Propagation
- Modeling
  - How to use these tools to model real world phenomena

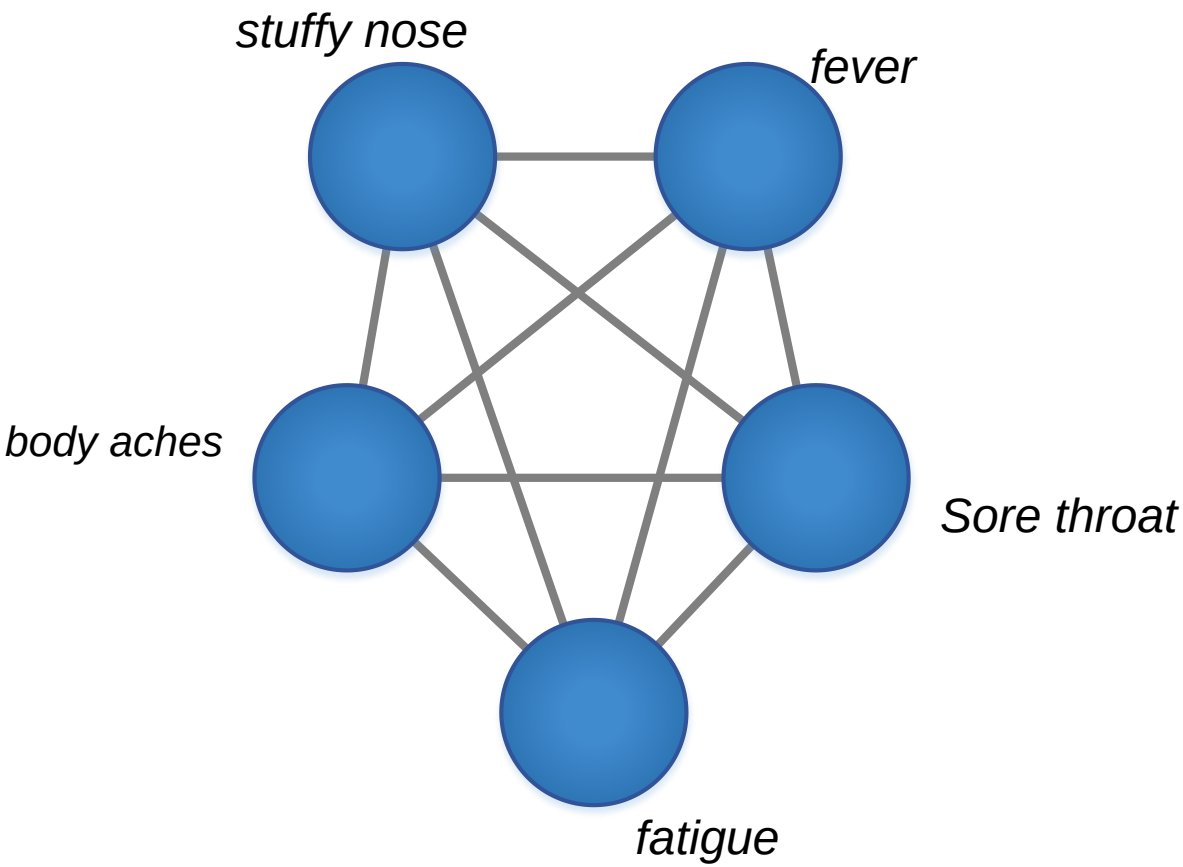


Is This A Good Model?



Not really, they are all dependent on one another regardless of what we condition on.

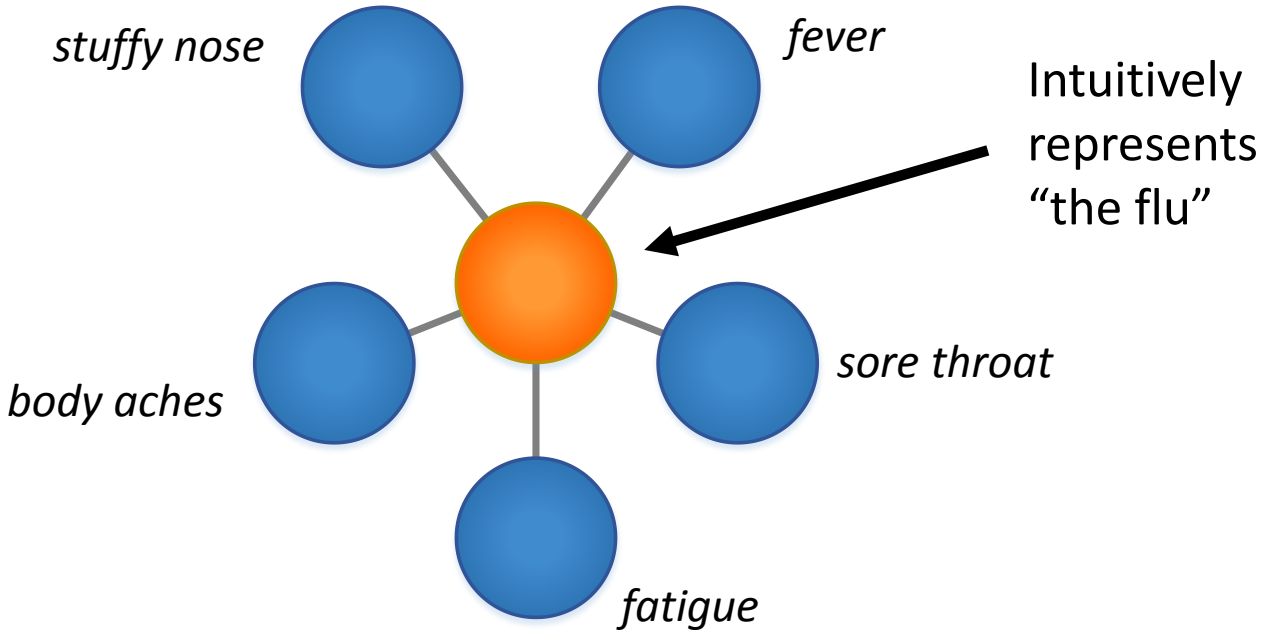
But This Model Seems Overly Complex



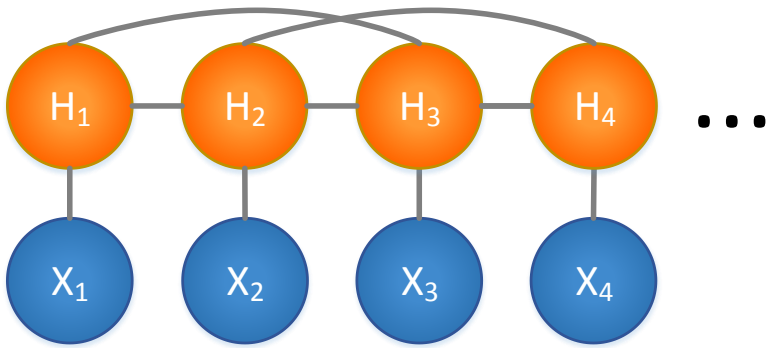
Latent Variables Can Help!

- There is a simpler explanation underlying the data.

- All the symptoms are conditionally independent given this explanation.

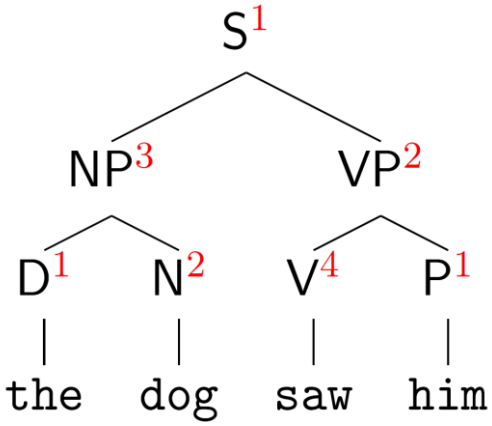


Latent Variable Models

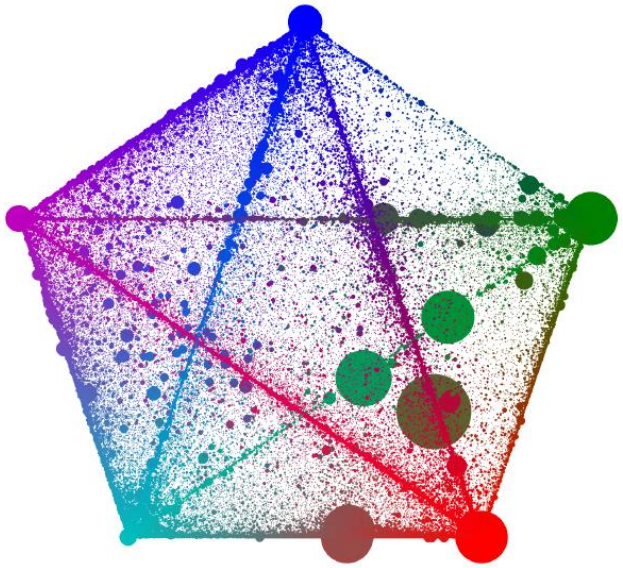


Sequence models

Parsing



Cohen et al. 2012

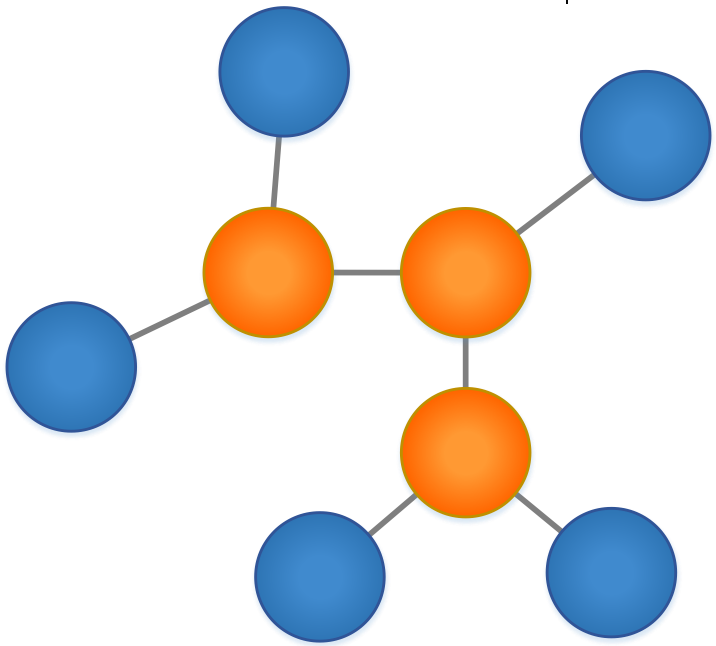


Ho. et al. 2012

Mixed membership models

But Learning And Inference Become Harder

- Structure Learning
  - Cannot compute mutual information among latents
- Parameter Learning
  - Likelihood is no longer convex
- Inference
  - marginal MAP is hard even for trees



Existing Approaches

- Most existing approaches formulate problem as nonconvex optimization (e.g. Expectation Maximization)

- From Computational Perspective
  - Can get trapped in local optima
  - Slow to converge
  - Hard to parallelize

- From Modeling Perspective
  - Unclear what the relationship is between model and difficulty of learning/inference problem

Linear Algebra and Graphical Models

- Approach these problems through the lens of linear algebra
- Latent variable models correspond to low rank factorizations of the observed variables
- Leverage linear algebra tools to derive new solutions
  - Rank
  - SVD / Matrix Factorization
  - Eigenvalues
  - Tensors

Previous Work

- Our work is inspired by recent theoretical results from different communities
  - Theoretical Computer Science (Dasgupta, 1999)
  - Dynamical Systems (Katayama, 2005)
  - Phylogenetics (Mossel and Roch 2005)
  - Statistical Machine Learning (Hsu et al. 2009, Bailly 2009)

- Spectral Learning Algorithms for HMMs
  - Hsu et al. 2009, Bailly 2009 (Spectral HMMs)
  - Siddiqi et al. 2009 (Reduced Rank HMMs)
  - Song et al. 2010 (Kernel HMMs)

(Informal) Thesis Statement

- Take a more general graphical models point of view.
- Aim to approach the key problems of **structure learning**, **parameter learning**, and **inference** from the linear algebra perspective.
- Use these insights to design new models and solutions for problems in Natural Language Processing
  - **Unsupervised Parsing**
  - **Language Models**

Proposed Contributions (ML)

- Spectral Parameter Learning for Latent Trees / Junction Trees
  - Key Theme: Parameter Learning
- Kernel Embeddings for Latent Trees
  - Key Themes: Parameter/Structure Learning
- Spectral Approximations for Inference (e.g. marginal MAP / collapsed sampling)
  - Key Theme: Inference

Proposed Contributions (NLP)

- A Conditional Latent Tree Model for Unsupervised Parsing
  - Key Themes: Structure Learning / Modeling

- *N*-gram language modeling with non-integer *n*
  - Key Theme: Modeling

Other Concurrent Work to this Thesis

- Predictive state representations (Boots et al., 2010/2013)
- Topic models (Anandkumar et al., 2012/2013)
- Latent variable supervised parsing (Luque et al., 2012, Cohen et al., 2012, Dhillon et al. 2012, Cohen and Collins, 2013)
- Spectral Learning via Local Loss Optimization (Balle et al. 2012)
- Word embeddings (Dhillon et al. 2011/2012)

Outline

- Linear Algebra View of Latent Variable Models
- Spectral Parameter Learning
- Spectral Approximations for Inference
- Language Modeling

(Kernels and unsupervised parsing omitted due to time)

Linear Algebra View of Latent Variable Models

Important Notation

- Probabilities

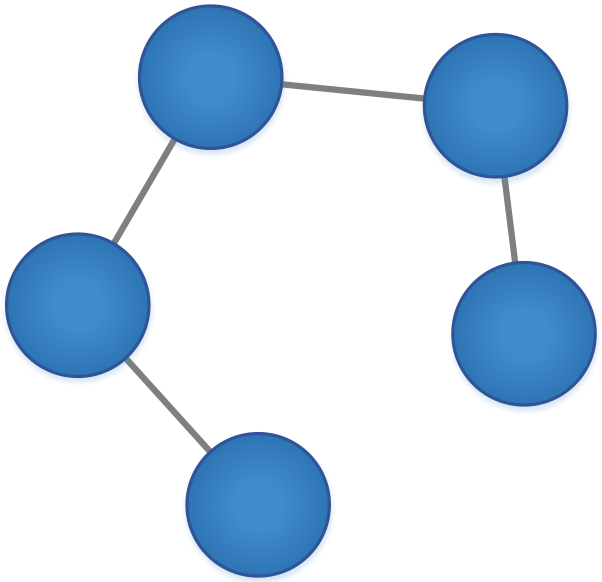
$$P(X,Y) = P(Y\,|X)\,P(X)$$

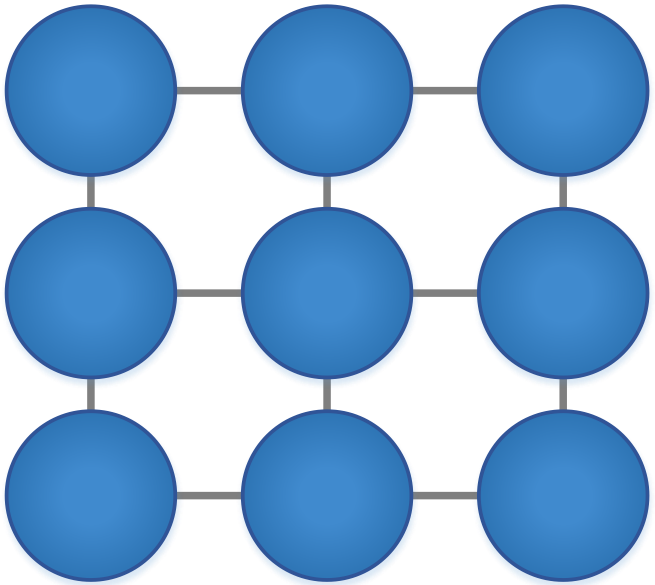
- Probability Vectors/Matrices/Tensors

$$\mathcal{P}(Y) = \mathcal{P}(Y\,|X)\mathcal{P}(X)$$

Traditional View Of Graphical Models

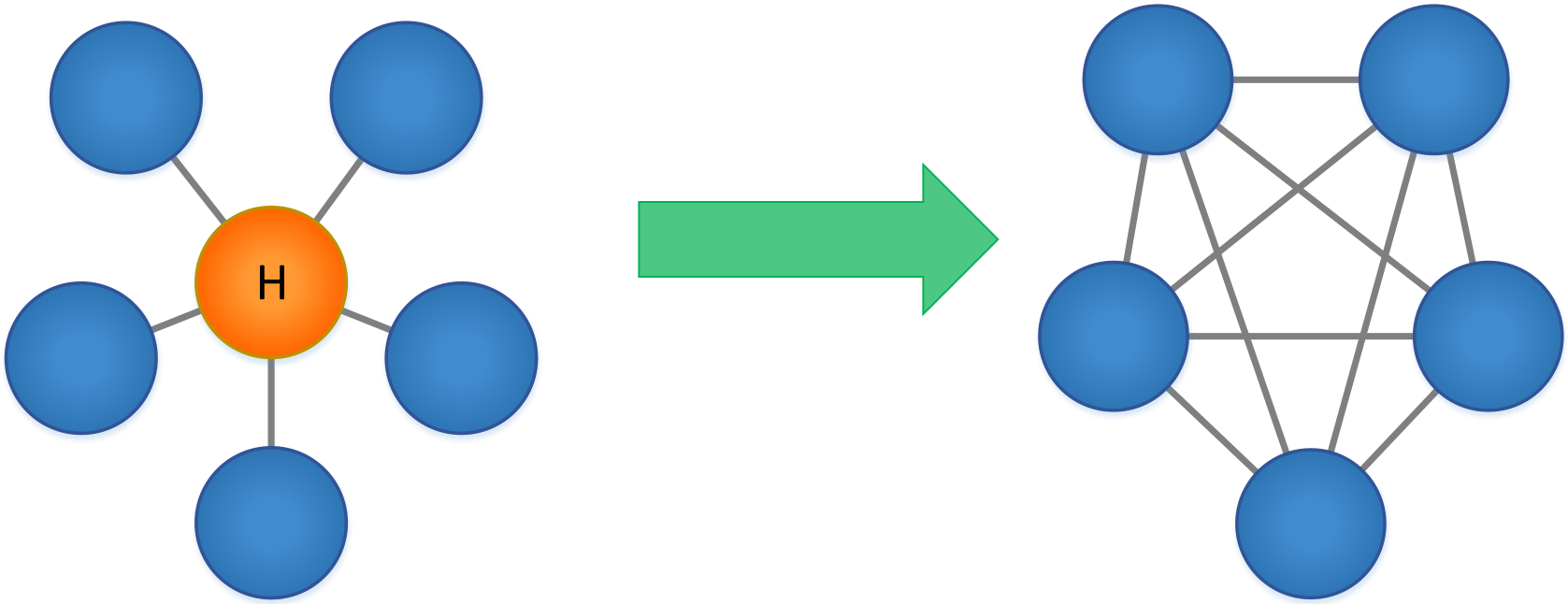
- Complexity of mostly characterized by structure or **treewidth**





Shortcoming Of This View

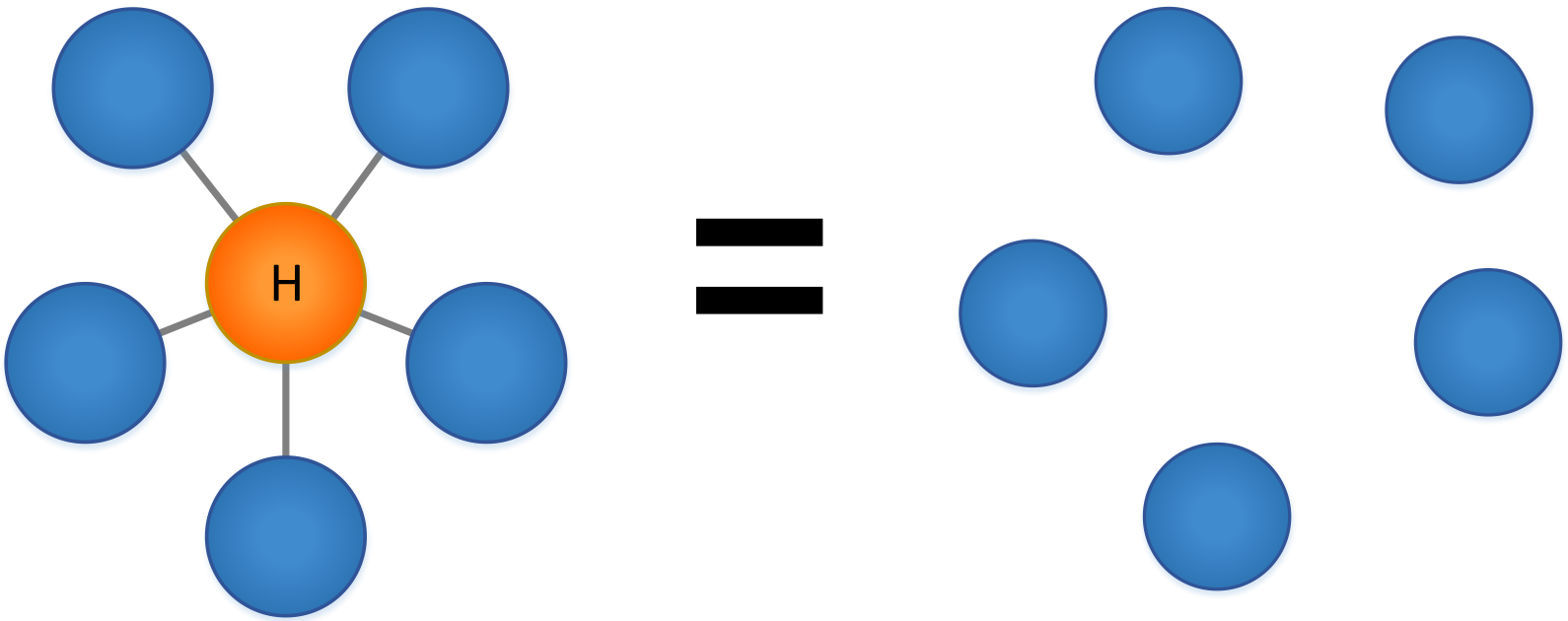
- Marginalizing out latent variable leads to a clique



- But aren't these models different?

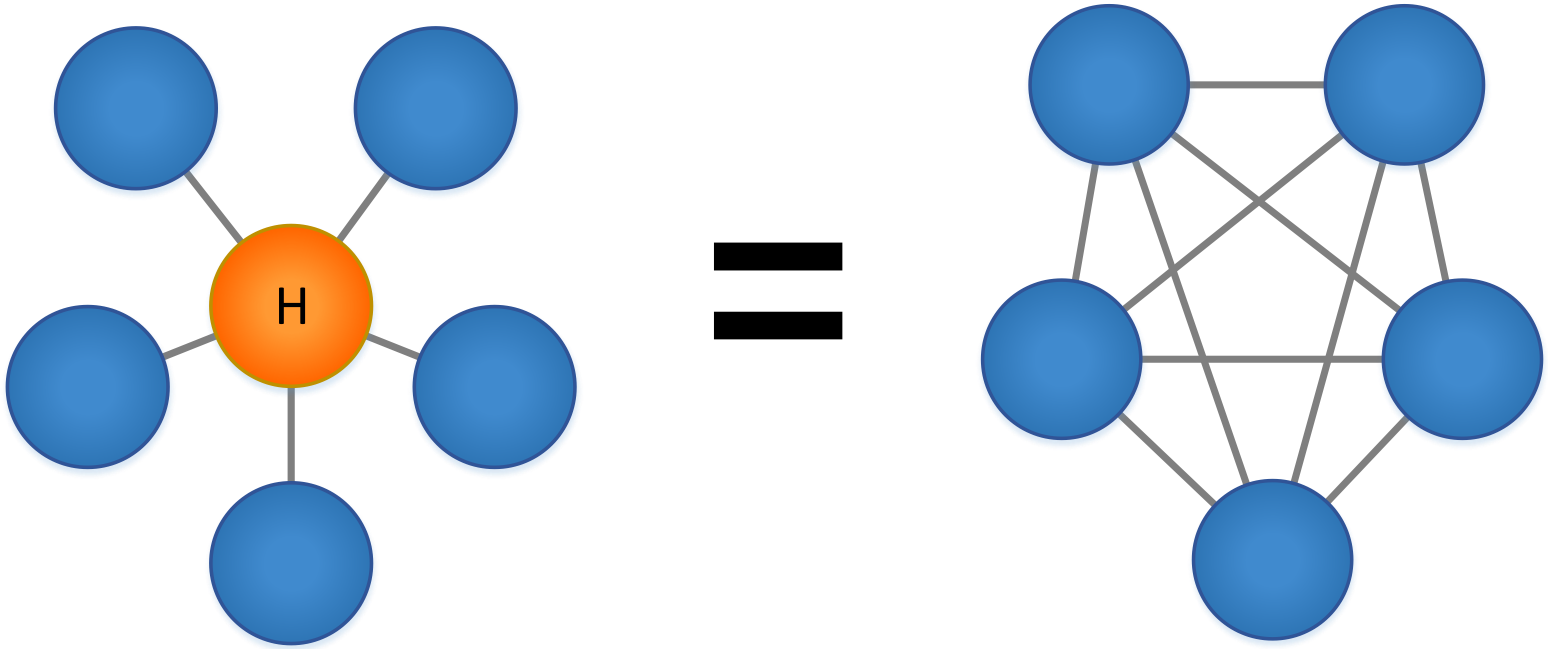
It Depends on The Number Of States

- If $\boldsymbol{H}$ takes on only one state then the variables are independent.



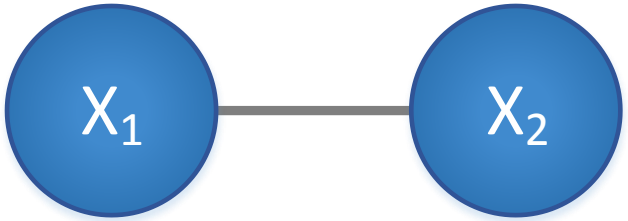
It Depends on the Number of States

- Assume observed variables take $\boldsymbol{m}$ states
- If $\boldsymbol{H}$ takes on $\boldsymbol{m}^{\mathbf{5}}$ states then model can be equivalent to clique



- But what about all the other cases?

Graphical Models: The Linear Algebra View



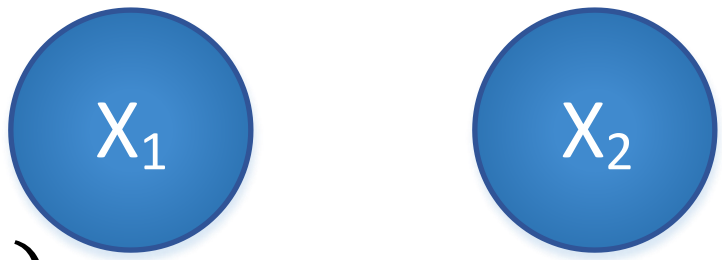
X_1 and X_2 have m states each.

$$\mathcal{P}(X_1, X_2)$$

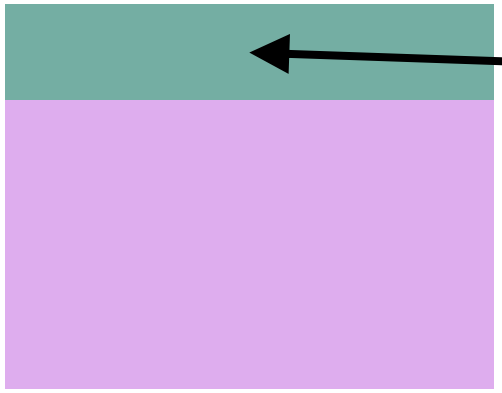
- In general, nothing we can say about the nature of this matrix.

Independence: The Linear Algebra View

- What if we know $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are independent?



$$\mathcal{P}(X_1, X_2)$$

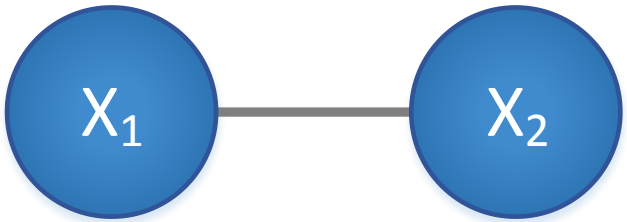


$$[\, P(X_1 = 1, X_2 = 1), \ldots, P(X_1 = 1, X_2 = m)\,]$$

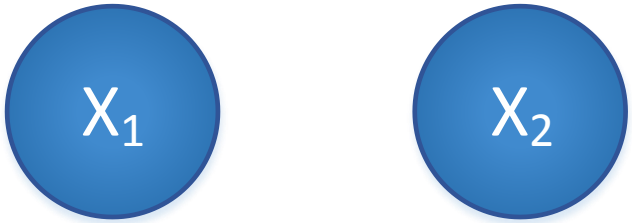
$$= [\, P(X_1 = 1)\,(P(X_2 = 1), \ldots, P(X_2 = m)\,)]$$

- Matrix is rank one!

Independence and Rank



$\mathcal{P}(X_1, X_2)$ **has rank m (at most)**

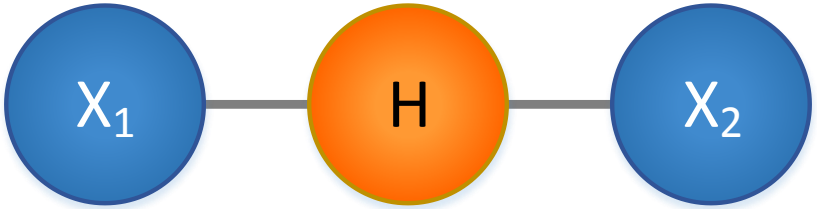


$\mathcal{P}(X_1, X_2)$ **has rank 1**

- What about rank in between ***1*** and ***m***?

Low Rank Structure

- $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are not marginally independent (They are only conditionally independent given $\boldsymbol{H}$).

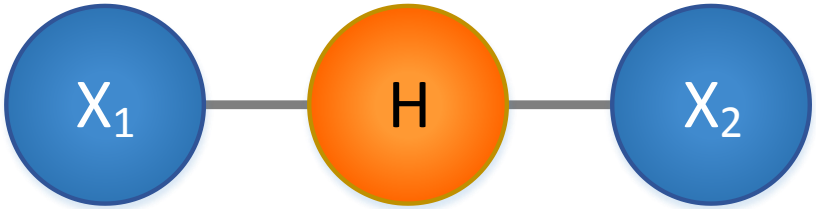


- Assume $\boldsymbol{H}$ has $\boldsymbol{k}$ states

- Then, $\boldsymbol{rank}(\mathcal{P}(\boldsymbol{X_1}, \boldsymbol{X_2})) \leq \boldsymbol{k}$

- Why?

Low Rank Structure



$\mathfrak{P}(X_1, X_2)$ $\mathfrak{P}(X_1 \mid H)$ $\mathfrak{P}(\oslash\, H)$ $\mathfrak{P}(X_2 \mid H)^{\top}$

$=$

$\boldsymbol{rank \leq k}$ $\boldsymbol{rank \leq k}$ $\boldsymbol{rank \leq k}$ $\boldsymbol{rank \leq k}$

The Linear Algebra View

- Latent variable models encode low rank dependencies among variables *(both marginal and conditional)*

- Use tools from linear algebra to exploit this structure.
  - Rank
  - Eigenvalues
  - SVD / Matrix Factorization
  - Tensors

Spectral Parameter Learning

Related Papers

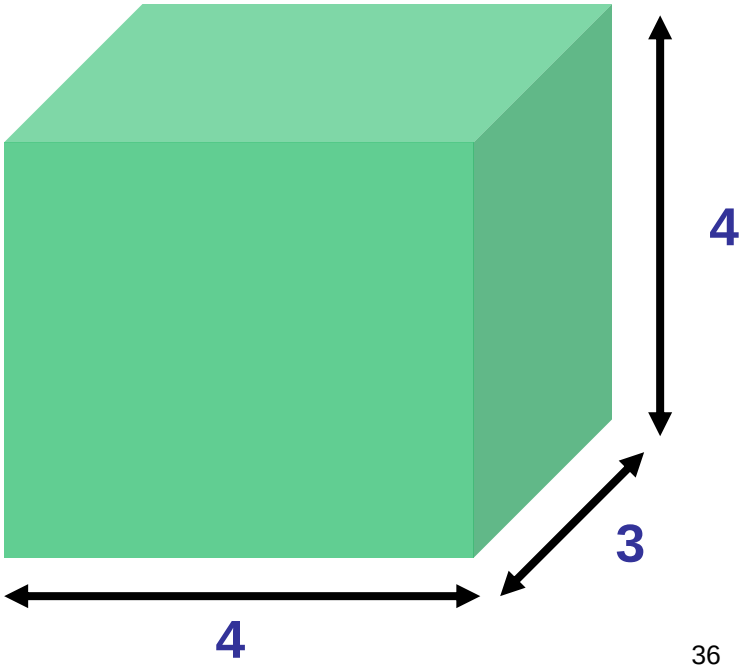
- **A.P. Parikh**, L. Song, E.P. Xing, *A Spectral Algorithm for Latent Tree Graphical Models* (ICML 2011)
- L.Song, **A.P. Parikh**, E.P. Xing, *Kernel Embeddings of Latent Tree Graphical Models* (NIPS 2011)
- **A.P. Parikh**, L. Song, M. Ishteva, G. Teodoru, E.P. Xing, *A Spectral Algorithm for Latent Junction Trees* (UAI 2012)
- L. Song, M. Ishteva, **A.P. Parikh**, H. Park, and E.P. Xing, *Hierarchical Tensor Decomposition for Latent Tree Graphical Models* (ICML 2013)

Tensors

- Multidimensional arrays
- A Tensor of order $\boldsymbol{N}$ has $\boldsymbol{N}$ modes ($\boldsymbol{N}$ indices):

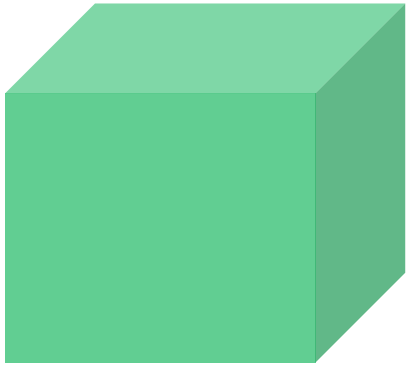
$$\mathcal{T}(i_1, \ldots, i_n)$$

- Each mode is associated with a dimension. In the example,
  - Dimension of mode 1 is 4
  - Dimension of mode 2 is 3
  - Dimension of mode 3 is 4



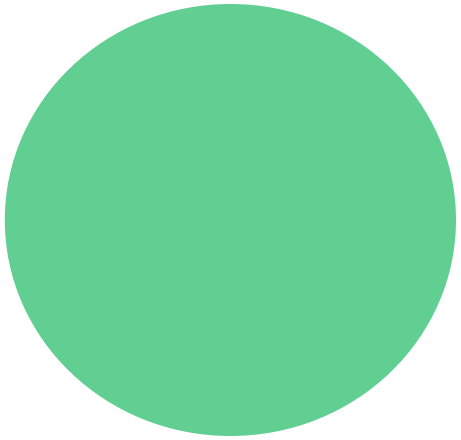
Tensors

1st order tensor (vector)



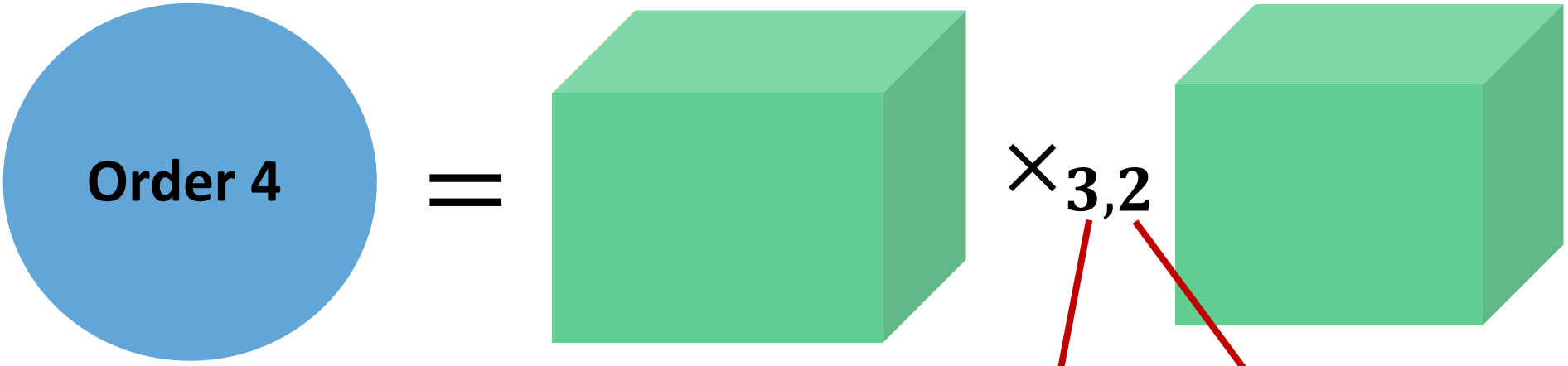
3rd order tensor (cube)

2nd order tensor (matrix)



Higher order tensors

Tensor Tensor Multiplication



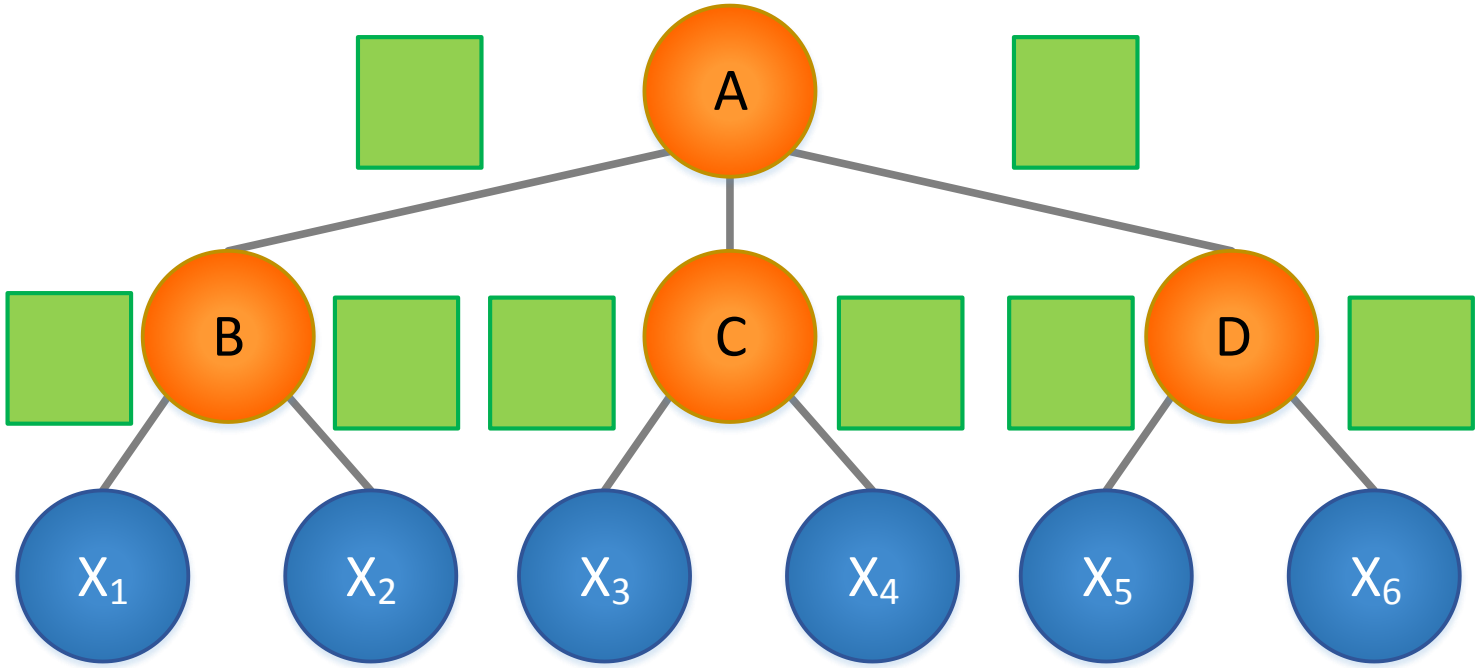
$$\mathcal{R}(i,j,k,l) = \sum_{m} \mathcal{T}_1(i,j,m)\mathcal{T}_2(k,m,l)$$

Tensor Tensor Multiplication

- Matrix multiplication is closed (i.e. product of matrices is a matrix).

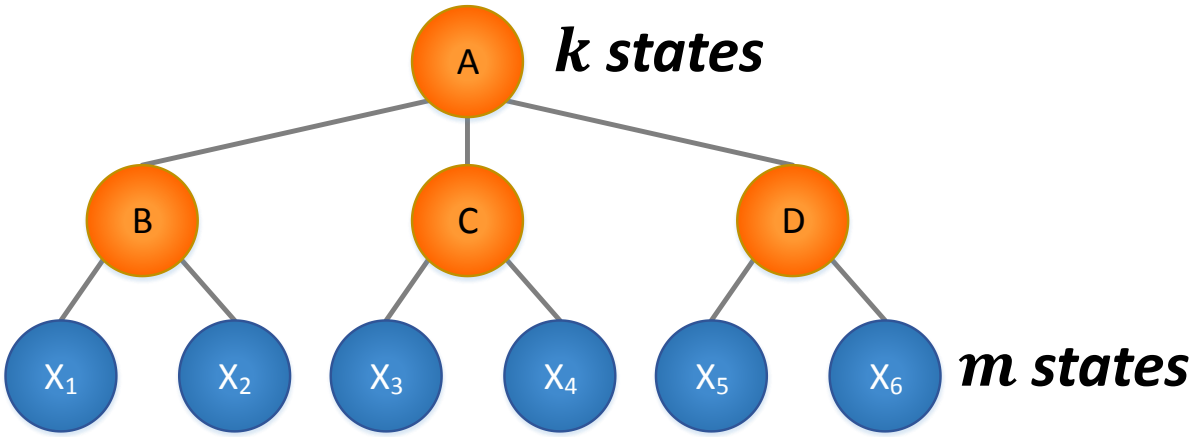
- Tensor multiplication is not closed (e.g. product of two tensors of order $\boldsymbol{n}$ is a tensor of order $\mathbf{2}\boldsymbol{n} - \mathbf{2}$)

Parameter Learning in Latent Trees

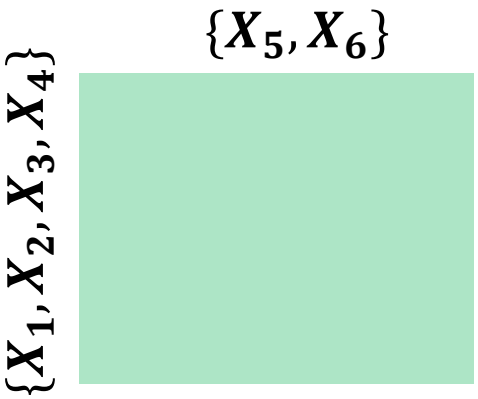


- Conditional probability tables depend on latent variables
- Common approach is Expectation Maximization (EM)
  - Local optima
  - Slow to converge

Low Rank Perspective



$$\mathcal{P}\left(X_{\{1,2,3,4\}}, X_{\{5,6\}}\right)$$



has rank k

Low Rank Matrices "Factorize"

$$M = LR$$

If M has rank k

m by n m by k k by n

We already know one factorization!!!

$$\mathfrak{P}\left(X_{\{1,2,3,4\}}, X_{\{5,6\}}\right) = \mathfrak{P}\left(X_{\{1,2,3,4\}} \mid A\right) \mathfrak{P}(\emptyset\ A)\, \mathfrak{P}\left(X_{\{5,6\}} \mid A\right)^{\mathrm{T}}$$

Factor of 6 variables Factor of 5 variables Factor of 3 variables

Factor of 1 variable

Alternate Factorizations

- The key insight is that this factorization is not unique.

- Consider Matrix Factorization. Can add any invertible transformation:

$$\boldsymbol{M} = \boldsymbol{L}\boldsymbol{R}$$
$$\boldsymbol{M} = \boldsymbol{L}\boldsymbol{S}\,\boldsymbol{S}^{-1}\boldsymbol{R}$$

- **The magic of spectral learning is that there exists an alternative factorization that only depends on observed variables!**

An Alternate Factorization

- Let us say we only want to factorize this matrix of 6 variables

$$\mathcal{P}\left(X_{\{1,2,3,4\}}, X_{\{5,6\}}\right)$$

such that it is product of matrices that contain at most five ***observed*** variables e.g.

$$\mathcal{P}\left(X_{\{1,2,3,4\}}, X_{\{5\}}\right)$$

$$\mathcal{P}\left(X_{\{4\}}, X_{\{5,6\}}\right)$$

An Alternate Factorization

- Note that

$$\mathfrak{P}\left(X_{\{1,2,3,4\}}, X_{\{5\}}\right) = \underline{\mathfrak{P}\left(X_{\{1,2,3,4\}} \mid A\right)}\ \underline{\mathfrak{P}(\emptyset\ A)}\ \underline{\mathfrak{P}\left(X_{\{5\}} \mid A\right)}^{T}$$

$$\mathfrak{P}\left(X_{\{4\}}, X_{\{5,6\}}\right) = \underline{\mathfrak{P}\left(X_{\{4\}} \mid A\right)}\ \underline{\mathfrak{P}(\emptyset\ A)}\ \underline{\mathfrak{P}\left(X_{\{5,6\}} \mid A\right)}^{T}$$

- Product of green terms (in some order) is

$$\mathfrak{P}\left(X_{\{1,2,3,4\}}, X_{\{5,6\}}\right)$$

- Product of red terms (in some order) is

$$\mathfrak{P}\left(X_{\{4\}}, X_{\{5\}}\right)$$

An Alternate Factorization

$$\underbrace{\mathfrak{P}\left(X_{\{1,2,3,4\}}, X_{\{5,6\}}\right)}_{\text{factor of 6 variables}} = \underbrace{\mathfrak{P}\left(X_{\{1,2,3,4\}}, X_{\{5\}}\right)}_{\text{factor of 5 variables}} \underbrace{\mathfrak{P}(X_4, X_5)^{-1}\, \mathfrak{P}\left(X_{\{4\}}, X_{\{5,6\}}\right)}_{\text{factor of 3 variables}}$$

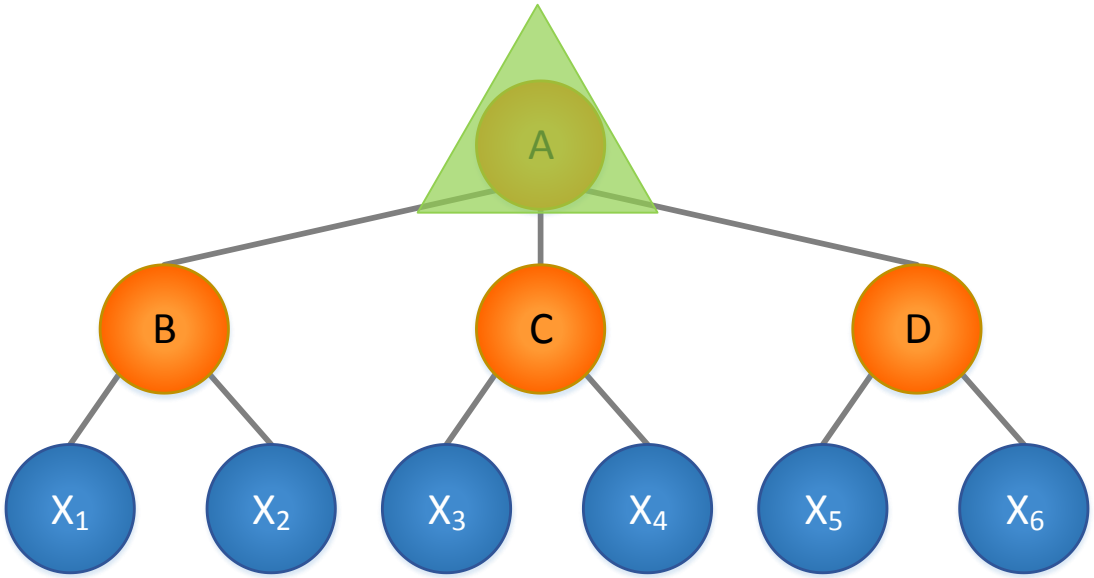
Advantage: Factors are only functions of observed variables! Can be directly computed from data without EM!!!!

Caveat: Factors are no longer probability tables

What Low Rank Relationship Did We Use

$$\mathfrak{P}\left(X_{\{1,2,3,4\}}, X_{\{5\}}\right) = \mathfrak{P}\left(X_{\{1,2,3,4\}} \mid A\right) \mathfrak{P}(\emptyset\ A)\, \mathfrak{P}\left(X_{\{5\}} \mid A\right)$$

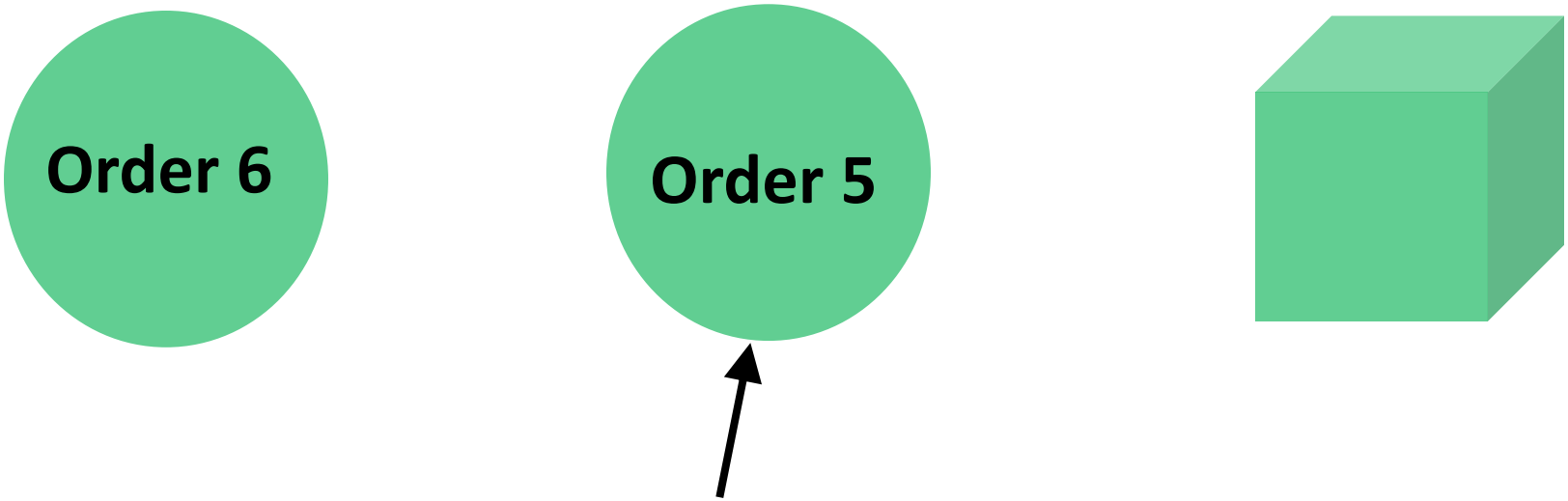
$$\mathfrak{P}\left(X_{\{4\}}, X_{\{5,6\}}\right) = \mathfrak{P}\left(X_{\{4\}} \mid A\right) \mathfrak{P}(\emptyset\ A)\, \mathfrak{P}\left(X_{\{5,6\}} \mid A\right)$$



The Tensor Point Of View

$$\mathcal{P}\left(X_{\{1,2,3,4\}}, X_{\{5,6\}}\right) = \mathcal{P}\left(X_{\{1,2,3,4\}}, X_{\{5\}}\right)\mathcal{P}(X_4, X_5)^{-1}\,\mathcal{P}\left(X_{\{4\}}, X_{\{5,6\}}\right)$$

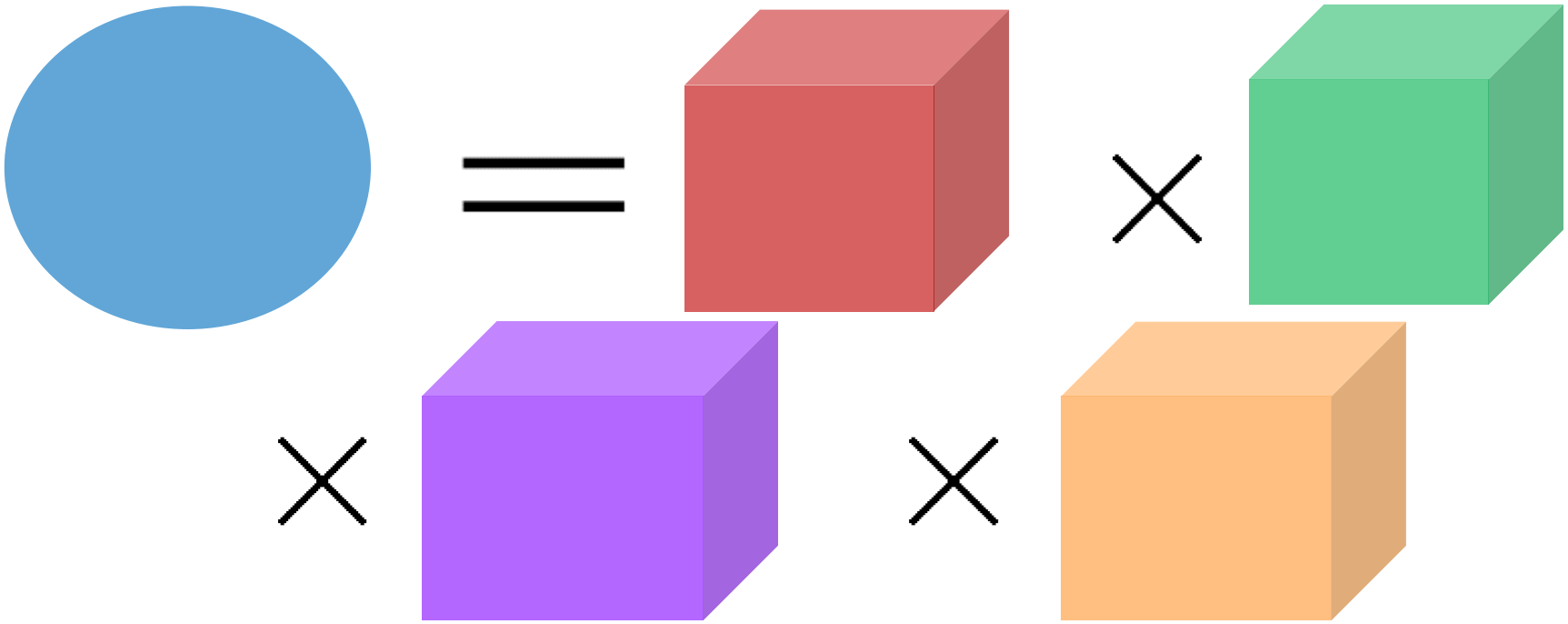
$$\mathcal{P}(X_1, X_2, X_3, X_4, X_5, X_6) = \underline{\mathcal{P}(X_1, X_2, X_3, X_4, X_5)} \times_5 \underline{\mathcal{P}(X_4, X_5)^{-1} \times_4 \mathcal{P}(X_4, X_5, X_6)}$$



Decompose recursively!!

Observable Factorization

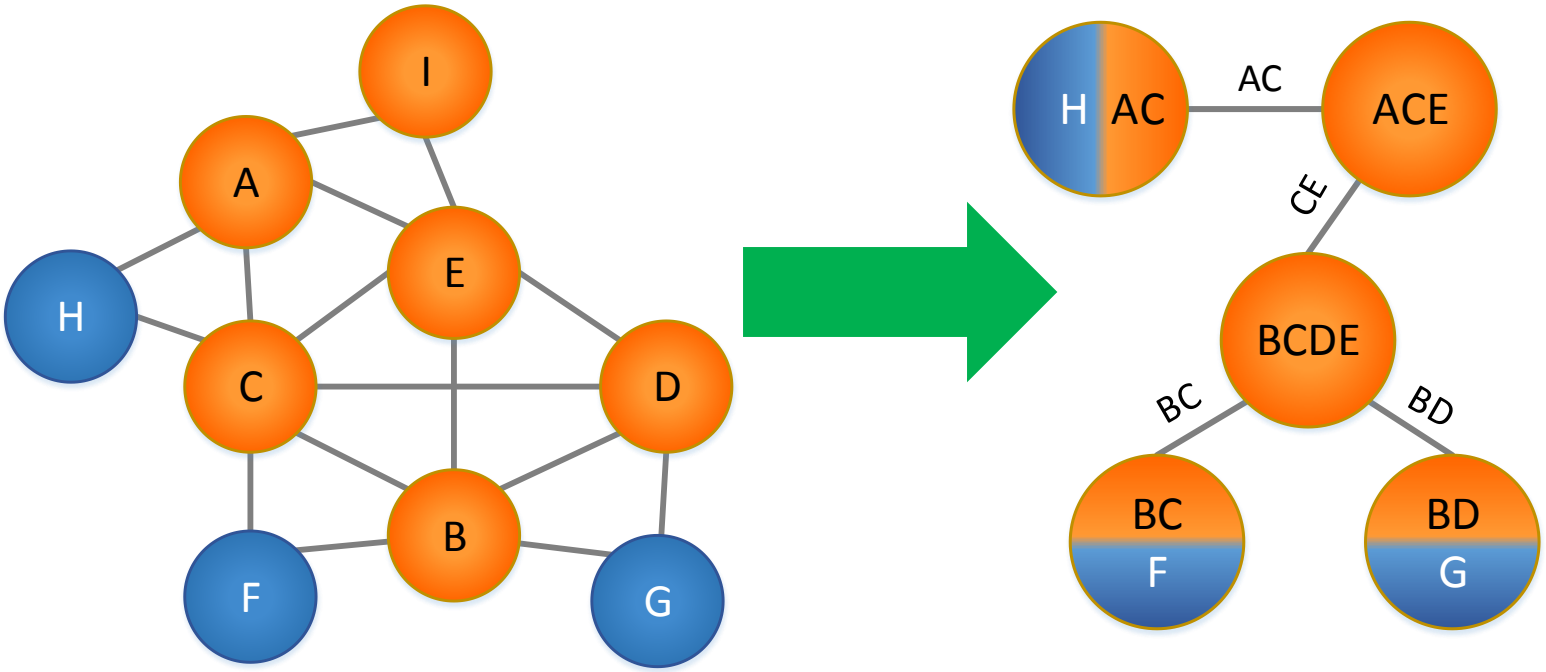
$\mathcal{P}(X_1, X_2, X_3, X_4, X_5, X_6)$ $\mathcal{P}(X_1, X_3, X_5)$ $\mathcal{P}(X_1, X_2, X_3) \times_3 \mathcal{P}(X_1, X_3)^{-1}$



$\mathcal{P}(X_3, X_4, X_5) \times_5 \mathcal{P}(X_3, X_5)^{-1}$ $\mathcal{P}(X_4, X_5, X_6) \times_4 \mathcal{P}(X_4, X_5)^{-1}$

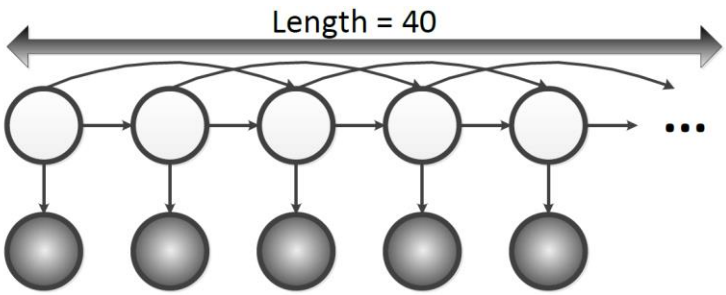
Also Works For Junction Trees

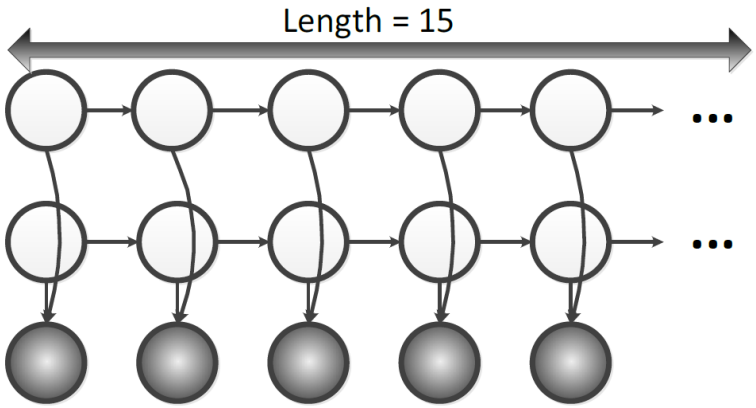
- Need to use higher order tensor operations since separator sets have size > 1
  - Multi-mode multiplication
  - Tensor inversion

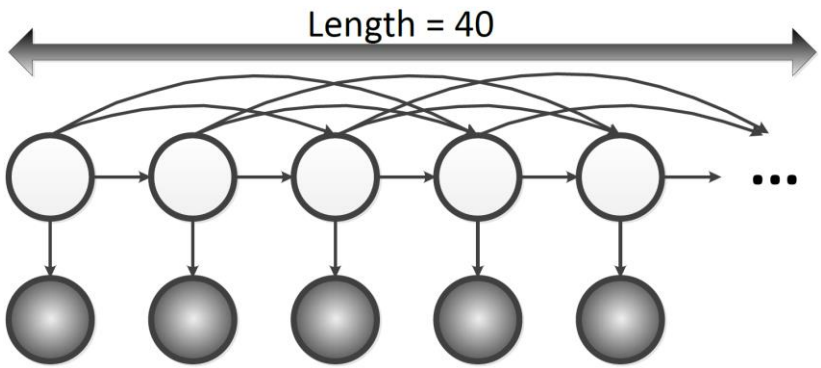


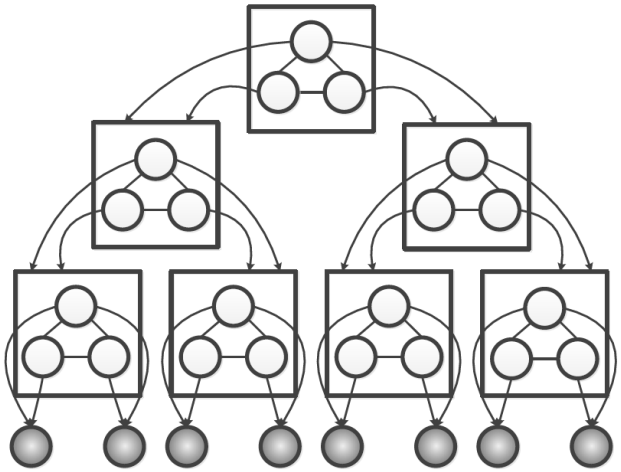
Empirical Results

- Some example results for junction trees



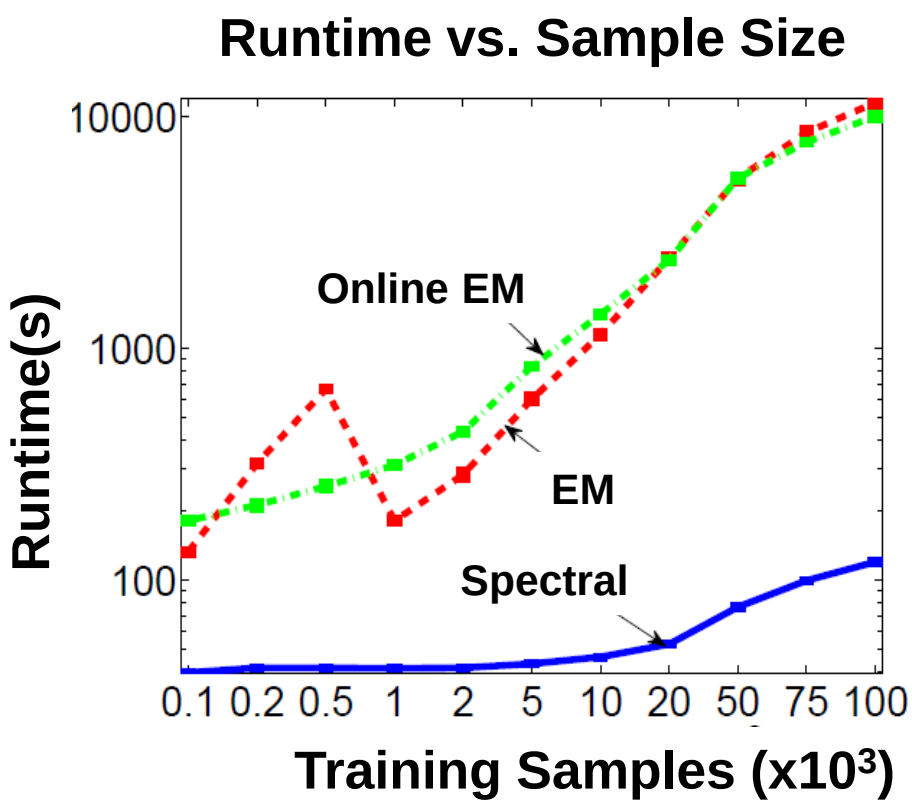


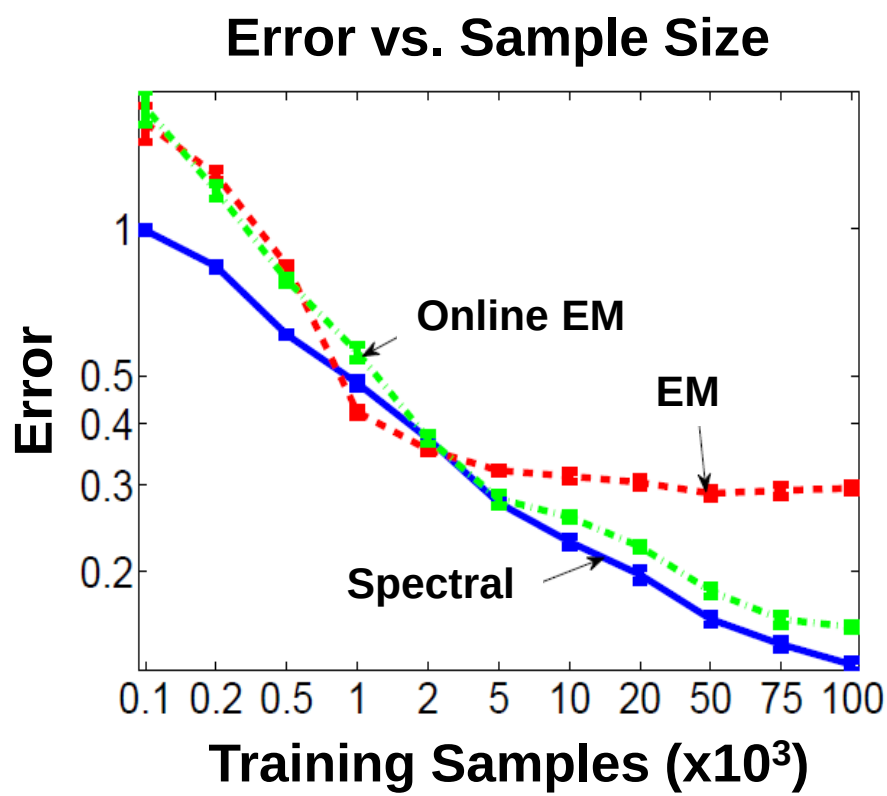




3rd Order NonHomogeneous HMM

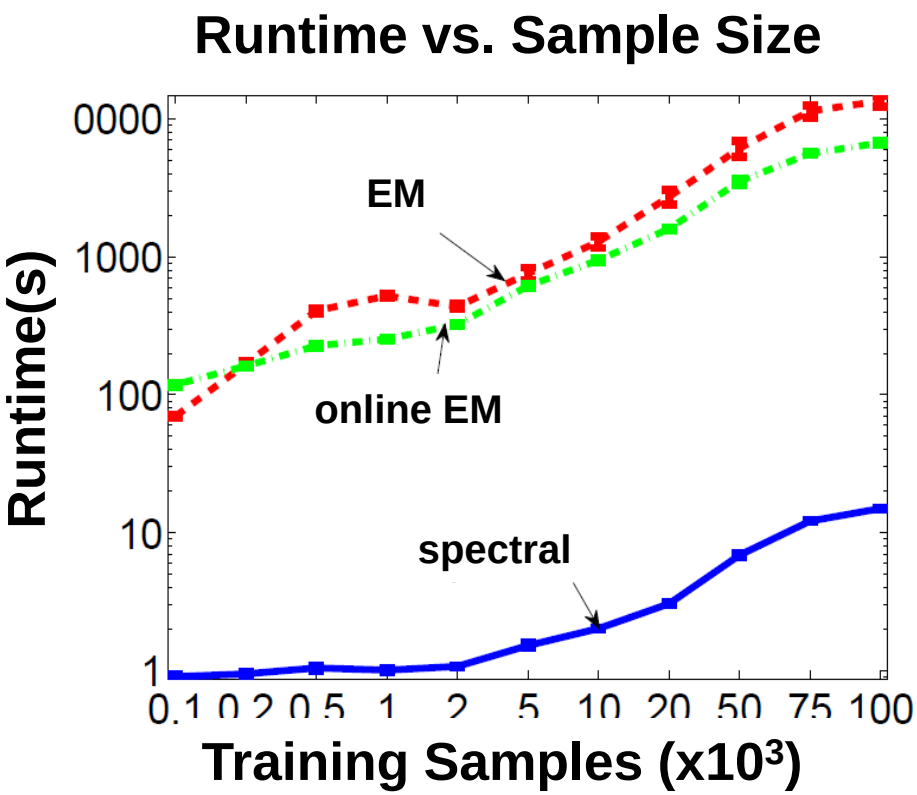
- (Spectral/EM/Online EM)

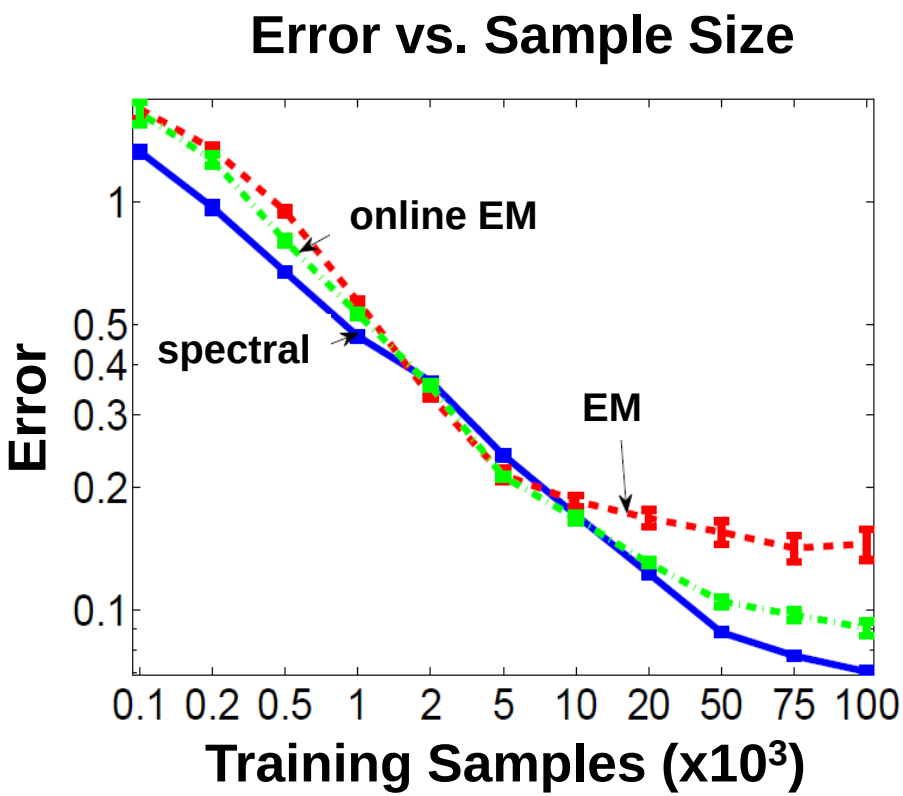




Results from Parikh et al. 2012

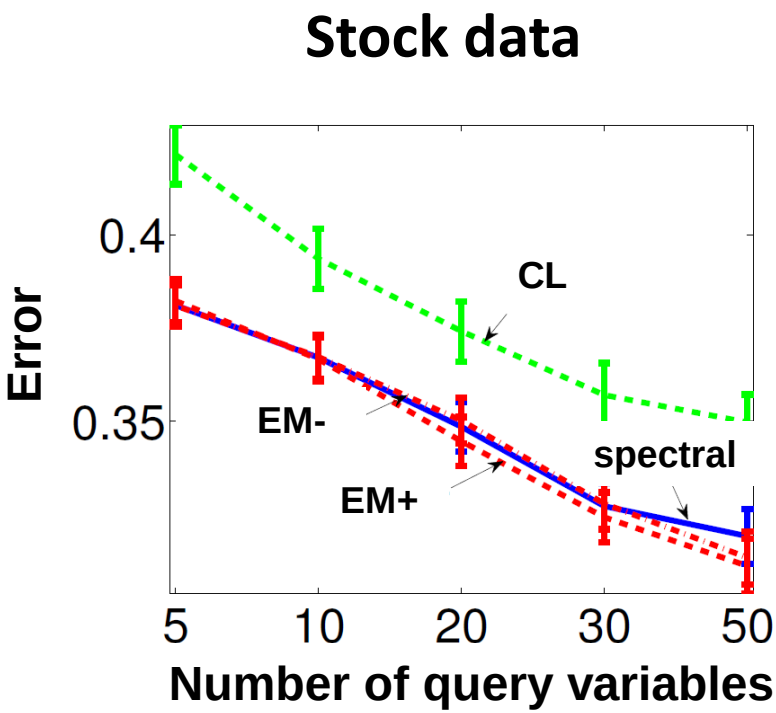
Synthetic Junction Tree

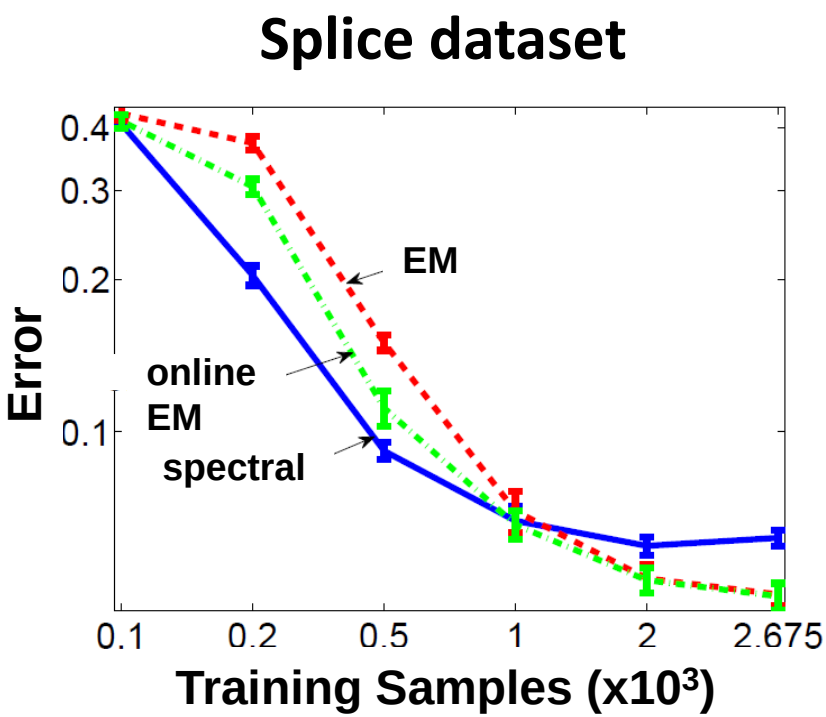




Results from Parikh et al. 2012

Real Data

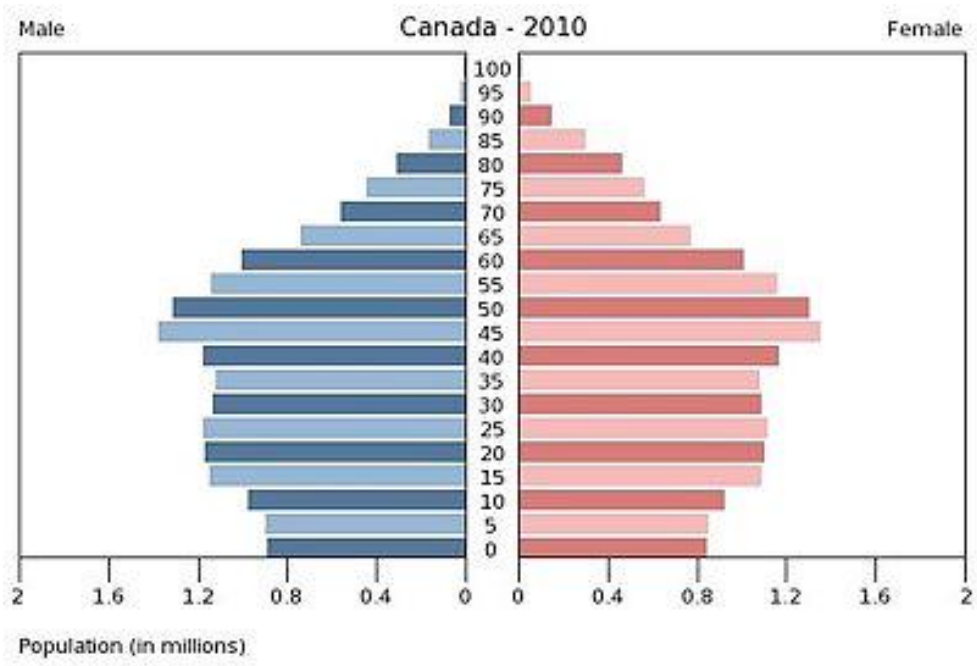


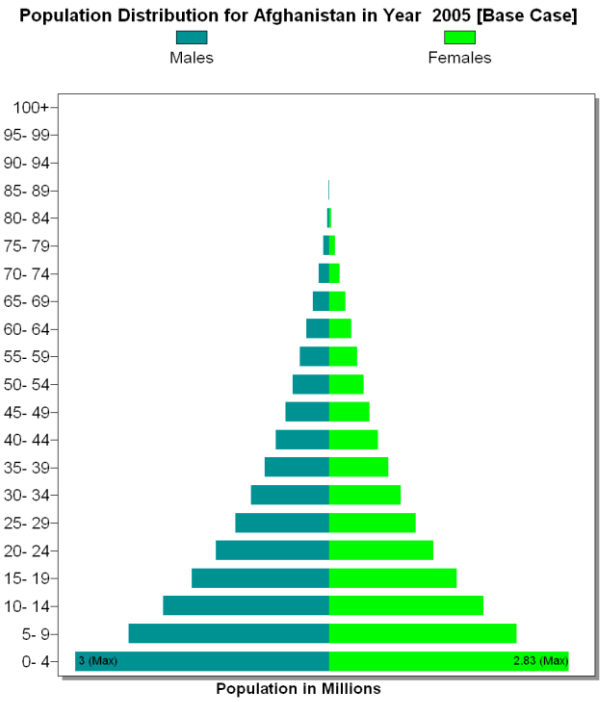


Results from Parikh et al. 2011/2012

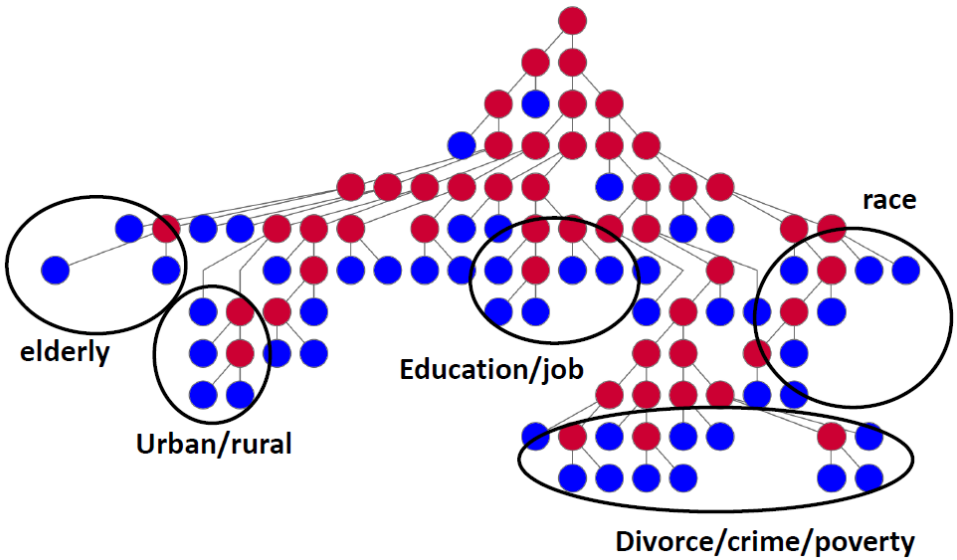
Kernel Embeddings of Latent Trees

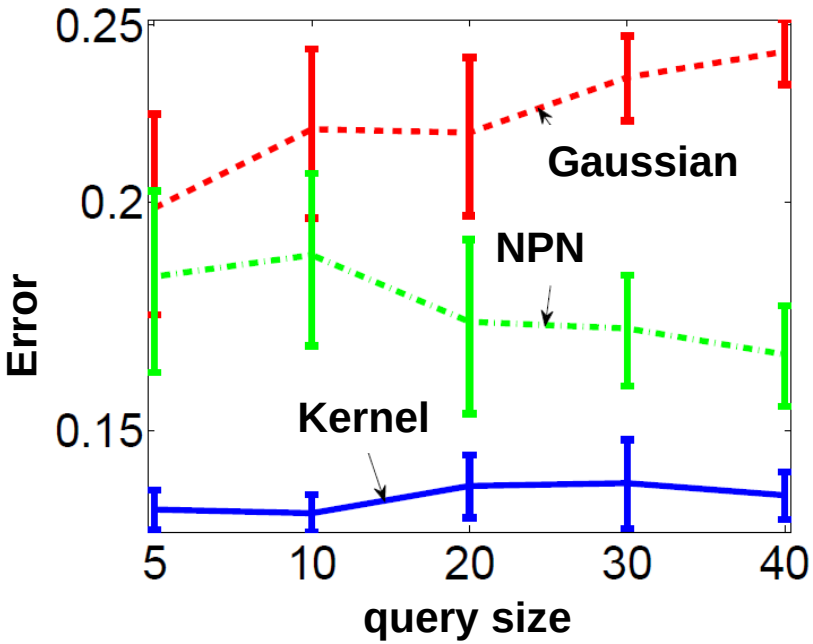
- Algorithm generalizes to non-Gaussian, continuous variables with Hilbert Space Embeddings [Smola et al. 2007, Song et al. 2010]





Kernel Embeddings of Latent Trees



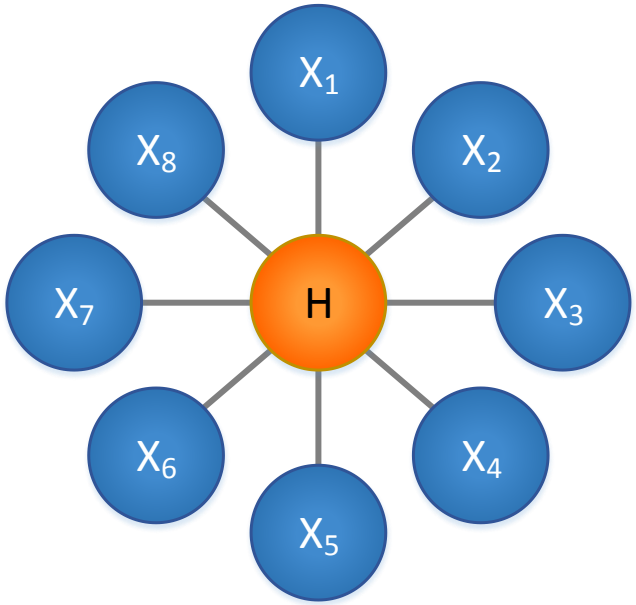


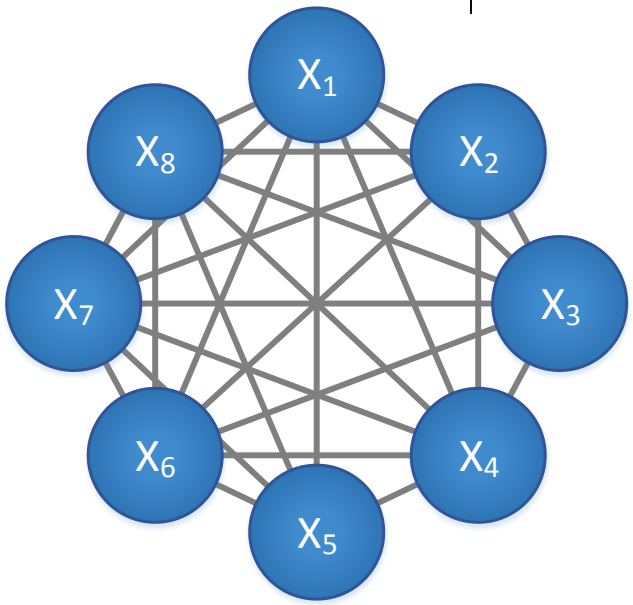
Spectral Approximations for Inference

Inference from the Spectral Perspective

- Traditionally, inference methods designed and analyzed from the perspective of the structure of the graphical model
- Difficulty of inference characterized by treewidth
- However, this paradigm can be limiting.

Example 1: Collapsed Sampling





$$x_i \sim P(X_i | h_i)$$

$$h_i \sim P(H | x_1, \dots, x_8)$$

Slow mixing

Easy parallelizable

$$x_i \sim P(X_i | x_1, \dots, x_{i-1}, x_{i+1}, x_8)$$

Faster mixing

Hard to parallelize

Parallel Collapsed Sampling

$$x_1' \sim P(X_1 | x_2, \ldots, x_8)$$
$$x_2 \sim P(X_2 | x_1', x_3, \ldots, x_8)$$

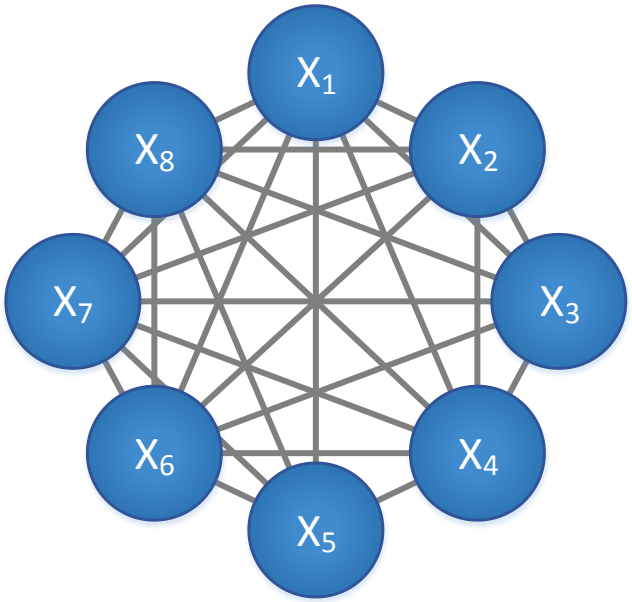
- Technically, must be in sequential.

- However, (partially) asynchronous sampling works well in practice
  - Fully Asynchronous (Symth et al. 2008)
  - Stale Synchronous (Ho et al. 2013)

- But a theoretical understanding is lacking
  - One exception is Ihler et al. 2011

Treewidth Perspective Only Gives Worst Case

- Assume $\boldsymbol{H}$ takes on $\boldsymbol{m^8}$ states. Then model is equivalent to the clique.

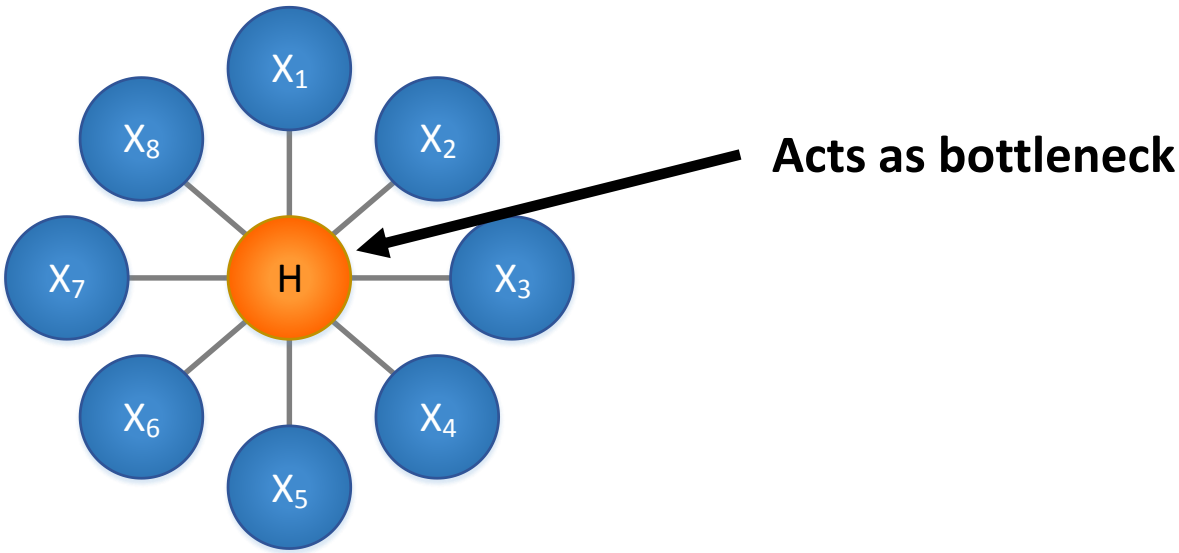


- In this case, there are $\boldsymbol{O(m^8)}$ parameters and thus

$$\boldsymbol{P(X_i|x_1,\ldots,x_{i-1},x_{i+1},x_8)} \not\approx \boldsymbol{P(X_i|x_1{}',\ldots,x_{i-1},x_{i+1},x_8)}$$

But If Model Is Low Rank

- If $\boldsymbol{H}$ takes on fewer states, then model is low rank.

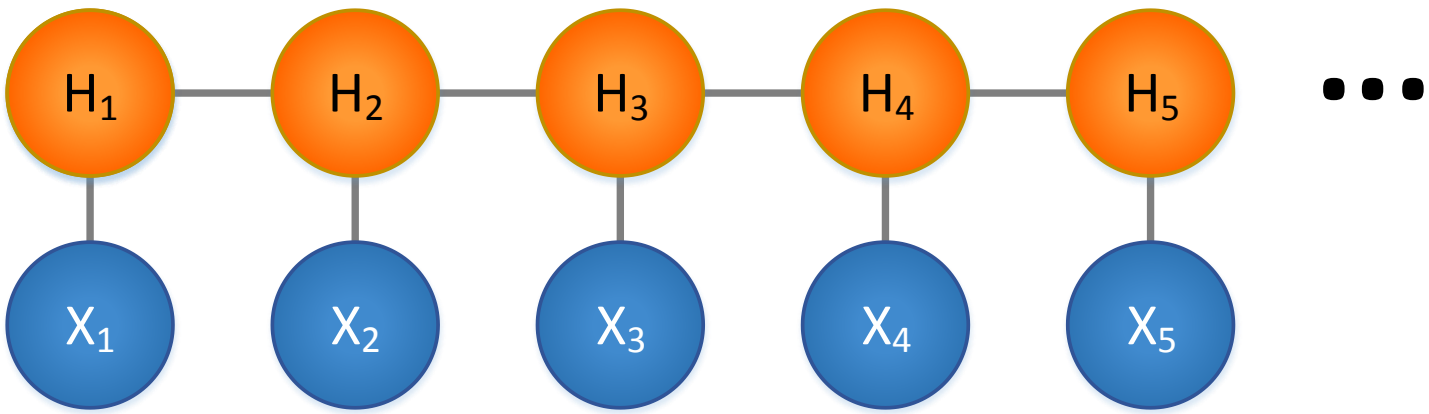


- It is now likely that

$$\boldsymbol{P(X_i|x_1,\ldots,x_{i-1},x_{i+1},x_8) \approx P(X_i|x_1{}',\ldots,x_{i-1},x_{i+1},x_8)}$$

- How can this help us justify approximate parallel sampling methods?

Example 2: Marginal MAP



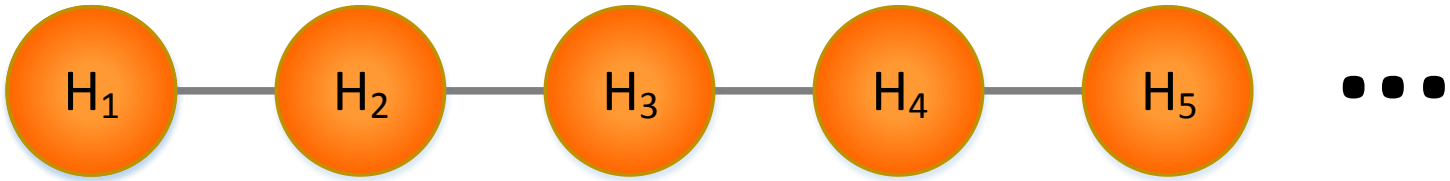
- Find most likely assignment of a set of variables (max nodes) while marginalizing out the rest (sum nodes).

Sometimes Easy

- Sum out $\boldsymbol{X_1, \ldots, X_5}$ while finding the most likely assignment to most likely assignment to $\boldsymbol{H_1, \ldots, H_5}$

$$\max_{H_1,\ldots,H_5} \sum_{X_1,\ldots,X_5} \prod_i P(X_i|H_i)\, P(H_i|\, H_{i-1})$$

- First run sum-product to marginalize out $\boldsymbol{X_1, \ldots, X_5}$



- Then run on max-product on remaining nodes $\boldsymbol{H_1, \ldots, H_5}$

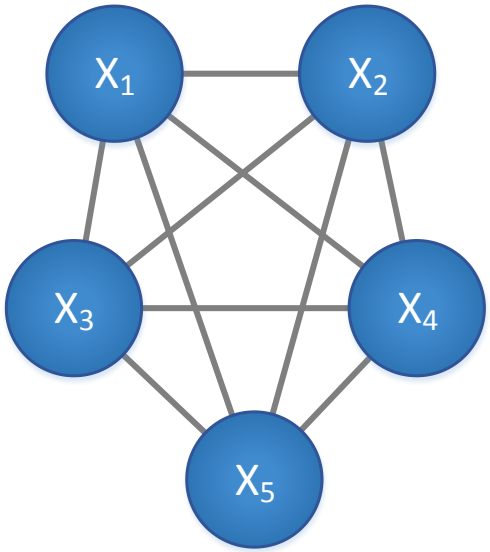
But Sometimes Hard

- Sum out $\boldsymbol{H_1, \ldots, H_5}$ while finding the most likely assignment to most likely assignment to $\boldsymbol{X_1, \ldots, X_5}$

$$\max_{X_1,\ldots,X_5} \sum_{H_1,\ldots,H_5} \prod_i P(X_i|H_i)\, P(H_i|\, H_{i-1})$$

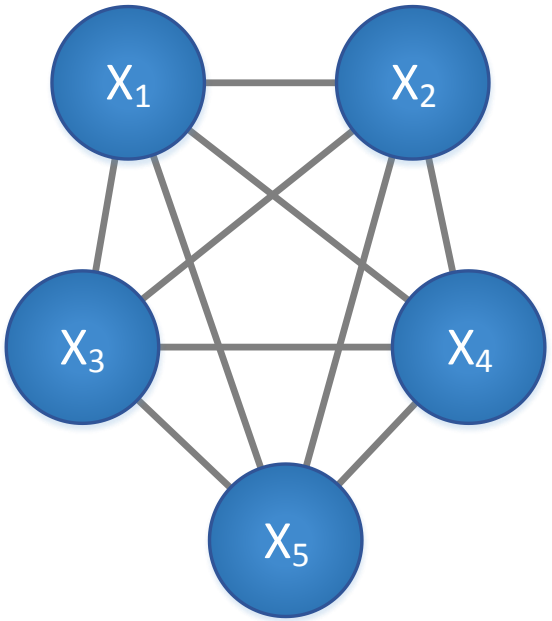
- First run sum-product to marginalize out $\boldsymbol{H_1, \ldots, H_5}$

- But now I have a clique ☹



But Again The Clique Is Low Rank....

- Can we somehow take advantage of this?



- A work in progress!

A Low Rank Framework For Language Modeling

Language Modeling

- Evaluate probabilities of sentences

Spectral methods are awesome $P(w_1, .., w_4) = 0.3648$

Spectral methods are boring $P(w_1, .., w_4) = 0.1922$

- Very useful in downstream applications such as machine translation and speech recognition.

N-Gram Language Model

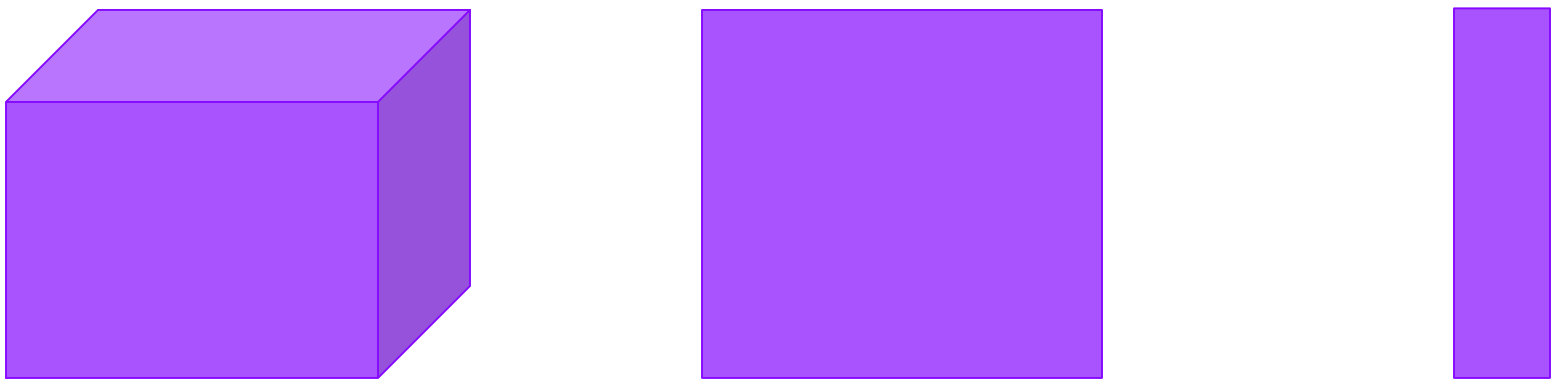
- Predominant language modeling technique for two decades

- Assume joint distribution of words follows $\boldsymbol{n^{th}}$ order Markov chain

$$P(w_1, .., w_\ell) = \prod_{i=1}^{\ell} P(w_i | \, w_1, \dots, w_{i-1})$$
$$\approx \prod_{i=1}^{\ell} P(w_i | \, w_{i-n+1}, \dots, w_{i-1})$$

But This Doesn't Solve the Problem

- Vocabulary is often very large $\approx 10^4$
  $(for\ relatively\ small\ datasets)$

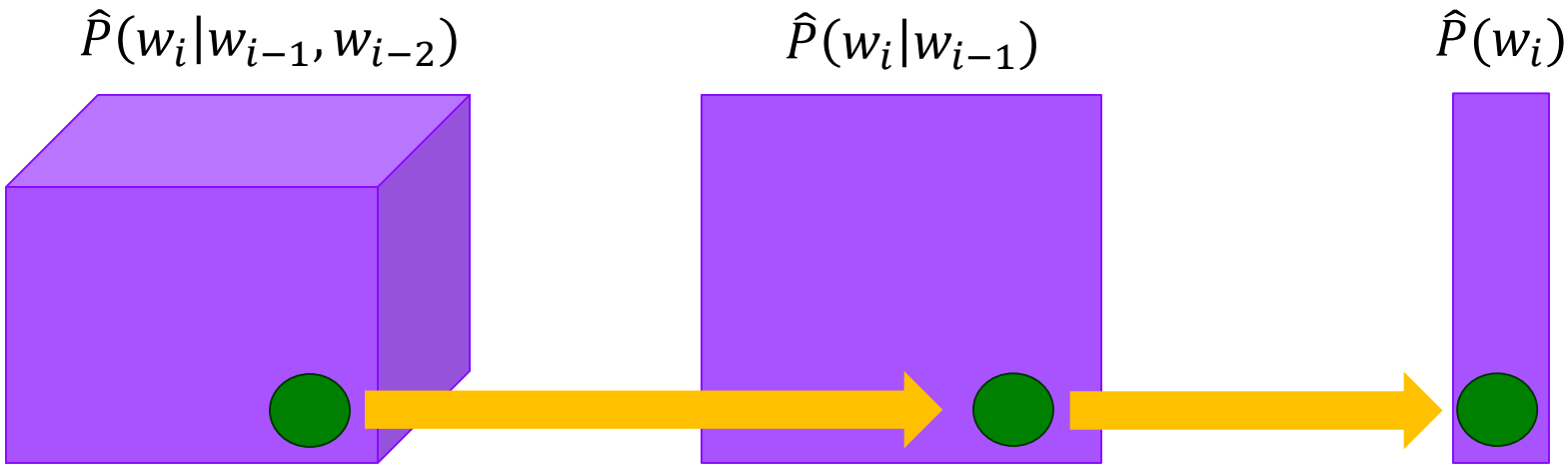


$\mathbf{10^{12}}$ **parameters** $\mathbf{10^{8}}$ **parameters** $\mathbf{10^{4}}$ **parameters**

- Extreme bias-variance tradeoff

N-gram Smoothing

- Essentially combine these as needed.



- Subtract probability out of higher order *n*-grams to make sure probability sums to one

Discounting

$$\hat{P}_d(w_i|w_{i-1}) = \frac{c(w_i, w_{i-1}) - d}{\sum_w c(w, w_{i-1})}$$

count

$$\hat{P}_{sm}(w_i|w_{i-1}) = \begin{cases} \hat{P}_d(w_i|w_{i-1}) \; if \;\; \hat{P}_d(w_i|w_{i-1}) = 0 \\ \gamma(w_{i-1})\, \hat{P}(w_i) \;\; otherwise \end{cases}$$

Where $\boldsymbol{\gamma(w_{i-1})}$ is the leftover probability

Shortcoming

$$\hat{P}(w_i) = \sum_{w_{i-1}} \hat{P}(w_i|w_{i-1})\hat{P}(w_i)$$

$$\hat{P}_{sm}(w_i) = \sum_{w_{i-1}} \hat{P}_{sm}(w_i|w_{i-1})\hat{P}(w_i)$$

- Lower order marginal doesn't align!!

$$\hat{P}(w_i) \neq \hat{P}_{sm}(w_i)$$

- **Intuition:** Lower order distribution needs to be adjusted.

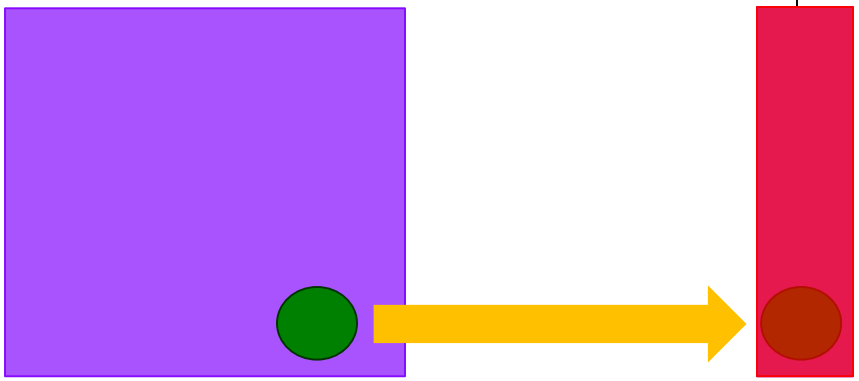
Kneser Ney - Intuition

- Consider first a unigram model

- Consider two words ***York*** and ***door***
  - ***York*** may have higher marginal probability than ***door***
  - $\boldsymbol{P(York) > P(door)}$

Kneser Ney - Intuition

- Consider bigram model



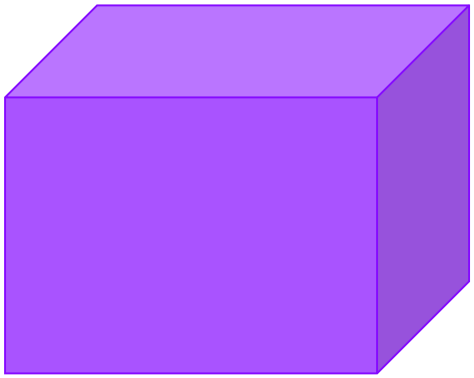
- Diversity of history
  - York only follows very few words i.e. New York
  - Door can follow many words i.e. “the door”, “red door”, “my door” etc.

- Therefore if we back off on $\boldsymbol{w_{i-1}}$, $\mathbf{w_i} = \mathbf{door}$ is more likely

$$P(w_i = door \mid backed - off\ on\ w_{i-1})$$
$$> P(w_i = York \mid backed - off\ on\ w_{i-1})$$

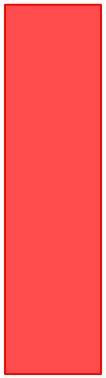
Kneser Ney

- Use altered lower order distributions to take this into account



MLE trigram

altered bigram



altered unigram

Kneser Ney

$$N_{-}(w_i) = |\{w : c(w_i, w) > 0\}|$$

Diversity of $w_i's$ history

$$\hat{P}_{kn-uni}(w_i) = \frac{N_{-}(w_i)}{\sum_w N_{-}(w)}$$

$$\hat{P}_{kney}(w_i|w_{i-1}) = \begin{cases} \hat{P}_d(w_i|w_{i-1}) & if \;\; \hat{P}_d(w_i|w_{i-1}) = 0 \\ \gamma(w_{i-1})\hat{P}_{kn-uni}(w_i) & otherwise \end{cases}$$

Lower Order Marginal Aligns!

$$\hat{P}(w_i) = \sum_{w_{i-1}} \hat{P}_{kney}(w_i|w_{i-1})\hat{P}(w_i)$$

Low Rank Approaches

- Consider ML problems with sparse matrices e.g.
  - Collaborate filtering
  - Matrix completion
- Many solutions involve low rank approximation
- These solutions have been attempted in NLP
  - Saul and Periera 1997
  - Hutchinson et al. 2011
- Unfortunately, not generally competitive Kneser Ney

Our Contribution

- We present a generalization of existing language models such as Kneser Ney to allow for non-integer n

- Key Insights:
  - Treat Kneser Ney as a special case of a low rank framework
  - Take low rank approximations of alternate quantities instead of MLE
  - Generalize discounting for this framework

Rank

- Rank distinguishes bigram and unigram

$$P(w_i|w_{i-1}) = \frac{P(w_i, w_{i-1})}{P(w_{i-1})}$$

Full rank

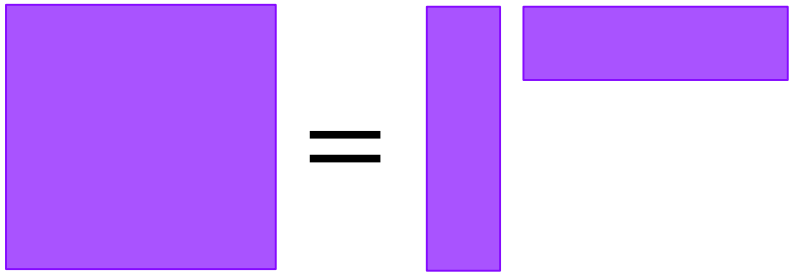
Rank

- Let $\boldsymbol{B}$ be the matrix such that

$$\boldsymbol{B}(w_i, w_{i-1}) = c(w_i, w_{i-1})$$

- Let

$$\boldsymbol{M}_1 = min_{\boldsymbol{M}:\boldsymbol{M}\geq 0, rank(\boldsymbol{M})=1} \|\boldsymbol{B} - \boldsymbol{M}\|_{KL}$$



- Then

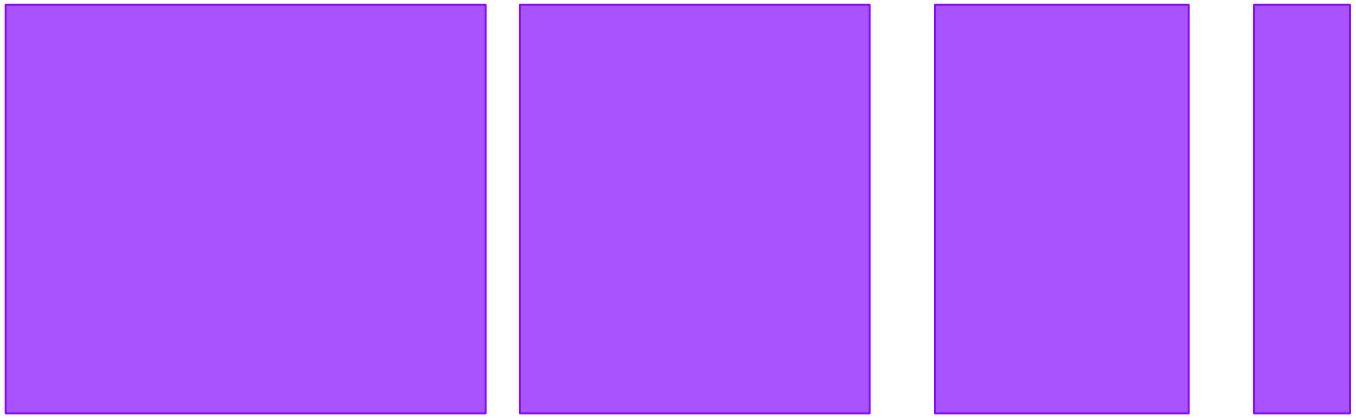
$$\boldsymbol{M}_1(w_i, w_{i-1}) \propto P(w_i)P(w_{i-1})$$

Rank

- MLE unigram is normalized rank 1 approx. of MLE bigram under KL:

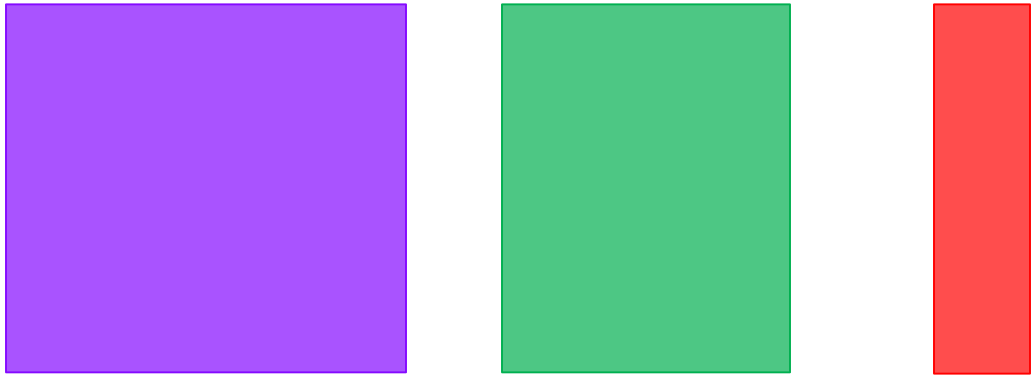
$$P(w_i) = \frac{\boldsymbol{M}_1(w_i, w_{i-1})}{\sum_w \boldsymbol{M}_1(w_i, w_{i-1})}$$

- Vary rank to obtain quantities between bigram and unigram



Generalizing The “Alteration”

- Kneser Ney uses alternate lower order matrices.
- Need continuous generalization of this “alteration”
- Take low rank approximations of these alternate quantities



Consider Elementwise Power

$$B \qquad\qquad B^{0.5} \qquad\qquad B^{0}$$

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 5 & 0 \\ 2 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1.4 & 1 \\ 0 & 2.2 & 0 \\ 1.4 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

row sum

$$\begin{bmatrix} 4 \\ 5 \\ 2 \end{bmatrix}$$

row sum

$$\begin{bmatrix} 3.4 \\ 2.2 \\ 1.4 \end{bmatrix}$$

row sum

$$\begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}$$

emphasis on diversity

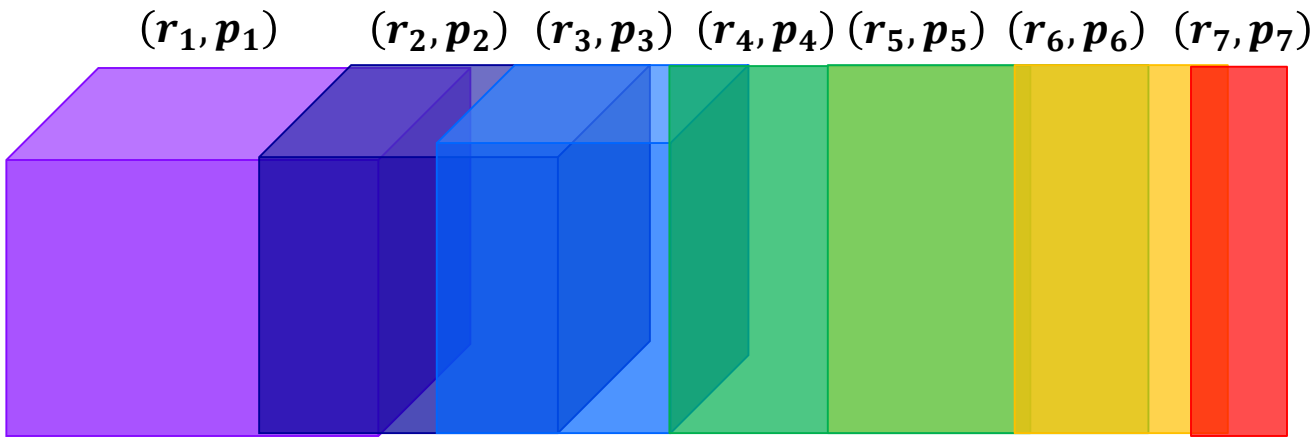
Consider Elementwise Power

$$\boldsymbol{M}_1^0 = min_{\boldsymbol{M}:\boldsymbol{M}\geq 0, rank(\boldsymbol{M})=1} \left\| \boldsymbol{B}^{\boldsymbol{0}} - \boldsymbol{M} \right\|_{KL}$$

$$\hat{P}_{kn-uni}(w_i) = \frac{\boldsymbol{M}_1^0(w_i, w_{i-1})}{\sum_w \boldsymbol{M}_1^0(w, w_{i-1})}$$

Low Rank Ensembles

- Intermediate matrices/tensors are both low rank and different power



What about the marginal constraint?

$$P(w_i) \overset{?}{=} \sum_{w_{i-1}} P_{low-rank}(w_i | w_{i-1})\, P(w_i)$$

- Not if fixed discounting is used
- **But we have developed a generalized discounting scheme such that the constraint holds**

Empirical results coming soon!

Outline

- Linear Algebra View of Latent Variable Models
- Spectral Parameter Learning
- Spectral Approximations for Inference
- Language Modeling

(Kernels and unsupervised parsing omitted due to time)

Conclusion

- Linear algebra can provide a different perspective on graphical models
- This viewpoint allows us to design solutions for probabilistic modeling that have both theoretical and practical value.
- Results on spectral parameter learning show promise
- Working on inference, language modeling, unsupervised parsing

(Tentative) Timeline

Contribution	Status and Schedule
Spectral Learning of Latent Trees	[ICML 2011, ICML 2013], journal (Summer 2014)
Spectral Learning of Latent Junction Trees	[UAI 2012], journal (Summer 2014)
Kernel Embeddings of Latent Trees	[NIPS 2011]
Power Low Rank Language Models	Fall 2013
Unsupervised Parsing	Fall 2013/Early Spring 2014
Spectral Approximations of Inference	Spring/Summer 2014
Scalable Low Rank Language Models	Fall 2014
Writing Thesis	Fall/Winter 2014

Acknowledgements

Thesis Committee Members

Eric Xing

Geoff Gordon

Noah Smith

Le Song

Ben Taskar

Acknowledgements

Thesis-Related Collaborators

Eric Xing

Le Song

Mariya Ishteva

Chris Dyer

Shay Cohen

Avneesh Saluja

Other Collaborators

Wei Wu, Qirong Ho, Mladen Kolar, Hoifung Poon, Kristina Toutanova, Asela Gunawardana, Chris Meek

Thanks!