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

© Ankur Parikh @ CMU, 2013
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

- Fever
- Sore throat
- Body aches
- Fatigue
- Stuffy nose

© Ankur Parikh @ CMU, 2013
Latent Variables Can Help!

- There is a simpler explanation underlying the data.

- All the symptoms are conditionally independent given this explanation.

Intuitively represents “the flu”

- stuffy nose
- fever
- body aches
- sore throat
- fatigue
Latent Variable Models

Sequence models

Parsing

Mixed membership models

Ho. et al. 2012
Cohen et al. 2012
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
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)
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)


- 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**

\[ P(X, Y) = P(Y \mid X) P(X) \]

- **Probability Vectors/Matrices/Tensors**

\[ P(Y) = P(Y \mid X) P(X) \]
Traditional View Of Graphical Models

- Complexity of mostly characterized by structure or \texttt{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 $H$ takes on only one state then the variables are independent.
It Depends on the Number of States

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

- But what about all the other cases?
In general, nothing we can say about the nature of this matrix. $X_1$ and $X_2$ have $m$ states each.

\[ P(X_1, X_2) \]

- In general, nothing we can say about the nature of this matrix.
What if we know $X_1$ and $X_2$ are independent?

Matrix is rank one!

$P(X_1, X_2) = \begin{bmatrix} P(X_1 = 1, X_2 = 1), \ldots, P(X_1 = 1, X_2 = m) \end{bmatrix} = \begin{bmatrix} P(X_1 = 1) (P(X_2 = 1), \ldots, P(X_2 = m) ) \end{bmatrix}$
What about rank in between 1 and \( m \)?
\(X_1\) and \(X_2\) are not marginally independent (They are only conditionally independent given \(H\)).

- Assume \(H\) has \(k\) states
- Then, \(\text{rank}(\mathcal{P}(X_1, X_2)) \leq k\)
- Why?
Low Rank Structure

\[ \mathcal{P}(X_1, X_2) \]

\[ \mathcal{P}(X_1 \mid H) \]

\[ \mathcal{P}(\emptyset \ H) \]

\[ \mathcal{P}(X_2 \mid H)^\top \]

\( rank \leq k \)  \hspace{1cm}  \( rank \leq k \)  \hspace{1cm}  \( rank \leq k \)  \hspace{1cm}  \( rank \leq k \)
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

Tensors

- Multidimensional arrays
- A Tensor of order $N$ has $N$ modes ($N$ 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

1\textsuperscript{st} order tensor (vector)

2\textsuperscript{nd} order tensor (matrix)

3\textsuperscript{rd} order tensor (cube)

Higher order tensors
Tensor Tensor Multiplication

\[ R(i, j, k, l) = \sum_{m} T_1(i, j, m) T_2(k, m, l) \]
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 $n$ is a tensor of order $2n - 2$).
Conditional probability tables depend on latent variables

Common approach is Expectation Maximization (EM)

- Local optima
- Slow to converge
Low Rank Perspective

\[ \mathbf{P}(\{X_1, X_2, X_3, X_4\}, \{X_5, X_6\}) \]

has rank \( k \)
Low Rank Matrices “Factorize”

\[ M = LR \]

If \( M \) has rank \( k \)

We already know one factorization!!!

\[
p(X_{\{1,2,3,4\}}, X_{\{5,6\}}) = p(X_{\{1,2,3,4\}} \mid A) \ p(\emptyset \mid A) \ p(X_{\{5,6\}} \mid A)^T
\]

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:

\[
M = LR \\
M = LS \, S^{-1} \, R
\]

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

\[ p(X\{1,2,3,4\}, X\{5,6\}) \]

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

\[ p(X\{1,2,3,4\}, X\{5\}) \]
\[ p(X\{4\}, X\{5,6\}) \]
An Alternate Factorization

- Note that

\[ p(X_{\{1,2,3,4\}, X_{\{5\}}}) = p(X_{\{1,2,3,4\} | A}) \ p(\emptyset | A) \ p(X_{\{5\} | A})^T \]

\[ p(X_{\{4\}, X_{\{5,6\}}}) = p(X_{\{4\} | A}) \ p(\emptyset | A) \ p(X_{\{5,6\} | A})^T \]

- Product of green terms (in some order) is

\[ p(X_{\{1,2,3,4\}, X_{\{5,6\}}} \]

- Product of red terms (in some order) is

\[ p(X_{\{4\}, X_{\{5\}}} \]
An Alternate Factorization

\[ p(X_{1,2,3,4}, X_{5,6}) = p(X_{1,2,3,4}, X_5) p(X_4, X_5)^{-1} p(X_4, X_{5,6}) \]

- Factor of 6 variables
- Factor of 5 variables
- 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

\[ p(X_{\{1,2,3,4\}}, X_{\{5\}}) = p(X_{\{1,2,3,4\}} | A) \ p(\emptyset A) \ p(X_{\{5\}} | A) \]

\[ p(X_{\{4\}}, X_{\{5,6\}}) = p(X_{\{4\}} | A) \ p(\emptyset A) \ p(X_{\{5,6\}} | A) \]
The Tensor Point Of View

\[ p(X_{\{1,2,3,4\}, X_{\{5,6\}}} = p(X_{\{1,2,3,4\}, X_{\{5\}}} p(X_{4}, X_{5})^{-1} p(X_{\{4\}}, X_{\{5,6\}}) \]

\[ p(X_{1}, X_{2}, X_{3}, X_{4}, X_{5}, X_{6}) = p(X_{1}, X_{2}, X_{3}, X_{4}, X_{5}) \times_{5} p(X_{4}, X_{5})^{-1} \times_{4} p(X_{4}, X_{5}, X_{6}) \]

Order 6

Order 5

Decompose recursively!!
Observable Factorization

\[
p(X_1, X_2, X_3, X_4, X_5, X_6) = p(X_1, X_3, X_5) \times_3 p(X_1, X_3)^{-1} \]

\[
\times \\
p(X_3, X_4, X_5) \times_5 p(X_3, X_5)^{-1} \times \\
p(X_4, X_5, X_6) \times_4 p(X_4, X_5)^{-1}
\]
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)

Runtime vs. Sample Size

Error vs. Sample Size

Results from Parikh et al. 2012
Results from Parikh et al. 2012
Real Data

Stock data

Splice dataset

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

Error

query size

Gaussian

NPN

Kernel

© Ankur Parikh @ CMU, 2013
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

\[ x_i \sim P(X_i|h_i) \]
\[ h_i \sim P(H|x_1, \ldots, x_8) \]

Slow mixing
Easy parallelizable

\[ x_i \sim P(X_i|x_1, \ldots, x_{i-1}, x_{i+1}, x_8) \]

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
Assume $H$ takes on $m^8$ states. Then model is equivalent to the clique.

In this case, there are $O(m^8)$ parameters and thus

$$P(X_i|x_1, \ldots, x_{i-1}, x_{i+1}, x_8) \approx P(X_i|x'_1, \ldots, x_{i-1}, x_{i+1}, x_8)$$
But If Model Is Low Rank

- If $H$ takes on fewer states, then model is low rank.

- It is now likely that

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

- 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).
- Sum out $X_1, \ldots, X_5$ while finding the most likely assignment to most likely assignment to $H_1, \ldots, H_5$

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

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

- Then run on max-product on remaining nodes $H_1, \ldots, H_5$
But Sometimes Hard

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

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

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

- 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  \[ P(w_1, \ldots, w_4) = 0.3648 \]
Spectral methods are boring  \[ P(w_1, \ldots, w_4) = 0.1922 \]

- 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 $n^{th}$ order Markov chain

$$P(w_1, \ldots, w_\ell) = \prod_{i=1}^{\ell} P(w_i | w_1, \ldots, w_{i-1})$$

$$\approx \prod_{i=1}^{\ell} P(w_i | w_{i-n+1}, \ldots, w_{i-1})$$
But This Doesn’t Solve the Problem

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

- Extreme bias-variance tradeoff
**N-gram Smoothing**

- Essentially combine these as needed.

\[ \hat{P}(w_i | w_{i-1}, w_{i-2}) \]

\[ \hat{P}(w_i | w_{i-1}) \]

\[ \hat{P}(w_i) \]

- Subtract probability out of higher order \( n \)-grams to make sure probability sums to one
Discounting

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

\[ \hat{P}_{sm}(w_i|w_{i-1}) = \begin{cases} \hat{P}_d(w_i|w_{i-1}) & \text{if } \hat{P}_d(w_i|w_{i-1}) = 0 \\ \gamma(w_{i-1}) \hat{P}(w_i) & \text{otherwise} \end{cases} \]

Where \( \gamma(w_{i-1}) \) 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*
  - $P(York) > P(door)$
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 $w_{i-1}, w_i = \text{door}$ is more likely
  $$ P(w_i = \text{door} \mid \text{backed off on } w_{i-1}) > P(w_i = \text{York} \mid \text{backed off on } w_{i-1}) $$
Kneser Ney

- Use altered lower order distributions to take this into account

MLE trigram  altered bigram  altered unigram
Kneseer Ney

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

Diversity of \( w_i \)'s history

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

\[ \hat{P}_{kney}(w_i|w_{i-1}) = \begin{cases} 
\hat{P}_d(w_i|w_{i-1}) & \text{if } \hat{P}_d(w_i|w_{i-1}) = 0 \\
\gamma(w_{i-1}) \hat{P}_{kn-uni}(w_i) & \text{otherwise}
\end{cases} \]
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
Let $B$ be the matrix such that

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

Let

$$M_1 = \min_{M: M \succeq 0, \text{rank}(M) = 1} \|B - M\|_{KL}$$

Then

$$M_1(w_i, w_{i-1}) \propto P(w_i)P(w_{i-1})$$
Rank

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

\[ P(w_i) = \frac{M_1(w_i, w_{i-1})}{\sum_w M_1(w_i, w_{i-1})} \]

- 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

\[
B = \begin{bmatrix}
1 & 2 & 1 \\
0 & 5 & 0 \\
2 & 0 & 0
\end{bmatrix}
\quad
B^{0.5} = \begin{bmatrix}
1 & 1.4 & 1 \\
0 & 2.2 & 0 \\
1.4 & 0 & 0
\end{bmatrix}
\quad
B^0 = \begin{bmatrix}
1 & 1 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0
\end{bmatrix}
\]

row sum

emphasis on diversity
Consider Elementwise Power

\[ M_1^0 = \min_{M: M \geq 0, \text{rank}(M) = 1} \| B^0 - M \|_{KL} \]

\[ \hat{P}_{kn-uni}(w_i) = \frac{M_1^0(w_i, w_{i-1})}{\sum_w M_1^0(w, w_{i-1})} \]
Low Rank Ensembles

- Intermediate matrices/tensors are both low rank and different power
What about the marginal constraint?

\[ P(w_i) = \sum_{w_{i-1}} P_{\text{low-rank}}(w_i|w_{i-1}) P(w_i) \]

- 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

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