Spectral Probabilistic Modeling and Applications to Natural Language Processing

Doctoral Thesis

Ankur Parikh
Machine Learning Department
Carnegie Mellon University

Thesis Committee:

Eric Xing (chair)
Geoff Gordon
John Platt (Microsoft Research)
Noah Smith
Le Song (GeorgiaTech)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Copyright © 2013, Ankur Parikh

May 26, 2015
Acknowledgements

This thesis would not have been possible without the help of a great number of people. I would first like to thank my undergraduate advisors Profs. Sharad Malik and Rob Schapire for introducing me to the world of research and inspiring me to commit several years of my life to being a PhD student. It was definitely worth it.

I am very grateful to have Prof. Eric Xing as my thesis advisor, who played an instrumental role in helping me to pick impactful problems and construct a long term vision for my thesis that spanned both theory and application. With his guidance and the breadth of expertise represented by the SAILING lab, I had the opportunity to be able to do research in a diverse range of fields spanning machine learning, natural language processing, and computational biology.

I’m also very thankful to my other thesis committee members Le Song, Noah Smith, Geoff Gordon, John Platt, and previously Ben Taskar for their time and feedback that have helped to improve many parts of this thesis. Le’s mentorship taught me how to tackle technically challenging problems and introduced me to the field of spectral learning. Noah has always made time to listen to my practice talks and give honest and useful feedback. Geoff and John have been really great to discuss machine learning ideas with, and Ben gave me some key suggestions on how to keep my thesis more focused.

In natural language processing (NLP), I have been fortunate to collaborate with and have learned a lot from Chris Dyer, Shay Cohen, Hoifung Poon, and Kristina Toutanova. Their guidance exposed me to an amazing area that I’ve become very passionate about. I’m also grateful to Avneesh Saluja who has been great to collaborate and discuss ideas with, and Brendan O’Connor who first sparked my interest in NLP by inviting me to Noah’s reading group.

Although not part of my thesis, I have benefited greatly from my computational biology collaborations with Prof. Wei Wu and Micol Marchetti-Bowick. I’m also grateful to Mladen Kolar who helped me write my first publication and my other collaborators Mariya Ishteva, Qirong Ho, Asela Gunawardana, and Chris Meek.

The Machine Learning Department has been an awesome experience, and I’m very thankful to Diane Stidle, Mallory Deptola, and Michelle Martin for all their help, support, and patience in the many years I’ve been here.

Lastly, I’m very grateful to my parents, Prashant and Manjari Parikh, whose unwavering support for me to pursue my passion has been crucial to being able to complete this thesis. Without their advice and support, it would have been difficult to navigate the ups and downs of a PhD.
Abstract

Probabilistic modeling with latent variables is a powerful paradigm that has led to key advances in many applications such natural language processing, text mining, and computational biology. Unfortunately, while introducing latent variables substantially increases representation power, learning and modeling can become considerably more complicated. Most existing solutions largely ignore non-identifiability issues in modeling and formulate learning as a nonconvex optimization problem, where convergence to the optimal solution is not guaranteed due to local minima.

In this thesis, we propose to tackle these problems through the lens of linear/multi-linear algebra. Viewing latent variable models from this perspective allows us to approach key problems such as structure learning and parameter learning using tools such as matrix/tensor decompositions, inversion, and additive metrics. These new tools enable us to develop novel solutions to learning in latent variable models with theoretical and practical advantages. For example, our spectral parameter learning methods for latent trees and junction trees are provably consistent, local-optima-free, and 1-2 orders of magnitude faster than EM for large sample sizes.

In addition, we focus on applications in Natural Language Processing, using our insights to not only devise new algorithms, but also to propose new models. Our method for unsupervised parsing is the first algorithm that has both theoretical guarantees and is also practical, performing favorably to the CCM method of Klein and Manning. We also developed power low rank ensembles, a framework for language modeling that generalizes existing n-gram techniques to non-integer n. It consistently outperforms state-of-the-art Kneser Ney baselines and can train on billion-word datasets in a few hours.
Contents

1 Introduction ......................................................... 7
   1.1 Challenges ...................................................... 8
   1.2 A Linear Algebra Approach To Graphical Models ........... 9
   1.3 Applications to Natural Language Processing .............. 9
   1.4 Thesis Statement ............................................... 11
   1.5 Related Work .................................................. 12
   1.6 Outline .......................................................... 12

2 Background: Modeling with Probabilistic Graphical Models .... 13
   2.1 Tasks in a graphical model ................................... 16
   2.2 Latent Variable Graphical Models ............................ 20
   2.3 Review of Hilbert Space Embedding of Distributions ... 23

3 A Linear Algebra View of Latent Variable Models ............... 32
   3.1 Tensor Notation .................................................. 32
   3.2 The Spectral View ............................................... 34

4 Spectral Learning Algorithms for Graphical Models ............ 37

4 A Spectral Algorithm for Latent Tree Graphical Models ....... 38
   4.1 Intuition .......................................................... 39
Chapter 1

Introduction

Probabilistic graphical models have become an indispensable framework in artificial intelligence (Pearl [1988]; Koller and Friedman [2009]; Murphy [2012]). Their ability to model and reason about complex uncertainty among large sets of variables has been an important driving force behind advances in many applications such as natural language processing (Manning and Schütze [1999]) and computational biology (Baldi et al. [2001]) in the last two decades.

Four central themes in graphical models are:

1. **Structure Learning**: How do we determine the structural dependencies among a set of random variables?

2. **Parameter Learning**: Once the model structure has been determined, how can the parameters be learned from the data?

3. **Inference**: How can we reason or make predictions about a set of variables conditioned on the values of other variables?

4. **Modeling**: How to leverage these tools to construct effective models for real world phenomena?

For graphical models where all variables are observed in the data, many of the solutions to these algorithmic challenges are well studied. Often for simple structures, many of these problems have tractable solutions e.g. the Chow-Liu algorithm for structure learning of tree graphical models, maximum likelihood estimation with fully observed data for parameter learning, max-product for inference on trees. For more complex graphical models, structure learning, parameter learning, and inference are generally NP-hard but approximate solutions have been developed (Wainwright and Jordan [2008]; Koller and Friedman [2009]).

However, limiting ourselves to only using observed variables can be quite restrictive from the perspective of modeling. This is because, in many cases, the observed variables alone may not suffice to provide a concise explanation of the data. Consider the machine translation example shown in Figure 1.1. Here the goal is to translate the simple English sentence “spectral learning is awesome” into Spanish, which is “aprendizaje espectral es impresionante”. At first this task can
seem daunting since the $i^{th}$ word in the English sentence does not necessarily correspond to the $i^{th}$ word in the Spanish sentence. For example, here “spectral”, the first word in the English sentence corresponds to “espectral”, the second word in the Spanish sentence. For more complicated sentences, whole phrases may be moved/inserted when translating to another language. Therefore, strictly modeling with only the observed variables, as informally shown in Figure 1.1(a), can produce very complicated dependencies.

Intuitively, the problem would be much simpler if we knew which English word mapped to each Spanish word i.e. the word alignment, as (informally) shown in Figure 1.1(b). Given the alignment function $a$, translation becomes much easier, since to get the $i^{th}$ Spanish word, we can simply look up the Spanish translation of the $a(i)^{th}$ English word. However, the alignment is not provided in the training data (which generally just consists of English/Spanish sentence pairs), and therefore is a latent (or hidden) variable.

1.1 Challenges

Unfortunately, while latent variables provide advantages for model design, they pose substantial challenges for structure learning, parameter learning, and modeling. There exist many problems that are easy for fully observed models but are considerably more difficult for latent variable models. For example, maximum likelihood estimation with latent variables for structure/parameter learning is generally NP-hard. Similarly, while observed models are always identifiable, the same is generally not true for latent variable models, making evaluation and interpretation complicated.

The vast majority of existing approaches addressing these challenges rely on local-search heuristics based on either greedy search or non-convex optimization. While in many cases these methods work well in practice, they often require careful initialization and problem-specific heuristics to yield good results. For example, unsupervised parsing performance with random initializations can vary greatly, making carefully crafted initializers essential to reliable results. In addition, the sequential nature of these existing methods can often make parallelization challenging and thus scalability non-trivial in today’s distributed computing paradigm.

From the theoretical perspective, these approaches often do not shed light into the nature and complexity of latent models making it difficult to ascertain what makes the problem difficult, and whether certain relaxations can lead to tractable solutions.
1.2 A Linear Algebra Approach To Graphical Models

In this thesis, we tackle these challenges from a different perspective revolving around linear algebra. From the linear algebra point of view, latent variables induce low rank dependencies among the observed variables. The rank of the latent space can then be theoretically used to quantify the complexity/hardness of learning in the latent variable model. As we will see, when the rank becomes very large, the problem will become intractable, and we will be forced to resort to heuristic search methods. However, in the low rank scenario, which we will argue occurs more often in practice, we can leverage tools from linear algebra to provide provably consistent solutions in many cases.

Our work is inspired by recent theoretical results from different communities such as dynamical systems (Katayama, 2005), theoretical computer science (Dasgupta, 1999), statistical machine learning (Hsu et al., 2009; Bailly et al., 2009), and phylogenetics (Mossel and Roch, 2005). Before this thesis however, the majority of spectral learning methods were limited to hidden Markov models (Hsu et al., 2009; Bailly et al., 2009; Siddiqi et al., 2010; Song et al., 2010a).

In this dissertation, we take a more general view, proposing not just to leverage low rank matrix factorization (although that is certainly a major component), but also higher order tensor operations, additive tree metrics, and other tools to present theoretically principled and practical spectral solutions to a broad class of problems. In particular, we seek to focus on parameter learning, structure learning, and modeling with latent variables from the linear algebra point of view.

1.3 Applications to Natural Language Processing

From the application standpoint, we seek to use these theoretical insights to develop new models and algorithms for Natural Language Processing (NLP) tasks. We focus on unsupervised parsing and language modeling, two probabilistic modeling problems which we believe can benefit substantially from the linear algebra perspective.

1.3.1 Unsupervised Parsing

Unsupervised parsing (also known as grammar induction) is the problem of discovering syntactic structure in sentences without the help of annotated training examples marked with syntactic trees. Solutions to the problem of grammar induction have been long sought after since the early days of computational linguistics and are interesting both from cognitive and engineering perspectives. Cognitively, it is more plausible to assume that children obtain only terminal strings of parse trees and not the actual parse trees, which means the unsupervised setting may be a better model for studying language acquisition.

From the engineering perspective, training data for unsupervised parsing exists in abundance (i.e. sentences and part-of-speech tags), and is much cheaper than data required for supervised training, which requires manual syntactic annotation.
Most of the solutions suggested treat the problem of unsupervised parsing by assuming a parametric model, which is then estimated by identifying a local maximum of an objective function such as the likelihood (Klein and Manning, 2004) or a variant of it (Cohen and Smith, 2009; Haddad et al., 2009; Spitkovsky et al., 2010b; Gillenwater et al., 2010; Golland and DeNero, 2012). Unfortunately, finding the global maximum for these objective functions is usually intractable (Cohen and Smith, 2010). As a result, many of these methods suffer from severe local-optima and initializers are crafted to obtain good solutions.

In this thesis, we take a very different approach to unsupervised parsing. We propose to formulate unsupervised parsing as a latent structure learning problem where the latent structure (parse tree) varies for each example. Our goal is to then leverage linear algebra tools to derive a provably correct learning algorithm that also works empirically well in practice.

1.3.2 Language Modeling

Language modeling - the task of assigning probabilities to sequences of words, is an important component of many NLP and speech systems. It is a seemingly simple, yet actually very challenging problem that is useful in a number of applications e.g. machine translation (Koehn, 2010) since it allows one to determine which candidate sequences are more likely than others. The predominant approach to language modeling is the $n$-gram language model, wherein the probability of a word sequence $P(w_1, \ldots, w_m)$ is first factored and then approximated (with the Markov assumption) as:

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i|w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i|w_{i-n+1}^{i-1})$$

In other words, one only needs to take into account the previous $n-1$ words when computing the probability of a word $w_i$ given its word history. This assumption reduces parameters significantly, but it is not enough. Due to the large vocabulary space, maximum likelihood approaches will often assign zero probability to word sequences that were unseen in the training data (but can be in the test data).

A rich literature in language model (LM) smoothing has thus arisen in response to this core issue, with the basic idea behind most approaches being to interpolate with or back off to lower order $n$-gram models (which are less sparse) as the need arises (Chen and Goodman, 1999). While surprisingly simple, these techniques, in particular Kneser-Ney smoothing (Kneser and Ney, 1995) and variants, have often been the state of the art for more than a decade. However, these approaches fail to exploit the semantic/syntactic relatedness among words that can intuitively be exploited to help alleviate the data sparsity problem in a more efficient manner.

As part of this dissertation, we examine how ensembles of specially constructed low rank matrices/tensors can be leveraged to provide a novel solution to the language modeling problem. Our proposed solution contains Absolute Discounting (Ney et al., 1994) and Kneser Ney (Kneser and Ney, 1995) as special cases of our general framework. Our approach can take incorporate semantic relatedness among words to improve performance while preserving scalability to large datasets.
1.4 Thesis Statement

The central theme of this thesis revolves around the following statement:

 Viewing latent variable models through the lens of linear/multi-linear algebra allows us to approach key problems such as structure and parameter learning using tools such as matrix/tensor decompositions, inversion, and hilbert space operators. These new tools enable us to develop novel solutions for learning and inference in latent variable models that have both theoretical and practical advantages. In addition, these insights aid us in designing new models and algorithms for language modeling and unsupervised parsing in Natural Language Processing.

Key Contributions

1. Chapter 4: A Spectral Algorithm for Latent Tree Graphical Models - We leverage tensor algebra to derive a provably consistent parameter learning algorithm for latent trees. Our approach gives comparable or better accuracy to Expectation Maximization, while being 10-100x faster in a variety of settings (Key theme: parameter learning). Preliminary version: Parikh et al. (2011).

2. Chapter 5: A Spectral Algorithm for Latent Junction Trees - We generalize the work from the previous chapter to low-treewidth graphical models via junction trees (Key theme: parameter learning). Preliminary version: Parikh et al. (2012).

3. Chapter 6: Kernel Embeddings of Latent Tree Graphical Models - Using Hilbert Space Embeddings (Smola et al., 2007), we develop algorithms for latent variable parameter and structure learning in latent tree graphical models with continuous, non-Gaussian variables (Key themes: parameter and structure learning). Preliminary version: Song et al. (2011b).

4. Chapter 7: Alternate Spectral Algorithm for Latent Tree Models - We show that the spectral algorithm derived in Chapter 4 is not unique and there exists another spectral representation that only requires tensors of order 3 regardless of the tree topology. Both factorizations are compared empirically in different settings (Key theme: parameter learning). Preliminary version: Parikh et al. (2011).

5. Chapter 8: Spectral Unsupervised Parsing with Additive Tree Metrics - We propose a novel model for unsupervised parsing that revolves around structure learning of projective latent trees where the latent structure changes for each example. Unlike existing approaches, our method has provable guarantees on structure recovery and also performs well empirically (Key themes: structure learning and modeling). Preliminary version: Parikh et al. (2014a).

6. Chapter 9: Language Modeling with Power Low Rank Ensembles - We develop a novel low rank framework of n-gram language models where existing methods such as Absolute Discounting and Kneser Ney are special cases. Our approach allows n-grams of non-integer n i.e. 2.5-grams, 1.5-grams etc. and consistently outperforms state-of-the-art Kneser Ney baselines without sacrificing the computational advantages of n-gram based methods (Key theme: modeling). Preliminary version: Parikh et al. (2014b).
1.5 Related Work

As mentioned previously, the basis for this dissertation is built on work from many different communities such as dynamical systems (Katayama, 2005), theoretical computer science (Dasgupta, 1999), statistical machine learning (Hsu et al., 2009; Bailly et al., 2009), and phylogenetics (Mossel and Roch, 2005). Before this thesis however, the majority of spectral learning methods were limited to hidden Markov models (Hsu et al., 2009; Bailly et al., 2009; Siddiqi et al., 2010; Song et al., 2010a).

In this work we take a more general graphical models point of view, leveraging tensor algebra to show that spectral approaches can lead to principled and practical solutions for a wide class of problems in graphical models and Natural Language Processing. Our approach allows us to gain a deeper understanding of spectral learning and its connections to tensor algebra. Furthermore, it allows us to use these insights to develop novel models and solutions for new domains.

Concurrently with this thesis, other researchers have also developed spectral learning methods for other problems such as predictive state representations (Boots et al., 2010), topic models (Anandkumar et al., 2012, 2013), latent probabilistic context free grammars (Cohen et al., 2012; Cohen and Collins, 2012) and other models (Bailly et al., 2010; Balle et al., 2011; Luque et al., 2012; Dhillon et al., 2012a; Balle et al., 2012; Boots and Gordon, 2013; Zhang et al., 2014; Chaganty and Liang, 2014; Quattoni et al., 2014; Wang and Zhu, 2014; Subakan et al., 2014; Saluja et al., 2014).

In particular, the works of Cohen et al. (2012); Cohen and Collins (2012) and Dhillon et al. (2012a); Melnyk and Banerjee (2014) leverage the tensor formulation we developed in this thesis.

Another line of work that seeks to develop alternative learning methods for latent variable models is that of anchor-based methods, first proposed by Arora et al. (2012b), and followed up by Arora et al. (2012a); Ding et al. (2013); Cohen and Collins (2014); Nguyen et al. (2014). Compared to spectral approaches, these methods make a rather different set of assumptions, assuming that for each latent state there exists a feature that can only be generated by that state. While it would be very interesting to compare these approaches directly with spectral methods, it is beyond the scope of the thesis.

1.6 Outline

Before describing the primary contributions of this thesis in more detail, some background on graphical models and kernel methods is presented in Chapter 2. We then discuss a linear algebra perspective on latent variable models in Chapter 3 which serves as a foundation for our work. Chapters 4, 5, 6, 7, 8, and 9 then present the main contributions of the thesis.
Chapter 2

Background: Modeling with Probabilistic Graphical Models

This chapter largely serves to give a brief background of probabilistic graphical models (Pearl 1988; Koller and Friedman 2009; Murphy 2012), a popular and elegant paradigm for modeling random variables, that is the basis for this thesis. We also give a brief tutorial on hilbert space embeddings (Smola et al. 2007), a concept central to Chapter 6.

2.0.1 Motivation

Consider $p$ random variables $X = \{X_1, ..., X_p\}$, where each variable is allowed to take on $m$ states i.e. $X_i \in \{0, ..., m - 1\} \ \forall X_i \in X$. Then there are $p^m$ different settings of the random variables, each of which can be assigned a different probability.

First consider making no assumptions on the underlying distribution $p(X_1, ..., X_p)$ so that there are no constraints among the probabilities of different assignments (with the exception that they must sum to one). For example, for $p = 3, m = 2$ we would have:

$$
\begin{align*}
P(X_1 = 0, X_2 = 0, X_3 = 0) &= \theta_{000} \\
P(X_1 = 0, X_2 = 0, X_3 = 1) &= \theta_{001} \\
P(X_1 = 0, X_2 = 1, X_3 = 0) &= \theta_{010} \\
P(X_1 = 0, X_2 = 1, X_3 = 1) &= \theta_{011} \\
P(X_1 = 1, X_2 = 0, X_3 = 0) &= \theta_{100} \\
P(X_1 = 1, X_2 = 0, X_3 = 1) &= \theta_{101} \\
P(X_1 = 1, X_2 = 1, X_3 = 0) &= \theta_{110} \\
P(X_1 = 1, X_2 = 1, X_3 = 1) &= \theta_{111}
\end{align*}
$$

(2.1)

where the parameters $\theta_{000}, ..., \theta_{111}$ must be estimated from data. Note that all but one of the parameters can be set arbitrarily, since the only constraints are that the parameters are non-negative and that the distribution must sum to one.
However, in this scenario, statistical estimation of parameters (learning) becomes very challenging. Consider the following lemma from Roy (2011):

**Lemma 1.** Let $p_\theta$ be a multinomial distribution with $B$ bins and parameters $\theta$ i.e. bin $i$ is associated with probability $\theta_i$. Let $\hat{\theta}$ be the maximum likelihood estimate (MLE) of the multinomial parameters from $N$ iid samples from $p_\theta$. Then,

$$E[KL(p_\theta | p_{\hat{\theta}})] \leq \frac{B - 1}{N}$$  \hspace{1cm} (2.2)

Since $B = m^p$, Lemma 1 implies that $N \in \Omega(m^p)$ for $E[KL(p_\theta | p_{\hat{\theta}})] \rightarrow 0$, making statistical estimation challenging for more than a handful of variables.

Similarly, this approach also leads to severe computational problems when attempting to do inference, e.g. computing lower order marginals. Let $S$ and $T$ be two sets of variables $X_S, X_T \subseteq \mathcal{X}$ s.t. $X_S \cup X_T = \mathcal{X}$ and let $x_S, x_T$ indicate a possible set of instantiations.

$$P(X_S = x_S) = \sum_{x_T} P(X_S = x_S, x_T)$$  \hspace{1cm} (2.3)

where the summation would have computational complexity $\Theta(m^{|X_T|})$ i.e. exponential with respect to the number of variables being summed out.

As a result, the brute force representation of probability distributions, while very expressive from the modeling point of view, is statistically and computationally prohibitive.

### 2.0.2 Leveraging Independencies

Let us now consider the other extreme, showing that we can severely restrict modeling power to increase computational and statistical efficiency.

Assume that all $X_1, ..., X_p$ are mutually independent, implying that the factorization below:

$$P(X_1, ..., X_p) = \prod_{i=1}^{p} P(X_i)$$  \hspace{1cm} (2.4)
Considering our \( m = 2, p = 3 \) example again we have that:

\[
\begin{align*}
p(X_1 = 0, X_2 = 0, X_3 = 0) &= \theta_{X_1=0} \theta_{X_2=0} \theta_{X_3=0} \\
p(X_1 = 0, X_2 = 0, X_3 = 1) &= \theta_{X_1=0} \theta_{X_2=0} \theta_{X_3=1} \\
p(X_1 = 0, X_2 = 1, X_3 = 0) &= \theta_{X_1=0} \theta_{X_2=1} \theta_{X_3=0} \\
p(X_1 = 0, X_2 = 1, X_3 = 1) &= \theta_{X_1=0} \theta_{X_2=1} \theta_{X_3=1} \\
p(X_1 = 1, X_2 = 0, X_3 = 0) &= \theta_{X_1=1} \theta_{X_2=0} \theta_{X_3=0} \\
p(X_1 = 1, X_2 = 0, X_3 = 1) &= \theta_{X_1=1} \theta_{X_2=0} \theta_{X_3=1} \\
p(X_1 = 1, X_2 = 1, X_3 = 0) &= \theta_{X_1=1} \theta_{X_2=1} \theta_{X_3=0} \\
p(X_1 = 1, X_2 = 1, X_3 = 1) &= \theta_{X_1=1} \theta_{X_2=1} \theta_{X_3=1}
\end{align*}
\]

(2.5)

where we have 3 constraints (in addition to non-negativity):

\[
\begin{align*}
\theta_{X_1=0} + \theta_{X_1=1} &= 1 \\
\theta_{X_2=0} + \theta_{X_2=1} &= 1 \\
\theta_{X_3=0} + \theta_{X_3=1} &= 1
\end{align*}
\]

(2.6)

Generally, \( p \) independent variables mean \( O(pm) \) parameters which is much fewer than the \( O(m^p) \) required previously. This results in statistical and computational tractability. Statistically, Lemma 1 means \( N = \Omega(mp) \) is sufficient for \( \mathbb{E}[KL(p_\theta|p_\hat{\theta})] \to 0 \).

Computationally, marginalization can be computed efficiently since the sum can be distributed inside the factors. Again let \( S \) and \( T \) be two sets of variables \( X_S, X_T \subseteq X \) s.t. \( X_S \cup X_T = X \) and let \( x_S, x_T \) indicate a possible set of instantiations. Then,

\[
P(X_S = x_S) = \sum_{x_T \in X_T} P(X_S = x_S, x_T) = \prod_{s \in S} P(X_s = x_s) \prod_{t \in T} \sum_{x_t} P(X_t = x_t)
\]

(2.7)

### 2.0.3 Probabilistic Graphical Models

Graphical models provide a formal framework for trading off model expressivity with statistical/computational tractability of learning and inference.

A graphical model is a graph \( \mathcal{G} = (V, E) \) where the vertices \( V \) correspond to random variables and the edges \( E \) indicate dependencies. For this work we will assume that the graph is directed and acyclic (a DAG) and thus is also known as a Bayesian network.

The connectivity of the graph indicates the independence assumptions made by the graph. No edges as shown in Figure 2.1(a) indicates that all the variables are independent while a clique (all edges present) as shown in Figure 2.1(b) implies no conditional independences.

Most of the rest of the thesis (with the exception of Chapter 5) is focused on graphical models where \( \mathcal{G} \) is a tree. In tree-shaped graphical models, two nodes \( X_i \) and \( X_j \) become conditionally independent conditioned on any variable in the path between \( X_i \) and \( X_j \). The rigorous definition
Figure 2.1: Example graphical model structures

is provided below:

**Definition 1.** Given a tree graphical model $G = (\mathcal{V}, \mathcal{E})$, let $X_i, X_j \in \mathcal{V}$ and let $\mathcal{Z}$ denote the set of variables on the (unique) path between $X_i$ and $X_j$. Let $\mathcal{A}$ be some subset of $\mathcal{V}$. Then $X_i \perp X_j | \mathcal{A}$ if and only if $\mathcal{A} \cap \mathcal{Z} \neq \emptyset$.

Note that the independences implied by the structure are unrelated to the direction of the edges in a tree graphical model. (For conditional independencies induced by more complex graph structures such as v-structures, please see Koller and Friedman (2009)). This gives the following corollary:

**Corollary 1.** Let $G = (\mathcal{V}, \mathcal{E})$ be a tree graphical model directed via an arbitrary root $X_r$. A particular distribution $p$ respects the independences in $G$ if and only if:

$$P(X_1, ..., X_p) = \prod_{i=1}^{p} P(X_i | X_{\pi(i)})$$  \hspace{1cm} (2.8)

where $X_{\pi(i)}$ denotes the parent of $X_i$ and $X_{\pi(r)} = \emptyset$ for notational convenience.

For example, consider the tree in Figure 2.1(c). Choosing $X_1$ as the root gives the following factorization:

$$P(X_1, X_2, X_3, X_4, X_5) = \prod_{i=1}^{5} P(X_i | X_{\pi(i)})$$

$$= P(X_1)P(X_2 | X_1)P(X_3 | X_1)P(X_5 | X_1)P(X_4 | X_5)$$  \hspace{1cm} (2.9)

### 2.1 Tasks in a graphical model

Structure learning, parameter learning, and inference are the three main algorithmic tasks in graphical models. As we will show below, when $X_1, ..., X_p$ are all observed in the training data, these tasks have elegant and principled solutions (at least for tree-shaped models).
2.1.1 Parameter Learning

Typically parameter learning is done via maximum likelihood estimation i.e. finding the set of parameters that maximize the log-likelihood of the data:

$$
\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(x^{(n)}, \theta)
$$  \hspace{1cm} (2.10)

where $n = 1, \ldots, N$ indexes the $N$ data points and $x^{(n)} = \{x_1^{(n)}, \ldots, x_p^{(n)}\}$ is the $n^{th}$ sample. Depending on the graph structure, $p(x^{(n)}, \theta)$ factorizes into local probabilities, making the problem statistically tractable. For instance, for tree shaped models, the log-likelihood decomposes as follows:

$$
\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{n=1}^{N} \sum_{i \in V} \sum_{a=1}^{m} \sum_{b=1}^{m} \mathbb{I}[x_i^{(n)} = a \wedge x_{\pi(i)}^{(n)} = b] \log \theta_{X_i=a|X_{\pi(i)}=b} \hspace{1cm} (2.11)
$$

with the constraints that $\theta$ is non-negative and that $\sum_a \theta_{X_i=a|X_{\pi(i)}=b} = 1 \forall i, b$. Taking the derivative and setting equal to zero will show that this reduces to simply computing the conditional probability estimates from the data:

$$
\hat{\theta}_{X_i=a|X_{\pi(i)}=b} = \hat{p}(X_i = a|X_{\pi(i)} = b) = \frac{\text{count}(X_i = a, X_{\pi(i)} = b)}{\text{count}(X_{\pi(i)} = b)}
$$  \hspace{1cm} (2.12)

where $\hat{p}$ denotes the empirical probability estimate and count($\cdot$) returns the number of data instances for which the argument $\cdot$ holds true.

2.1.2 Structure Learning

Given $N$ samples of $p$ variables $X_1, \ldots, X_p$ the goal of structure learning is to find the structure (set of edges $E$) that maximizes the log-likelihood of the data $D$. This problem is in NP-hard if arbitrary structures are permitted \cite{Koller2009}. However, if we limit the class of structures considered to trees, then a provably optimal algorithm, typically called the Chow-Liu algorithm, exists \cite{Edmonds1967, Chow1968}. We
Figure 2.2: Example graphical models for belief propagation.

derive it below.

\[
\log P(D|\hat{\theta}, G) = \sum_{n=1}^{N} \sum_{i=1}^{p} \log \hat{P}(x^{(n)}_i | x^{(n)}_{\pi(i)})
\]

\[
\propto \sum_{x_i, x_{\pi(i)}} \sum_{i=1}^{p} \hat{P}(x_i, x_{\pi(i)}) \log \hat{P}(x_i | x_{\pi(i)})
\]

\[
= \sum_{i=1}^{p} \sum_{x_i, x_{\pi(i)}} \hat{P}(x_i, x_{\pi(i)}) \log \left( \frac{\hat{P}(x_i, x_{\pi(i)})}{\hat{P}(x_{\pi(i)})} \right)
\]

\[
= \sum_{i=1}^{p} \sum_{x_i, x_{\pi(i)}} \hat{P}(x_i, x_{\pi(i)}) \log \left( \frac{\hat{P}(x_i, x_{\pi(i)})\hat{P}(x_i)}{\hat{P}(x_{\pi(i)})\hat{P}(x_i)} \right) + \sum_{i=1}^{p} \sum_{x_i, x_{\pi(i)}} \hat{P}(x_i, x_{\pi(i)}) \log \hat{P}(x_i)
\]

\[
= \sum_{i=1}^{p} MI(x_i, x_{\pi(i)}) - \sum_{i=1}^{p} H(x_i)
\]

(2.13)

where \(H(\cdot)\) denotes entropy and \(MI(\cdot)\) denotes mutual information. Since \(H(x_i)\) does not depend on the graph structure, the best structure is the one that maximizes \(MI(x_i, x_{\pi(i)})\). This can be done using a maximum spanning tree algorithm since we are limiting the structure to a tree.

### 2.1.3 Inference

Mathematically, the goal of inference is to compute the probabilities a subset of variables conditioned on the values of other variables. As mentioned earlier, attempting to brute force this summation can result in exponential computational complexity.

However, for trees the inference problem is tractable and can be solved exactly via message passing / belief propagation [Pearl, 1988]. The key insight is to use the conditional independence statements of the model to push some of the factors outside certain summations that they do not depend on. For simplicity, let us focus on the graphical model in Figure 2.2(a) and consider computing the marginal probability of all the leaf nodes fixed to certain evidence values while
summing out all the internal variables.

The algorithm works bottom up from the leafs, with each node gathering all the messages from its children and passing a new message to its parent. The message at the root equals the desired probability.

The outgoing message from the leaf to its parent is:

\[ m_{i \rightarrow \pi(i)}(x_{\pi(i)}) = P(\bar{x}_i | X_{\pi}) = \sum_{x_i} P(X_i | X_{\pi(i)}) \mathbb{I}[x_i = \bar{x}_i] \]  

(2.14)

where \( \mathbb{I}[x_i = \bar{x}_i] \) (indicator) function is 1 if and only if \( x_i \) equals the corresponding evidence value \( \bar{x}_i \).

**Example**: Assume all the leaf nodes in Figure 2.2(a) are evidence variables gives:

\[ m_{4 \rightarrow 2}(x_2) = P(\bar{x}_4 | x_2) \]
\[ m_{5 \rightarrow 3}(x_3) = P(\bar{x}_5 | x_3) \]
\[ m_{6 \rightarrow 3}(x_3) = P(\bar{x}_6 | x_3) \]  

(2.15)

At internal nodes, the outgoing message is:

\[ m_{i \rightarrow \pi(i)}(x_{\pi(i)}) = P(\bar{x}_i | X_{\pi}) = \sum_{x_i} P(X_i | X_{\pi(i)}) \prod_{j \in c(i)} m_{j \rightarrow i}(x_i) \]  

(2.16)

where \( c(i) \) is the set of the children of \( i \) and \( x_{\pi(i)} \) is the set of leaves in the subtree rooted at \( X_i \). Note that the message from node \( X_i \) to its parent \( X_{\pi(i)} \) can only be computed after all the messages of its children have been computed.

**Example**: Assuming the internal non-root nodes in Figure 2.2(a) are non-evidence variables gives:

\[ m_{2 \rightarrow 1}(x_1) = \sum_{x_2} P(x_2 | x_1)m_{4 \rightarrow 2}(x_2) = \sum_{x_2} P(x_2, \bar{x}_4 | x_2) \]
\[ m_{3 \rightarrow 1}(x_1) = \sum_{x_3} P(x_3 | x_1)m_{5 \rightarrow 3}(x_3)m_{6 \rightarrow 3}(x_3) = \sum_{x_3} P(x_3, \bar{x}_5, \bar{x}_6 | x_1) \]  

(2.17)

Finally at the root,

\[ P(x_E = \bar{x}_E) = \sum_{X_i} P(X_i) \prod_{j \in c(i)} m_{j \rightarrow i}(x_i) \]  

(2.18)

where \( E \) denotes the set of evidence variables (assumed to be all the leaves in this case for simplicity).

**Example**: For the root node in Figure 2.2(a) we have:

\[ P(\bar{x}_4, \bar{x}_5, \bar{x}_6) = \sum_{x_1} P(x_1)m_{2 \rightarrow 1}(x_1)m_{3 \rightarrow 1}(x_1) = \sum_{x_1} P(x_1, \bar{x}_4\bar{x}_5\bar{x}_6) \]  

(2.19)
Note here that each message can be computed in polynomial time: \( O(m^2) \) if all the variables take on \( m \) states.

### 2.1.4 Consistency

The procedures we have described above have a very desirable statistical property called consistency. In particular, if the data is generated by assumed model, then structure learning, parameter learning, and inference (as described in the previous sections for trees) will return the correct structures/parameters/probabilities respectively as the training data size goes to infinity. This can be mathematically expressed as follows:

**Lemma 2.** Assume that underlying distribution \( p(x) \) is generated by a tree graphical model with graph structure \( G = (\mathcal{V}, \mathcal{E}) \) and parameters \( \theta \). Given \( N \) samples generated from \( p(x) \), we learn \( \hat{G} \) from the method in 2.1.2, \( \hat{\theta} \) from 2.1.1, and estimate the probability \( \hat{P}(x) \) from message passing from 2.1.3. Then,

\[
\hat{G} \rightarrow G \quad \text{as} \quad N \rightarrow \infty
\]

\[
\hat{\theta} \rightarrow \theta \quad \text{as} \quad N \rightarrow \infty
\]

\[\hat{P}(x) \rightarrow P(x) \quad \forall x \quad \text{as} \quad N \rightarrow \infty\] (2.20)

### 2.2 Latent Variable Graphical Models

Let us now attempt to model a real world scenario using these techniques. Assume that there are patients entering a doctor’s office with various symptoms e.g. sneezing, coughing etc. These symptoms can be modeled as observed variables since they are easily measured. Consider reasoning about these variables under the graphical model structures shown in Figure 2.1(a)-(c) where we let each of the \( X_i \)'s indicate a different symptom. The model in Figure 2.1(a) is clearly unreasonable since it assumes all the symptoms are independent. Figure 2.1(b) doesn’t make any modeling assumptions, but its computationally and statistically intractable. Figure 2.1(c) also seems suboptimal since in reality all the variables are quite dependent even if we know what a few of them are.

However, there is intuitively a simple explanation underlying the data. In particular if we know the the underlying illness the patient has then it is reasonable to assume the symptoms are conditionally independent given this illness. This “illness” is a latent (hidden) variable since it is not something that appears in the data but something we assume to exist.

Incorporating this latent variable leads to the graphical model in Figure 2.1(d) where the orange node \( H \) intuitively represents the illness, and all the symptoms are conditionally independent given \( H \). Note that the model is rather compact since the number of edges is linear in the number of variables.

More generally, latent variables offer expressive modeling power without sacrificing statistical efficiency since they can model complicated dependencies with a compact number of edges. However, as we will see below, the cost usually comes in the form of computational efficiency. While
learning in observed variables is often based on strong theoretical foundations (e.g. consistency), learning in latent variable models is challenging and largely tackled with local search methods that do generally have any guarantees on consistency.

### 2.2.1 Parameter Learning

The challenge in learning is most easily apparent in parameter learning of latent variable models. Since a subset of the variables are no longer observed, the goal in parameter learning is now to learn the set of parameters that maximizes the marginal log-likelihood:

\[
\hat{\theta}_{\text{MLE}} = \max_{\theta} \sum_{n=1}^{N} \log p(x^{(n)}, \theta) = \max_{\theta} \sum_{n=1}^{N} \log \sum_{h} p(x^{(n)}, h, \theta)
\] (2.21)

where \(x\) denotes the observed variables and \(h\) denotes the latent variables. Unlike fully observed models, the function above is non-concave due to the summation inside the logarithm, which makes the optimization problem NP-hard. Existing optimization strategies mostly revolve around local search and are therefore prone to being trapped in a local optima. One popular approach is Expectation Maximization (EM) ([Dempster et al., 1977]).

EM aims to maximize a surrogate objective function that is a lower bound on the original objective. Let \(Q(h|x^{(n)})\) be some probability distribution. Then the surrogate function can be derived via Jensen’s Inequality:

\[
\sum_{n=1}^{N} \log \sum_{h} p(x^{(n)}, h, \theta) = \sum_{n=1}^{N} \log \sum_{h} Q(h|x^{(n)}) \frac{p(x^{(n)}, h, \theta)}{Q(h|x^{(n))}}
\geq \sum_{n=1}^{N} \sum_{h} Q(h|x^{(n)}) \log \frac{p(x^{(n)}, h, \theta)}{Q(h|x^{(n)})} \quad \text{(using Jensen’s Inequality)}
\leq \sum_{n=1}^{N} \sum_{h} Q(h|x^{(n)}) \log p(x^{(n)}, h, \theta) - \sum_{n=1}^{N} \sum_{h} Q(h|x^{(n)}) \log Q(h|x^{(n)})
\] (2.22)

EM performs coordinate descent on Eq. 2.22, alternatively maximizing \(\theta\) where \(Q\) is fixed (called the M-step) and maximizing \(Q\) when \(\theta\) is fixed (called the E-step).

**M-step** (optimize in terms of \(\theta\)): The second term in Eq. 2.22 is fixed as a function of \(\theta\) so we are left with only optimizing the first term, which is essentially a weighted log-likelihood that is concave:

\[
\hat{\theta}_{\text{MLE}} \leftarrow \max_{\theta} \sum_{n=1}^{N} \sum_{h} Q(h|x^{(n)}) \log p(x^{(n)}, h, \theta)
\] (2.23)
E-step (optimize in terms of $Q$): Eq. 2.22 can be rewritten as

$$\sum_{n=1}^{N} \sum_{h} Q(h|x^{(n)}) \log \frac{p(x^{(n)}, h, \theta)}{Q(h|x^{(n)})} = \sum_{n=1}^{N} \sum_{h} Q(h|x^{(n)}) \log \frac{p(h|x^{(n)}, \theta)p(x^{(n)}, \theta)}{Q(h|x^{(n)})}$$

$$= \sum_{n=1}^{N} \sum_{h} Q(h|x^{(n)}) \log p(x^{(n)}, \theta) - \sum_{n=1}^{N} \sum_{h} Q(h|x^{(n)}) \log \frac{p(h|x^{(n)}, \theta)}{Q(h|x^{(n)})}$$

$$= \sum_{n=1}^{N} \log p(x^{(n)}, \theta) - \sum_{n=1}^{N} KL(Q(h|x^{(n)}) || p(h|x^{(n)}, \theta)) \tag{2.24}$$

where $KL$ indicates KL divergence. Thus, maximizing Eq. 2.24 in terms of $Q$ means setting $Q(h|x^{(n)})$ to $p(h|x^{(n)}, \theta)$. The latter term requires performing inference (e.g. message passing) to compute.

**Advantages/Shortcomings:** It can be shown that EM never decreases the objective function, and thus will eventually converge to a local optimum. Moreover, it was recently shown that given a proper initialization, EM will converge to the maximum likelihood estimate (Balakrishnan et al., 2014). This makes EM highly statistically efficient since it aims to find the maximum likelihood solution, which is the most statistically efficient estimator (Casella and Berger, 2002).

Unfortunately since the search space typically contains many local optima, and thus often EM will get trapped in sub-optimal solutions as with any other local search method. With arbitrary initialization, there is no guarantee that EM will reach the optimal (or even a good) solution, and therefore is not consistent.

Moreover, EM can be quite slow since the E-step requires performing inference over all the training examples which can become very computationally expensive for large datasets. While there do exist methods to speed up EM such as online EM, these approaches introduce additional parameters such as the learning rate, that performance is often sensitive to.

### 2.2.2 Structure Learning

Structure learning of latent variable models is also considerably more challenging than the fully observed case. Recall from the previous section that when learning observed trees we can write the likelihood as a sum of mutual information and entropy (Eq. 2.13).

Unfortunately when doing structure learning with latent variable models it is difficult to compute mutual information among pairs of variables when at least one is not observed. Thus it has largely been tackled by heuristics that do not have any consistency guarantees. Examples from the phylogenetic community include maximum parsimony and maximum likelihood methods (Semple and Steel, 2003). In the machine learning community, Zhang (2004) proposed a search heuristic for hierarchical latent class models by defining a series of local search operations and using EM to compute the likelihood of candidate structures. Harmeling and Williams (2011) proposed a greedy algorithm to learn binary trees by joining two nodes with a high mutual information and iteratively performing EM to compute the mutual information among newly added hidden nodes. Alternatively, Bayesian hierarchical clustering (Heller and Ghahramani, 2005) is an agglomerative clustering technique that merges clusters based on a statistical hypothesis test.
There do exist a class of methods called distance-based approaches (Saitou and Nei, 1987; Semple and Steel, 2003; Choi et al., 2010; Anandkumar et al., 2011) that leverage the concept of additive tree metrics (Buneman, 1971; Erdős et al., 1999) from the phylogenetics community to give structure learning algorithms with theoretical guarantees. We will be using these ideas in Chapters 6 and 8.

2.2.3 Inference

Once the latent variable model has been learned, then inference can proceed as described in 2.1.3. There are some interesting variants like marginal MAP (Jiang et al., 2011; Liu and Ihler, 2013) that are more challenging but they are outside the scope of this thesis.

2.2.4 Motivation for Linear Algebra Point of View

As a result, while latent variable models provide expensive modeling power, existing learning methods, in an attempt to match the statistical efficiency of learning in observed models (i.e. maximum likelihood estimation), suffer from severe computational problems and do not guarantee statistical consistency. By formulating the problem using linear algebra, we will show in this thesis that it is possible to trade-off statistical and computational complexity while preserving modeling power, thus providing a different approach to latent variable learning.

2.3 Review of Hilbert Space Embedding of Distributions

So far the focus has been on modeling discrete variables. However, in Chapter 6 we will discuss generalizing some of our proposed approaches to continuous, non-Gaussian settings where the variables do not easily fit into a parametric family. For example, demographics data is highly skewed while image data often is multimodal.

To do this, we will leverage a recent idea called hilbert space embeddings of distributions (Smola et al., 2007; Song et al., 2009) and its applications to kernel graphical models (Song et al., 2010a,b, 2011a). These techniques provide an elegant way to do probabilistic inference for continuous, nonparametric distributions. We give a brief review of these ideas below.

2.3.1 Intuition

For intuition, consider the Gaussian distribution, which is one continuous distribution that is easily handled by conventional probabilistic modeling methods. Gaussians are easy to model for two reasons:

1. They have a compact set of sufficient statistics: the mean and variance ($\mu, \sigma^2$)
Hilbert space embeddings provide a means to do the same for a variable $X$ from an arbitrary (smooth) distribution $D$, by constructing a sufficient statistic $\mu_X$ for this distribution that allows for efficient probabilistic inference.

One strategy would be to try to take just the mean or the first order moment of the distribution:

$$\mu^1_X = \left( \mathbb{E}[X] \right)$$

However, this statistic is not sufficient since there are many distributions that can have the same mean. We could also take more moments e.g.

$$\mu^2_X = \left( \begin{array}{c} \mathbb{E}[X] \\ \mathbb{E}[X^2] \end{array} \right) \quad \quad \quad \mu^3_X = \left( \begin{array}{c} \mathbb{E}[X] \\ \mathbb{E}[X^2] \\ \mathbb{E}[X^3] \end{array} \right)$$

This allows us to distinguish among a wider range of distributions. For example if distribution $D_1$ and $D_2$ do not share the same mean, variance, or skewness then $\mu^3_X \neq \mu^3_Y$.

Note that we can keep adding moments, possibly even constructing an infinite dimensional statistic $\mu^\infty_X$ that would enable us to distinguish among a large class of distributions. But how will we perform computations with this statistic! This is where kernels come in. While $\mu^\infty_X$ may be infinite dimensional it can be constructed so that taking inner products is easy to compute (i.e. the “kernel trick”).

Before formally defining hilbert space embeddings, we review relevant concepts in functional analysis and reproducing kernel hilbert spaces that are key to the exposition. Table 2.1 provides some informal notation analogs for easier understanding.

### 2.3.2 Hilbert Spaces

**Vector space**: A vector space is a set of objects $\mathcal{V}$ that are closed under linear operations i.e. $v, w \in \mathcal{V} \rightarrow \alpha v + \beta w \in \mathcal{V}$. In elementary linear algebra $\mathcal{V}$ is typically $\mathbb{R}^p$ and each element is a $p$-dimensional real valued vector. However, when discussing hilbert space embeddings we will

---

1Hilbert space embeddings actually do not use the moments to construct sufficient statistics. This is merely an example for intuition.
take a more general point of view and $\mathcal{V}$ may be a space of functions (which can be informally be thought of as infinite dimensional vectors).

**Hilbert space:** A hilbert space is a complete vector space endowed with an inner product $\langle \cdot , \cdot \rangle$. This inner product $\langle f, g \rangle$ satisfies the following properties:

- **Symmetry:** $\langle f, g \rangle = \langle g, f \rangle$
- **Linearity** $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \langle f_1, g \rangle + \langle f_2, g \rangle$
- **Nonnegativity** $\langle f, f \rangle \geq 0$
- **Zero** $\langle f, f \rangle = 0 \implies f = 0$

Some examples of inner products are the vector dot product $\langle v, w \rangle = v^\top w$, or the integral $\langle f, g \rangle = \int f(x)g(x)\,dx$.

**Operator:** An operator $C : \mathcal{F} \rightarrow \mathcal{G}$ maps a function $f \in \mathcal{F}$ in one Hilbert Space to a function $g$ in another i.e. $g = Cf$. For the special case when $\mathcal{F}$ is the space of $p$ dimensional vectors in $\mathbb{R}^p$ and $\mathcal{G}$ is the space of $q$ dimensional vectors in $\mathbb{R}^q$, an operator would be a $p \times q$ matrix.

**Adjoint:** The adjoint $C^\top : \mathcal{G} \rightarrow \mathcal{F}$ of an operator $C : \mathcal{F} \rightarrow \mathcal{G}$ is defined such that $\langle g, Cf \rangle = \langle C^\top g, f \rangle \forall f \in \mathcal{F}, g \in \mathcal{G}$. This is the generalization of transpose/conjugate transpose for matrices i.e. $w^\top Mv = (M^\top w)^\top v$.

**Outer product:** $f \otimes g$ is defined implicitly such that $f \otimes g(h) = \langle g, h \rangle f$. It is the generalization of the traditional vector outer product $v \otimes w = vw^\top$.

**Reproducing Kernel Hilbert Space:** A reproducing kernel hilbert space is a special type of hilbert space that is characterized using a mercer kernel. A mercer kernel $K(x, y)$ is a function of two variables such that

$$\int \int K(x, y)f(x)f(y)\,dx\,dy > 0 \ \forall f \in \mathcal{F} \quad (2.27)$$

Note that this a generalization of a positive definite matrix. The most common kernel we will use is the Gaussian RBF kernel:

$$K(x, y) = \exp\left(\frac{||x - y||^2}{\sigma^2}\right) \quad (2.28)$$

Consider holding one element of the kernel fixed i.e. $\phi_x = K(x, \cdot)$, which is a function of one variable that we call the feature function. The collection of feature functions is called the **feature map**.

For a Gaussian kernel, the feature functions are unnormalized Gaussians:

$$\phi_1(y) = \exp\left(\frac{||1 - y||^2}{\sigma^2}\right) \quad \phi_{1.5}(y) = \exp\left(\frac{||1.5 - y||^2}{\sigma^2}\right) \quad (2.29)$$

and the inner product is defined as $\langle \phi_x, \phi_y \rangle = \langle K(x, \cdot), K(y, \cdot) \rangle := K(x, y)$. Note that $\phi_x(y) = \phi_y(x) = K(x, y)$.
Consider the set of functions that can be formed with linear combinations of these feature functions:

\[
\mathcal{F}_0 = \left\{ f(z) : \sum_{j=1}^{k} \alpha_j \phi_{x_j}(z), \forall k \in \mathbb{N}_+ \text{ and } \forall x_j \in \mathcal{X} \right\} \tag{2.30}
\]

The RKHS \( \mathcal{F} \) is defined as the completion of \( \mathcal{F}_0^2 \). Intuitively one can think of the feature functions as an overcomplete basis for \( \mathcal{F} \).

**Reproducing Property**: An RKHS has the key property that taking the inner product of a function \( f \) with \( \phi_x \) evaluates \( f \) at point \( x \).

\[
\langle f, \phi_x \rangle = \langle \sum_j \alpha_j \phi_{x_j}, \phi_x \rangle = \sum_j \alpha_j \langle \phi_{x_j}, \phi_x \rangle = \sum_j \alpha_j K(x_j, x) = f(x) \tag{2.31}
\]

This is the generalization for \( v(i) = v^\top \delta_i \) in the vector space of \( p \)-dimensional vectors in \( \mathbb{R}^p \) where \( \delta_i \) is an indicator vector that has a 1 in position \( i \) and 0 in all other positions.

### 2.3.3 Hilbert Space Embeddings

Given a kernel \( K \) and associated feature map \( \phi \) and RKHS \( \mathcal{F} \), the hilbert space embedding of a univariate random variable \( X \) is defined as [Smola et al., 2007]:

\[
\mu_X(\cdot) = \mathbb{E}_X[\phi_X] = \int P(x) \phi_x(dx) \tag{2.32}
\]

If the kernel is universal then the map from distributions to embeddings is one-to-one. Examples of universal kernels include the Gaussian RBF Kernel and the Laplace Kernel.

The empirical estimate is defined as

\[
\hat{\mu}_X = \frac{1}{N} \sum_{n=1}^{N} \phi_{x^{(n)}} \tag{2.33}
\]

However, note that since \( \phi \) is infinite dimensional, it is not possible to directly compute this estimate from data (this will be handled in 2.3.5).

**Example (discrete)**: Let us show how this reduces to the traditional probability vector in the discrete case. Consider a random variable \( X \) that takes the values in the set \( \{1, 2, 3, 4\} \). We want to

---

\(^2\)A metric space \( \mathcal{F} \) is complete if for any cauchy sequence of elements in \( \mathcal{F} \) the limit of this sequence is also in \( \mathcal{F} \)
embed it into an RKHS of 4 dimensional vectors in $\mathbb{R}^4$. The feature functions in this RKHS are:

$$
\begin{align*}
\phi_1 &= \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \\
\phi_2 &= \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \\
\phi_3 &= \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \\
\phi_4 &= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}
\end{align*}
$$

(2.34)

Thus,

$$
\mu_X = \mathbb{E}_X[\phi_X] = \sum_{i} P(X = i) \phi_i
$$

(2.35)

**Evaluating the embedding:** $\mu_X$ is a function in $\mathcal{F}$ that characterizes the distribution of $X$. But how do we obtain probability estimates from $\mu_X$? Consider again the discrete example above. In this case, we can obtain $P(X = i)$ by taking the inner product of $\mu_X$ with $\phi_i$ in Eq. (2.34) (the indicator vector) i.e.

$$
P(X = i) = \mu_X(i) = \langle \mu_X, \phi_i \rangle
$$

(2.36)

For the general case, we follow the same strategy leveraging the reproducing property:

$$
\mu_X(i) = \langle \mu_X, \phi_i \rangle = \mathbb{E}[\langle \phi_X, \phi_i \rangle] = \mathbb{E}[K(X, i)]
$$

(2.37)

For a gaussian RBF kernel, this is equal to:

$$
\mathbb{E}\left[ \exp\left(\frac{||i - X||^2}{\sigma^2}\right) \right] \approx \frac{1}{N} \sum_{n=1}^{N} \exp\left(\frac{||i - x^{(n)}||^2}{\sigma^2}\right)
$$

(2.38)

Note this is exactly equal to the (unnormalized) kernel density estimate under a gaussian RBF kernel! Thus, in one dimension the hilbert space embedding of a distribution is equivalent to kernel density estimation. Now let us consider embedding multiple variables.

**Cross-covariance operator:** $C_{XY} = \mathbb{E}_{XY}[\psi_Y \otimes \phi_X]$ for $\phi \in \mathcal{F}, \psi \in \mathcal{G}$. This is the embedding of the joint distribution among two variables. Just like with the one dimensional case, we can evaluate this operator at $(i, j)$ using $\psi_i, \phi_j$ to give the the unormalized (2-dimensional) kernel density estimate of $P(Y = i, X = j)$:

$$
\langle \psi_i, C_{XY} \phi_j \rangle = \mathbb{E}_{XY}[K_G(Y, i)K_F(X, j)]
$$

$$
\approx \frac{1}{N} \sum_{n=1}^{N} \exp\left(\frac{||i - y^{(n)}||^2}{\sigma^2}\right) \exp\left(\frac{||j - x^{(n)}||^2}{\sigma^2}\right) \quad \text{(empirical estimate under gaussian RBF kernels)}
$$

(2.39)

Note that we could represent $p$ variables using a huge cross-covariance operator $C_{1..p} = \mathbb{E}_{X_1..X_p}[\phi_{X_1} \otimes ... \otimes \phi_{X_p}]$. This would be analogous to $p$-dimensional kernel density estimation.
Consequently, this would be pointless since due to the curse of dimensionality, kernel density estimation works poorly when $p$ is even moderately large.

However, the key advantage of Hilbert space embeddings over kernel density estimation in high dimensions is that we will be able to “factorize” the joint cross-covariance operator into smaller operators analogous to how a large discrete probability table can be factorized into smaller factors based on conditional independence assumptions. To do this, we need to define the conditional embedding operator.

**Conditional embedding operator** (Song et al., 2009): The conditional embedding operator is defined as

$$C_{X|Y} = C_{XY}C_{YY}^{-1}$$  \hspace{1cm} (2.40)

and has the following property:

$$E_{X|y} [\phi_X | y] = C_{X|Y} \phi_y$$  \hspace{1cm} (2.41)

This is analogous to “slicing” the conditional probability table in the discrete case i.e. $P_{X|Y=1} = P_{X|Y} \delta_1$ where $P_{X|Y}$ is the conditional probability matrix of $X|Y$ and $\delta_i$ is the indicator vector. Furthermore, using this property it is possible to derive two key operations that are essential to probabilistic inference:

**Lemma 3.** The sum rule: $C_{YX} = C_{Y|X} \mu_X$

*Proof.* We first show the proof for the discrete case in terms of expectations since it directly generalizes to the kernel scenario. Let $P_{X|Y}$ be the conditional probability matrix of $X|Y$ i.e. $P_{X|Y}(i, j) = P(X = i | Y = j)$ and $P_Y$ be the marginal probability vector of $Y$ i.e. $P_Y(i) = P(Y = i)$. Then,

$$P_{X|Y}P_Y = P_{X|Y}E_Y[\delta_Y]$$

$$= E_Y[P_{X|Y} \delta_Y]$$

$$= E_Y[E_{X|Y}[\delta_X]]$$

$$= E_{XY}[\delta_X]$$

$$= P_X$$  \hspace{1cm} (2.42)

Now the kernel case follows analogously:

$$C_{X|Y} \mu_Y = C_{X|Y}E_Y[\phi_Y]$$ \hspace{1cm} (definition of $\mu_Y$)

$$= E_Y[C_{X|Y} \phi_Y]$$

$$= E_Y[E_{X|Y}[\phi_X]]$$ \hspace{1cm} (property of conditional embedding operator)

$$= E_{XY}[\phi_X]$$

$$= \mu_X$$  \hspace{1cm} (2.43)

**Lemma 4.** Product rule: $C_{YX} = C_{Y|X}C_{XX}$. (The proof follows similarly to the one for the sum rule so is not shown)
2.3.4 Kernel Graphical Models

The tools from the previous section allow us to factorize arbitrary multivariate continuous distributions in an analogous manner to discrete probability tables, thus generalizing graphical models to the continuous case as done by Song et al. [2010c,b, 2011a].

Consider the graphical model in Figure 2.2(b) with 4 variables, \( A, B, C, D \). We represent the conditional probability distributions as follows:

- \( C_{AA} \) marginal distribution of \( A \)
- \( C_{B|A} \) conditional distribution of \( B|A \)
- \( C_{C|B} \) conditional distribution of \( C|B \)
- \( C_{D|C} \) conditional distribution of \( D|C \)

Now by applying the sum and product rules we obtain:

\[
C_{A,D} = C_{AA} C_{B|A}^T C_{C|B}^T C_{D|C}^T
\]  

(Song et al. [2010c, 2011a]) generalized this strategy to derive kernel message passing algorithms to perform nonparametric inference in tree and non-tree models.

2.3.5 Empirical computation of hilbert space embeddings

Note that empirical estimates require

\[
\hat{\mu}_X = \frac{1}{N} \sum_{n=1}^{N} \phi(x^{(n)})
\]

\[
\hat{C}_{YX} = \frac{1}{N} \sum_{n=1}^{N} \phi(y^{(n)}) \otimes \phi(x^{(n)})
\]  

These empirical estimates are difficult to directly compute since \( \phi \) is not required to be a finite dimensional vector. However, note that because of the kernel trick it is possible to evaluate these operators at specific points:

\[
\hat{\mu}_X(\bar{x}) = \langle \hat{\mu}_X, \phi_{\bar{x}} \rangle = \frac{1}{N} \sum_{n=1}^{N} K(x^{(n)}, \bar{x})
\]

\[
\hat{C}_{YX}(\bar{y}, \bar{x}) = \langle \phi_{\bar{y}}, \hat{C}_{YX} \phi_{\bar{x}} \rangle = \frac{1}{N} \sum_{n=1}^{N} K(y^{(n)}, \bar{y})K(x^{(n)}, \bar{x})
\]  

However, we will find this unwieldy and slow in practice because performing the sum/chain rule will require manipulation using the kernel matrix \( K \) which is \( N \times N \). Below we will show it is possible to build a low rank approximation of the kernel matrix and build approximate features \( \tilde{\phi} \).
that are much smaller than $N \times 1$.

**Approximate feature maps using incomplete cholesky decomposition**

It is more common (and more computationally efficient) to construct a finite dimensional feature map via incomplete cholesky decomposition \cite{FineScheinberg2002, BachJordan2003}. Note that using QR decomposition,

$$K = \Phi^\top \Phi = R^\top Q R = R^\top R$$  \hspace{1cm} (2.47)

where $\Phi := \{\phi_{x(1)}, ..., \phi_{x(N)}\}$ is the feature map, $Q = \{q_1, ..., q_n\}$ is an orthogonal matrix and $R = \{r_1, ..., r_n\}$ is a matrix of coordinates in the new basis. Note that each $q_i$ may be infinite-dimensional, but each $r_i$ is $N \times 1$.

Interestingly, $R$ presents an alternate feature map (just transformed to be in the basis of $Q$). In particular we can set $\tilde{\phi}_{x(i)} := r_i$. While the original $\phi_{x(i)}$ was infinite dimensional $r_i$ must be finite because the kernel matrix $K$ is $N \times N$ and therefore has at most rank $N$.

Thus, our goal will be to compute $R$. A first strategy would just be to use standard QR decomposition. In this procedure, $q_1 := \frac{\phi_{x(1)}}{\|\phi_{x(1)}\|}$ and then each subsequent $q_i$ is updated as follows:

$$q_j = \phi_{x(j)} - \sum_{t=1}^{j-1} \langle q_t, \phi_{x(t)} \rangle q_t$$ \hspace{1cm} (2.48)

followed by setting $q_j = \frac{q_j}{\|q_j\|}$. This can be rewritten as:

$$q_j = \frac{1}{v_j} \left( \phi_{x(j)} - \sum_{t=1}^{j-1} \langle q_t, \phi_{x(t)} \rangle q_t \right)$$ \hspace{1cm} (2.49)

where $v_j$ is the normalization constant. $R$ can then be recovered by noting that $R(j,i) = \langle q_j, \phi_{x(i)} \rangle$. However, this procedure is not feasible to run since the feature function $\phi_{x(i)}$ (and subsequently $q$) can be infinite dimensional.

Below we show one can recover $R$ directly without $Q$! By definition of $R(j,i)$ we can recover it by taking the inner product with $\phi_{x(i)}$ on both sides:

$$R(j,i) = \langle q_j, \phi_{x(i)} \rangle = \frac{1}{v_j} \left( \langle \phi_{x(j)}, \phi_{x(i)} \rangle - \sum_{t=1}^{j-1} \langle q_t, \phi_{x(i)} \rangle \langle q_t, \phi_{x(j)} \rangle \right)$$

\begin{equation}
R(j,i) = \frac{1}{v_j} \left( K(j,i) - \sum_{t=1}^{j-1} R(t,j) R(t,i) \right), i = j + 1, ..., N \hspace{1cm} (2.50)
\end{equation}

Since $t < j$, $R(t,j)$ and $R(t,i)$ have already been computed. The only question is how to compute $v_j$. To do this recall that $d_i := K(i,i) = \langle \phi_{x(i)}, \phi_{x(i)} \rangle = \|\phi_{x(i)}\|^2$. If we interpret $d_t$ to be the residual

\footnote{We follow the exposition in http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/incompleteCholesky.pdf}
norm squared, then at each step we update $d_i$ by subtracting $R(j, i) = \langle q_j, \phi_i \rangle$ i.e.:

$$d_i \leftarrow d_i - R(j, i)^2$$  \hspace{1cm} (2.51)

Then we set $v_j = \sqrt{d_j}$.

Note that when $\|d\|_2 \leq \epsilon$, then all the data points are approximately represented by the existing basis, suggesting that this criterion presents a principled way to find an approximate solution that can result in considerable computational gains. In particular it can reduce the dimension $\phi$ to be considerably smaller than $N$. This strategy is called the incomplete cholesky decomposition and is shown in Algorithm 1. Using this approximate feature map, we can directly compute the

---

**Algorithm 1** Calculate approximate feature map $\hat{\Phi}$ via incomplete cholesky decomposition of $K$

**In:** Kernel matrix $K$ (assumed to be $N \times N$), tolerance $\epsilon$

**Out:** approximate feature map $\hat{\Phi} := \{\hat{\phi}_{x(1)}, ..., \hat{\phi}_{x(N)}\}$

1. $d \leftarrow \text{diag}(K)$
2. $j \leftarrow 1$
3. while $\|d\|_2 \leq \epsilon$ do
   4. $v_j = \sqrt{d(j)}$
   5. for $i = j + 1; i \leq N; + +$ do
      6. $R(j, i) = \frac{1}{v_j} \left( K(j, i) - \sum_{t=1}^{j-1} R(t, j)R(t, i) \right)$
      7. $d_i \leftarrow d_i - R(j, i)^2$
   8. end for
   9. $j \leftarrow j + 1$
4. end while
5. $\hat{\Phi} \leftarrow R$

---

embedding operators:

$$\hat{\mu}_X = \frac{1}{N} \sum_{n=1}^{N} \hat{\phi}_{x(n)}$$

$$\hat{C}_{YX} = \frac{1}{N} \sum_{n=1}^{N} \hat{\phi}_{y(n)} \otimes \hat{\phi}_{x(n)}$$  \hspace{1cm} (2.52)

Note that this is much more efficient for inference since $\mu_X$ will have length smaller much for $N$ if the value of $\epsilon$ in Algorithm 1 is chosen wisely.
Chapter 3

A Linear Algebra View of Latent Variable Models

Before, delving into the rest of the thesis, we describe latent variable models from the point of view of linear algebra, a formulation that is key to the rest of the thesis.

3.1 Tensor Notation

We first give an introduction to the tensor notation tailored to this thesis.

An Nth order tensor is a multiway array with N “modes”, i.e., N indices $[i_1, i_2, \ldots, i_N]$ are needed to access its entries. Subarrays of a tensor are formed when a subset of the indices is fixed, and we use a colon to denote all elements of a mode. For instance, $A(i_1, \ldots, i_{n-1}:, i_{n+1}, \ldots, i_N)$ are all elements in the nth mode of a tensor $A$ with indices from the other $N-1$ modes fixed to $\{i_1, \ldots, i_{n-1}, i_{n+1}, \ldots, i_N\}$ respectively. Furthermore, we also use the shorthand $i_{pq} = \{i_p, i_{p+1}, \ldots, i_{q-1}, i_q\}$ for consecutive indices, e.g., $A(i_1, \ldots, i_{n-1}:, i_{n+1}:N) = A(i_{1:n-1}, :i_{n+1:N})$.

Furthermore, let $P(X)$ denote probability vectors/matrices/tensors. For example $P(X)$ is the marginal probability vector of $X$ i.e. $P(X)_i = P(X = i)$. Similarly $P(X, Y|Z)$ encodes the conditional probability tensor of $X, Y$ given $Z$ i.e. $P(X, Y|Z)_{i,j,k} = P(X = i, Y = j|Z = k)$.

Mode-specific diagonal matrices/tensors. We use $\delta$ to denote an N-way relation: its entry $\delta(i_{1:N})$ at position $i_{1:N}$ equals 1 when all indexes are the same ($i_1 = i_2 = \ldots = i_N$), and 0 otherwise. We will use $\otimes_d$ to denote repetition of an index $d$ times. For instance, we use $P(\otimes_d X)$ to denote a $d$th order tensor where its entries at $(i_{1:d})$th position are specified by $\delta(i_{1:d})P(X = x_{i_1})$. A diagonal matrix with its diagonal equal to $P(X)$ is then denoted as $P(\otimes_2 X)$. Similarly, we can define a $(d + d')$th order tensor $P(\otimes_d X \otimes_{d'} Y)$ where its $(i_{1:d} j_{1:d'})$th entry corresponds to $\delta(i_{1:d})\delta(j_{1:d'})P(X = x_{i_1}|Y = y_{j_1})$. By default $\otimes$ is equivalent to $\otimes_2$.

Matrix Sum Rule: Note that $P(Y) = P(Y|X)P(X)$ since the matrix multiplication marginalizes out $X$. This is the matrix formulation of $P(Y) = \sum_X P(Y|X)P(X)$. 

32
**Matrix Chain Rule:** If we put $X$ on the diagonal i.e. $\mathcal{P}(\emptyset_2 X)$ then $X$ will not be marginalized out. This gives us $\mathcal{P}(Y, X) = \mathcal{P}(Y|X)\mathcal{P}(\emptyset_2 X)$, which is the matrix formulation of $P(Y, X) = P(Y|X)P(X)$.

**Matricization:** Sometimes we will find it more convenient to rearrange a tensor into a matrix by placing a set of modes on the rows and the rest on the columns. For example, consider the probability tensor $\mathcal{P}(X_1, X_2, X_3, X_4)$ where each of the $X_i$ take on $n$ states. Then one matricization may be the $n^2 \times n^2$ matrix $\mathcal{P}([X_1, X_2], [X_3, X_4])$ where $X_1, X_2$ are on the rows and $X_3, X_4$ are on the columns. Other possible matricizations include $\mathcal{P}([X_1, X_3], [X_2, X_4])$ and $\mathcal{P}([X_1], [X_2, X_3, X_4])$.

**Labeled Tensors:** In contrast to the conventional tensor notation such as the one described in [Kolda and Bader](2009), the ordering of the modes of a tensor will generally not be essential in this thesis. We define labeled tensors to be tensors such that the modes of the tensor are labeled with random variables. Each mode will correspond to a random variable and what is important is to keep track of this correspondence. Therefore, two labeled tensors are equivalent if they have the same set of labels and they can be obtained from each other by a permutation of the modes for which the labels are aligned.

In the matrix case this translates to $A$ and $A^\top$ being equivalent in the sense that $A^\top$ carries the same information as $A$, as long as we remember that the rows of $A^\top$ are the columns of $A$ and vice versa. We will use the following notation to denote this equivalence

$$ A \cong A^\top \quad (3.1) $$

Under this notation, the dimension (or the size) of a mode labeled by variable $X$ will be the same as the number of possible values for variable $X$. Furthermore, when we multiply two labeled tensors together, we will always carry out the operation along (a set of) modes with matching labels.

**Tensor multiplication for labeled tensors.** Let $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ be an $N$th order tensor and $\mathcal{B} \in \mathbb{R}^{J_1 \times J_2 \times \cdots \times J_M}$ be an $M$th order tensor. If $X$ is a common mode label for both $\mathcal{A}$ and $\mathcal{B}$ (w.l.o.g. we assume that this is the first mode, implying also that $I_1 = J_1$), multiplying along this mode will give

$$ C = \mathcal{A} \times_X \mathcal{B} \in \mathbb{R}^{I_2 \times \cdots \times I_N \times J_2 \times \cdots \times J_M}, \quad (3.2) $$

where the entries of $C$ is defined as

$$ C(i_{2:N}, j_{2:M}) = \sum_{i=1}^{I_1} \mathcal{A}(i, i_{2:N}) \mathcal{B}(i, j_{2:M}) $$

Further notation/operations for labeled tensors such as multi-mode multiplication and tensor inversion will be introduced in Chapter 5. Table 3.1 gives provides a quick reference for the notation conventions.
Consider Figure 3.1(a) where there are four observed variables $X_1, X_2, X_3, \text{ and } X_4$ (indicated in blue) and two latent variables $G$ and $H$ (in yellow). Let $S_O$ be the number of observed states (i.e. the number of states each of $X_1, X_2, X_3, \text{ and } X_4$ take on) and $S_H$ be the number of latent states (i.e. the number of states that each of $G, H$ can take on).

Consider the problem of structure learning: recovering the structural relationship among these variables given only samples of $X_1, X_2, X_3, \text{ and } X_4$. One strategy may be to greedily merge the observed variables (Harmeling and Williams, 2011). While this may work sometimes, it will not lead to a consistent solution in general. Similarly, once the structure is known, learning parameters (i.e. the conditional probability tables) is commonly done with the Expectation Maximization (EM) algorithm. EM is essentially coordinate descent on a nonconvex objective and is therefore not guaranteed to return the optimal answer.

However, are these problems really intractable, or are they simply difficult from the likelihood optimization from the point of view? For intuition, let us first examine parameter learning. Consider setting $S_G = S_H = 1$. In this case, $X_1, X_2, X_3, \text{ and } X_4$ are independent as shown in Figure 3.1(b), since no information can travel through the latent variables. The learning problem becomes trivial in this scenario.

On the other side of the spectrum, let $S_H = S_O^4$. In this case the problem becomes equivalent to learning the clique model in Figure 3.1(c) which has $O(S_O^4)$ parameters. In general, if we had $p$ observed variables all connected to one latent variable, then setting $S_H = S_O^p$ would be equivalent to $p$-way clique. This leads to an exponential increase in parameters and defeats the point of a graphical model since there are no conditional independence statements.

However, what about $1 < S_H < S_O^p$. From the optimization point of view these values of $S_H$ lead
to nonconvex objectives and therefore are difficult. However, how much harder is $S_H = S_O^p - 1$ than $S_H = 2$? These are the types of questions that likelihood optimization cannot quantify.

Now, let us approach the problem from a linear algebra point of view. First, consider two variables $X_1$ and $X_2$. In Figure 3.1(c) there is in general nothing we can say is special about the matrix $P(X_1, X_2)$. However, in the case of Figure 3.1(b) $P(X_1, X_2) = P(X_1)P(X_2)^T$. Therefore $P(X_1, X_2)$ is rank one.

More generally, for Figure 3.1(a) $P(X_1, X_2, X_3) = P(X_1 | G)P(G)P(X_2 | G)^T$. Assuming all the matrices on the right hand side are full rank, this implies that $P(X_1, X_2)$ has rank $S_H$, implying that our factorization is a low rank factorization.

Let us now extend this logic to all 4 variables in Figure 3.1(a). Let $P((X_1, X_2), (X_3, X_4))$ be the $S_O^2 \times S_O^2$ joint probability matrix of $X_1, X_2, X_3, X_4$ where the values of $X_1, X_2$ are on the rows and $X_3, X_4$ is on the columns. This matrix has the following low rank factorization:

$$P((X_1, X_2), (X_3, X_4)) = P((X_1, X_2) | G)P(G)P((X_3, X_4) | H)^T$$

Moreover, different matricizations lead to different factorizations i.e.

$$P((X_1, [X_2, X_3, X_4]) = P(X_1 | G)P(G)P([X_2, X_3, X_4] | G)^T$$

$$P((X_1, X_2, X_3), X_4) = P((X_1, X_2, X_3) | H)P(G)P(X_4 | H)^T$$

However, note all these low rank factorizations still have rank $S_H$. Thus, while connections among observed variables in a graphical model specify “hard” conditional independences/dependencies, latent variables induce low rank dependencies among observed variables. Small $S_H$ translates into small rank, implying simple dependencies among observed variables. As $S_H$ gets larger, the rank of the model increases and the dependencies become more complicated. As we will see later, the cardinality of $S_H$ plays a key role in determining the difficulty of learning with latent variables.

While this linear algebra point of view may seem to be just a simple reformulation, it leads to an interesting perspective that inspires solutions to a variety of problems that we will explore in later chapters:

- **Parameter Learning**: The particular low rank factorization we showed above is special in that it is composed of conditional probability matrices. However, low rank factorizations are in general not unique. For any matrix factorization $M = LR$, $M = LS^{-1}SR$ is also a low rank factorization where $S$ can be any invertible matrix. Thus, a natural question to ask is do there exist other factorizations that only depend on observed variables and therefore do not require EM to learn?

- **Structure Learning**: The fact that $P((X_1, X_2), (X_3, X_4))$ is low rank but $P((X_1, X_3), (X_2, X_4))$ is not reveals aspects about the latent structure of the model. How can this intuition be used to learn the latent structure underlying a set of continuous, non-Gaussian variables? In particular, we will leverage the notion of additive tree metrics [Saitou and Nei, 1987; Lake 1994; Choi et al. 2011].
- **Modeling**: Can this connection between linear algebra and probabilistic modeling motivate new models and solutions for language modeling and unsupervised parsing? For example, in language modeling can we create $n$-grams for non-integer values of $n$ i.e. 1.5-grams, 2.5-grams etc.?
Part I

Spectral Learning Algorithms for Graphical Models
Chapter 4

A Spectral Algorithm for Latent Tree Graphical Models

Using the insights from the previous chapter, we first tackle parameter learning in latent variable models. As mentioned earlier, conventional approaches have primarily relied on likelihood maximization and local search heuristics such as expectation maximization (EM) \cite{Dempster1977}. In addition to the problem of local optima, EM can require many iterations to reach a prescribed training precision, and high dimensional problems can dramatically slow down EM.

While EM tries to recover the full set of parameters in latent variable models, in many applications it is the inference task that is most interesting. For instance, in speech classification, we are interested in estimating the likelihood of a test sequence under different models; in quantitative finance, we are interested in predicting the price of one stock given the prices of other stocks; or in biological analysis, we are interested in forecasting the expression of one gene given perturbations to other genes. In all these examples, the inference task involves estimating either the joint or conditional distribution of a set of observed variables. Ideally, we want to avoid explicitly recovering the parameters related to latent variables (which leads to non-convex problems), and proceed directly to the interested quantities.

Recently, \cite{Hsu2009} and \cite{Bailly2009} proposed spectral algorithms for learning hidden Markov models (HMM) which directly estimates the joint distribution of the observed variables without recovering the HMM model parameters. The major computation of the algorithms involves a singular value decomposition (SVD) of small marginal probability matrices involving pairs of observed variables. Compared to EM, this spectral algorithm does not have the problem of local optima, and one can formally study its statistical properties. However, this spectral algorithm is specific to HMMs, and it is not clear whether their techniques can be extend to latent variable models with other topologies. \cite{Mossel2006} also proposed a spectral algorithm for latent variable models which applies to arbitrary tree topologies, but they made very restrictive assumptions: all variables (observed and latent) have exactly the same number of states, and all conditional probability tables (CPT) are invertible. Under these conditions, they derived a spectral algorithm that can explicitly recovers all CPTs from marginals of triples of observed variables. In many applications, however, latent variables can represent factors simpler than the noisy observations, and the number of hidden states can be smaller than that of the observed
states. In these cases, the CPTs are no longer invertible, which renders this spectral algorithm no longer applicable. Moreover, as we show in Chapter 7, even when in settings when this method is applicable it performs poorly in practice.

**Contribution of this chapter:** In this paper, we propose a novel spectral algorithm for latent variable models with arbitrary tree topologies that allows the number of hidden states to be smaller or larger than the number of observed states. Instead of learning the original conditional probability tables (like Mossel and Roch (2006) do), we learn a linear transformation of these parameters that still enables efficient inference. This enables a provably consistent and local-optima-free learning algorithm with sample complexity analysis. Empirical results indicate that our spectral approach performs comparably or better than EM, while being 1-2 orders of magnitude faster.

Key to approach is our extensive use of tensor algebra that enables generalization of spectral learning to models beyond HMMs. Later works such as Cohen et al. (2012); Cohen and Collins (2012); Dhillon et al. (2012a) have leveraged our tensor formulation for supervised parsing with latent variables. Our work is also closely related with tensor decomposition (Kolda and Bader, 2009b), that we discuss at the end of the chapter.

**Outline:** Some intuition is presented using a small example. We then establish a tensor message passing scheme for latent tree graphical models. Subsequently the spectral representation is derived and different aspects are discussed including sample complexity analysis and practical considerations. Finally, empirical results are presented.

**Prerequisites:** This chapter assumes a general understanding of latent variable models as presented in 2.2, the connection between latent variable models and low rank factorization in Chapter 3, and the tensor notation in 3.1.

### 4.1 Intuition

Recall that in Chapter 3 we had established that latent variables introduce low rank factorizations, i.e. $M = LR$, over the marginal probability of the observed variables. For the example in Figure 3.1 we can set,

$$M := \mathcal{P}([X_1, X_2], [X_3, X_4])$$
$$L := \mathcal{P}([X_1, X_2]|H)\mathcal{P}(\varnothing_2 H) = \mathcal{P}([X_1, X_2], H)$$
$$R := \mathcal{P}([X_3, X_4]|H)$$

(4.1)

However, low rank factorizations are in general not unique. For any matrix factorization, $M = LR$, we also have that

$$M = \begin{pmatrix} L & S^{-1} \end{pmatrix} \begin{pmatrix} \tilde{L} & \tilde{R} \end{pmatrix}$$

(4.2)

and thus an *alternate factorization* $M = \tilde{L}\tilde{R}$. However, note that while $L$ and $R$ may be probability tables, while $\tilde{L}$ and $\tilde{R}$ can may have negative values (since even if $S$ is non-negative, $S^{-1}$ may have negative entries).
The natural question to ask is that does there exist a low rank factorization that is only a function of the observed variables $X_1, X_2, X_3, X_4$? Interestingly, the answer is yes!

To see why, consider the following expansions:

$$P([X_1, X_2], X_3) = P([X_1, X_2]|H)P(\emptyset H)P(X_3|H)^T$$

$$P(X_2, [X_3, X_4]) = P(X_2|H)P(\emptyset H)P([X_3, X_4]|H)^T$$

(4.3)

The product of the green terms (underlined), in some order, is $P([X_1, X_2], [X_3, X_4])$ (what we want!). The product of the red terms (not underlined), in some order, is $P(X_2, X_3)$ (what we need to eliminate). This leads us the following alternate factorization that only depends on observed variables:

$$P([X_1, X_2], [X_3, X_4]) = P([X_1, X_2], X_3)P(X_2, X_3)^{-1}P(X_2, [X_3, X_4])$$

(4.4)

Note that while the original (CPT) factorization was a product of conditional probability matrices, the alternate factorization is a product of marginal probability matrices and their inverses.

How does this relate to our original factorization? Consider setting $S = P(X_3|H)$. Then we can prove that

$$LS = P([X_1, X_2], X_3)$$

$$S^{-1}R = P(X_2, X_3)^{-1}P(X_2, [X_3, X_4])$$

(4.5)

### 4.2 Notation for Latent Tree Graphical Models

We now generalize this intuition to arbitrary latent tree graphical models. A latent tree model defines a joint probability distribution over a set of $O$ observed variables $\mathcal{O} = \{X_1, \ldots, X_O\}$ and a set of $H$ hidden variables $\mathcal{H} = \{X_{O+1}, \ldots, X_{O+H}\}$. The complete set of variables is denoted by $\mathcal{X} = \mathcal{O} \cup \mathcal{H}$. For simplicity, we assume that all observed variables have $S_O$ states and all hidden variables have $S_H$ states. For now assume $S_H = S_O$ and all conditional probability tables have full rank (we address the more general case later).

The joint distribution of $\mathcal{X}$ in a latent tree model is fully characterized by a set of conditional probability tables (CPTs). More specifically, we can select an arbitrary node in the tree as the root, and sort the nodes in the tree in topological order. Then the set of CPTs between nodes and their parents $P(X_i|X_{\pi(i)}$) are sufficient to characterize the joint distribution (the root node $X_r$ has no
parent, i.e., \( P(X_i|X_{\pi(i)}) = P(X_i) \),

\[
P(x_1, \ldots, x_{O+H}) = \prod_{i=1}^{O+H} P(x_i|x_{\pi(i)}) \quad (4.6)
\]

Compared to tree models which are defined solely on observed variables (e.g., models obtained from the (Chow and Liu, 1968b) algorithm), latent tree models encompass a much larger classes of models, allowing more flexibility in modeling observed variables. This is evident if we compute the marginal distribution of the observed variables by summing out the latent ones, which gives us a clique over the observed nodes:

\[
P(x_1, \ldots, x_O) = \sum_{x_{O+1}} \cdots \sum_{x_{O+H}} \prod_{i=1}^{O+H} P(x_i|x_{\pi(i)}) \quad (4.7)
\]

For simplicity, assume that all leaves are observed variables and all internal nodes are latent. For further notation, let \( X_i^r \) be some leaf in the subtree rooted at \( X_i \), and let \( T_i \) denote the set of all such leaves. and \( X_{-i}^r \) be a leaf not in the subtree rooted at \( X_i \) and let \( T_{-i} \) denote the set of all those leaves. Let \( c_j(i) \) denote the \( j^{th} \) child of node \( X_i \), and where \( X_i \) has \( \alpha_i \) children.

### 4.3 Derivation of Spectral Algorithm

Our spectral derivation has three main components:

1. Showing how the marginal probability tensor \( P(X_1, \ldots, X_O) \) can be factorized into a collection of lower order tensors where the maximum tensor order is equal to the maximum degree of the tree. This can be shown to be equivalent to a tensor message passing scheme.

2. Inserting the invertible transformations \( F \) and \( F^{-1} \) into the message passing scheme to define an alternate low rank factorization that still yields the same marginal probabilities as the original representation.

3. Setting \( F \) such that the factors in this alternate factorization only depend on small groups of observed variables.

For simplicity of exposition, we assume all internal nodes have exactly three neighbors. The factorization generalizes to trees of arbitrary topology, but this comes at the cost of memory since the maximum order of the tensor in the factorization is equal to the maximum degree of a node in the tree. However, we show that in Chapter \( \text{[7]} \) that there exists a representation where the maximum tensor order is 3 regardless of the order of the tree.\footnote{While it is possible to binarize a tree while adding additional latent variables and setting the factors to enforce certain constraints, these constraints are difficult to enforce in the spectral algorithm.}
4.3.1 Factorizing the Marginal Probability Tensor

Consider Figure 4.1. First let us exploit the low rank structure implied by the root $A$. We can get the following factorizations of the marginal probability tensor $P(E, F, G, H, I, J)$:

$$P(E, F, G, H, I, J) = P(E,F|A)P(\otimes_2 A)P(G,H,I,J|A)^T$$


$$P(E, F, G, H, I, J) = P(E, F, I,J|A)P(\otimes_2 A)P(G,H|A)^T$$

But how to combine all of these into one factorization? With tensors! Let the root probability $P(A)$ be embedded in a third order labeled tensor: $P(\otimes_3 A)$. One can see that by using (labeled) tensor-matrix multiplication,

$$P(E, F, G, H, I, J) = P(\otimes_3 A) \times_A P(E,F|A) \times_A P(G,H|A) \times_A P(I,J|A)$$

We then proceed to factorize recursively e.g. $P(E,F|A) = P(\otimes_2 B|A) \times_B P(E|B) \times_B P(F|B)$.

4.3.2 Tensor Message Passing

The recursive factorization procedure described above can be expressed as message passing, a form that will more be notationally convenient for our subsequent derivation. Instead of attempting to reconstruct the entire marginal probability tensor let us simply focus on a single element of this tensor (e.g. $P(\bar{e}, f, \bar{g}, h, \bar{i}, \bar{j})$). Again Figure 4.1 will be used as a running example. Associate each node with the following labeled tensor (which are essentially the original conditional probability tables (CPTs) embedded into higher-order tensors):

- root: $R := P(\otimes_3 X_r)$
- internal node: $T_i := P(\otimes_2 X_i|X_{\pi(i)})$
- leaf: $L_i = P(X_i|X_{\pi(i)})$

Assuming the leaf node is an evidence variable, the message it passes to its parent is the following vector:

$$m_i = L_i \times_i \delta_{\bar{x}_i}$$

**Example:** $M_E = P(\bar{e}|B) = P(E|B) \times_E \delta_e$

Similarly, the message from a non-root internal node sends to its parent is:

$$m_i = T_i \times_i m_{c1(i)} \times_i m_{c2(i)}$$

**Example:** $m_B = P(\otimes_2 B|A) \times_B m_E \times_B m_F$

and finally the root agglomerates the messages to give the final probability:

$$P(\bar{x}_1, ..., \bar{x}_O) = R \times_r m_{c1(r)} \times_r m_{c2(r)} \times_r m_{c3(r)}$$
Example: \( P(\bar{e}, \bar{f}, \bar{g}, h, \bar{i}, \bar{j}) = P(\varnothing_3 A) \times_A m_B \times_A m_C \times_A m_D. \)

4.3.3 Transformed Representation

Next, note we do not need to recover the tensor representation explicitly if our focus is to perform inference using the message passing algorithm as in (7.6)–(7.8). As long as we can recover the running example in Figure 4.1,

\[
P(\bar{e}, \bar{f}, \bar{g}, h, \bar{i}, \bar{j}) = P(\varnothing_3 A) \times_A m_B \times_A m_C \times_A m_D
\]

Next, expand \( I \) as a matrix inversion pair \( F \) and \( F^{-1} \) and regroup the terms:

\[
P(\bar{e}, \bar{f}, \bar{g}, h, \bar{i}, \bar{j}) = (P(\varnothing_3 A) \times_A F_B \times_A F_C \times_A F_D)
\]

\[
\times_{\omega_B} \left( m_B \times_A F_B^{-1} \right)
\]

\[
\times_{\omega_C} \left( m_C \times_A F_C^{-1} \right)
\]

\[
\times_{\omega_D} \left( m_D \times_A F_D^{-1} \right)
\]

where \( \omega_B, \omega_C, \omega_D \) are mode labels that depend on the definitions of \( F_B, F_C, F_D \) respectively and are defined in the next section. This proceeds recursively e.g.

\[
m_D = P(\bar{i}, \bar{j}|A) = (P(\varnothing_2 D|A) \times_D F_1 \times_D F_2) \times_{\omega_D} \left( P(\bar{j}|D) \times_D F_j^{-1} \right)
\]

In general, we can define the following transformed tensor representation:

- **root:** \( \bar{R} = P(\varnothing_3 X_r) \times_r F_{c_1(r)} \times_r F_{c_2(r)} \times_r F_{c_3(r)} \)
- **internal:** \( \bar{T}_i = P(\varnothing_2 X_l|X_{\pi(i)}) \times_{\pi(i)} F_i^{-1} \times_i F_{c_1(i)} \times_i F_{c_2(i)} \)
- **leaf:** \( \bar{L}_i = P(X_l|X_{\pi(i)}) \times_{\pi(i)} F_i^{-1} \)

4.3.4 Observable Representation

We now derive the observable representation by choosing \( F \) and \( F^{-1} \) systematically, so that we can recover each transformed parameter using the marginal probability of a small set of observed variables. This is summarized by the lemma below, generalizing our intuition from Section 4.1.

**Lemma 5.** For each \( X_i \), set \( F_i = P(X_i|X_{\pi(i)}) \) with mode labels \( \omega_i = X_r, X_{\pi(i)}. \) Then we have the following:

- \( \bar{R} = P(X_{c_1(r)}, X_{c_2(r)}, X_{c_3(r)}) \)
We will describe the purpose of $U_i$ and the thin SVD in the next section.

### 4.3.5 Training and Test

**Training:** In training, we compute the observable (spectral) parameters as shown in Algorithm 2. We now prove the following relation:

- $\mathcal{T}_i = \mathcal{P}(X_{c_1(i)'}, X_{c_2(i)'}, X_{-r}) \times_{-r} \mathcal{P}(X_{-r}, X_r)^{-1}$
- $\mathcal{L}_i = \mathcal{P}(X_i, X_{-r}) \times_{-r} \mathcal{P}(X_{-r}, X_r)^{-1} = I$

**Proof:** Below is the derivation for internal nodes (the root and leaf are just special cases). Recall that $X_r$ is some leaf in the subtree rooted at $X_i$ and $X_{-r}$ is a leaf not in the subtree rooted at $X_i$. First note that,

$$\tilde{T}_i = \mathcal{P}(X_{c_1(i)'}, X_{c_2(i)'}, X_{-r}) \times_{-r} \mathcal{P}(X_{-r}, X_r)^{-1}$$

We now prove the following relation:

$$\mathcal{T}_i \times_{r} \mathcal{P}(X_r, X_{-r}) = \mathcal{P}(X_{c_1(i)'}, X_{c_2(i)'}, X_{-r})$$

(4.13)

Note that $\mathcal{P}(X_r, X_{-r}) = \mathcal{P}(X_r|X_{\pi(i)})\mathcal{P}(\otimes X_{\pi(i)})\mathcal{P}(X_{-r}|X_{\pi(i)})^\top$. Thus,

$$\tilde{T}_i \times_{r} \mathcal{P}(X_r, X_{-r}) = \mathcal{P}(X_{c_1(i)'}, X_{c_2(i)'}, X_{-r}) \times_{r} \mathcal{P}(X_r|X_{\pi(i)})^\top \mathcal{P}(\otimes X_{\pi(i)}) \mathcal{P}(X_{-r}|X_{\pi(i)}) $$

$$= \mathcal{P}(X_{c_1(i)'}, X_{c_2(i)'}, X_{-r}) \times_{r} \mathcal{P}(\otimes X_{\pi(i)}) \mathcal{P}(X_{-r}|X_{\pi(i)})^\top$$

(4.14)

From here, one can conclude that

$$\tilde{T}_i = \mathcal{P}(X_{c_1(i)'}, X_{c_2(i)'}, X_{-r}) \times_{-r} \mathcal{P}(X_{-r}, X_r)^{-1}$$

(4.14)

**Example:** In the example in Figure 4.1, one possible value of $\tilde{T}_B$ is $\mathcal{P}(E, F, G) \times_G \mathcal{P}(G, E)^{-1}$. Table 4.1 shows a table comparing the original, tensor, and observable (spectral) parameters.

#### Table 4.1: Original, tensor, and observable parameters for nodes $A$, $B$, and $E$ in the example in Figure 4.1 for the case where $S_H = S_O$. Note that the observable parameters are not unique, but this is one valid set.

<table>
<thead>
<tr>
<th>Original</th>
<th>Tensor</th>
<th>Observable(Spectral)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(A)$</td>
<td>$\mathcal{P}(\otimes A)$</td>
<td>$\mathcal{P}(E, G, I)$</td>
</tr>
<tr>
<td>$P(B</td>
<td>A)$</td>
<td>$\mathcal{P}(\otimes_2 B</td>
</tr>
<tr>
<td>$P(E</td>
<td>B)$</td>
<td>$\mathcal{P}(E</td>
</tr>
</tbody>
</table>

44
**Test:** In test time, the goal is that given a set of evidence variables $E = \{x_{e1}, ..., x_{eg}\}$ to compute the probability $P(x_{e1}, ..., x_{eg})$. We can do this using the tensor message passing scheme shown before except with the transformed parameters instead of the original parameters. The only difference is that depending on if $x_i$ is an evidence variable or not the leaf message changes.

If leaf is an evidence variable multiply by the indicator vector to select the appropriate row:

$$m_i = \tilde{L}_i \times_i \delta_{x_i}$$

(4.15)

If leaf is not an evidence variable sum over $X_i$:

$$m_i = \tilde{L}_i \times_i 1$$

(4.16)

The general algorithm is in given in Algorithm 3.

4.3.6 Expectation Form

We will find it useful later in this chapter, as well as in Chapter 6 to consider an equivalent form of the spectral algorithm in terms of expectations. We simply give the algorithm below again in this

---

**Algorithm 2** Spectral learning algorithm for latent tree graphical models

**In:** Tree topology and $N$ i.i.d. samples $\{x_1^0, ..., x_n^0\}_{n=1}^N$

**Out:** Estimated observable root, internal, and leaf parameters, $\hat{R}, \hat{T}_i$ for each internal node, $\tilde{L}_i$ for each leaf node,

1. For each node $X_i$, perform a “thin” singular value decomposition of $\hat{P}(X_r, X_{-r}) = U \Sigma V^\top$; let $\hat{U}_i = U(:,1:S_i)$ be the first $S_i$ principal left singular vectors.

2. Compute observable parameters as:

   $\hat{R} = \hat{P}(X_{c1}, X_{c2}, X_{c3}) \times_{c1} U_{c1} \times_{c2} U_{c2} \times_{c3} U_{c3}$

   $\hat{T}_i = \hat{P}(X_{r}, X_{-r}) \times_{-r} (\hat{P}(X_r, X_{-r}) \times_r \hat{U}_i) \times_{c1} U_{c1} \times_{c2} U_{c2}$

   $\tilde{L}_i = \hat{P}(X_i, X_{-r}) \times_{-r} (\hat{P}(X_r, X_{-r}) \times_r \hat{U}_i) \times_r \hat{U}_i = \hat{U}_i$

---

**Algorithm 3** Inference with Spectral Parameters

**In:** Tree topology, set of spectral parameters $\hat{R}, \hat{T}_i$ for each internal node, $\tilde{L}_i$ for each leaf node, and set of evidence $E = \{\bar{x}_{e1}, ..., \bar{x}_{eg}\}$

**Out:** Estimated probability $P(\bar{x}_{e1}, ..., \bar{x}_{eg})$

1. In reverse topological order, each node accumulates at message at leaf and sends to parent
   - Evidence Leaf: $m_i = \tilde{L}_i \times_i \delta_{x_i}$
   - Non-Evidence Leaf: $m_i = \tilde{L}_i \times_i 1$
   - Internal Node: $m_i = \hat{T}_i \times_i \hat{m}_{c1} \times_i \hat{m}_{c2}$
   - Root: $P(\bar{x}_{e1}, ..., \bar{x}_{eg}) = \hat{R} \times_i m_{c1} \times_i m_{c2} \times_i \hat{m}_{c3}$
form for future use. As noted in [2, 3] probability matrices can be written in expectation form:

\[ R = \mathcal{P}(\otimes_3 X_r) = \mathbb{E}_{X_r}[\delta_{X_r} \otimes \delta_{X_o} \otimes \delta_{X_r}] \]

\[ T_i = \mathcal{P}(\otimes_2 X_i|X_{\pi(i)}) = \mathbb{E}_{X_i|X_{\pi(0)}}[\delta_{X_i} \otimes \delta_{X_i}|X_{\pi(0)}] \]

\[ L_i = \mathcal{P}(X_i|X_{\pi(0)}) = \mathbb{E}_{X_i|X_{\pi(0)}}[\delta_{X_i}|X_{\pi(0)}] \]

where \( \delta_{X_i} \) is the indicator vector with a one in the position of the value of \( X_i \).

Setting \( F_i = \mathcal{P}(X_i|H) = \mathbb{E}[\delta_{X_i}|H] \) gives the expectation forms of the observable representation:

- \( \tilde{R} = \mathbb{E}[\delta_{X_{1,0}} \otimes \delta_{X_{2,0}} \otimes \delta_{X_{3,0}}] \)
- \( T_i = \mathbb{E}[\delta_{X_{1,0}} \otimes \delta_{X_{2,0}} \otimes \delta_{X_{r,0}}] \times_{-r} \mathbb{E}[\delta_{X_{r,0}} \otimes \delta_{X_{r,0}}]^{-1} \)
- \( L_i = \mathbb{E}[\delta_{X_{r,0}} \otimes \delta_{X_{r,0}}] \times_{-r} \mathbb{E}[\delta_{X_{r,0}} \otimes \delta_{X_{r,0}}]^{-1} \)

### 4.4 Dealing with \( S_H \neq S_O \)

In the above derivations we have assumed \( S_H = S_O \) and all underlying probability tables are full rank. This guarantees that the inverse of \( \mathcal{P}(X_r, X_{-r}) \) exists and that:

\[ \mathcal{P}(X_r, X_{-r})^{-1} = (\mathcal{P}(X_{r,0}|X_{\pi(0)})^{-1})^\top \mathcal{P}(\otimes X_{\pi(0)})^{-1} \mathcal{P}(X_r|X_{\pi(0)})^{-1} \]  

(4.17)

which is essential for the correctness of the algorithm (Lemma 5). However, in general, \( S_H \neq S_O \) causing either the inverse to not exist or the above equality to not hold. We detail these two cases below.

#### 4.4.1 What if \( S_H < S_O \)?

Based on our intuition in Chapter 3, rank (equivalent to the number of latent states if we assume full rank CPTs) is a measure of the amount of long range dependency in a latent model. Thus, \( S_H < S_O \) actually means shorter range dependencies and should be easier to solve than the \( S_H = S_O \) case. However, the algorithm described in the previous section does not directly apply since \( \mathcal{P}(X_r, X_{-r}) \) is no longer invertible. The solution is to simply project all the matrices/tensors to the \( S_H \) dimensional space where the inverse exists. As a result, instead of choosing \( F = \mathcal{P}(X_r|X_{\pi(0)}) \) we choose \( F_i = U_i^\top \mathcal{P}(X_r|X_{\pi(0)}) \) where \( U_i \) is the top \( S_H \) left singular vectors of \( \mathcal{P}(X_r, X_{-r}) \) and replacing \( F^{-1} \) with \( F^\top \) where \( \top \) indicates pseudo-inverse. We show why this works below using the derivation of the internal node (root and leaf are special cases):

\[ \tilde{T}_i = \mathcal{P}(\otimes_2 X_i|X_{\pi(0)}) \times_{-i}^\top \times_i F_{c_1(i)} \times_i F_{c_2(i)} \]

\[ \tilde{T}_i = \mathcal{P}(\otimes_2 X_i|X_{\pi(0)}) \times_{-i}^\top \times_i \left(U_i^\top \mathcal{P}(X_r|X_{\pi(0)})\right) \times_i U_{c_1(i)}^\top \mathcal{P}(X_{c_1(i)}|X_i) \times_i U_{c_2(i)}^\top \mathcal{P}(X_{c_2(i)}|X_i) \]

\[ \tilde{T}_i = \mathcal{P}(X_{c_1(i)}^\top, X_{c_2(i)}^\top|X_{\pi(0)}) \times_{-i} \left(U_i^\top \mathcal{P}(X_r|X_{\pi(0)})\right) \times_{c_1(i)}^\top U_{c_1(i)}^\top \times_{c_2(i)}^\top U_{c_2(i)}^\top \]

(4.18)
Again we prove the following relation:

\[
\tilde{T}_i \times_p u_i^T \mathcal{P}(X_{p'}, X_{-p'}) = \mathcal{P}(X_{c_1(i)'}, \ldots, X_{c_2(i)'}, X_{-p'}) \times_{c_1(i)'} u_{c_1(i)'}^T c_{2(i)} u_{c_2(i)}^T
\]

(4.19)

where again, \(X_{-p'}\) is an observed leaf that is not a descendant of \(X_i\).

The proof follow similarly as before by expanding \(\mathcal{P}(X_{p'}, X_{-p'})\):

\[
\tilde{T}_i \times_p u_i^T \mathcal{P}(X_{p'}, X_{-p'}) = \mathcal{P}(X_{c_1(i)'}, \ldots, X_{c_2(i)'}, X_{-p'}) \times_{c_1(i)'} u_{c_1(i)'}^T c_{2(i)} u_{c_2(i)}^T
\]

\[
\begin{align*}
&= \mathcal{P}(X_{c_1(i)'}, \ldots, X_{c_2(i)'}) \times_{c_1(i)'} u_{c_1(i)'}^T c_{2(i)} u_{c_2(i)}^T \\
&\quad \times_\pi u_i^T (\mathcal{P}(X_{p'}|X_{pi(i)})(\mathcal{P}(X_{-p'}|X_{pi(i)})^\top c_{1(i)} c_{2(i)} u_{c_2(i)}^T \\
&\quad \times_\pi (\mathcal{P}(X_{c_1(i)'}, \ldots, X_{c_2(i)'}) \times_{c_1(i)'} u_{c_1(i)'}^T c_{2(i)} u_{c_2(i)}^T )^\dagger \\
&\quad \times_\pi (\mathcal{P}(X_{c_1(i)'}, \ldots, X_{c_2(i)'}, X_{-p'}) \times_{c_1(i)'} u_{c_1(i)'}^T c_{2(i)} u_{c_2(i)}^T )^\dagger
\end{align*}
\]

(4.20)

From here, one can conclude that

\[
\tilde{T}_i = \mathcal{P}(X_{c_1(i)'}, \ldots, X_{c_2(i)'}, X_{-p'}) \times_{c_1(i)'} u_{c_1(i)'}^T c_{2(i)} u_{c_2(i)}^T
\]

(4.21)

This explains the presence of \(U_i\) in Algorithm 2.

4.4.2 What if \(S_H > S_O\)?

On the other hand, when \(S_H > S_O\), \(\mathcal{P}(X_{p'}, X_{-p'})\) is full rank but the relation \(\mathcal{P}(X_{p'}, X_{-p'})^{-1} = (\mathcal{P}(X_{p'}|X_{pi(i)})^{-1} \mathcal{P}(X_{-p'}|X_{pi(i)})^{-1}\) no longer holds. This scenario indicates longer range dependencies and therefore a more challenging learning scenario. For intuition on why this is the case, consider the following generative process:

1. With probability 0.5, let \(S = X\), and with probability 0.5 let \(S = Y\).
2. Print \(A m\) times.
3. Print \(S\)
4. Go back to step (2)

With \(m = 1\) we either generate AXAXAXA or AYAYAYA... With \(m = 2\) we either generate AAXAXAXA.. or AAYAYAYA..

Note that an HMM, which is a natural way to model this generative process, needs \(2m\) states to be able to correctly capture the long range dynamics. This is because it needs to remember the count as well as whether we picked \(S = X\) or \(S = Y\) in Step 1. However, the number of observed states \(m\) does not change, so spectral learning as we have described it so far, will break for \(m > 2\). How to deal with this in the spectral framework?

[1] Siddiqi et al. (2010) and Cohen et al. (2012) proposed solutions to this problem by constructing features out of groups of observations. In our case, this would mean constructing features out of all the observed descendants in the subtree rooted at \(X_i\) (denoted as \(T_i\)) instead of using just one descendant \(X_i\).
This can be accomplished using the expectation form we derived in 4.3.6 and setting $T_i = \mathbb{E}[\phi_{T_i}|X_{n(i)}]$ where $\phi_{T_i}$ is an arbitrary feature vector that can be a function of any of the variables in $T_i$. Note that $U_i$ would now be the SVD of $\mathbb{E}[\phi_{T_i} \otimes \phi_{T_{-i}}]$ instead of $\mathbb{E}[\delta_{X_i'} \otimes \delta_{X_{-i}'}]$. For the example in Figure 4.1, we could set $T_B = \mathbb{E}[\phi_{T_B}|A]$ where $\phi_{T_B} := [\delta_E; \delta_I]$ (; represents vertical concatenation) and $\phi_{T_{-B}} = [\delta_C; \delta_H; \delta_I; \delta_I]$. This would give the following observable representation:

\begin{itemize}
  \item $\tilde{R} = \mathbb{E}[\phi_{T_{c(1)}} \otimes \phi_{T_{c(0)}} \otimes \phi_{T_{c(1)}}] \times_{T_{c(0)}} U_{c(0)} \times_{T_{c(2)}} U_{c(2)} \times_{T_{c(0)}} U_{c(0)}$
  \item $\tilde{T}_i = \mathbb{E}[\phi_{T_{c(1)}} \otimes \phi_{T_{c(2)}} \otimes \phi_{T_{-c}'}] \times_{T_{-c}'} (U_i^\top \mathbb{E}[\phi_{T_i} \otimes \phi_{T_{-i}}])^\dagger \times_{T_{c(0)}} U_{c(0)} \times_{T_{c(2)}} U_{c(2)}$
  \item $L_i = \mathbb{E}[\phi_{T_{-c}'} \otimes \phi_{T_{-c}'}] \times_{T_{-c}'} (U_i^\top \mathbb{E}[\phi_{T_i} \otimes \phi_{T_{-i}}])^\dagger$
\end{itemize}

4.4.3 Linear Systems to Improve Stability

There is another technique we can use to improve performance in practice. Consider the below relation for non-root internal nodes, which is key to the derivation of the algorithm (same as Eq. 4.13):

$$\tilde{T}_i \times_{T'} \mathcal{P}(X_{T'}, X_{-T'}) = \mathcal{P}(X_{c(1)'}, X_{c(2)'}, X_{-T'})$$

(4.22)

which then implies that

$$\tilde{T}_i = \mathcal{P}(X_{c(1)'}, X_{c(2)'}, X_{-T'}) \times_{T'} \mathcal{P}(X_{-T'}, X_{T'})^{-1}$$

(4.23)

However, there may be many choices of $X_{-T'}$ (which we denote with $X_{-T'}^{(1)}, \ldots, X_{-T'}^{(Z)}$) for which the above equality is true:

$$\tilde{T}_i \times_{T'} \mathcal{P}(X_{T'}, X_{-T'}^{(1)}) = \mathcal{P}(X_{c(1)'}, X_{c(2)'}, X_{-T'}^{(1)})$$

$$\tilde{T}_i \times_{T'} \mathcal{P}(X_{T'}, X_{-T'}^{(2)}) = \mathcal{P}(X_{c(1)'}, X_{c(2)'}, X_{-T'}^{(2)})$$

$$\tilde{T}_i \times_{T'} \mathcal{P}(X_{T'}, X_{-T'}^{(Z)}) = \mathcal{P}(X_{c(1)'}, X_{c(2)'}, X_{-T'}^{(Z)})$$

This defines an over-constrained linear system, and we can solve for $\tilde{T}_i$ using least squares. In the case where $S_O > S_H$, and the projection matrix $U_i$ is needed, our system of equations becomes:

$$\tilde{T}_i \times_{T'} (U_i^\top \mathcal{P}(X_{T'}, X_{-T'}^{(1)})) = \mathcal{P}(X_{c(1)'}, X_{c(2)'}, X_{-T'}^{(1)}) \times_{T_{c(1)'}} U_{c(1)} \times_{T_{c(2)'}} U_{c(2)}$$

$$\tilde{T}_i \times_{T'} (U_i^\top \mathcal{P}(X_{T'}, X_{-T'}^{(2)})) = \mathcal{P}(X_{c(1)'}, X_{c(2)'}, X_{-T'}^{(2)}) \times_{T_{c(1)'}} U_{c(1)} \times_{T_{c(2)'}} U_{c(2)}$$

$$\tilde{T}_i \times_{T'} (U_i^\top \mathcal{P}(X_{T'}, X_{-T'}^{(Z)})) = \mathcal{P}(X_{c(1)'}, X_{c(2)'}, X_{-T'}^{(Z)}) \times_{T_{c(1)'}} U_{c(1)} \times_{T_{c(2)'}} U_{c(2)}$$

In this case, one good choice of $U_i$ is the top singular vectors of the matrix formed by the horizontal concatenation of $\mathcal{P}(X_{T'}, X_{-T'}^{(1)}), \ldots, \mathcal{P}(X_{T'}, X_{-T'}^{(Z)})$.

This linear system method allows for more robust estimation, especially in smaller sample sizes.
(at the cost of more computation). It can be applied to the leaf nodes as well. One does not need to set up the linear system with all the valid choices; a subset is also acceptable.

4.5 Sample Complexity Analysis

We analyze the sample complexity of Algorithm 2 and show that it depends on the tree topology and the spectral properties of the true model. Let $d_i$ be the degree of node $i$ in the tree.

**Theorem 1.** Let $d_{\text{max}} = \max_i d_i$. Then, for any $\epsilon > 0, 0 < \delta < 1$, if

$$N \geq \Theta \left( \frac{4S^2_H \ln |O + H| |O + H|^2}{\epsilon^2 \alpha^4} \right)$$

where $\sigma_1(*)$ returns the $\tau^{th}$ largest singular value and

$$\alpha = \min_i \sigma_{S_H}(P(X_r, X_{-r})), \quad \beta = \min_i \sigma_{S_H}(F_i)$$

Then with probability $1 - \delta$,

$$\sum_{x_1, \ldots, x_O} |\hat{P}_{\text{spectral}}(x_1, \ldots, x_O) - P(x_1, \ldots, x_O)| \leq \epsilon,$$

where $\hat{P}_{\text{spectral}}$ indicates the probability returned by the spectral algorithm.

**Proof.** The proof is a special case of Theorem 3 in Chapter 5 for latent junction trees. □

Note the dependence on the singular values of certain probability tensors. In fully observed models, the accuracy of the learned parameters depends only on how close the empirical estimates of the factors are to the true factors. However, our spectral algorithm also depends on how close the inverses of these empirical estimates are to the true inverses, which depends on the spectral properties of the matrices (Stewart and Sun, 1990).

4.6 Empirical Results

We present a set of synthetic experiments to evaluate our method in a variety of settings. We compare our method to the Expectation Maximization (EM) algorithm (Dempster et al., 1977). Convergence of EM is determined by measuring the change in the log likelihood at iteration $t$ (denoted by $f(t)$) over the average: $\frac{|f(t) - f(t-1)|}{\text{avg}(|f(t) - f(t-1)|)} \leq \gamma$. To explore the computational / accuracy trade-off we consider a variety of choices of $\gamma$. We denote $EM-$ to indicate convergence threshold $\gamma = 1e-3$ (low precision EM), $EM$ to indicate convergence threshold $\gamma = 1e-4$ (standard precision as used in Murphy (2005)), and $EM+$ to indicate convergence threshold $\gamma = 1e-5$ (high precision EM). All variants given 5 restarts by default, and our method is given a linear system size of 10 (as discussed in 4.4.3) unless specified otherwise.
We compare the methods on both accuracy and training time (All approaches have similar test runtime). For accuracy, we measure the performance of joint estimation using \( \epsilon = \frac{|\hat{P}(x_1, ..., x_O) - P(x_1, ..., x_O)|}{P(x_1, ..., x_O)} \) averaged over 1000 test points randomly drawn from the underlying model. For runtime, the training time of each method is reported in seconds. Results are reported as the training set size size is varied from 100 to 100,000.

The first set of experiments demonstrates performance as a function of tree depth (Figure 4.2). All trees generated were binary, and we experimented with depths 4, 5, 6, and 7. \( S_O \) was set to 4 and \( S_H \) was set to 2. In terms of accuracy, we generally observe 3 distinct regions, low-sample size, mid-sample size, and large sample size. In the low sample size region, EM and EM+ tend to overfit to the training data, and our spectral algorithm and EM- perform the best. In the mid-sample size region, EM- does poorly since it does not converge to a high enough precision. Our method does well but EM and EM+ tend to do better since they benefit from a smaller number of parameters and lack of dependence on singular values. However, once a certain sample size is reached (the large sample size region), our spectral algorithm consistently outperforms all the EM variants which suffer from local minima and convergence issues. From the perspective of runtime our method is considerably faster than all the variants, and is almost two orders of magnitude faster than EM+.

The second set of experiments (Figure 4.3) evaluates the effect of the number of restarts on EM. The tree structure used is a binary tree with depth 6 (64 total nodes), and as in the previous experiment \( S_O \) was set to 4 and \( S_H \) was set to 2. The results are as expected. More restarts tends to improve the performance of EM while also increasing runtime. However, we see that the effect of restarts is considerably smaller than the effect of the precision threshold for this family of models.

Lastly we evaluate the effect of the linear system size on the spectral algorithm as shown in Figure 4.4. In some sense this is the parameter that trades off accuracy and runtime for our method analogous to how the convergence threshold / number of restarts trade off accuracy and speed for EM. We fix the tree topology to a binary tree of depth 6 (64 nodes) and fix \( S_O = 5 \). \( S_H \) is varied from 2 to 5 and we compare linear system sizes of 1, 2, 5, and 10. When \( S_H = 2 \) all the variants perform similarly, but as the number of hidden states is increased, increasing the size of the linear system produces a considerably more accurate solution. Runtime is not considerably different across the different variants due to the efficacy of MATLAB linear algebra libraries (although of course a larger linear system involves more computational cost).

4.7 Connections with Tensor Decomposition

Before moving on to the next chapter, we briefly discuss how spectral learning and EM are related to tensor decomposition methods in the applied mathematics community.

**PARAFAC Decomposition** (Harshman, 1970): Given an \( n^{th} \) order tensor \( \mathcal{T} \) of dimensions \( D_1 \times ... \times D_n \), PARAFAC decomposition of rank \( \bar{K} \) consists of one \( K \times 1 \) vector \( z \), and \( n \) matrices \( M_1, ..., M_n \), each of dimension \( D_i \times K \). \( \mathcal{T} \) is then approximated as:

\[
\mathcal{T}_{\text{parafac}}(i_1, ..., i_n) = \sum_{k=1}^{K} z(k)M_1(i_1, k)....M_n(i_n, k)
\]

(4.24)
Figure 4.2: Comparison of our spectral algorithm (blue) to standard EM with convergence set at $1e - 4$ (red), low precision EM with convergence set at $1e - 3$ (green) and high precision EM with converge set at $1e - 5$ (magenta) for binary trees of various depth. All variants of EM are given 5 random restarts and we use a max linear system size of 10 for the spectral algorithm. Number of observed states ($S_O$) is fixed to 4 and number of latent states ($S_H$) is fixed to 2. Both errors and runtimes in log scale.

Figure 4.3: Comparison of our spectral algorithm (blue) to standard EM with convergence set at $1e - 4$ (red), low precision EM with convergence set at $1e - 3$ (green) and high precision EM with converge set at $1e - 5$ (magenta) for different numbers of restarts provided given to EM. The latent tree structure is a binary tree with depth 6 (64 total nodes). Number of observed states ($S_O$) is fixed to 4 and number of latent states ($S_H$) is fixed to 2. For the spectral algorithm we use a max linear system size of 10. Both errors and runtimes in log scale.
### Figure 4.4: Evaluation of the spectral algorithm for different values of the max linear system size: 1-blue, 2-red, 5-green, 10-magenta. The latent tree structure is a binary tree with depth 6 (64 total nodes). Number of observed states \( S_0 \) is fixed to 5 and number of latent states \( S_H \) is varied from 2 to 5 across (a), (b), (c), (d). Both errors and runtimes in log scale.

### Figure 4.5: Example of a latent tree model with three observed nodes

### Tucker Decomposition [Hitchcock, 1927; Tucker, 1966]:
We have seen that PARAFAC decomposition is related to EM. We now describe a tensor decomposition that is related to spectral learning.

Given an \( n^{th} \) order tensor \( T \) of dimensions \( D_1 \times ... \times D_n \), TUCKER decomposition of rank \( K \) consists of one \( K \times K \times ... \times K \) core tensor \( Z \), and \( n \) matrices \( M_1, ..., M_n \), each of dimension \( D_i \times K \). \( T \) is then approximated as:

\[
T_{\text{tucker}}(i_1, ..., i_n) = \sum_{k_1=1}^{K} ... \sum_{k_n=1}^{K} Z(k_1, ..., k_n)M_1(i_1, k_1)....M_n(i_n, k_n)
\]  

(4.25)

Note that unlike the PARAFAC Decomposition, the Tucker decomposition has a provably consis-
tent learning algorithm using SVD. Recall that $T_{\{A\},\{B\}}$ is the unfolding of $T$ such that the modes in the set $A$ are on the rows and the modes of set $B$ are on the columns. Then the following algorithm recovers the Tucker decomposition:

- Set $M_i$ to the top $k$ singular vectors of $T_{\{i\},\{1,\ldots,i-1,i+1,\ldots,n\}}$ for each $i$
- Define $Z := \mathcal{T} \times_1 M_1 \ldots \times_n M_n$

Note that this is equivalent to the spectral algorithm for the graphical model shown in Figure 4.5.

### 4.8 Conclusion

This chapter presents a spectral learning algorithm for latent tree graphical models that is provably consistent unlike traditional methods. Empirical results show the algorithm gives a favorable statistical/computational trade-off to the EM algorithm across a variety of settings often giving a 1-2 order of magnitude speed up.

This chapter also serves a foundation for the rest of Part I of this thesis. The following 3 chapters will extend this work in a variety of ways.
Chapter 5

A Spectral Algorithm for Latent Junction Trees

The previous chapter proposed a provably consistent spectral learning algorithm for latent tree models. However, latent structures beyond trees, such as higher order HMMs (Kundu et al., 1989), factorial HMMs (Ghahramani and Jordan, 1997) and Dynamic Bayesian Networks (Murphy, 2002), are needed and have been proven useful in many real world problems. The challenges for generalizing spectral algorithms to general latent structured models include the larger factors, more complicated conditional independence structures, and the need to sum out multiple variables simultaneously.

Contribution of this chapter: In this thesis, we take one step toward developing spectral methods for more general graphical models by proposing a spectral algorithm for latent junction trees. The key idea of our approach is to embed the clique potentials of the junction tree into higher order tensors such that the computation of the marginal probability of observed variables can be carried out via tensor operations. While this novel representation leads only to a moderate increase in the number parameters for junction trees of low treewidth, it allows us to design an algorithm that can recover a transformed version of the tensor parameterization and ensure that the joint probability of observed variables are computed correctly and consistently. As with the spectral algorithm described in the previous chapter, the main computation of the algorithm involves only tensor operations and singular value decompositions which can be orders of magnitude faster than EM algorithms for large datasets. To our knowledge, this is the first provably consistent parameter learning technique for latent variable models beyond trees. In our experiments, we show that our spectral algorithm can be almost 2 orders of magnitude faster than EM while at the same achieving comparable or better accuracy. Our spectral algorithm also achieves comparable accuracy to EM on real data.

Outline: A high level overview of our approach is given in Figure 5.1. We first provide some additional notation and background on latent junction trees. We then derive the spectral algorithm by representing junction tree message passing with tensor operations, and then transform this representation into one that only depends on observed variables. Finally, we analyze the sample complexity of our method and evaluate it on synthetic and real datasets.
Figure 5.1: Our algorithm for local-minimum-free learning of latent variable models consist of four major steps. (1) First, we transform a model into a junction tree, such that each node in the junction tree corresponds to a maximal clique of variables in the triangulated graph of the original model. (2) Then we embed the clique potentials of the junction tree into higher order tensors and express the marginal distribution of the observed variables as a tensor-tensor/matrix multiplication according to the message passing algorithm. (3) Next we transform the tensor representation by inserting a pair of transformations between those tensor-tensor/matrix operations. Each pair of transformations is chosen so that they are inversions of each other. (4) Lastly, we show that each transformed representation is a function of only observed variables. Thus, we can estimate each individual transformed tensor quantity using samples from observed variables.

Prerequisites: This chapter assumes a general understanding of latent variable models as presented in Chapter 2, the connection between latent variable models and low rank factorization in Chapter 3, and the tensor notation in Section 3.1. It is also recommended to read Chapter 4 first.

5.1 Additional Tensor Notation

In addition to the notation in Section 3.1, we define some additional operations that are unique to this chapter.

Tensor multi-mode multiplication for labeled tensors. Let $\sigma = \{X_1, \ldots, X_k\}$ be an arbitrary set of $k$ modes ($k$ variables) shared by $\mathcal{A}$ and $\mathcal{B}$ (w.l.o.g. we assume these labels correspond to the first $k$ modes, and $I_1 = I_1, \ldots, I_k = I_k$ holds for the corresponding dimensions). Then multiplying two labeled tensors $\mathcal{A}$ and $\mathcal{B}$ along $\sigma$ results in

$$\mathcal{D} = \mathcal{A} \times_\sigma \mathcal{B} \in \mathbb{R}^{i_1 \times \ldots \times i_N \times i_{k+1} \times \ldots \times i_M}, \quad (5.1)$$

where the entries of $\mathcal{D}$ are defined as

$$\mathcal{D}(i_{k+1:N}, j_{k+1:M}) = \sum_{i_1:k} \mathcal{A}(i_{1:k}, i_{k+1:N}) \mathcal{B}(i_{1:k}, j_{k+1:M}).$$

Multi-mode multiplication can also be interpreted as reshaping the $\sigma$ modes of $\mathcal{A}$ and $\mathcal{B}$ into a single mode and doing single-mode tensor multiplication. Furthermore, tensor multiplication with labels is symmetric in its arguments, i.e., $\mathcal{A} \times_\sigma \mathcal{B} \equiv \mathcal{B} \times_\sigma \mathcal{A}$.

Mode-specific labeled identity tensor. We now define our notion of identity tensor with respect to a set of modes $\sigma = \{X_1, \ldots, X_k\}$. Let $\mathcal{A}$ be a labeled tensor with mode labels containing $\sigma$, and $I_\sigma$ be a tensor with $2K$ modes with mode labels $\{X_1, \ldots, X_K, X_1, \ldots, X_K\}$. Then $I_\sigma$ is an identity tensor.
tensor with respect to modes $\sigma$ if

$\mathcal{A} \times_\sigma I_\sigma \cong \mathcal{A}$. \hspace{1cm} (5.2)

One can also understand $I_\sigma$ using its matrix representation: flattening $I_\sigma$ with respect to $\sigma$ (the first $\sigma$ modes mapped to rows and the second $\sigma$ modes mapped to columns) results in an identity matrix.

**Mode-specific labeled tensor inversion.** Let $\mathcal{F}, \mathcal{F}^{-1} \in \mathbb{R}^{c_1(i) \times \cdots \times l_k \times l_{k+1} \times \cdots \times l_{k+k'}}$ be labeled tensors of order $K + K'$, and both have two sets of mode labels $\sigma = \{X_1, \ldots, X_K\}$ and $\sigma' = \{X_{K+1}, \ldots, X_{K+K'}\}$. Then $\mathcal{F}^{-1}$ is the inverse of $\mathcal{F}$ w.r.t. modes $\omega$ if and only if

$\mathcal{F} \times_\omega \mathcal{F}^{-1} \cong I_\sigma$. \hspace{1cm} (5.3)

Multimode inversion can also be interpreted as reshaping $\mathcal{F}$ with respect to $\omega$ into a matrix of size $(c_1(i) \ldots l_k) \times (l_{k+1} \ldots l_{k+k'})$, taking the inverse, and then rearranging back into a tensor. Thus the existence and uniqueness of this inverse can be characterized by the rank of the matricized version of $\mathcal{F}$.

### 5.2 Latent Junction Trees

In this chapter, we will focus on discrete latent variable models where the number of states, $S_H$, for each hidden variable is much smaller than the number of states, $S_O$, for each observed variable. A latent variable model defines a joint probability distribution over a set of variables $\mathcal{X} = \mathcal{O} \cup \mathcal{H}$. Let $O := |\mathcal{O}|$ and $H := |\mathcal{H}|$. Here, $\mathcal{O}$ denotes the set of observed variables, $\{X_1, \ldots, X_O\}$ and $\mathcal{H}$ denotes the set of hidden variables, $\{X_{O+1}, \ldots, X_{O+H}\}$.

We will focus on latent variable models where the structure of the model is a junction tree of low treewidth $[1999]$. Each node $C_i$ in a junction tree corresponds to a subset (clique) of variables from the original graphical model. We will also use $C_i$ to denote the collection of variables contained in the node, i.e. $C_i \subseteq \mathcal{X}$. Let $\mathcal{C}$ denote the set of all clique nodes. The treewidth is then the size of a largest clique in a junction tree minus one, that is $t = \max_{C_i \in \mathcal{C}} |C_i| - 1$. Furthermore, we associate each edge in a junction tree with a separator set $S_{ij} := C_i \cap C_j$ which contains the common variables of the two cliques $C_i$ and $C_j$ it is connected to. If we condition on all variables in any $S_{ij}$, the variables on different sides of $S_{ij}$ will become independent.

As in the previous chapter, without loss of generality, we assume that each internal clique node in the junction tree has exactly 3 neighbors. Then we can pick a clique $C_r$ as the root of the tree and reorient all edges away from the root to induce a topological ordering of the clique nodes. Given the ordering, the root node will have 3 children nodes, denoted as $C_{c_1(r)}$, $C_{c_2(r)}$, and $C_{c_3(r)}$. Each other internal node $C_i$ will have a unique parent node, denoted as $C_{\pi(i)}$, and 2 children nodes denoted as $C_{c_1(i)}$ and $C_{c_2(i)}$. Each leaf node $C_j$ is only connected with its unique parent node $C_{\pi(j)}$. Furthermore, we can simplify the notation for the separator set between a node $C_i$ and its parent $C_{\pi(i)}$ as $S_i = C_i \cap C_{\pi(i)}$, omitting the index for the parent node. Then the remainder set of a node is defined as $R_i = C_i \setminus S_i$. We also assume w.l.o.g. that if $C_i$ is a leaf in the junction tree, $R_i$ consists of

---

1If this is not the case, the derivation is similar but notationally much heavier.
only observed variables. We will use \( r_i \) to denote an instantiation of the set of variables in \( R_i \). See Figure 5.2 for an illustration of notation.

Given a root and a topological ordering of the nodes in a junction tree, the joint distribution of all variables \( \mathcal{X} \) can be factorized according to

\[
P(\mathcal{X}) = \prod_{i=1}^{|\mathcal{X}|} P(R_i|S_i),
\]

where each CPT \( P(R_i|S_i) \), also called a clique potential, corresponds to a node \( C_i \). The number of parameters needed to specify the model is \( O(|C_i|S_i') \), linear in the number of cliques but exponential in the tree width \( t \). Then the marginal distribution of the observed variables can be obtained by summing over the latent variables,

\[
P(x_1, \ldots, x_O) = \sum_{x_{O+1}} \cdots \sum_{x_{O+H}} \left[ \prod_{i=1}^{O} P(R_i|S_i) \right],
\]

Note that each (non-leaf) remainder set \( R_i \) contains a small subset of all latent variables. The presence of latent variables introduces complicated dependency between observed variables, while at the same time only a small number of parameters corresponding to the entries in the CPTs are needed to specify the model.

The process of eliminating the latent variables can be carried out efficiently via message passing. Each node only needs to sum out a small number of variables and then the intermediate result, called the message, is passed to its parent for further processing. In the end the root node incorporates all messages from its children and produces the final result \( P(x_1, \ldots, x_O) \). The local summation step, called the message update, can be generically written as

\[
M(S_i) = \sum_{R_i} P(R_i|S_i) M(S_{c_1(i)}) M(S_{c_2(i)})
\]

where we use \( M(S_i) \) to denote the intermediate results of eliminating variables in the remainder set \( R_i \). This message update is then carried out recursively according the reverse topological order of the junction tree until we reach the root node. The local summation step for the leaf nodes and root node can be viewed as special cases of (5.6).

Example. The message update at the internal node \( C_{BCDE} \) in Figure 5.2 is

\[
M([C,E]) = \sum_{B,D} P(B,D|C,E) P(f|B,C) P(g|B,D).
\]

### 5.3 Tensor Representation for Message Passing

Although the parametrization of latent junction trees using CPTs is very compact and inference (message passing) can be carried out efficiently, parameters in this representation can be difficult to learn. Since the likelihood of the observed data is no longer concave in the latent parameters,
Figure 5.2: Example latent variable models with variables $\mathcal{X} = \{A, B, C, D, E, F, G, H, I, \ldots\}$, the observed variables are $\mathcal{O} = \{F, G, H, \ldots\}$ (only partially drawn). Its corresponding junction tree is shown in the middle panel. Corresponding to this junction tree, we also show the general notation for it in the rightmost panel.

local search heuristics, such as EM, are often employed to learning the parameters. Therefore, our goal is to design a new representation for latent junction trees, such that subsequent learning can be carried out in a local-optima-free fashion.

In this section, we will develop a new representation for the message update in (5.6) by embedding each CPT $P(R_i|S_i)$ into a higher order tensor $P(C_i)$. This form is a strict generalization of the tensor representation proposed in Chapter 4.

### 5.3.1 Embed CPTs to higher order tensors

As we can see from (7.4), the joint probability distribution of all variables can be represented by a set of conditional distributions over just subsets of variables. Each one of this conditionals is a low order tensor. For example in Figure 5.2, the CPT corresponding to the clique node $C_{BCDE}$ would be a 4th order tensor $P(B, D|C, E)$ where each variable corresponds to a different mode of the tensor. However, this representation is not suitable for deriving the observable representation since message passing cannot be defined easily using the tensor multiplication/sum connection shown above. Instead we will embed these tensors into even higher order tensors to facilitate the computation. The key idea is to introduce duplicate indexes using the mode-specific identity tensors, such that the sum-product steps in message updates can be expressed as tensor multiplications.

More specifically, the number of times a mode of the tensor is duplicated will depend on how many times the corresponding variable in the clique $C_i$ appears in the separator sets incident to $C_i$. We can define the count for a variable $X_j \in C_i$ as

$$d_{ji} = \mathbb{I}[X_j \in S_i] + \mathbb{I}[X_j \in S_{c_1(i)}] + \mathbb{I}[X_j \in S_{c_2(i)}],$$

where $\mathbb{I}[\cdot]$ is an indicator function taking value 1 if its argument is true and 0 otherwise. Then the
labeled tensor representation of the node $C_i$ is

$$\mathcal{P}(C_i) := \mathcal{P}(\ldots, (\otimes_{d_{ji}} X_j), \ldots, (\otimes_{d_{rj}} X_r), \ldots),$$

$$\forall X_j \in R_i \quad \forall X_r \in S_i$$

(5.8)

where the labels for the modes of the tensor are the combined labels of the separator sets, i.e., $\{S_i, S_{c_1(0)}, S_{c_2(0)}\}$. The number of times a variable is repeated in the label set is exactly equal to $d_{ji}$.

Essentially, the labeled tensor $\mathcal{P}(C_i)$ contains exactly the same information as the original CPT $P(R_i|S_i)$. Furthermore, $\mathcal{P}(C_i)$ has a lot of zero entries, and the entries from $P(R_i|S_i)$ simply embedded in the higher order tensor $\mathcal{P}(C_i)$. Suppose all variables in node $C_i$ are latent variables each taking $S_H$ values. Then the number of entries needed to specify $P(R_i|S_i)$ is $S_H^{|C_i|}$, while the high order tensor $\mathcal{P}(C_i)$ has $S_H^{d_{ji}}$ entries where $d_{ji} := \sum_{i \in C_i} d_{ji}$ which is never smaller than $S_H^{|C_i|}$. In a sense, the parametrization using higher order tensor $\mathcal{P}(C_i)$ is less compact than the parametrization using the original CPTs. However, constructing the tensor $\mathcal{P}$ this way allows us to express the junction tree message update step in (5.6) as tensor multiplications (more details in the next section), and then we can leverage tools from tensor analysis to design a local-minimum-free learning algorithm.

The tensor representation for the leaf nodes and the root node are special cases of the representation in (5.8). The tensor representation at a leaf node $C_i$ is simply equal to $\mathcal{P}(C_i) = \mathcal{P}(R_i|S_i)$. The root node $C_r$ has no parent, so $\mathcal{P}(C_r) = \mathcal{P}(\ldots, (\otimes_{d_{ji}} X_j), \ldots), \forall X_j \in C_r$. Furthermore, since $d_{ji}$ is simply a count of how many times a variable in $C_i$ appears in each of the incident separators, the size of each tensor does not depend on which clique node was selected as the root.

**Example.** In Figure 5.2, node $C_{BCDE}$ corresponds to CPT $P(B, D|C, E)$. Its high order tensor representation is $\mathcal{P}(C_{BCDE}) = P(\otimes_2 B, D|\otimes_2 C, E)$, since both $B$ and $C$ occur twice in the separator sets incident to $C_{BCDE}$. Therefore the tensor $\mathcal{P}([B, C, D, E])$ is a 6th order tensor with mode labels $\{B, B, D, C, E\}$.

### 5.3.2 Tensor message passing

With the higher order tensor representation for clique potentials in the junction tree as in (5.8), we can express the message update step in (5.6) as tensor multiplications. Consequently, we can compute the marginal distribution of the observed variables $\mathcal{O}$ in equation (5.5) recursively using a sequence of tensor multiplications. More specifically the general message update equation for a node in a junction tree can be expressed as

$$\mathcal{M}(S_i) = \mathcal{P}(C_i) \times_{S_{c_1(0)}} \mathcal{M}(S_{c_1(0)}) \times_{S_{c_2(0)}} \mathcal{M}(S_{c_2(0)}).$$

(5.9)

Here the modes of the tensor $\mathcal{P}(C_i)$ are labeled by the variables, and the mode labels are used to carry out tensor multiplications as explained in Section 5.1.

The tensor message passing steps in leaf nodes and the root node are special cases of the tensor message update in equation (5.9). The outgoing message $\mathcal{M}(S_i)$ at a leaf node $C_i$ can be computed by simply setting all variables in $R_i$ to the actual observed values $r_i$, i.e.,

$$\mathcal{M}(S_i) = \mathcal{P}(C_i)_{R_i=r_i} = P(R_i = r_i|S_i).$$

(5.10)
At the root, we obtain the marginal probability of the observed variables by aggregating all incoming messages from its 3 children, \( i.e. \)
\[
P(x_1, ..., x_O) = P(C_r) \times S_{c_1(r)} M(S_{c_1(r)}) \times S_{c_2(r)} M(S_{c_2(r)}) \times S_{c_3(r)} M(S_{c_3(r)})
\] (5.11)

**Example.** For Figure 5.2, using the following tensors
\[
P(\{B, C, D, E\}) = P(\varnothing B, D | \varnothing C, E)
M(\{B, C\}) = P(f|B, C)
M(\{B, D\}) = P(g|B, D),
\]
we can write the message update for node \( C_{BCDE} \) in the form of equation (5.9) as
\[
M(\{C, E\}) = P(\{B, C, D, E\}) \times_{\{B, C\}} M(\{B, C\})
\times_{\{B, D\}} M(\{B, D\}).
\]

Note how the tensor multiplication sums out \( B \) and \( D \): \( P(\{B, C, D, E\}) \) has two \( B \) labels, and it appears in the subscripts of tensor multiplication twice; \( D \) appears once in the label and in the subscript of tensor multiplication respectively. Similarly, \( C \) is not summed out since there are two \( C \) labels but it appears only once in the subscript of tensor multiplication.

### 5.4 Transformed Representation

Explicitly learning the tensor representation in (5.8) is still an intractable problem. Our key observation, as that in Chapter 4 is that we do not need to recover the tensor representation explicitly if our focus is to perform inference using the message passing algorithm as in (5.9)–(5.11). As long as we can recover the tensor representation up to some invertible transformation, we can still obtain the correct marginal probability \( P(x_1, ..., x_O) \).

More specifically, we can insert a mode-specific identity tensor \( I_\sigma \) into the message update equation in (5.9) without changing the outgoing message. Subsequently, we can then replace the mode-specific identity tensor by a pair of labeled tensors, \( F \) and \( F^{-1} \), which are mode-specific inversions of each other (\( F \times_{\omega} F^{-1} \equiv I_\sigma \)). Then we can group these inserted tensors with the representation \( P(C) \) from (5.8), and obtain a transformed version \( \tilde{P}(C) \) (also see Figure 5.1). Furthermore, we have the freedom in choosing these collections of tensor inversion pairs. We will show that if we choose them systematically, we will be able to estimate each transformed tensor \( \tilde{P}(C) \) using the marginal probability of a small set of observed variables (observable representation).

In this section, we will first explain the transformed tensor representation.

Let us first consider a node \( C_i \) and its parent node \( C_{\pi(i)} \). Then the outgoing message of \( C_{\pi(i)} \) can be computed recursively as
\[
M(S_{\pi(i)}) = P(C_{\pi(i)}) \times S_i M(S_i) \times \ldots
\]
\[
P(C_i) \times S_{c_1(i)} M(S_{c_1(i)}) \times S_{c_2(i)} M(S_{c_2(i)})
\]
Inserting a mode specific identity tensor \( I_{S_i} \) with labels \( \{S_i, S_j\} \) and similarly defined mode specific
identity tensors $I_{S_{c_{1(0)}}}$ and $I_{S_{c_{2(0)}}}$ into the above two message updates, we obtain

$$M(S_{c_{1(0)}}) = \mathcal{P}(C_{c_{1(0)}}) \times S_{c_{1(0)}} (I_{S_{c_{1(0)}}} \times S_{c_{2(0)}} M(S_{c_{2(0)}})) \times \ldots$$

$$\mathcal{P}(C_{c_{1(0)}} \times S_{c_{1(0)}} (I_{S_{c_{1(0)}}} \times S_{c_{2(0)}} M(S_{c_{2(0)}})) \times \ldots$$

Then we can further expand $I_{S_{c_{1(0)}}}$ using tensor inversion pairs $F_{i}, F_{i}^{-1}$, i.e., $I_{S_{c_{1(0)}}} = F_{i} \times \omega_{i} F_{i}^{-1}$. Note that both $F_{i}$ and $F_{i}^{-1}$ have two sets of mode labels, $S_{i}$ and another set $\omega_{i}$ which is related to the observable representation and explained in the next section. Similarly, we expand $I_{S_{c_{1(0)}}}$ and $I_{S_{c_{2(0)}}}$ using their corresponding tensor inversion pairs.

After expanding tensor identities $I$, we can regroup terms, and at node $C_{i}$ we have

$$M(S_{c_{1(0)}}) = (\mathcal{P}(C_{c_{1(0)}}) \times S_{c_{1(0)}} F_{c_{1(0)}} \times S_{c_{2(0)}} F_{c_{2(0)}} \times \omega_{c_{1(0)}} (F_{c_{1(0)}}^{-1} \times S_{c_{1(0)}} M(S_{c_{1(0)}})) \times \omega_{c_{2(0)}} (F_{c_{2(0)}}^{-1} \times S_{c_{2(0)}} M(S_{c_{2(0)}}))$$

Now we can define the transformed tensor representation for $\mathcal{P}(C_{i})$ as

$$\overline{\mathcal{P}}(C_{i}) := (\mathcal{P}(C_{i}) \times S_{c_{1(0)}} F_{c_{1(0)}} \times S_{c_{2(0)}} F_{c_{2(0)}} \times S_{c_{1(0)}} F_{c_{1(0)}}^{-1}$$

where the two transformations $F_{c_{1(0)}}$ and $F_{c_{2(0)}}$ are obtained from the children side and the transformation $F_{c_{1(0)}}^{-1}$ is obtained from the parent side. Similarly, we can define the transformed representation for a leaf node and for the root node as

$$\overline{\mathcal{P}}(C_{i}) = \mathcal{P}(C_{i}) \times S_{c_{1(0)}} F_{c_{1(0)}}^{-1}$$

$$\overline{\mathcal{P}}(C_{r}) = \mathcal{P}(C_{r}) \times S_{c_{1(0)}} F_{c_{1(0)}}^{-1} \times S_{c_{2(0)}} F_{c_{2(0)}} \times S_{c_{3(0)}} F_{c_{3(0)}}$$

Applying these definitions of the transformed representation recursively, we can perform message passing based purely on these transformed representations

$$\overline{M}(S_{c_{(0)}}) = \overline{\mathcal{P}}(C_{c_{(0)}}) \times \omega_{c} \overline{M}(S_{c_{(0)}}) \times \ldots$$

5.5 Observable Representation

We now derive the observable representation by choosing the collection of tensor pairs $F_{i}$ and $F_{i}^{-1}$ systematically, so that we can recover each transformed tensor $\overline{\mathcal{P}}(C)$ using the marginal probability of a small set of observed variables.

We will focus on the transformed tensor representation in (5.14) for an internal node $C_{i}$ (other cases follow as special cases). Generalizing the idea in Chapter 4, we choose $F_{i} = \mathcal{P}(\theta_{i} \mid S_{i})$ where
\( \mathcal{O}_i \) is a set of observed variables in the subtree rooted at \( C_i \). Plugging this in gives,

\[
\mathcal{P}(C_i) = \mathcal{P}_i \times S_{c_i(i)} \mathcal{P}(\mathcal{O}_i|S_{c_i(i)}) \times S_{c_2(i)} \mathcal{P}(\mathcal{O}_c|S_{c_2(i)}) \times S_{l_i} \mathcal{P}(\mathcal{O}_i|S_{l_i})^{-1}
\]

where the first two tensor multiplications essentially eliminate the latent variables in \( S_{c_1(i)} \) and \( S_{c_2(i)} \). With these choices, we also fix the mode labels \( \omega_i, \omega_{c_1(i)} \) and \( \omega_{c_2(i)} \) in (5.12) (5.13) and (5.17). That is \( \omega_i = \mathcal{O}_i, \omega_{c_1(i)} = \mathcal{O}_c(i) \) and \( \omega_{c_2(i)} = \mathcal{O}_c(2) \).

To remove all dependencies on latent variables in \( \mathcal{P}(C_i) \) and relate it to observed variables, we need to eliminate the latent variables in \( S_i \) and the tensor \( \mathcal{P}(\mathcal{O}_i|S_i) \). For this, we multiply the transformed tensor \( \mathcal{P}(C_i) \) by \( \mathcal{P}(\mathcal{O}_i, \mathcal{O}_c) \), where \( \mathcal{O}_c \) denotes some set of observed variables which do not belong to the subtree rooted at node \( C_i \). Furthermore, \( \mathcal{P}(\mathcal{O}_i, \mathcal{O}_c) \) can be re-expressed using the conditional distribution of \( \mathcal{O}_i \) and \( \mathcal{O}_c \), respectively, conditioning on the separator set \( S_i \), \( i.e., \)

\[
\mathcal{P}(\mathcal{O}_i, \mathcal{O}_c) = \mathcal{P}(\mathcal{O}_i|S_i) \times S_i \mathcal{P}(\mathcal{O}_c|S_i) \times S_i \mathcal{P}(\mathcal{O}_c|S_i).
\]

Therefore, we have

\[
\mathcal{P}(\mathcal{O}_i|S_i) \times S_i \mathcal{P}(\mathcal{O}_c|S_i) \times S_i \mathcal{P}(\mathcal{O}_c|S_i),
\]

and plugging this into (5.19), we have

\[
\mathcal{P}(C_i) \times S_{c_i(i)} \mathcal{P}(\mathcal{O}_i|S_{c_i(i)}) \times S_{c_2(i)} \mathcal{P}(\mathcal{O}_c|S_{c_2(i)}) \times S_{l_i} \mathcal{P}(\mathcal{O}_c|S_{l_i}) \times S_{l_i} \mathcal{P}(\mathcal{O}_c|S_{l_i})^{-1} \times S_{l_i} \mathcal{P}(\mathcal{O}_i|S_{l_i})^{-1} \times S_i \mathcal{P}(\mathcal{O}_c|S_i) \times S_i \mathcal{P}(\mathcal{O}_c|S_i),
\]

where \( \mathcal{P}(C_i) \) is now related to only marginal probabilities of observed variables. From the equivalent relation, we can inverting \( \mathcal{P}(\mathcal{O}_i, \mathcal{O}_c) \), and obtain the observable representation for \( \mathcal{P}(C_i) \)

\[
\mathcal{P}(C_i) = \mathcal{P}(\mathcal{O}_c, \mathcal{O}_c, \mathcal{O}_c) \times \mathcal{P}(\mathcal{O}_c, \mathcal{O}_c) \times \mathcal{P}(\mathcal{O}_c, \mathcal{O}_c)^{-1}.
\]

**Example.** For node \( C_{BCDE} \) in Figure 5.2, the choices of \( \mathcal{O}_i, \mathcal{O}_c(i), \mathcal{O}_c(2) \) and \( \mathcal{O}_c \) are \{F, G\}, G, F and H respectively.

There are many valid choices of \( \mathcal{O}_c \). As in Chapter 4, these different choices can be combined via a linear system using Eq. 5.19. This can substantially increase performance.

For the leaf nodes and the root node, the derivation for their observable representations can be viewed as special cases of that for the internal nodes. We provide the results for their observable representation below:

\[
\mathcal{P}(C_i) = \mathcal{P}(\mathcal{O}_c(i), \mathcal{O}_c(2), \mathcal{O}_c(3)),
\]

\[
\mathcal{P}(C_i) = \mathcal{P}(\mathcal{O}_c, \mathcal{O}_c) \times \mathcal{P}(\mathcal{O}_c, \mathcal{O}_c)^{-1}.
\]

If \( \mathcal{P}(\mathcal{O}_c, \mathcal{O}_c) \) is invertible, then \( \mathcal{P}(C_i) = I \mathcal{O}_c \). Otherwise we need to project \( \mathcal{P}(\mathcal{O}_c, \mathcal{O}_c) \) using a tensor \( \mathcal{U}_i \) to make it invertible, as discussed in the next section. The overall training and test algorithms are given in Algorithms 4 and 5. Given \( N \) i.i.d. samples of the observed nodes, we simply replace \( \mathcal{P}(\cdot) \) by the empirical estimate \( \mathcal{P}(\cdot) \).

---

3If a latent variable in \( S_{c_1(i)} \cup S_{c_2(i)} \) is also in \( S_i \), it is not eliminated in this step but in another step.
5.6 Discussion

The observable representation exists only if there exist tensor inversion pairs $\mathcal{F}_i = P(\theta_i|S_i)$, and $\mathcal{F}_i^{-1}$. This is equivalent to requiring that the rank of the matricized version of $\mathcal{F}_i$ (rows corresponds to modes $\theta_i$ and column to modes $S_i$) has rank $\tau_i := S_H \times |S_i|$. Similarly, the matricized version of $P(\theta_i|S_i)$ also needs to have rank $\tau_i$, so that the matricized version of $\mathcal{P}(\theta_i, \theta_-)$ has rank $\tau_i$ and is invertible. Thus, it is required that $\#\text{states}(\theta_i) \geq \#\text{states}(S_i)$. This can be achieved by either making $\theta_i$ consist of a few high dimensional observations, or of many smaller dimensional ones. In the case when $\#\text{states}(\theta_i) > \#\text{states}(S_i)$, we need to project $\mathcal{F}_i$ to a lower dimensional space using a tensor $\mathcal{U}_i$, so that it can be inverted. In this case, we define $\mathcal{F}_i := \mathcal{P}(\theta_i|S_i) \times_{\theta_i} \mathcal{U}_i$. For example, following this through the computation for the leaf gives us that $\mathcal{P}(C_i) = \mathcal{P}(\theta_i, \theta_-) \times_{\theta_i} (\mathcal{P}(\theta_i, \theta_-) \times_{\theta_i} \mathcal{U}_i)^{-1}$. A good choice of $\mathcal{U}_i$ can be obtained by performing a singular value decomposition of the matricized version of $\mathcal{P}(\theta_i, \theta_-)$ (variables in $\theta_i$ are arranged to rows and those in $\theta_-$ to columns).

For HMMs and latent trees, this rank condition can be expressed simply as requiring the conditional probability tables of the underlying model to not be rank-deficient. However, junction trees encode more complex latent structures that introduce subtle considerations. A general characterization of the existence condition for observable representation with respect to the graph topology will be our future work.

The other practical aspects of spectral learning such as features, the expectation form, and linear systems as discussed in Chapter 4 also apply to our junction tree algorithm.

---

**Algorithm 4** Spectral learning algorithm for latent junction trees

**In**: Junction tree topology and $N$ i.i.d. samples $\{x_1^{(0)}, \ldots, x_N^{(0)}\}$

**Out**: Spectral parameters $\mathcal{P}(C_i)$, $\mathcal{P}(C)$ for all leaves, $\mathcal{P}(C_i)$ for all non-root internal nodes

1. For each node $C_i$, perform a “thin” singular value decomposition of $\mathcal{P}(\theta_i, \theta_-) = \mathcal{U}_i \Sigma \mathcal{V}_i^\top$; let $\hat{\mathcal{U}}_i = \mathcal{U}_i(:, 1 : S_H)$ be the first $S_H$ principal left singular vectors where $\tau_i = |S_i|$

2. Estimate $\mathcal{P}(C_i)$ for the root, leaf and internal nodes

   $\mathcal{P}(C_i) = \mathcal{P}(\theta_i, \theta_2, \theta_3) \times_{\theta_2} \mathcal{U}_i \times_{\theta_3} \mathcal{U}_i \times_{\theta_4} \mathcal{U}_i \times_{\theta_5} \mathcal{U}_i

   \mathcal{P}(C_i) = \mathcal{P}(\theta_i, \theta_1) \times_{\theta_1} (\mathcal{P}(\theta_i, \theta_1) \times_{\theta_1} \mathcal{U}_i)^{-1}

   \mathcal{P}(C_i) = \mathcal{P}(\theta_2, \theta_3, \theta_4) \times_{\theta_2} \mathcal{U}_i \times_{\theta_3} \mathcal{U}_i \times_{\theta_4} \mathcal{U}_i

   \times_{\theta_5} (\mathcal{P}(\theta_i, \theta_-) \times_{\theta_i} \mathcal{U}_i)^{-1}$

---

**Algorithm 5** Inference with spectral parameters for latent junction trees

**In**: Spectral parameters $\mathcal{P}(C_i), \mathcal{P}(C)$ for all leaves, $\mathcal{P}(C_i)$ for all non-root internal nodes and evidence $\{x_1, \ldots, x_O\}$

**Out**: Estimated marginal $\mathcal{P}(x_1, \ldots, x_O)$

1. In reverse topological order, leaf and internal nodes send messages

   $\mathcal{M}(S_i) = \mathcal{P}(C_i) \times_{\theta_i} \mathcal{M}(S_i)$

   $\hat{\mathcal{M}}(S_i) = \mathcal{P}(C_i) \times_{\theta_i} \hat{\mathcal{M}}(S_i)$

2. At the root, obtain $\mathcal{P}(x_1, \ldots, x_O)$ by

   $\mathcal{P}(C_i) \times_{\theta_i} \hat{\mathcal{M}}(S_i) \times_{\theta_i} \hat{\mathcal{M}}(S_i) \times_{\theta_i} \hat{\mathcal{M}}(S_i)$
5.7 Sample Complexity

We analyze the sample complexity of Algorithm 4 and show that it depends on the junction tree topology and the spectral properties of the true model. Let $d_i$ be the order of $\mathcal{P}(C_i)$ and $e_i$ be the number of modes of $\mathcal{P}(C_i)$ that correspond to observed variables.

**Theorem 2.** Let $\tau_i = S_h \times |S_i|$, $d_{\text{max}} = \max_i d_i$, and $e_{\text{max}} = \max_i e_i$. Then, for any $\epsilon > 0, 0 < \delta < 1$, if

$$N \geq O \left( \frac{4S_H^2 \delta}{\beta} d_{\text{max}} \frac{e_{\text{max}}}{\epsilon} \ln \frac{|C|}{\delta} |C|^2 \right)$$

where $\sigma_\tau(*)$ returns the $\tau$th largest singular value and

$$\alpha = \min_i \sigma_\tau(P(O_i, \partial)) \quad \beta = \min_i \sigma_\tau(F_i)$$

Then with probability $1 - \delta$,

$$\sum_{x_1, \ldots, x_\theta} \left| \mathbb{P}(x_1, \ldots, x_\theta) - \mathbb{E}(x_1, \ldots, x_\theta) \right| \leq \epsilon.$$

See the supplementary for a proof. The result implies that the estimation problem depends exponentially on $d_{\text{max}}$ and $e_{\text{max}}$, but note that $e_{\text{max}} \leq d_{\text{max}}$. Furthermore, $d_{\text{max}}$ is always greater than or equal to the treewidth. Note the dependence on the singular values of certain probability tensors. In fully observed models, the accuracy of the learned parameters depends only on how close the empirical estimates of the factors are to the true factors. However, our spectral algorithm also depends on how close the inverses of these empirical estimates are to the true inverses, which depends on the spectral properties of the matrices (Stewart and Sun, 1990).

5.8 Experiments

We now evaluate our method on synthetic and real data and compare it with both standard EM (Dempster et al., 1977) and stepwise online EM (Liang and Klein, 2009). All methods were implemented in C++, and the matrix library Eigen (Guennebaud et al., 2010) was used for computing SVDs and solving linear systems. For all experiments, standard EM is given 5 random restarts. Online EM tends to be sensitive to the learning rate, so it is given one restart for each of 5 choices of the learning rate $\{0.6, 0.7, 0.8, 0.9, 1\}$ (the one with highest likelihood is selected). Convergence is determined by measuring the change in the log likelihood at iteration $t$ (denoted by $f(t)$) over the average:

$$\frac{|f(t) - f(t-1)|}{\text{avg}(f(t), f(t-1))} \leq 10^{-4}$$

(roughly the same precision as used in Murphy (2005)).

For large sample sizes our method is almost two orders of magnitude faster than both EM and online EM. This is because EM is iterative and every iteration requires inference over all the training examples which can become expensive. On the other hand, the computational cost of our method is dominated by the SVD/linear system. Thus, it is primarily dependent only on the number of observed states and maximum tensor order, and can easily scale to larger sample sizes.

In terms of accuracy, we generally observe 3 distinct regions, low-sample size, mid-sample size, and large sample size. In the low sample size region, EM/online EM tend to overfit to the training data and our spectral algorithm usually performs better. In the mid-sample size region EM/online EM tend to perform better since they benefit from a smaller number of parameters. However, once
a certain sample size is reached (the large sample size region), our spectral algorithm consistently outperforms EM/online EM which suffer from local minima and convergence issues.

### 5.8.1 Synthetic Evaluation

We first perform a synthetic evaluation. 4 different latent structures are used (see Figure 5.3): a second order nonhomogenous (NH) HMM, a third order NH HMM, a 2 level NH factorial HMM, and a complicated synthetic junction tree. The second/third order HMMs have \(S_I = 2\) and \(S_O = 4\), while the factorial HMM and synthetic junction tree have \(S_I = 2\), and \(S_O = 16\). For each latent structure, we generate 10 sets of model parameters, and then sample \(N\) training points and 1000 test points from each set, where \(N\) is varied from 100 to 100,000. For evaluation, we measure the accuracy of joint estimation using \(\text{error} = \frac{|\hat{P}(x_1,...,x_O) - P(x_1,...,x_O)|}{P(x_1,...,x_O)}\). We also measure the training time of both methods.

Figure 5.3 shows the results. As discussed earlier, our algorithm is between one and two orders of magnitude faster than both EM and online EM for all the latent structures. EM is actually slower for very small sample sizes than for mid-range sample sizes because of overfitting. Also, in all cases, the spectral algorithm has the lowest error for large sample sizes. Moreover, critical sample size at which spectral overtakes EM/online EM is largely dependent on the number of parameters in the observable representation compared to that in the original parameterization of the model. In higher order/factorial HMM models, this increase is small, while in the synthetic junction tree it is larger.
5.8.2 Splice dataset

We next consider the task of determining splicing sites in DNA sequences (Asuncion and Newman, 2007). Each example consists of a DNA sequence of length 60, where each position in the sequence is either an \( A \), \( T \), \( C \), or \( G \). The goal is to classify whether the sequence is an Intron/Exon site, Exon/Intron site, or neither. During training, for each class a different second order nonhomogeneous HMM with \( S_H = 2 \) and \( S_O = 4 \) is trained. At test, the probability of the test sequence is computed for each model, and the one with the highest probability is selected (which we found to perform better than a homogeneous one).

Figure 5.4 shows our results, which are consistent with our synthetic evaluation. Spectral performs the best in low sample sizes, while EM/online EM perform a little better in the mid-sample size range. The dataset is not large enough to explore the large sample size regime. Moreover, we note that spectral algorithm is much faster for all the sample sizes.

5.9 Conclusion

We have developed an alternative parameterization that allows fast, local minima free, and consistent parameter learning of latent junction trees that generalizes the algorithm we presented in Chapter 4. Our approach generalizes spectral algorithms to a much wider range of structures such as higher order, factorial, and semi-hidden Markov models. Unlike traditional nonconvex optimization formulations, spectral algorithms allow us to theoretically explore latent variable models in more depth. The spectral algorithm depends not only on the junction tree topology but also on the spectral properties of the parameters. Thus, two models with the same structure may pose different degrees of difficulty based on the underlying singular values. This is very different from learning fully observed junction trees, which is primarily dependent on only the topology/treewidth. Future directions include learning discriminative models and structure learning.
5.10 Appendix

5.10.1 Sample Complexity Theorem

We prove the sample complexity theorem.

Notation

For simplicity, the main text describes the algorithm in the context of a binary tree. However, in the sample complexity proof, we adopt a more general notation. Let \( C_i \) be a clique, and \( C_{c_1(i)}, \ldots, C_{c_{\alpha_i}(i)} \) denote its \( \alpha_i \) children. Let \( \mathcal{C} \) denote the set of all the cliques in the junction tree, and \(|\mathcal{C}|\) denote its size. Define \( d_{\text{max}} \) to be maximum degree of a tensor in the observable representation and \( e_{\text{max}} \) to be maximum number of observed variables in any tensor. Furthermore, let \( \tau_i = |S_i| \times |S_i| \) (i.e. the number of states associated with the separator).

We can now write the transformed representation for junction trees with more than 3 neighbors as:

**Root:**
\[
\tilde{\mathcal{P}}(C_i) = \mathcal{P}(C_i) \times S_{c_1(i)} \mathcal{F}_{c_1(i)} \times \ldots \times S_{c_{\alpha_i}(i)} \mathcal{F}_{c_{\alpha_i}(i)}
\]

**Internal nodes:**
\[
\tilde{\mathcal{P}}(C_i) = \mathcal{P}(C_i) \times S_i \mathcal{F}_i^\dagger \times S_{c_1(i)} \mathcal{F}_{c_1(i)} \times \ldots \times S_{c_{\alpha_i}(i)} \mathcal{F}_{c_{\alpha_i}(i)}
\]

**Leaf:**
\[
\tilde{\mathcal{P}}(C_i) = \mathcal{P}(C_i) \times S_i \mathcal{F}_i^\dagger
\]

and the observable representation as:

- **root:** \( \mathcal{P}(C_i) = \mathcal{P}(\theta_{c_1(i)}, \ldots, \theta_{c_{\alpha_i}(i)}) \times \theta_{c_1(i)} \mathcal{U}_{c_1(i)} \times \ldots \times \theta_{c_{\alpha_i}(i)} \mathcal{U}_{c_{\alpha_i}(i)} \)
- **internal:** \( \mathcal{P}(C_i) = \mathcal{P}(\theta_{c_1(i)}, \ldots, \theta_{c_{\alpha_i}(i)}, \theta_{-i}) \times \theta_{-i} (\mathcal{P}(\theta_i, \theta_{-i}) \times \theta_i \mathcal{U}_i) \times \theta_{c_1(i)} \mathcal{U}_{c_1(i)} \times \ldots \times \theta_{c_{\alpha_i}(i)} \mathcal{U}_{c_{\alpha_i}(i)} \)
- **leaf:** \( \mathcal{P}(C_i) = \tilde{\mathcal{P}}(R_i, \theta_{-i}) \times \theta_{-i} (\tilde{\mathcal{P}}(\theta_i, \theta_{-i}) \times \theta_i \mathcal{U}_i) \times \theta_{c_1(i)} \mathcal{U}_{c_1(i)} \times \ldots \times \theta_{c_{\alpha_i}(i)} \mathcal{U}_{c_{\alpha_i}(i)} \)

Sometimes when the multiplication indices are clear, we will omit it to make things simpler.

**Rearranged version:**
We will often find it convenient to rearrange the tensors into lower order tensors for the purpose of taking some norms (such as the spectral norm). We define \( R(\cdot) \) as the “rearranging” operation. For example, \( R(\tilde{\mathcal{P}}(\theta_i, \theta_{-i})) \) is the matricized version of \( \tilde{\mathcal{P}}(\theta_i, \theta_{-i}) \) with \( \theta_i \) being mapped to the rows and \( \theta_{-i} \) being mapped to the columns.

More generally, if \( \mathcal{P}(C_i) \) (which has order \( d_i \)) then \( R(\mathcal{P}(C_i)) \) is a rearrangement of \( \mathcal{P}(C_i) \) such that it has order equal to the number of neighbors of \( C_i \). The set of modes of \( \mathcal{P}(C_i) \) corresponding to a single separator of \( C_i \) get mapped to a single mode in \( R(\mathcal{P}(C_i)) \). We let \( R(S_{\alpha_i(j)}) \) denote the mode that \( S_{\alpha_i(j)} \) maps to in \( R(\mathcal{P}(C_i)) \).
In the case of the root, \( R(\mathcal{P}(C_i)) \) rearranges \( \mathcal{P}(C_i) \) into a tensor of order \( \alpha_i \). For other clique nodes, \( R(\mathcal{P}(C_i)) \) rearranges \( \mathcal{P}(C_i) \) into a tensor of order \( \alpha_i + 1 \).

**Example:** In Figure 5.2 \( \mathcal{P}(C_{BCDE}) = \mathcal{P}(\varnothing \cup B, D \mid \varnothing C, E) \), \( C_{BCDE} \) has 3 neighbors and so \( R(\mathcal{P}(C_{BCDE})) \) is of order 3 where each of the modes correspond to \( \{B, C\}, \{B, D\}, \{C, E\} \) respectively.

This rearrangement can be applied to the other quantities. For example, \( R(\mathcal{P}(\mathcal{O}_{c_{1}(i)}, \ldots, \mathcal{O}_{c_{n}(i)}) \) is a rearranging of \( \mathcal{P}(\mathcal{O}_{c_{1}(i)}, \ldots, \mathcal{O}_{c_{n}(i)}) \) into a tensor of order \( \alpha_i \) where the \( \mathcal{O}_{c_{1}(i)}, \ldots, \mathcal{O}_{c_{n}(i)} \) correspond to one mode each. \( R(\hat{\mathcal{P}}(\mathcal{O}_{i}, \mathcal{O}_{-i})) \) is the matricized version of \( \hat{\mathcal{P}}(\mathcal{O}_{i}, \mathcal{O}_{-i}) \) with \( \mathcal{O}_{i} \) being mapped to the rows and \( \mathcal{O}_{-i} \) being mapped to the columns. Similarly, \( R(\mathcal{F}_{i}) \) is the matricized version of \( \mathcal{F}_{i} \) and \( R(\mathcal{U}_{i}) \) is the matricized version of \( \mathcal{U}_{i} \).

Thus, we can define the rearranged quantities:

**Root:**
\[
R(\tilde{\mathcal{P}}(C_i)) = R(\mathcal{P}(C_i)) \times_{R(S_{c_{1}(i)})} R(\mathcal{F}_{c_{1}(i)}) \times \ldots \times_{R(S_{c_{n}(i)})} R(\mathcal{F}_{c_{n}(i)})
\]

**Leaf:**
\[
R(\tilde{\mathcal{P}}(C_i)) = R(\mathcal{P}(C_i)) \times_{R(S_i)} R(\mathcal{F}_{i})
\]

and the observable representation as:

**root:**
\[
R(\mathcal{P}(C_i)) = R(\mathcal{P}(\mathcal{O}_{c_{1}(i)}, \ldots, \mathcal{O}_{c_{n}(i)})) \times_{R(S_{c_{1}(i)})} R(\mathcal{U}_{c_{1}(i)}) \times \ldots \times_{R(S_{c_{n}(i)})} R(\mathcal{U}_{c_{n}(i)})
\]

**internal:**
\[
R(\mathcal{P}(C_i)) = R(\mathcal{P}(\mathcal{O}_{S_{c_{1}(i)}}, \ldots, \mathcal{O}_{c_{n}(i)}, \mathcal{O}_{-i})) \times_{R(S_{c_{1}(i)})} R(\mathcal{P}(\mathcal{O}_{i}, \mathcal{O}_{-i}) \times_{R(S_{c_{n}(i)})} R(\mathcal{U}_{c_{n}(i)})
\]

**leaf:**
\[
R(\mathcal{P}(C_i)) = R(\hat{\mathcal{P}}(\mathcal{O}_{i}, \mathcal{O}_{-i})) \times_{R(\mathcal{O}_{i}, \mathcal{O}_{-i})} R(\hat{\mathcal{P}}(\mathcal{O}_{i}, \mathcal{O}_{-i}) \times_{R(\mathcal{O}_{i}, \mathcal{O}_{-i})} R(\mathcal{U}_{i})
\]

Furthermore define the following:

\[
\alpha = \min_{C_i \in \mathcal{C}} \sigma_{\mathcal{C}, i}(\mathcal{P}[\mathcal{O}_{i}, \mathcal{O}_{-i}])
\]

\[
\beta = \min_{C_i \in \mathcal{C}} \sigma_{\mathcal{C}, i}(\mathcal{F}_{i})
\]

The proof is based on the technique of HKZ \cite{hsu2009} but has differences due to the junction tree topology and higher order tensors.

**Tensor Norms**

We briefly define several tensor norms that are a generalization of matrix norms. For more information about matrix norms see \cite{horn1990}.  

68
Frobenius Norm: Just like for matrices, the Frobenius norm of a tensor of order $N$ is defined as:

$$
\|T\|_F = \sqrt{\sum_{c_1(\cdot),...,i_N} T(c_1(\cdot),...,i_N)^2} \quad (5.26)
$$

Elementwise One Norm: Similarly, the Elementwise one norm of a tensor of order $N$ is defined as:

$$
\|T\|_1 = \sum_{c_1(\cdot),...,i_N} |T(c_1(\cdot),...,i_N)| \quad (5.27)
$$

Spectral Norm: For tensors of order $N$ the spectral norm (Nguyen et al., 2010) can be defined as:

$$
\|T\|_2 = \sup_{v \text{ s.t.} \|v\|_2 \leq 1, \forall 1 \leq i \leq N} T \times_N v_N, ..., \times_2 v_2 \times_1 v_1 \quad (5.28)
$$

In our case, we will find it more convenient to use the rearranged spectral norm, which we define as:

$$
\|\|T\||_2R = \|R(T)\|_2 \quad (5.29)
$$

where $R(\cdot)$ was defined in the previous section.

Induced One Norm: For matrices the induced one norm is defined as the max column sum of the matrix: $\|M\|_1 = \sup_{v \text{ s.t.} \|v\|_1 \leq 1} \|Mv\|_1$. We can generalize this to be the max slice sum of a tensor on a tensor where some modes are fixed and others are summed over. Let $\sigma$ denote the modes that will be maxed over. Then:

$$
\|\|T\||_\sigma^q = \sup_{v \text{ s.t.} \|v\|_1 \leq 1, \forall 1 \leq i \leq |\sigma|} \|T \times_{\sigma_1} v_1, ..., \times_{\sigma_{|\sigma|}} v_{|\sigma|}\|_1 \quad (5.30)
$$

In the Appendix, we prove several lemmas regarding these tensor norms.

Concentration Bounds

$$
\epsilon(\theta_1, ..., \theta_d) = \left\| \mathcal{P}(\theta_1, ..., \theta_d) - \mathcal{P}(\tilde{\theta}_1, ..., \tilde{\theta}_d) \right\|_F \quad (5.31)
$$

$$
\epsilon(\tilde{\theta}_1, ..., \tilde{\theta}_{d-e}, \theta_{d-e+1}, ..., \theta_d) = \left\| \mathcal{P}(\tilde{\theta}_1, ..., \tilde{\theta}_{d-e}, \theta_{d-e+1}, ..., \theta_d) - \mathcal{P}(\tilde{\theta}_1, ..., \tilde{\theta}_{d-e}, \theta_{d-e+1}, ..., \theta_d) \right\|_F \quad (5.32)
$$

1, ..., $d - e$ denote the $d - e$ non-evidence variables while $d - e + 1, ..., d$ denote the $e$ evidence variables. $d$ indicates the total number of modes of the tensor. As the number of samples $N$ gets large, we expect these quantities to be small.

Lemma 6 (variant of HKZ (Hsu et al., 2009)). If the algorithm independently samples $N$ of the observa-
tions, then with probability at least $1 - \delta$.

$$e(O_1, \ldots, O_d) \leq \sqrt{\frac{1}{N} \ln \frac{2|C|}{\delta}} + \frac{\sqrt{1}}{N}$$

(5.33)

$$\sum_{o_{d-e+1}, \ldots, o_d} e(O_1, \ldots, O_{d-e}, O_{d-e+1}, \ldots, O_d) \leq \sqrt{\frac{Se_{\max}}{N}} \ln \frac{2|C|}{\delta} + \sqrt{\frac{Se_{\max}}{N}}$$

(5.34)

for all tuples $(O_1, \ldots, O_d)$ that are used in the spectral algorithm.

The proof is the same as that of HKZ [Hsu et al., 2009] except the union bound is taken over $2|C|$ instead of $3$ (since each transformed quantity in the spectral algorithm is composed of at most two such terms). The last bound can be made tighter, identical to HKZ, but for simplicity we do not pursue that approach here.

**Singular Value Bounds**

 Basically this is the generalized version of Lemma 9 in HKZ [Hsu et al., 2009], which is stated below for completeness:

**Lemma 7** (variant of HKZ [Hsu et al., 2009]). Suppose $e(O_i, O_{-i}) \leq \epsilon \times \sigma_i(P(O_i, O_{-i}))$ for some $\epsilon < 1/2$. Let $\epsilon_0 = \epsilon(O_i, O_{-i})^2 / (1 - \epsilon) \sigma_i(P(O_i, O_{-i}))^2$. Then:

1. $\epsilon_0 < 1$
2. $\sigma_i(P(O_i, O_{-i}) \times O_i \hat{U}_i) \geq (1 - \epsilon_0) \sigma_i(P(O_i, O_{-i}))$
3. $\sigma_i(P(O_i, O_{-i}) \times O_i \hat{U}_i) \geq \sqrt{1 - \epsilon_0} \sigma_i(P(O_i, O_{-i}))$
4. $\sigma_i(P(O_i|S_i) \times O_i \hat{U}_i) \geq \sqrt{1 - \epsilon_0} \sigma_i(P(O_i|S_i))$

It follows that if $e(O_i, O_{-i}) \leq \sigma_i(P(O_i, O_{-i}))/3$ then this implies that $\epsilon_0 \leq \frac{1/9}{4/9} = \frac{1}{4}$. It then follows that,

1. $\sigma_i(P(O_i, O_{-i}) \times O_i \hat{U}_i) \geq \frac{3}{4} \sigma_i(P(O_i, O_{-i}))$
2. $\sigma_i(P(O_i, O_{-i}) \times O_i \hat{U}_i) \geq \frac{\sqrt{3}}{2} \sigma_i(P(O_i, O_{-i}))$
3. $\sigma_i(P(O_i|S_i) \times O_i \hat{U}_i) \geq \frac{\sqrt{3}}{2} \sigma_i(P(O_i|S_i))$

**Bounding the Transformed Quantities**

Define,
1. root: $\tilde{P}(C_i) := P(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}) \times \tilde{U}(\alpha_{c_1(i)} \ldots \times \alpha_{c_n(i)})$ 

2. leaf: $\tilde{P}_{\tau_i}(C_i) = P(R_i = r_i, \theta_{\tau_i}) \times \tilde{U}(\theta_{\tau_i})$ 

3. internal: $\tilde{P}(C_i) = P(\theta_{S_1(i)}, \ldots, \theta_{s_n(i)}, \theta_{\tau_i}) \times \tilde{U}(\theta_{\tau_i})$ 

(which are the observable quantities with the true probabilities, but empirical $\tilde{U}$'s). Similarly, we can define $\tilde{F}_i = P[\theta_i | S_i] \times \tilde{U}_i$. We have also abbreviated $P(\theta_{\tau_i}, \theta_{\tau_i}) \times \tilde{U}_i$ with $P_{\tilde{U}}(\theta_{\tau_i}, \theta_{\tau_i})$. 

We seek to bound the following three quantities: 

$\delta_i^{\text{root}} := ||(P(C_i) - \tilde{P}(C_i)) \times \theta_{c_1(i)}^{-1} \times \ldots \times \theta_{c_n(i)}^{-1} ||_1$ \hspace{1cm} (5.35) 

$\delta_i^{\text{internal}} := ||(P(C_i) - \tilde{P}(C_i)) \times \theta_{c_1(i)} \times \ldots \times \theta_{c_n(i)}^{-1} ||_1$ \hspace{1cm} (5.36) 

$\xi_i^{\text{leaf}} := \sum_{r_i} ||(P_{\tau_i}(C_i) - \tilde{P}_{\tau_i}(C_i)) \times \theta_{\tau_i} ||_1$ \hspace{1cm} (5.37) 

**Lemma 8.** If $\epsilon(\theta_{\tau_i}, \theta_{\tau_i}) \leq \sigma_{\tau_i}(P(\theta_{\tau_i}, \theta_{\tau_i}))/3$ then 

$\delta_i^{\text{root}} \leq \frac{2d_{\text{max}}S_{\text{max}}^{\text{max}} \epsilon(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)})}{3d^{2/3}d^d}$ \hspace{1cm} (5.38) 

$\delta_i^{\text{internal}} \leq \frac{2d_{\text{max}}^{\text{max}}S_{\text{H}}^{\text{max}}}{3 \sqrt{3}d^{2d} \sigma_{\tau_i}(P(\theta_{\tau_i}, \theta_{\tau_i}))^2} \left( \frac{\epsilon(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}, \theta_{\tau_i})}{\sigma_{\tau_i}(P(\theta_{\tau_i}, \theta_{\tau_i}))^2} + \frac{\epsilon(\theta_{\tau_i}, \theta_{\tau_i})}{\sigma_{\tau_i}(P(\theta_{\tau_i}, \theta_{\tau_i}))^2} \right)$ \hspace{1cm} (5.39) 

$\xi_i^{\text{leaf}} \leq \frac{8d_{\text{max}}S_{\text{H}}^{\text{max}}}{3} \left( \frac{\epsilon(\theta_{\tau_i}, \theta_{\tau_i})}{\sigma_{\tau_i}(P(\theta_{\tau_i}, \theta_{\tau_i}))^2} + \frac{\sum_{R_i} \epsilon(R_i = r_i, \theta_{\tau_i})}{\sigma_{\tau_i}(P_{\tilde{U}}(\theta_{\tau_i}, \theta_{\tau_i}))^2} \right)$ \hspace{1cm} (5.40) 

We define $\Delta = \max(\delta_i^{\text{root}}, \delta_i^{\text{internal}}, \xi_i^{\text{leaf}})$ (over all $i$). 

**Proof.** 

**Case $\delta_i^{\text{root}}$:** 

For the root, $\tilde{P}(C_i) = \tilde{P}(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}) \times \tilde{U}(\alpha_{c_1(i)} \ldots \times \alpha_{c_n(i)})$ and similarly $\tilde{P}(C_i) = P(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}) \times \tilde{U}(\alpha_{c_1(i)} \ldots \times \alpha_{c_n(i)})$. 

71
\[ \delta_{\text{root}} = \left\| \mathbf{P}(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}) \times \mathbf{O}_{c_1(i)} \mathbf{\tilde{u}}_{c_1(i)} \times \ldots \times \mathbf{O}_{c_n(i)} \mathbf{\tilde{u}}_{c_n(i)} \right\|_{1} \\
\leq S_{H}^{d_{\text{max}}} \left\| \mathbf{P}(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}) - \mathbf{P}(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}) \times \mathbf{O}_{c_1(i)} \mathbf{\tilde{u}}_{c_1(i)} \times \ldots \times \mathbf{O}_{c_n(i)} \mathbf{\tilde{u}}_{c_n(i)} \right\|_{2R} \\
\times \left\| \mathbf{f}_{c_1(i)}^{-1} \right\|_{2R} \ldots \left\| \mathbf{f}_{c_1(i)}^{-1} \right\|_{2R} \\
\leq \frac{S_{H}^{d_{\text{max}}}}{\prod_{j=1}^{n_i} \sigma_{\alpha_{i,j}}(\mathbf{f}_{a,i}(j))} \left\| \mathbf{P}(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}) - \mathbf{P}(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}) \right\|_{2R} \times \left\| \mathbf{\tilde{u}}_{c_1(i)} \right\|_{2R} \left\| \mathbf{\tilde{u}}_{c_2(i)} \right\|_{2R} \ldots \left\| \mathbf{\tilde{u}}_{c_n(i)} \right\|_{2R} \\
= \frac{S_{H}^{d_{\text{max}}}}{\prod_{j=1}^{n_i} \sigma_{\alpha_{i,j}}(\mathbf{f}_{a,i}(j))} \epsilon(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)}) \\
\text{Between the first and second line we use Lemma 13 to convert from elementwise one norm to spectral norm, and Lemma 11 (submultiplicativity). Lemma 11 (submultiplicativity) is applied again to get to the second-to-last line. We also use the fact that } \left\| \mathbf{\tilde{u}}_{c_1(i)} \right\|_{2R} = 1. \]

Thus, by Lemma 7

\[ \delta_{i}^{\text{root}} \leq \frac{2^{d_{\text{max}}} S_{H}^{d_{\text{max}}} \epsilon(\theta_{c_1(i)}, \ldots, \theta_{c_n(i)})}{\mathcal{P}^{d_{\text{max}}/2 \mathcal{P}^{d_{\text{max}}}}} \quad (5.41) \]

**Case \( \xi_{i}^{\text{leaf}} \):**

We note that \( \mathbf{P}_{r_{i}}(C_{i}) = \mathbf{P}(R_{i} = r_{i}, \theta_{-i}) \times_{\theta_{i}} (\mathbf{P}_{\mathbf{u}}(\theta_{i}, \theta_{-i}))^{+} \) and similarly \( \mathbf{P}_{r_{i}}(C_{i}) = \mathbf{P}(R_{i} = r_{i}, \theta_{-i}) \times_{\theta_{i}} (\mathbf{P}_{\mathbf{u}}(\theta_{i}, \theta_{-i}))^{+} \).

Again, we use Lemma 13 to convert from the one norm to the spectral norm, and Lemma 11 for submultiplicativity.

\[ \xi_{i}^{\text{leaf}} = \sum_{r_{i}} \left\| \mathbf{P}_{r_{i}}(C_{i}) - \mathbf{P}_{r_{i}}(C_{i}) \right\|_{1}^{S_{i}} \leq \sum_{r_{i}} \left\| \mathbf{P}_{r_{i}}(C_{i}) - \mathbf{P}_{r_{i}}(C_{i}) \times_{\theta_{i}} \mathbf{\tilde{u}} \right\|_{1}^{S_{i}} \leq \sum_{r_{i}} S_{H}^{d_{\text{max}}} \left\| \mathbf{P}_{r_{i}}(C_{i}) - \mathbf{P}_{r_{i}}(C_{i}) \right\|_{2R} \left\| \mathbf{\tilde{u}} \right\|_{2R} \leq \sum_{r_{i}} S_{H}^{d_{\text{max}}} \left\| \mathbf{P}_{r_{i}}(C_{i}) - \mathbf{P}_{r_{i}}(C_{i}) \right\|_{2R} \quad (5.42) \]
Note that $\|\mathcal{P}[\theta_i|S]\|_1^\delta = 1$, and $\|\mathcal{U}\|_{2R} = 1$.

$$
\left\|\mathcal{P}_r(C_i) - \mathcal{P}_r(C_i)\right\|_{2R} = \left\|\mathcal{P}(R_i = r_i, \theta_i) \times {\mathcal{P}_\tilde{u}(\theta_i, \theta_i) }^\dagger - \mathcal{P}(R_i = r_i, \theta_i) \times {\mathcal{P}_\tilde{u}(\theta_i, \theta_i) }^\dagger \right\|_{2R} \\
\leq \left\|\mathcal{P}(R_i = r_i, \theta_i) \times {\mathcal{P}_\tilde{u}(\theta_i, \theta_i) }^\dagger - \mathcal{P}(R_i = r_i, \theta_i) \right\|_{2R} + \left\|\mathcal{P}(R_i = r_i, \theta_i) \right\|_{2R} \\
\leq \left\|\mathcal{P}(R_i = r_i)\right\|_{2R} \left\|\mathcal{P}_\tilde{u}(\theta_i, \theta_i) \right\|_{2R} + \left\|\mathcal{P}_\tilde{u}(\theta_i, \theta_i) \right\|_{2R} \\
\leq P(R_i = r_i) \frac{1 + \sqrt{5}}{2} e(\theta_i, \theta_i) - \min(\sigma_{\tau_i}(\mathcal{P}_\tilde{u}(\theta_i, \theta_i)), \sigma_{\tau_i}(\mathcal{P}_\tilde{u}(\theta_i, \theta_i))) + \frac{e(R_i = r_i, \theta_i)}{\sigma_{\tau_i}(\mathcal{P}_\tilde{u}(\theta_i, \theta_i))} \\
$$

The last line follows from Eq. 5.55. We have also used the fact that $\|\mathcal{P}(R_i = r_i, \theta_i)\|_{2R} \leq \|\mathcal{P}(R_i = r_i, \theta_i)\|_{2R} \leq P(R_i = r)$. Thus, using Lemma 7.

$$
\xi^\text{leaf}_i \leq \sum_{r_i} S_{\text{max}} \left( P(R_i = r_i) \frac{1 + \sqrt{5}}{2} e(\theta_i, \theta_i) - \min(\sigma_{\tau_i}(\mathcal{P}_\tilde{u}(\theta_i, \theta_i)), \sigma_{\tau_i}(\mathcal{P}_\tilde{u}(\theta_i, \theta_i))) + \frac{e(R_i = r_i, \theta_i)}{\sigma_{\tau_i}(\mathcal{P}_\tilde{u}(\theta_i, \theta_i))} \right) \\
$$

$$
\xi^\text{leaf}_i \leq \sum_{r_i} S_{\text{max}} \left( \frac{1 + \sqrt{5}}{2} \frac{16 P(R_i = r_i) e(\theta_i, \theta_i)}{9 \sigma_{\tau_i}(\mathcal{P}_\tilde{u}(\theta_i, \theta_i))^2} + \frac{2 e(R_i = r_i, \theta_i)}{\sqrt{3} \sigma_{\tau_i}(\mathcal{P}_\tilde{u}(\theta_i, \theta_i))} \right) \\
\leq \frac{8 S_{\text{max}}}{3} \left( \frac{e(\theta_i, \theta_i)}{\sigma_{\tau_i}(\mathcal{P}(\theta_i, \theta_i))^2} + \frac{\sum_{r_i} e(R_i = r_i, \theta_i)}{\sigma_{\tau_i}(\mathcal{P}_\tilde{u}(\theta_i, \theta_i))} \right) \\
$$

$$
\delta_{\text{internal}}_i = \mathcal{P}(C_i) = \mathcal{P}(\theta_{c_1}(i), \ldots, \theta_{c_q}(i), \theta_{-i}) \times \theta_{-i} \left( \mathcal{P}_\tilde{u}(\theta_i, \theta_i) \right)^\dagger \times \theta_{c_1}(i) \times \ldots \times \theta_{c_q}(i) \mathcal{U}_{c_1}(i) \times \ldots \times \theta_{c_q}(i) \mathcal{U}_{c_q}(i). \\
$$

Similarly, $\mathcal{P}(C_i) = \mathcal{P}(\theta_{c_1}(i), \ldots, \theta_{c_q}(i), \theta_{-i}) \times \theta_{-i} \left( \mathcal{P}_\tilde{u}(\theta_i, \theta_i) \right)^\dagger \times \theta_{c_1}(i) \times \ldots \times \theta_{c_q}(i) \mathcal{U}_{c_1}(i) \times \ldots \times \theta_{c_q}(i) \mathcal{U}_{c_q}(i).$

Again, we use Lemma 13 to convert from one norm to spectral norm and Lemma 11 for submul-
Note that \( |||\mathcal{P}[\Theta|S]\||_1^S = 1\), and \( \|\hat{\mathbf{u}}\|_{2R} = 1\).
Bounding the Propagation of Error

We now show all these errors propagate on the junction tree. For this section, assume the clique nodes are numbered 1, 2, ..., |C| in breadth first order (such that 1 is the root). Moreover let $\Phi_{1:c}(C)$ be the transformed factors accumulated so far if we computed the joint probability from the root down (instead of the bottom up). For example,

$$\Phi_{1:1}(C) = \mathcal{P}(C_1)$$

$$\Phi_{1:2}(C) = \mathcal{P}(C_1) \times S_2 \mathcal{P}(C_2)$$

$$\Phi_{1:2}(C) = \mathcal{P}(C_1) \times S_2 \mathcal{P}(C_2) \times S_3 \mathcal{P}(C_3)$$

(Note how this is very computationally inefficient: the tensors get very large. However, it is useful for proving statistical properties). Then the modes of $\Phi_{1:c}$ can be partitioned into mode groups, $M_1, ..., M_d$, (where each mode group consists of the variables on the corresponding separator edge).

We now prove the following lemma,

**Lemma 9.** Define $\Delta = \max(\delta_i^{\text{root}}, \delta_i^{\text{internal}}, \delta_i^{\text{leaf}})$. Then,

$$\sum_{x_{1:c}} \| (\Phi_{1:c}(C) - \Phi_{1:c}(C)) \times M_1 \tilde{F}_1^{-1} \times \cdots \times M_d \tilde{F}_d^{-1} \|_1 \leq (1 + \Delta)^{c-1} \delta_i^{\text{root}} + (1 + \Delta)^{c-1} - 1 \tag{5.51}$$

$x_{1:c}$ is all the observed variables in cliques 1 through $c$. Note that when $c = |C|$ then this implies that $\Phi_{1:c}(C) = \tilde{P}[x_1, ..., x_O]$ and thus,

$$\sum_{x} | \tilde{P}(x_1, ..., x_O) - P(x_1, ..., x_O) | \leq (1 + \Delta)^{|C|-1} \delta_i^{\text{root}} + (1 + \Delta)^{|C|-1} - 1 \tag{5.52}$$

**Proof.** The proof is by induction on $c$. The base case follows trivially from the definition of $\delta_i^{\text{root}}$. For the induction step, assume the claim holds for $c \geq 1$. Then we prove it holds for $c + 1$. 

\[ \delta_i^{\text{internal}} \leq \frac{(2S_H)^{d_{\text{max}}}}{(\beta \sqrt{3})^{d_{\text{max}}}} \left( \frac{2 \epsilon(\Theta_{c_1(i)}, ..., \Theta_{c_{|C|}(i)}, \Theta_{-i})}{\sqrt{3} \sigma_i(P(\Theta_i, \Theta_{-i}))} + \frac{8 \epsilon(\Theta_i, \Theta_{-i})}{3(\sigma_i(P(\Theta_i, \Theta_{-i})))^2} \right) \]

\[ \leq \frac{8(2S_H)^{d_{\text{max}}}}{3(\beta \sqrt{3})^{d_{\text{max}}}} \left( \frac{\epsilon(\Theta_{c_1(i)}, ..., \Theta_{c_{|C|}(i)}, \Theta_{-i})}{\sigma_i(P(\Theta_i, \Theta_{-i}))} + \frac{\epsilon(\Theta_i, \Theta_{-i})}{(\sigma_i(P(\Theta_i, \Theta_{-i})))^2} \right) \tag{5.47} \]
Now for the second term, break up the three terms:

Note here the summation over $x_{1:c+1}$ is irrelevant since there is no evidence. We use Lemma $10$ to break up the three terms:

The first term,

$$\| (\Phi_{1:c}(C) \times (\hat{P}(C_{c+1}) - P(C_{c+1})) \times_{\mathcal{H}_i} \tilde{F}_1^{-1} \times_{\mathcal{H}_{d+1}} \tilde{F}_{d+1}^{-1} \|_1$$

$$\leq \left\| (\hat{P}(C_{c+1}) - P(C_{c+1})) \times_{\mathcal{H}_i} \tilde{F}_1^{-1} \times_{\mathcal{H}_{d+1}} \tilde{F}_{d+1}^{-1} \right\|_1$$

The first term above is simply $\delta_i^{\text{internal}} \leq \Delta$ while the second equals one since it is a joint distribution.

Now for the second term,

$$\| (\Phi_{1:c}(C) - \Phi_{1:c}(C)) \times (\hat{P}(C_{c+1}) - P(C_{c+1})) \times_{\mathcal{H}_i} \tilde{F}_1^{-1} \times_{\mathcal{H}_{d+1}} \tilde{F}_{d+1}^{-1} \|_1$$

$$\leq (1 + \Delta)^{\delta_{\text{root}}_{c+1}} + (1 + \Delta)^{\delta_{\text{internal}}_{c+1} - 1} \times \| (\hat{P}(C_{c+1}) - P(C_{c+1})) \times_{\mathcal{H}_i} \tilde{F}_1^{-1} \times_{\mathcal{H}_{d+1}} \tilde{F}_{d+1}^{-1} \|_1$$

( Dor induction hypothesis)

$$\leq (1 + \Delta)^{\delta_{\text{root}}_{c+1} + (1 + \Delta)^{\delta_{\text{internal}}_{c+1} - 1} \times \Delta$$

Now we consider two cases, when $c + 1$ is a leaf clique and when it is an internal clique.

**Case 1: Internal Clique**

Note here the summation over $x_{1:c+1}$ is irrelevant since there is no evidence. We use Lemma $10$ to break up the three terms:

The first term,

$$\| (\Phi_{1:c}(C) \times (\hat{P}(C_{c+1}) - P(C_{c+1})) \times_{\mathcal{H}_i} \tilde{F}_1^{-1} \times_{\mathcal{H}_{d+1}} \tilde{F}_{d+1}^{-1} \|_1$$

$$\leq \left\| (\hat{P}(C_{c+1}) - P(C_{c+1})) \times_{\mathcal{H}_i} \tilde{F}_1^{-1} \times_{\mathcal{H}_{d+1}} \tilde{F}_{d+1}^{-1} \right\|_1$$

The first term above is simply $\delta_i^{\text{internal}} \leq \Delta$ while the second equals one since it is a joint distribution.

Now for the second term,

$$\| (\Phi_{1:c}(C) - \Phi_{1:c}(C)) \times (\hat{P}(C_{c+1}) - P(C_{c+1})) \times_{\mathcal{H}_i} \tilde{F}_1^{-1} \times_{\mathcal{H}_{d+1}} \tilde{F}_{d+1}^{-1} \|_1$$

$$\leq (1 + \Delta)^{\delta_{\text{root}}_{c+1}} + (1 + \Delta)^{\delta_{\text{internal}}_{c+1} - 1} \times \| (\hat{P}(C_{c+1}) - P(C_{c+1})) \times_{\mathcal{H}_i} \tilde{F}_1^{-1} \times_{\mathcal{H}_{d+1}} \tilde{F}_{d+1}^{-1} \|_1$$

(via induction hypothesis)

$$\leq (1 + \Delta)^{\delta_{\text{root}}_{c+1} + (1 + \Delta)^{\delta_{\text{internal}}_{c+1} - 1} \times \Delta$$
The first term above equals 1 because it is a joint distribution and the second is the bound on the transformed quantity we had proved earlier (since \( r_i = x_{c+1} \)).

The second term,

\[
\sum_{x_{c+1}} \left\| (\Phi_{1:x}(C) - \tilde{\Phi}_{1:x}(C)) \times \left( \tilde{P}_{r_1}(C_{c+1}) - \tilde{P}_{r_1}(C_{c+1}) \right) \times \mathcal{M}_1 \tilde{f}_{d_1}^{-1} \times \ldots \times \mathcal{M}_{d_{c+1}} \tilde{f}_{d_{c+1}}^{-1} \right\|_1 \\
\leq \sum_{x_{c+1}} \left\| (\Phi_{1:x}(C) - \tilde{\Phi}_{1:x}(C)) \times \mathcal{M}_1 \tilde{f}_{d_1}^{-1} \times \ldots \times \mathcal{M}_{d_{c+1}} \tilde{f}_{d_{c+1}}^{-1} \right\|_1 \sum_{x_{c+1}} \left\| (\tilde{P}_{r_1}(C_{c+1}) - \tilde{P}_{r_1}(C_{c+1})) \times \theta^{-}_{c+1} \tilde{f}_{c+1} \right\|_1 \\
\leq (1 + \Delta)^{c-1} \delta^{\text{root}} + (1 + \Delta)^{c-1} - 1 \times \Delta
\]

The third term,

\[
\sum_{x_{c+1}} \left\| (\Phi_{1:x}(C) - \tilde{\Phi}_{1:x}(C)) \times \tilde{P}_{r_1}(C_{c+1}) \times \mathcal{M}_1 \tilde{f}_{d_1}^{-1} \times \ldots \times \mathcal{M}_{d_{c+1}} \tilde{f}_{d_{c+1}}^{-1} \right\|_1 \\
\leq \sum_{x_{c+1}} \left\| (\Phi_{1:x}(C) - \tilde{\Phi}_{1:x}(C)) \times \mathcal{M}_1 \tilde{f}_{d_1}^{-1} \times \ldots \times \mathcal{M}_{d_{c+1}} \tilde{f}_{d_{c+1}}^{-1} \right\|_1 \sum_{x_{c+1}} \left\| \tilde{P}_{r_1}(C_{c+1}) \times \theta^{-}_{c+1} \tilde{f}_{c+1} \right\|_1 \\
\leq (1 + \Delta)^{c-1} \delta^{\text{root}} + (1 + \Delta)^{c-1} - 1 \times 1
\]

Combining these terms proves the induction step. \( \square \)
Putting it all together

We use the fact from HKZ (Hsu et al., 2009) that $(1 + a/t)^t \leq 1 + 2a$ for $a \leq 1/2$. Now $\Delta$ is the main source of error. We set $\Delta \leq O(\varepsilon_{\text{total}}/|C|)$.

Note that

$$\Delta \leq \frac{2d_{\text{max}}^3 + 3d_{\text{max}}}{3 \sqrt{3d_{\text{max}}} \beta d_{\text{max}}} \left( \frac{\sum_{o_{d-e+1}, \ldots, o_d} e(\theta_1, \ldots, \theta_{d-e}, o_{d-e+1}, \ldots, o_d)}{\alpha^2} + \frac{\sum_{o_{d-e+1}, \ldots, o_d} e(\theta_1, \ldots, \theta_{d-e}, o_{d-e+1}, \ldots, o_d)}{\alpha^2} \right)$$

This gives,

$$\frac{2d_{\text{max}}^3 + 3d_{\text{max}}}{3 \sqrt{3d_{\text{max}}} \beta d_{\text{max}}} \left( \frac{\sum_{o_{d-e+1}, \ldots, o_d} e(\theta_1, \ldots, \theta_{d-e}, o_{d-e+1}, \ldots, o_d)}{\alpha^2} + \frac{\sum_{o_{d-e+1}, \ldots, o_d} e(\theta_1, \ldots, \theta_{d-e}, o_{d-e+1}, \ldots, o_d)}{\alpha^2} \right) \leq K\varepsilon_{\text{total}}/|C|$$

where $K$ is some constant.

This implies,

$$\sum_{o_{d-e+1}, \ldots, o_d} e(\theta_1, \ldots, \theta_{d-e}, o_{d-e+1}, \ldots, o_d) \leq K^{\frac{d_{\text{max}}}{2} + 1}\varepsilon_{\text{total}} \frac{\alpha^2 d_{\text{max}} \beta d_{\text{max}}}{2}$$

Now using the concentration bound (Lemma 6) will give,

$$K^{\frac{d_{\text{max}}}{2} + 1}\varepsilon_{\text{total}} \frac{\alpha^2 d_{\text{max}} \beta d_{\text{max}}}{2} \leq \sqrt{\frac{S^\text{max}_O}{N}} \ln \frac{2|C|}{\delta} + \sqrt{\frac{S^\text{max}_O}{N}}$$

Solving for $N$:

$$N \geq O \left( \frac{4S^2_H}{3\beta^2} d_{\text{max}} \sum_{o_{d-e+1}, \ldots, o_d} e(\theta_1, \ldots, \theta_{d-e}, o_{d-e+1}, \ldots, o_d) \right)$$

and this completes the proof.

Matrix Perturbation Bounds

This is Theorem 3.8 from pg. 143 in Stewart and Sun, 1990 (Stewart and Sun, 1990). Let $A \in \mathbb{R}^{m \times n}$, with $m \geq n$ and let $\tilde{A} = A + E$. Then

$$\left\| \tilde{A}^\dagger - A^\dagger \right\|_2 \leq \frac{1 + \sqrt{5}}{2} \max(\|A^\dagger\|_2, \|\tilde{A}\|_2) \|E\|_2$$
Tensor Norm Bounds

For matrices it is true that $\|Mv\|_1 \leq \|M\|_1 \|v\|_1$. We prove the generalization to tensors.

**Lemma 10.** Let $T_1$ and $T_2$ be tensors $\sigma$ a set of (labeled) modes.

$$\|T_1 \times_\sigma T_2\|_1 \leq \|\|T_2\|_1^\sigma \|T_1\|_1$$

(5.56)

**Proof.**

$$\|T_1 \times_\sigma T_2\|_1 = \sum_{i_{1:N}, \sigma} \sum_{j_{1:M}, \sigma} T_1(i_{1:N}, \sigma, \sigma = x) T_2(j_{1:N}, \sigma, \sigma = x)$$

$$= \sum_{i_{1:N}, \sigma} \sum_{j_{1:M}, \sigma} T_1(i_{1:N}, \sigma, \sigma = x) T_2(j_{1:N}, \sigma, \sigma = x)$$

$$= \sum_{i_{1:N}, \sigma} T_1(i_{1:N}, \sigma, \sigma = x) \sum_{j_{1:M}, \sigma} T_2(j_{1:N}, \sigma, \sigma = x)$$

$$\leq \max_x \left( \sum_{j_{1:M}, \sigma} T_2(j_{1:N}, \sigma, \sigma = x) \right) \|T_1\|_1$$

$$= \|\|T_2\|_1^\sigma \|T_1\|_1$$

\[\square\]

We prove a restricted analog of the fact that spectral norm is submultiplicative for matrices i.e. $\|AB\|_2 \leq \|A\|_2 \|B\|_2$.

**Lemma 11.** Let $T$ be a tensor of order $N$ and let $M$ be a matrix. Then,

$$\|\|T \times_1 M\|_2 \leq \|\|T\|_2 \|\|M\|_2$$

(5.57)

**Proof.**

$$\|\|T \times_1 M\|_2 = \sup_{v_1, v_2, \ldots, v_N} \sum_{c_1(1), \ldots, i_N} T(c_1(i), c_2(i), \ldots, i_N) M(c_1(i), m) v_1(m) c_2(i) v_3(i_3) \ldots v_n(i_n)$$

$$= \sup_{v_1, v_2, \ldots, v_N} \sum_{c_1(1), \ldots, i_N} T(c_1(i), c_2(i), \ldots, i_N) \sum_m M(c_1(i), m) v_1(m)$$

$$\leq \sup_{v_1, v_2, \ldots, v_N} \left( \sup_{c_1(1)} \left\| \sum_m M(c_1(i), m) v_1(m) \right\|_2 \right) \times$$

$$\sum_{c_1(1), \ldots, i_N} T(c_1(i), c_2(i), \ldots, i_N) \frac{1}{\left\| \sum_m M(c_1(i), m) v_1(m) \right\|_2} \left( \sum_m M(c_1(i), m) v_1(m) \right) c_2(i) v_3(i_3) \ldots v_n(i_n)$$

$$\leq \|\|M\|_2 \|\|T\|_2$$

\[\square\]
Lemma 12. Let $T$ be a tensor of order $N$. Then,

$$
|||T|||_2 \leq ||T \times_1 v_1, ..., \times_{n-1}v_{n-1}|||_F \leq ||T \times_1 v_1, ..., \times_{n-2} v_{n-2}|||_F \leq ... \leq ||T|||_F (5.58)
$$

Proof. It suffices to show that $\sup_{v \text{ s.t.} ||v||_2 \leq 1} ||T \times_1 v|||_F \leq ||T|||_F$. By submultiplicativity of the Frobenius norm: $||T \times_1 v|||_F \leq ||T|||_F ||v||_F \leq ||T|||_F$, since $||v||_F = ||v||_2 \leq 1$. □

Lemma 13. Let $T$ be a tensor of order $N$, where each mode is of dimension $k$. Then,

$$
|||T|||_1^\sigma \leq k^N |||T|||_2
$$

For any $\sigma$.

Proof. We simply prove this for $\sigma = \emptyset$ (which corresponds to elementwise one norm) since $|||T|||_1^{\sigma_1} \leq |||T|||_1^{\sigma_2}$ if $\sigma_2 \subseteq \sigma_1$. Note that $||T||_1 \leq k^N \max(|T|)$ (where $\max(|T|)$ is the maximum element of $|T|$). Similarly, $\max(|T|) \leq ||T||_2$ which implies that $||T||_1 \leq k^N ||T||_2$. □
Chapter 6

Nonparametric Latent Trees with Kernel Embeddings

In Chapter 4, we proposed linear-algebra based solutions to parameter learning in latent tree graphical models where the variables are discrete. However, in many real world scenarios, the variables under study take on continuous values. One example is demographics where variables like income, crime rate, and age are highly skewed and rarely follow bell curves. For example, Figure 6.1 shows the age distributions of two countries, Egypt, and Japan neither of which follow Gaussian distributions.

It is often difficult to apply traditional approaches to these datasets, since most algorithms for probabilistic modeling assume that the marginal/conditional distributions easily fit into a parametric family. Conventional nonparametric approaches like kernel density estimation severely suffer from the curse of dimensionality, making them impractical for modeling more than a handful of variables.

As a result, typical parameter learning and inference methods approximate the underlying distributions with Gaussians (or mixtures of Gaussians) so that traditional learning methods like Expectation Maximization can be used (Dempster et al., 1977). Structure learning of latent trees is even more challenging and has largely been tackled by heuristics since the search space of structures is intractable. Examples from the phylogenetic community include maximum parsimony and maximum likelihood methods (Semple and Steel, 2003). In the machine learning community, Zhang

Figure 6.1: Examples of real world non-Gaussian distributions: Age distributions of Egypt (left) and Japan (right). (Photo Source: www.census.gov)
(2004) proposed a search heuristic for hierarchical latent class models by defining a series of local search operations and using EM to compute the likelihood of candidate structures. Harmeling and Williams (2011) proposed a greedy algorithm to learn binary trees by joining two nodes with a high mutual information and iteratively performing EM to compute the mutual information among newly added hidden nodes. Alternatively, Bayesian hierarchical clustering (Heller and Ghahramani, 2005) is an agglomerative clustering technique that merges clusters based on a statistical hypothesis test. In addition to lacking theoretical guarantees, these methods do not apply to non-Gaussian continuous distributions.

Contributions of this chapter: In this work, we propose a method for latent tree models with continuous and non-Gaussian observations based on the concept of kernel embedding of distributions (Smola et al., 2007). Kernel embeddings enable efficient representation of large multivariate continuous distributions avoiding the curse of dimensionality inherent in naive kernel density estimation. We present principled methods for parameter learning/inference as well as structure learning:

- **Parameter learning/inference**: We generalize the spectral algorithm from Chapter 4 to the continuous scenario using kernel embeddings. Previous work on nonparametric latent models was limited to hidden Markov models (Song et al., 2010a).

- **Structure learning**: We define a distance measure between variables based on the singular value decomposition of covariance operators. This allows us to generalize some of the distance based latent tree learning procedures from the phylogenetics community such as neighbor joining (Saitou and Nei, 1987) and recursive grouping methods (Choi et al., 2010) to the nonparametric setting. These distance based methods come with strong statistical guarantees, but have previously been unexplored in the nonparametric setting.

Outline: We first review multivariate kernel density estimation to motivate our approach. We then assume the latent structure is known and propose a spectral algorithm for parameter learning (and inference), followed by a structure learning algorithm. Lastly, we demonstrate the efficacy of our approach on synthetic and real datasets.

Prerequisites: This chapter assumes a general understanding of latent variable models as presented in 2.2, knowledge of Hilbert space embeddings as discussed in 2.3, the connection between latent variable models and low rank factorization in Chapter 3, and the tensor notation in 3.1.

6.1 Notation

So that the chapter is standalone, we briefly review notation although it is the same as that in Chapter 4. A latent tree model defines a joint probability distribution over a set of $O$ observed variables $\mathcal{O} = \{X_1, \ldots, X_O\}$ and a set of $H$ hidden variables $\mathcal{H} = \{X_{O+1}, \ldots, X_{O+H}\}$. The complete set of variables is denoted by $\mathcal{X} = \mathcal{O} \cup \mathcal{H}$. We assume that all hidden variables have dimension $S_H$ (the observed variables are continuous). The joint distribution of $\mathcal{X}$ in a latent tree model is
fully characterized by the following equation:

\[ P(x_1, \ldots, x_{O+H}) = \prod_{s=1}^{O+H} P(x_s|x_{\pi(s)}). \]  

(6.1)

where \( X_{\pi(s)} \) denotes the parent of \( X_s \). For simplicity, assume that all leaves are observed variables and all internal nodes are latent. For further notation, let \( X_{s'} \) be some observed node in the subtree rooted at \( X_s \). Let \( c_j(s) \) denote the \( j^{th} \) child of node \( X_s \) where \( X_s \) has \( a_s \) children. For ease of exposition we will assume that all internal nodes have exactly 3 neighbors and that all internal nodes are latent while all leaf nodes are observed.

### 6.2 Kernel Density Estimation

To motivate our approach we first discuss how kernel density estimation (KDE) could be used to tackle our problem. KDE is a nonparametric way of fitting the density of continuous random variables with non-Gaussian statistical features such as multi-modality and skewness. Given a set of \( N \) i.i.d. samples \( \{(x_1^{(n)}, \ldots, x_O^{(n)})\}_{i=1}^N \) from \( p(X_1, \ldots, X_O) \), KDE estimates the density via

\[ \hat{P}(\bar{x}_1, \ldots, \bar{x}_O) \propto b_r := \mathbb{E}_{X_1,\ldots,X_O} \left[ \prod_{j=1}^{O} k(\bar{x}_j, X_j) \right] \approx \frac{1}{N} \sum_{n=1}^{N} \prod_{j=1}^{O} k(\bar{x}_j, x_j^{(n)}) \]  

(6.2)

where \( k(x, x') = \langle \phi_x, \phi_{x'} \rangle \) is a kernel function and \( \phi_x \in \mathcal{F} \) where \( \mathcal{F} \) is the underlying reproducing kernel hilbert space (RKHS). A commonly used kernel function, which we will focus on, is the Gaussian RBF kernel \( k(x, x') = \frac{1}{\sqrt{2\pi} \sigma} \exp(-\|x - x'\|^2 / 2\sigma^2) \). However, traditional KDE cannot model the underlying latent dependency structure among \( X_1, \ldots, X_O \), and therefore scales poorly to beyond a handful of variables due to the curse of dimensionality.\(^1\)

However, incorporating latent structure into KDE could greatly increase its statistical feasibility in high dimensions. In the example in Figure 6.2 consider computing the marginal probability

---

\(^1\)In particular the integrated square error of traditional KDE goes to zero at a rate of \( O(N^{\frac{1}{2}}) \) Casella and Berger, 2002.
\[ P(A = \bar{a}, B = \bar{b}, C = \bar{c}) : \]

\[ \tilde{P}(\bar{a}, \bar{b}, \bar{c}) \propto b_r := \mathbb{E}_{A,B,C,H}[k(\bar{a}, A)k(\bar{b}, B)k(\bar{c}, C)] \]

\[ = \mathbb{E}_H[\mathbb{E}_{A,B,C,H}[k(\bar{a}, A)k(\bar{b}, B)k(\bar{c}, C)]] \]

\[ = \mathbb{E}_H[\mathbb{E}_{A|H}[k(\bar{a}, A)]\mathbb{E}_{B|H}[k(\bar{b}, B)]\mathbb{E}_{C|H}[k(\bar{c}, C)] \quad (6.3) \]

where in the last line we have used the conditional independence implied by the graphical model structure. For general latent tree models, this immediately suggests a message passing scheme to compute the marginal probability of the evidence \((\bar{x}_1, ..., \bar{x}_O)\) at the leaf nodes:

- A leaf node \(X_s\) passes the following message to its parent: \(m_s(x_{\pi(s)}) = \mathbb{E}_{X_s|x_{\pi(s)}}[k(\bar{x}_s, X_s)]\).
- An internal latent node \(X_s\) aggregates incoming messages from its two children and then sends an outgoing message to its own parent \(m_s(x_{\pi(s)}) = \mathbb{E}_{X_s|x_{\pi(s)}}[m_{c1(s)}(X_s)m_{c2(s)}(X_s)]\).
- Finally, at the root node \(X_r\), all incoming messages are multiplied together and the root variable is integrated out \(b_r := \mathbb{E}_{\theta \cup \mathcal{F}} \left[ \prod_{j=1}^{O} k(\bar{x}_j, X_j) \right] = \mathbb{E}_{X_r}[m_{c1(r)}(X_r)m_{c2(r)}(X_r)m_{c3(r)}(X_r)]\).

Unfortunately, under the conventional KDE formulation it is difficult to define or manipulate these messages in practice. This motivates us to use Hilbert Space Embeddings, which provide us with the representation power we need (Smola et al. 2007).

### 6.3 Connection to Hilbert Space Embeddings

Consider Eq. (6.2) again. Products of kernels are also kernels, which allow us to rewrite \(\prod_{j=1}^{O} k(\bar{x}_j, X_j)\) as a single inner product \(\bigotimes_{j=1}^{O} \phi(X_j)\). Here \(\bigotimes_{j=1}^{O} \ast\) denotes the tensor product of \(O\) feature vectors which results in a rank-1 tensor of order \(O\).

Now from the background in Chapter 2.3, \(C_{X_1,..,X_O} := \mathbb{E}_{\theta} \left[ \bigotimes_{j=1}^{O} \phi(X_j) \right] \) is the cross-covariance operator that serves as the Hilbert space embedding of distribution \(p(X_1, ..., X_O)\) with tensor features \(\bigotimes_{j=1}^{O} \phi(X_j)\). Moreover, \(b_r\) is simply evaluating this operator using \(\phi_{\bar{x}_1}, ..., \phi_{\bar{x}_O}\):

\[ b_r := \mathbb{E}_{X_1,..,X_O} \left[ \prod_{j=1}^{O} k(\bar{x}_j, X_j) \right] = C_{X_1,..,X_O} \times_1 \phi_{\bar{x}_1} \times_2 \phi_{\bar{x}_O} \quad (6.4) \]

Thus the message passing scheme in Section 6.2 also serves to evaluate the cross covariance operator \(C_{X_1,..,X_O}\) at a particular value. Furthermore, following Song et al. (2010c), we can recast this message passing scheme in terms of (smaller) cross-covariance and conditional operators. Define the following parameters for each node in the latent tree:

- **root** \((X_r)\): \(C_{rrr} := \mathbb{E}[\phi_r \otimes \phi_r \otimes \phi_r] \) (embedding of \(X_r\) into the RKHS \(\mathcal{F} \otimes \mathcal{F} \otimes \mathcal{F}\))
- **internal node** \((X_s)\): \(C_{ss\pi(s)} := C_{ss\pi(s)}^{-1} = \mathbb{E}[\phi_s \otimes \phi_s \otimes \phi_{\pi(s)}] \mathbb{E}[\phi_{\pi(s)} \otimes \phi_{\pi(s)}]^{-1} \) (conditional embedding of \(X_s|X_{\pi(s)}\) into the RKHS \(\mathcal{F} \otimes \mathcal{F} \otimes \mathcal{F}\))
Then, we can reformulate the message passing scheme. First at the leaf:

\[ m_s(x_{\pi(s)}) = \mathbb{E}_{X_s|\pi_0}[k(\bar{x}_s, X_s)] = \mathbb{E}_{X_s|\pi_0}[(\phi_{x_s}, \phi_{X_s})] = \langle \phi_{x_s}, \mathbb{E}_{X_s|\pi_0}[(\phi_{X_s})] \rangle = C_{s|\pi(s)} \times_s \phi_{x_s} \times_{\pi(s)} \phi_{x_{\pi(s)}} \quad (6.5) \]

where in the last line we have used Eq. \[2.41\]. Thus, since \[ m_s(x_{\pi(s)}) = \langle m_s(\cdot) \rangle \] via the reproducing property, we can conclude that \[ m_s(\cdot) = C_{s|\pi(s)} \times_s \phi_{x_s} \].

For the internal node,

\[ m_s(x_{\pi(s)}) = \mathbb{E}_{X_s|\pi_0}[m_{c_1(s)}(X_s)m_{c_2(s)}(X_s)] = \mathbb{E}_{X_s|\pi_0}[(m_{c_1(s)} \phi_{X_s}) (m_{c_2(s)} \phi_{X_s})] = \mathbb{E}_{X_s|\pi_0}[\phi_{X_s} \times \phi_{X_s}] \times_s m_{c_1(s)} \times_s m_{c_2(s)} = C_{s|\pi(s)} \times_s m_{c_1(s)} \times_s m_{c_2(s)} \times_{\pi(s)} \phi_{x_{\pi(s)}} \quad (6.6) \]

Thus, \[ m_s(\cdot) = C_{s|\pi(s)} \times_s m_{c_1(s)} \times_s m_{c_2(s)} \]. Similarly for the root, we have that

\[ b_r = C_{rr} \times_r m_{c_1(r)} \times_r m_{c_2(r)} \times_r m_{c_3(r)} \quad (6.7) \]

Note this shows how hilbert space embeddings enable us to factorize a kernel density estimate according to a conditional independence structure analogous to conditional probability tables in discrete graphical models.

### 6.4 Deriving the Spectral Algorithm

The drawback of the representations in \[6.5\], \[6.6\] and \[6.7\] is that they require exact knowledge of conditional embedding operators associated with latent variables, but none of these are available in training. Next we will show that we can still make use of the tensor decomposition representation without the need for recovering the latent variables explicitly.

To derive our alternate factorization, we now insert invertible transformations \( F \). At the root,

\[
\begin{align*}
    b_r &= C_{rr} \times_r I \times_r m_{c_1(s)} \times_r I \times_r m_{c_2(s)} \times_r I \times_r m_{c_3(s)} \\
    &= C_{rr} \times_r (F_{c_1(s)} \times_{\omega_{c_1(s)}} F_{c_1(s)}^t) \times_r m_{c_1(s)} \times_r (F_{c_2(s)} \times_{\omega_{c_2(s)}} F_{c_2(s)}^t) \times_r m_{c_2(s)} \times_r (F_{c_3(s)} \times_{\omega_{c_3(s)}} F_{c_3(s)}^t) \times_r m_{c_3(s)} \\
    &= (C_{rr} \times_r F_{c_1(s)} \times_r F_{c_2(s)} \times_r F_{c_3(s)}) \times_{\omega_{c_1(s)}} (m_{c_1(s)} \times_r F_{c_1(s)}^t) \times_{\omega_{c_2(s)}} (m_{c_2(s)} \times_r F_{c_2(s)}^t) \times_{\omega_{c_3(s)}} (m_{c_3(s)} \times_r F_{c_3(s)}^t)
\end{align*}
\]
This continues recursively e.g.

\[ m_{c_1} = C_{ss|π(s)} \times_s (F_{c_1} \times_{ω_1} F^+_1) \times_s m_{c_1(s)} \times_s (F_{c_2} \times_{ω_2} F^+_2) \times_s m_{c_2(s)} \]

\[ = (C_{ss|π(s)} \times_s F_{c_1(s)} \times_s F_{c_2(s)}) \times_{ω_1} (m_{c_1(s)} \times_s F^+_1) \times_{ω_2} (m_{c_2(s)} \times_s F_{c_2(s)})^\dagger \]

where \( ω_1, ω_2, ω_3 \) depend on the definition of \( F \) and will be defined in the section. This leads to the transformed representation:

- **root**: \( \tilde{R} = C_{rrr} \times_r F_{c_1(s)} \times_r F_{c_2(s)} \times_r F_{c_3(s)} \)
- **internal**: \( \tilde{T}_s = C_{ss|π(s)} \times_s F_{c_1(s)} \times_s F_{c_2(s)} \times_{π(s)} F^+_s \)
- **leaf**: \( \tilde{L}_s = C_{s|π(s)} \times_{π(s)} F^+_s \)

### 6.4.1 Observable Representation

**Lemma 14.** If we set \( F_s = U_r^T C_{s^r|π(s)} \) where \( U_s \) are the top left singular vectors of \( C_{s^r|s} \) and \( U_s \) has mode labels \( \{s, ω_s\} \) then we have that

- **root**: \( \tilde{R} = C_{c_1(s)^r c_2(s)^r c_3(s)^r} \times_{c_1(s)^r} U_{c_1(s)} \times_{c_2(s)^r} U_{c_2(s)} \times_{c_3(s)^r} U_{c_3(s)} \)
- **internal**: \( \tilde{T}_s = C_{c_1(s)^r c_2(s)^r c_3(s)^r} \times_{c_1(s)^r} U_{c_1(s)} \times_{c_2(s)^r} U_{c_2(s)} \)
- **leaf**: \( \tilde{L}_s = C_{s^r|π(s)} \times_{π(s)} U^+_s \)

**Proof.** **Root:** We first prove that \( C_{rrr} \times_r C_{c_1(r)^r} \times_r C_{c_2(r)^r} \times_r C_{c_3(r)^r} = C_{c_1(r)^r c_2(r)^r c_3(r)^r} \). Consider any \( f, g, h \in F \). Then,

\[
\begin{align*}
C_{rrr} \times_r C_{c_1(r)^r} \times_r C_{c_2(r)^r} \times_r C_{c_3(r)^r} h \times_{c_2(r)^r} g \times_{c_1(r)^r} f &= \langle f \otimes g \otimes h, C_{rrr} \times_r C_{c_1(r)^r} \times_r C_{c_2(r)^r} \times_r C_{c_3(r)^r} \rangle \\
&= \mathbb{E}_{X_r} \left[ \langle C^T_{c_1(r)^r} \mid f, \phi(X_r) \rangle \langle C^T_{c_2(r)^r} \mid g, \phi(X_r) \rangle \langle C^T_{c_3(r)^r} \mid h, \phi(X_r) \rangle \right] \\
&= \mathbb{E}_{X_r} \left[ \langle f, C^T_{c_1(r)^r} \mid φ(X_r) \rangle \langle g, C^T_{c_2(r)^r} \mid φ(X_r) \rangle \langle h, C^T_{c_3(r)^r} \mid φ(X_r) \rangle \right] \\
&= \mathbb{E}_{X_r} \left[ \mathbb{E}_{X_{c_1(r)^r}} \left[ f(X_{c_1(r)^r}) \right] \mathbb{E}_{X_{c_2(r)^r}} \left[ g(X_{c_2(r)^r}) \right] \mathbb{E}_{X_{c_3(r)^r}} \left[ h(X_{c_3(r)^r}) \right] \right] \\
&= \mathbb{E}_{X_r} \left[ \mathbb{E}_{X_{c_1(r)^r}} \left[ f(X_{c_1(r)^r}) \right] g(X_{c_2(r)^r}) h(X_{c_3(r)^r}) \right] \\
&= \langle f \otimes g \otimes h, \mathbb{E}_{X_{c_1(r)^r}} \left[ f(X_{c_1(r)^r}) \right] \mathbb{E}_{X_{c_2(r)^r}} \left[ g(X_{c_2(r)^r}) \right] \mathbb{E}_{X_{c_3(r)^r}} \left[ h(X_{c_3(r)^r}) \right] \rangle \\
&= C_{c_1(r)^r c_2(r)^r c_3(r)^r} h \times_{c_2(r)^r} g \times_{c_1(r)^r} f
\end{align*}
\]

(6.8)

We can thus conclude that

\[
\tilde{R} = C_{c_1(s)^r c_2(s)^r c_3(s)^r} U_{c_1(s)} \times_{c_1(s)^r} U_{c_2(s)} \times_{c_2(s)^r} U_{c_3(s)}
\]

(6.9)
Leaf: Consider expanding the related quantity $\widetilde{\mathcal{L}}_s(U_s^TC_{\pi\gamma(s)^*}):$

\[
\begin{align*}
\widetilde{\mathcal{L}}_s(U_s^TC_{\pi\gamma(s)^*}) &= C_{s|\pi(s)}(U_s^TC_{s|\pi(s)})^\dagger(U_s^TC_{\pi\gamma(s)^*}) \\
&= C_{s|\pi(s)}(U_s^TC_{s|\pi(s)})^\dagger(U_s^TC_{s|\pi(s)}C_{\pi(s)|\pi(s)}^TC_{\pi(s)^*|\pi(s)}) \\
&= C_{s|\pi(s)}C_{\pi(s)|\pi(s)}^TC_{\pi(s)^*|\pi(s)} \\
&= C_{s|\pi(s)}^T 
\end{align*}
\]

(6.10)

where we have used the fact that $C_{s|\pi(s)}C_{\pi(s)^*|\pi(s)}^TC_{\pi(s)^*|\pi(s)} = C_{s|\pi(s)}^T$ (which is proved using the same technique as used for the proof of the root).

This implies that $\widetilde{\mathcal{L}}_s = (C_{\pi(s)^*|s}U_s)^\dagger C_{\pi(s)^*|s} = U_s$.

Intermediate Node: Consider expanding the quantity $\widetilde{\mathcal{T}}_s \times_{\alpha_s} (C_{\pi(s)^*|s}U_s):$

\[
\begin{align*}
\widetilde{\mathcal{T}}_s \times_{\alpha_s} (C_{\pi(s)^*|s}U_s) &= C_{ss|\pi(s)} \times_s U_{c_1(s)}^TC_{c_1(s)^*|s} \times_s U_{c_2(s)}^TC_{c_2(s)^*|s} \times_{\pi(s)} (C_{s'|\pi(s)}^T U_s)^\dagger \times_{\alpha_s} (C_{\pi(s)^*|s}U_s) \\
&= C_{ss|\pi(s)} \times_s C_{c_1(s)^*|s} \times_s C_{c_2(s)^*|s} \times_{\pi(s)} (C_{s'|\pi(s)}^T U_s)^\dagger \times_{c_1(s)^*} U_{c_1(s)}^T \times_{c_2(s)^*} U_{c_2(s)}^T \\
&= C_{ss|\pi(s)} \times_s C_{c_1(s)^*|s} \times_s C_{c_2(s)^*|s} \times_{\pi(s)} (C_{\pi(s)^*|\pi(s)}C_{\pi(s)|\pi(s)}) (C_{s'|\pi(s)}^T U_s)^\dagger \times_{c_1(s)^*} U_{c_1(s)}^T \times_{c_2(s)^*} U_{c_2(s)}^T \\
&= C_{c_1(s)^*}C_{c_2(s)^*|\pi(s)^*} \times_{c_1(s)^*} U_{c_1(s)}^T \times_{c_2(s)^*} U_{c_2(s)}^T 
\end{align*}
\]

(6.11)

where in the last line we have claimed that $C_{c_1(s)^*}C_{c_2(s)^*|\pi(s)^*} = C_{ss|\pi(s)} \times_s C_{c_1(s)^*|s} \times_s C_{c_2(s)^*|s} \times_{\pi(s)} C_{\pi(s)^*|\pi(s)}C_{\pi(s)|\pi(s)}$. To prove this assertion, first consider the $C_{ss|\pi(s)} \times_s C_{c_1(s)^*|s} \times_s C_{c_2(s)^*|s}$ part. For any $f, g \in \mathcal{F}$:

\[
\begin{align*}
\langle f \otimes g, C_{ss|\pi(s)} \times_s C_{c_1(s)^*|s} \times_s C_{c_2(s)^*|s} \times_{\pi(s)} \phi(x_{\pi(s)}) \rangle &= \langle (C_{c_1(s)^*}f) \otimes (C_{c_2(s)^*}g), C_{ss|\pi(s)} \times_{\pi(s)} \phi(x_{\pi(s)}) \rangle \\
&= \langle (C_{c_1(s)^*}f) \otimes (C_{c_2(s)^*}g), \mathbb{E}_{X_s|x_{\pi(s)}}[\phi(X_s) \otimes \phi(X_s)] \rangle \\
&= \mathbb{E}_{X_s|x_{\pi(s)}} \left[ \langle (C_{c_1(s)^*}f) \otimes (C_{c_2(s)^*}g), \phi(X_s) \otimes \phi(X_s) \rangle \right] \\
&= \mathbb{E}_{X_s|x_{\pi(s)}} \left[ \langle f, C_{c_1(s)^*} \phi(X_s) \rangle \langle g, C_{c_2(s)^*} \phi(X_s) \rangle \right] \\
&= \mathbb{E}_{X_s|x_{\pi(s)}} \left[ \mathbb{E}_{X_{c_1(s)^*}|X_s}[f(X_{c_1(s)^*})] \mathbb{E}_{X_{c_2(s)^*}|X_s}[g(X_{c_2(s)^*})] \right] \\
&= \mathbb{E}_{c_1(s)^*|s}C_{c_2(s)^*|\pi(s)^*} \times_{\pi(s)} \phi(x_{\pi(s)}) \\
&= \langle f \otimes g, C_{c_1(s)^*}C_{c_2(s)^*|\pi(s)^*} \times_{\pi(s)} \phi(x_{\pi(s)}) \rangle 
\end{align*}
\]

(6.12)

Thus, $C_{c_1(s)^*}C_{c_2(s)^*|\pi(s)^*} = C_{ss|\pi(s)} \times_s C_{c_1(s)^*|s} \times_s C_{c_2(s)^*|s}$. We can then conclude (using a similar derivation to that for the root) that $C_{c_1(s)^*}C_{c_2(s)^*|\pi(s)^*} = C_{c_1(s)^*}C_{c_2(s)^*|\pi(s)^*} \times_{\pi(s)} C_{\pi(s)^*|\pi(s)^*}C_{\pi(s)|\pi(s)}$. Thus,

\[
C_{c_1(s)^*}C_{c_2(s)^*|\pi(s)^*} = C_{ss|\pi(s)} \times_s C_{c_1(s)^*|s} \times_s C_{c_2(s)^*|s} \times_{\pi(s)} C_{\pi(s)^*|\pi(s)^*}C_{\pi(s)|\pi(s)} 
\]

(6.13)
Now, returning to Eq. 6.11 we get that
\[
\mathbf{\tilde{T}}_s = \mathbf{C}_{c_1(s)^*c_2(s)^*c_1(s)^*} \times \mathbf{U}^{\mathsf{T}}_{c_1(s)} \times \mathbf{U}^{\mathsf{T}}_{c_2(s)} \times \mathbf{C}_{\pi(s)^*s^*} \mathbf{U}_{s^*}^\dagger
\]  
(6.14)
where one valid choice for \(s^*\) is \(c_1(s)^*\).

### 6.4.2 Training and Test

Empirically, during training, our method needs to compute estimates of the following cross-covariance operators:
\[
\widehat{\mathbf{C}}_{s,t,\mu} = \frac{1}{N} \sum_{n=1}^{N} \phi_{x_s^{(n)}} \otimes \phi_{x_t^{(n)}} \otimes \phi_{x_u^{(n)}} \\
\widehat{\mathbf{C}}_{s,t} = \frac{1}{N} \sum_{n=1}^{N} \phi_{x_s^{(n)}} \otimes \phi_{x_t^{(n)}}
\]  
(6.15)

Please see 2.3.5 for how to do this efficiently using incomplete cholesky decompositions.

The training and test algorithms are shown in Algorithms 6 and 7. As one can see, although the derivation is different, the end result is very similar to Algorithms 9 and Algorithms 10 in Chapter 4 (the only difference being the probability tensors are replaced with hilbert space operators). Thus, spectral approaches elegantly generalize to nonparametric settings unlike traditional nonconvex-optimization based methods.

---

**Algorithm 6** Kernel spectral learning algorithm for latent tree graphical model

**In:** Tree topology and \(N\) i.i.d. samples \(\{x_1^{(n)}, \ldots, x_N^{(n)}\}_{n=1}^{N}\)

**Out:** Estimated observable root, internal, and leaf parameters, \(\widehat{\mathbf{R}}, \mathbf{\tilde{T}}_s\) for each non-root internal node, \(\widehat{\mathbf{L}}_c\) for each leaf node

1: For each node \(X_s\), compute the kernel matrix \(\mathbf{K}_s\) where \(K_s(m,n) = \frac{1}{2\alpha^2} \exp(-||x_s^{(m)} - x_s^{(n)}||^2/2\alpha^2)\).
2: For each node \(X_s\), recover the approximate feature map \(\widehat{\mathbf{F}}_{x_s^{(m)}}, \ldots, \widehat{\mathbf{F}}_{x_s^{(n)}}\) via incomplete cholesky decomposition on \(\mathbf{K}_s\).
3: Compute empirical estimates of covariance operators for all observed pairs and triples:
\[
\widehat{\mathbf{C}}_{s,t,\mu} = \frac{1}{N} \sum_{n=1}^{N} \widehat{\mathbf{F}}_{x_s^{(n)}} \otimes \widehat{\mathbf{F}}_{x_t^{(n)}} \otimes \widehat{\mathbf{F}}_{x_u^{(n)}} \\
\widehat{\mathbf{C}}_{s,t} = \frac{1}{N} \sum_{n=1}^{N} \widehat{\mathbf{F}}_{x_s^{(n)}} \otimes \widehat{\mathbf{F}}_{x_t^{(n)}}
\]  
(6.16)
4: For each node \(X_s\), perform a “thin” singular value decomposition of \(\mathbf{C}_{s\rightarrow s'} = \mathbf{U}\Sigma\mathbf{V}^\dagger\); let \(\widehat{\mathbf{U}}_i = \mathbf{U}_{(i,1:S)}\) be the the first \(S\) principal left singular vectors.
5: Compute observable parameters as:
\[
\widehat{\mathbf{R}} = \widehat{\mathbf{C}}_{c_1(s)^*c_2(s)^*c_3(s)^*} \times \mathbf{U}_{c_1(s)} \times \mathbf{U}_{c_2(s)} \times \mathbf{U}_{c_3(s)} \\
\mathbf{\tilde{T}}_s = \mathbf{C}_{c_1(s)^*c_2(s)^*c_3(s)^*} \times \mathbf{U}_{c_1(s)} \times \mathbf{U}_{c_2(s)} \times \mathbf{U}_{c_3(s)} \\
\widehat{\mathbf{L}}_c = \mathbf{C}_{c_{s\rightarrow s'}} \times \mathbf{U}_{c_{s\rightarrow s'}}^\dagger = \mathbf{U}_c
\]
6.5 Structure Learning of Latent Tree Graphical Models

The last section focused on parameter learning and inference where the structure of the latent tree is known. In this section, we focus on learning the structure of the latent tree. Structure learning of latent trees is a challenging problem that has largely been tackled by heuristics since the search space of structures is intractable. The additional challenge in our case is that the observed variables are continuous and non-Gaussian, which we are not aware of any existing methods for this problem.

Structure learning algorithm

We develop a distance based method for constructing latent trees of continuous, non-Gaussian variables. The idea is that if we have a tree metric (distance) between distributions on observed nodes, we can use the property of the tree metric to reconstruct the latent tree structure using algorithms such as neighbor joining (Saitou and Nei, 1987) and the recursive grouping algorithm (Choi et al., 2010). These methods take a distance matrix among all pairs of observed variables as input and output a tree by iteratively adding hidden nodes. While these methods are iterative, they have strong theoretical guarantees on structure recovery when the true distance matrix forms an additive tree metric. However, most previously known tree metrics are defined for discrete and Gaussian variables. The additional challenge in our case is that the observed variables are continuous and non-Gaussian. We propose a tree metric below which works for continuous non-Gaussian cases.

Tree metric and pseudo-determinant

We will first explain some basic concepts of a tree metric. If the joint probability distribution $P(X)$ has a latent tree structure, then a distance measure $d(s, t)$ between an arbitrary variables pairs $X_s$ and $X_t$ are called tree metric if it satisfies the following path additive condition: $d(s, t) = \sum_{(u, v) \in \text{Path}(s, t)} d(u, v)$. For discrete and Gaussian variables, tree metric can be defined via the determinant $|\cdot|$ (Choi et al., 2010)

$$d(s, t) = -\frac{1}{2} \log |P(X_s, X_t)P(X_s, X_t)^\top| + \frac{1}{4} \log |P(\varnothing X_s)P(\varnothing X_t)^\top| + \frac{1}{4} \log |P(\varnothing X_t)P(\varnothing X_t)^\top|$$  (6.17)

However, this definition of tree metric is restricted in the sense that it requires all discrete variables to have the same number of states and all Gaussian variables have the same dimension. This is because determinant is only defined (and non-zero) for square and non-singular matrices. For our more general scenario, where the observed variables are continuous non-Gaussian but the hidden variables have dimension $S_H$, we will define a tree metric based on pseudo-determinant which works for our operators.

Algorithm 7 Inference with Kernel Spectral Parameters

In: Tree topology, set of spectral parameters $\hat{\mathcal{R}}, \hat{T}$ for all non-root internal nodes, $\hat{L}$, for all leaf nodes, evidence $\{\bar{x}_1, ..., \bar{x}_O\}$, and associated feature functions $\{\hat{\phi}_{\bar{x}_1}, ..., \hat{\phi}_{\bar{x}_O}\}$

Out: estimated unnormalized probability $b_r \propto \hat{P}(\bar{x}_1, ..., \bar{x}_O)$

1: In reverse topological order, each node accumulates at message at leaf and sends to parent

- Leaf: $\bar{m}_s = \hat{L}_s \times \hat{\phi}_{\bar{x}_s}$
- Internal Node: $\bar{m}_s = \hat{T}_s \times_{s} \bar{m}_{c_1}(s) \times_{s} \bar{m}_{c_2}(s)$
- Root: $b_r = \hat{\mathcal{R}} \times_{r} \bar{m}_{c_1}(r) \times_{r} \bar{m}_{c_2}(r) \times_{r} \bar{m}_{c_3}(r)$
**Nonparametric tree metric** The pseudo-determinant is defined as the product of non-zero singular values of an operator $|C|_* = \prod_{i=1}^{S_H} \sigma_i(C)$. In our case, since we assume that the dimension of the hidden variables is $S_H$, the pseudo-determinant is simply the product of top $S_H$ singular values. Then we define the distance metric between as two continuous non-Gaussian variables $X_s$ and $X_t$ as follows:

$$d(s, t) = -\frac{1}{2} \log |C_{st}C_{st}^T|_* + \frac{1}{4} \log |C_{ss}C_{st}^T|_* + \frac{1}{4} \log |C_tC_{tt}^T|_*.$$  \hfill (6.18)

**Lemma 15.** The distance measure in Eq. (6.18) is an additive tree metric.

**Proof.** Here we just show why the lemma holds for the simple path $s \leftarrow u \rightarrow t$ where $u$ is latent and $s$ and $t$ are observed. The full proof is in the Appendix. Note that

$$C_{st} = C_{slu}C_{uu}C_{tu}^T$$  \hfill (6.19)

Thus,

$$C_{st}C_{st}^T = C_{slu}C_{uu}C_{tu}^TC_{tu}C_{uu}C_{slu}C_{tu}^T$$  \hfill (6.20)

Combining this definition with Sylvester’s Determinant Theorem ([Akritas et al. 1996]), gives us that:

$$|C_{st}C_{st}^T|_* = |C_{slu}C_{uu}C_{tu}^TC_{tu}C_{uu}C_{slu}C_{tu}^T|_* = |C_{tu}^TC_{slu}C_{uu}C_{tu}|_*$$  \hfill (6.21)

(i.e. we can move $C_{slu}^T$ the to the front).

Now $C_{slu}^TC_{slu}C_{uu}C_{tu}^TC_{tu}C_{uu}$ has rank $S_H$ so the pseudo-determinant equals the normal determinant in this case. Using the fact that $|AB| = |A||B|$ if $A$ and $B$ are square, we get

$$|C_{st}C_{st}^T|_* = |C_{slu}^TC_{slu}C_{uu}C_{tu}^TC_{tu}C_{uu}|_* = |C_{uu}||C_{tu}^TC_{slu}C_{uu}|_* = \frac{|C_{uu}||C_{tu}^TC_{slu}C_{uu}|_*}{|C_{uu}|}.$$  \hfill (6.22)

Furthermore, note that

$$|C_{uu}C_{slu}^TC_{uu}C_{uu}C_{tu}^T|_* = |C_{slu}C_{uu}C_{uu}C_{tu}^T|_* = |C_{uu}C_{st}^T|_*.$$  \hfill (6.23)

This gives,

$$|C_{st}C_{st}^T|_* = \frac{|C_{uu}C_{st}^T|_*}{|C_{uu}|} \times \frac{|C_{tu}^T|_*}{|C_{uu}|}.$$  \hfill (6.24)
Substituting back into Eq. (6.18) proves that

\[
d(s, t) = \frac{1}{2} \log |C_{su}C_{su}^\top| + \frac{1}{2} \log |C_{ut}C_{ut}^\top| + \frac{1}{4} \log |C_{ss}C_{ss}^\top| + \frac{1}{4} \log |C_{tt}C_{tt}^\top|
\]

\[
= d(s, u) + d(u, t)
\]  

(6.25)

To reconstruct the tree we use this metric along with the neighbor joining algorithm (Saitou and Nei 1987) that is described in the Appendix.

6.6 Experiments

We evaluate our method on synthetic data as well as a real-world crime/communities dataset (Asuncion and Newman 2007; Redmond and Baveja 2002). We primarily compare to 2 existing approaches. The first is to assume the data is multivariate Gaussian and use the tree metric defined in Choi et al. (2010) (which is essentially a function of the correlation coefficient). The second existing approach we compare to is the Nonparanormal (NPN) (Liu et al. 2009) which assumes that there exist marginal transformations \( f_1, \ldots, f_p \) such that \( f(X_1), \ldots, f(X_p) \sim N(\mu, \Sigma) \). If the data comes from a Nonparanormal distribution, then the transformed data are assumed to be multivariate Gaussian and the same tree metric as the Gaussian case can be used on the transformed data. Our approach makes much fewer assumptions about the data than either of these two methods which can be more favorable in practice.

To perform inference in our approach, we use the spectral algorithm described earlier in the paper. For inference in the Gaussian (and nonparanormal) cases, we use the technique in Choi et al. (2010) to learn the model parameters (covariance matrix). Once the covariance matrix has been estimated, marginalization in a Gaussian graphical model reduces to solving a linear equation of one variable if we are only computing the marginal of one variable given a set of evidence (Bickson 2008).

6.6.1 Synthetic data: density estimation

Before moving on to larger experiments, we first show that our model assumptions can result in more accurate density estimation of non-Gaussian distributions. The underlying data is generated as a two dimensional mixture of exponentials:

\[
P(x_1, x_2) \propto \exp(||x_1 - \mu_1|| + ||x_2 - \mu_2||) + \exp(||x_1 + \mu_1|| + ||x_2 + \mu_2||)
\]  

(6.26)

Note that the first component has mean \((\mu_1, \mu_2)\) while the second component has mean \((-\mu_1, -\mu_2)\). We experiment with the different values \((\mu_1, \mu_2) = (2, 2), (\mu_1, \mu_2) = (4, 4), \) and \((\mu_1, \mu_2) = (6, 6)\). For all methods we evaluate the density on a grid \( G \) of evenly spaced points in \([-2\mu_1, 2\mu_1] \times [-2\mu_2, 2\mu_2]\). The total error is measured as:

\[
err = \sqrt{\sum_{(x_1, x_2) \in G} ||P(x_1, x_2) - \hat{P}(x_1, x_2)||^2}
\]

91
Figure 6.3: Density estimation of 2-dimensional mixture of laplace distributions.

Figure 6.3 shows the results where we compare our approach with the Gaussian and Gaussian mixture distributions. As expected, the problem is more difficult when the components are closer together. Our method performs the best for all the cases.

### 6.6.2 Synthetic data: structure recovery.

The second experiment is to demonstrate how our method compares to the Gaussian and Nonparanormal methods in terms of structure recovery for larger trees. We experiment with 3 different tree types (each with 64 leaves or observed variables): a balanced binary tree, a completely binary skewed tree (like an HMM), and randomly generated binary trees. Furthermore we explore with two types of underlying distributions: (1) A multivariate Gaussian with mean zero and inverse covariance matrix that respects the tree structure. (2) A highly non-Gaussian distribution that uses the following generative process to generate the n-th sample from a node \( s \) in the tree (denoted \( x_s^{(n)} \)): If \( s \) is the root, sample from a mixture of 2 Gaussians. Else, with probability \( \frac{1}{2} \) sample from a Gaussian with mean \( -x_s^{(n)} \pi_s \) and with probability \( \frac{1}{2} \) sample from a Gaussian with mean \( x_s^{(n)} \pi_s \).

We vary the training sample size from 200 to 100,000. Once we have computed the empirical tree distance matrix for each algorithm, we use the neighbor joining algorithm [Saitou and Nei 1987] to learn the trees. For evaluation we compare the number of hops between each pair of leaves in the true tree to the estimated tree. For a pair of leaves \( i, j \) the error is defined as:

\[
\text{error}(i, j) = \frac{|\text{hops}^*(i, j) - \text{hops}(i, j)|}{\text{hops}^*(i, j)} + \frac{|\text{hops}^*(i, j) - \text{hops}(i, j)|}{\text{hops}(i, j)},
\]

where \( \text{hops}^* \) is the true number of hops and \( \text{hops} \) is the estimated number of hops. The total error is then computed by adding the error for each pair of leaves.

The performance of our method depends on the number of singular values chosen and we experimented with 2, 5 and 8 singular values. Furthermore, we choose the bandwidth \( \sigma \) for the Gaussian RBF kernel needed for the covariance operators using median distance between pairs of training points.

When the underlying distribution is not Gaussian, our method performs better than the Gaussian and Nonparanormal methods for all the tree structures. This is to be expected, since the non-Gaussian data we generated is neither Gaussian or Nonparanormal, yet our method is able to learn the structure correctly. We also note that balanced binary trees are the easiest to learn while
Figure 6.4: Comparison of our kernel structure learning method to the Gaussian and Nonpara-
normal methods on different tree structures. Top row: data points are generated from Gaussian
distributions with latent variables connected by a tree structure. Bottom row: data points are
generated from mixture of Gaussian distributions with latent variables connected by a tree struc-
ture. Especially in the latter case, our kernel structure learning method is able to adapt the data
distributions and recover the structure in a much more accurate way.

the skewed trees are the hardest (Figure 6.4).

Even when the underlying distribution is Gaussian, our method still performs very well com-
pared to the Gaussian and NPN approaches and outperforms them for the binary and balanced
trees. It performs worse for the skewed case likely due to the fact that the eigenvalues (dependence)
decay along the length of the tree leading to larger errors in the empirical distance matrix.

Although it would be interesting to compare to the pouch latent tree model (Poon et al., 2010),
their model assumes multiple observed variables can exist in the same leaf of the latent tree (unlike
our approach) which makes a direct structure comparison difficult.

6.6.3 Crime Dataset.

Finally, we explore the performance of our method on a communities and crime dataset from the
UCI repository (Asuncion and Newman, 2007; Redmond and Baveja, 2002). In this dataset several
real valued attributes are collected for several communities, such as ethnicity proportions, income,
poverty rate, divorce rate etc., and the goal is to predict the number of violent crimes (proportional
to size of community) that occur based on these attributes. In general these attributes are highly
skewed and therefore not well characterized by a Gaussian model.
We divide the data into 1400 samples for training, 300 samples for model selection (held-out likelihood), and 300 samples for testing. We pick the first 50 of these attributes, plus the violent crime variable and construct a latent tree using our tree metric and neighbor joining algorithm \cite{Saitou1987}. We depict the tree in Figure 6.5 and highlight a few coherent groupings. For example, the “elderly” group attributes are those related to retirement and social security (and thus correlated). The large clustering in the center is where the class variable (violent crimes) is located next to the poverty rate, and the divorce rate among other relevant variables. Other groupings include type of occupation and education level as well as ethnic proportions. Thus, overall our method is able to capture sensible relationships.

For a more quantitative evaluation, we condition on a set of $\mathcal{E}$ evidence variables where $|\mathcal{E}| = 30$ and predict the violent crimes class label. We experiment with a varying number of sizes of the training set from 200 to 1400. At test, we evaluate on all the 300 test examples for 10 randomly chosen evidence sets of evidence variables. Since the crime variable is a number between 0 and 1, our error measure is simply $err(\hat{c}) = |\hat{c} - c^*|$ (where $\hat{c}$ is the predicted value and $c^*$ is the true value).

In this experiment, in addition to comparing with the Gaussian and the Nonparanormal, we also compare with two standard classifiers, (Gaussian) Naive Bayes and Linear Regression. Although Naive Bayes can easily handle missing values, linear regression cannot. To deal with this problem, we simply use the mean value of a variable if it is not in $\mathcal{E}$ (this performed much better than setting it to zero).

As one can see in Figure 6.5 our method outperforms all the other approaches. We find that our method performs similarly for different choices of $\mathcal{E}$. Moreover, the accuracy of the Gaussian and nonparanormal vary widely for different evidence sets (and thus the more erratic overall performance). Thus, in this case our method is better able to capture the skewed distributions of the variables than the other methods.
6.7 Conclusion

We present a distribution embedding framework for nonparametric latent tree graphical models. Our approach be used to recover the latent tree structures, and perform local-minimum-free spectral learning and inference for continuous and non-Gaussian variables. Both simulation and results on real datasets show the advantage of our proposed approach for non-Gaussian data.

6.8 Appendix

6.8.1 Proof of Lemma 15

Proof. For conciseness, we simply prove the property for paths of length 2. The proof for more general cases follows similarly (e.g. see [Anandkumar et al., 2011]). For paths of length 2, $s - u - t$, there are three cases:

- $s \leftarrow u \rightarrow t$
- $s \leftarrow u \leftarrow t$
- $s \rightarrow u \rightarrow t$

Case 1 was already shown in the main text so we just show Cases 2 and 3 here.

**Case 2:** $s \leftarrow u \leftarrow t$ Since only leaf nodes can be observed, $u$ and $t$ must be latent but $s$ can be either observed or latent. We assume it is observed, the latent case follows similarly.

Using the conditional independence relationship,

$$C_{st} = C_{sl|u} C_{u|t} C_{tt}$$  \hspace{1cm} (6.27)

Thus, via Sylvester’s Determinant Theorem (Ákritas et al., 1996) as before,

$$|C_{st} C_{st}^T|_+ = |C_{sl|u} C_{u|t} C_{tt} C_{u|t} C_{u|t} C_{tt}^T C_{sl|u}^T|_+$$

$$= |C_{sl|u} C_{sl|u} C_{u|t} C_{tt} C_{tt}^T C_{tt}^T|_+$$  \hspace{1cm} (6.28)

Now $C_{sl|u}^T C_{sl|u} C_{tt} C_{tt} C_{tt}^T$ has rank $S_H$, the pseudo-determinant equals the normal determinant in this case. Using the fact that $|AB| = |A||B|$ if $A$ and $B$ are square, we get

\[\text{2 although the additive distance metric is undirected, the conditional embedding operators are defined with respect to parent-child relationships, so we must consider direction.}\]
Furthermore, note that

\[
|C_{st}C_{st}^T| = |C_{slu}C_{slu}^T|C_{ut}C_{ut}^T|C_{slu}C_{slu}^T|
\]

\[
= |C_{slu}C_{slu}^T||C_{ut}C_{ut}^T|
\]

\[
= |C_{slu}||C_{ut}C_{ut}^T|	imes |C_{slu}|
\]

\[
= |C_{slu}^TC_{slu}C_{uu}|C_{uu}|	imes |C_{ut}C_{ut}^T|
\]

(6.29)

This gives,

\[
|C_{st}C_{st}^T| = |C_{slu}C_{slu}^T||C_{uu}|	imes |C_{ut}C_{ut}^T|
\]

(6.30)

Substituting back into Eq. (6.18) proves that

\[
d(s, t) = -\frac{1}{2} \log |C_{ss}C_{ss}^T| + \frac{1}{2} \log |C_{uu}C_{uu}^T| + \frac{1}{4} \log |C_{uu}C_{uu}^T| + \frac{1}{4} \log |C_{uu}C_{uu}^T|
\]

\[
= d(s, u) + d(u, t)
\]

(6.32)

Case 3: s → u → t  The same argument as case 2 holds here.

\[\square\]

6.8.2 Neighbor Joining Algorithm

Neighbor joining (Saitou and Nei, 1987) is a distance based algorithm that given a distance matrix among pairs of observed (leaf) nodes, returns a binary latent tree. Although it is greedy, if the distance matrix satisfies the tree additivity property, neighbor joining is provably consistent (Atteson, 1997). We describe it below.

Let \(W\) be the active working set. Initially \(W = \{X_1, ..., X_O\}\). To decide which two nodes to merge, neighbor joining creates a \(Q\) matrix defined as follows:

\[
Q(i, j) = (|W| - 2)d(i, j) - \sum_{k \in W} d(i, k) - \sum_{k \in W} d(j, k)
\]

(6.33)

3We follow the exposition in http://en.wikipedia.org/wiki/Neighbor_joining
Algorithm 8 Neighbor Joining Algorithm (Saitou and Nei, 1987)

Input: Pairwise distances $d(s, t)$ between observed variables in $\mathcal{O}$

Output: Latent tree structure $\mathcal{T} = (\mathcal{V}, \mathcal{E})$

Initialize the latent tree structure: $\mathcal{W} = \{1, ..., O\}, \mathcal{E} = \emptyset$

Define a working set: $\mathcal{W} = \{1, ..., O\}$, $u = |\mathcal{O}| + 1$

while $|\mathcal{W}| \geq 2$ do

Create $Q$ matrix according to Eq. 6.33

$(f^*, g^*) = \text{argmin}_{(f, g) \in \mathcal{W}, f \neq g} Q(f, g)$

Create a new node with index $u$ to join node $f^*$ and $g^*$

Update distances to the new node $u$ according to Eqs. 6.36

$\mathcal{W} \leftarrow \mathcal{W} \setminus \{f^*, g^*\} \cup \{u\}$

$\mathcal{V} \leftarrow \mathcal{V} \cup \{u\}$

$\mathcal{E} \leftarrow \mathcal{E} \cup \{(u, f^*), (u, g^*)\}$

$u \leftarrow u + 1$

end while

It then selects the pair of nodes $f^*, g^*$ such that $Q(f^*, g^*)$ is the lowest.

If $f^*$ and $g^*$ are joined to a parent $u$, compute the distances to $u$ as follows.

\[
\begin{align*}
    d(f^*, u) &= \frac{1}{2} d(f^*, g^*) + \frac{1}{2(|\mathcal{W}| - 2)} \left[ \sum_{k \in \mathcal{W}} d(f^*, k) - \sum_{k \in \mathcal{W}} d(g^*, k) \right] \\
    d(g^*, u) &= d(f^*, g^*) - d(f^*, u) \\
    d(k, u) &= \frac{1}{2} [d(f^*, k) + d(g^*, k) - d(f^*, g^*]) \quad \forall k \in \mathcal{W} \setminus \{f^*, g^*\}
\end{align*}
\] (6.34, 6.35, 6.36)

After these distances are computed, $f^*$ and $g^*$ are merged into a node $u$. $f^*, g^*$ are removed from $\mathcal{W}$ and $u$ is added to $\mathcal{W}$. $Q$ is then recomputed and the algorithm picks another pair of nodes to merge until the tree is complete. The algorithm is shown in Figure 6.
Chapter 7

Alternative Spectral Representation of Latent Tree Graphical Models

The spectral representation that we had derived in Chapter 4 has the disadvantage that it is not suitable for trees where certain nodes have large numbers of children, because the order of the parameter tensor required is equal to the degree of the associated node. This limits the applicability of this method in practice, a problem that convention learning methods such as Expectation Maximization (EM) (Dempster et al., 1977) do not face.

Moreover, Mossel and Roch (Mossel and Roch, 2006) also proposed a spectral algorithm for latent variable models which applies to arbitrary tree topologies and only requires tensors of order 3. While they made very restrictive assumptions, and their method does not perform well empirically, it still hints that a more efficient observable representation may be possible.

**Contribution of this chapter:** We derive an alternate spectral learning algorithm for latent tree graphical models that only requires tensors of order 3 regardless of the topology of the tree. Key to our derivation is the use of an alternate representation for tensor message passing that is more compact. Sample complexity results are provided and empirically we show that this new representation performs favorably than the representation in Chapter 4, especially for trees with larger degree. We also compare with the algorithm of Mossel and Roch (2006) showing our method gives considerably more stable results across a variety of tree topologies. Unfortunately, this representation doesn’t extend easily to junction trees or kernel embeddings in Chapters 5 and 6.

**Outline:** We first provide an example for intuition and then derive the spectral algorithm by representing message passing in a compact tensor form, followed by transforming this representation into one that only depends on observed variables. Finally, we analyze the sample complexity of our method and compare it empirically to the spectral algorithm derived in Chapter 4 and that of Mossel and Roch (2006).

**Prerequisites:** This chapter assumes a general understanding of latent variable models as presented in 2.2, the connection between latent variable models and low rank factorization in Chapter 3, and the tensor notation in 3.1. It is also recommended to read Chapter 4 first.
7.1 Intuition

We first present a small example for intuition. Consider the graphical model in Figure 7.1. The original spectral algorithm in Chapter 4 first proposes the following factorization of the marginal probability tensor $\mathcal{P}(X_1, X_2, X_3, X_4, X_5)$:

$$
\mathcal{P}(X_1, X_2, X_3, X_4, X_5) = \mathcal{P}(\emptyset_5 H) \times_H \mathcal{P}(X_1|H) \times_H \mathcal{P}(X_2|H) \times_H \mathcal{P}(X_3|H) \times_H \mathcal{P}(X_4|H) \times_H \mathcal{P}(X_5|H)
$$

Although $\mathcal{P}(\emptyset_5 H)$ is a fifth order tensor, it is diagonal, and therefore very compact. However, the observable representation is:

$$
\mathcal{P}(X_1, X_2, X_3, X_4, X_5) = \mathcal{P}(X_1, X_2, X_3, X_4, X_5) \times X_1^T \times X_2^T \times X_3^T \times X_4^T \times X_5^T
$$

Now the fifth order tensor is fully dense, which is the central cause of the parameter explosion in the original spectral algorithm.

Let us consider an alternate factorization of the marginal probability tensor. Associate the conditional probability table of each child $X_i$ with the 3rd order labeled tensor $\mathcal{P}(X_i|\emptyset_2 H)$. Instead of directly multiplying each of these tensors with the root tensor that is a function of $H$, we will instead first agglomerate them:

$$
\mathcal{P}(X_1, X_2, X_3, X_4, X_5|\emptyset_2 H) = \mathcal{P}(X_1|\emptyset_2 H) \times_H \mathcal{P}(X_2|\emptyset_2 H) \times_H \mathcal{P}(X_3|\emptyset_2 H) \times_H \mathcal{P}(X_4|\emptyset_2 H) \times_H \mathcal{P}(X_5|\emptyset_2 H)
$$

(7.1)

The $\emptyset_2 H$ prevents $H$ from being marginalized out while the tensors are multiplied. We then let the root be associated with a vector $\mathcal{P}(H)$ and recover the marginal probability tensor as follows:

$$
\mathcal{P}(H) \times_H (\mathcal{P}(X_1, X_2, X_3, X_4, X_5|\emptyset_2 H) \times_H \mathbf{1})
$$

(7.2)

This gives us the following tensor representation for the graphical model in Figure 7.1

$$
\mathcal{P}(X_1, X_2, X_3, X_4, X_5) = \mathcal{P}(H) \times_H (\mathcal{P}(X_1|\emptyset_2 H) \times_H \mathcal{P}(X_2|\emptyset_2 H)
\times_H \mathcal{P}(X_3|\emptyset_2 H) \times_H \mathcal{P}(X_4|\emptyset_2 H)
\times_H \mathcal{P}(X_5|\emptyset_2 H) \times_H \mathbf{1})
$$

(7.3)

Note that the maximum tensor order of the right hand side is 3 since $\mathcal{P}(H)$ is a vector. Moreover, in the following sections we will demonstrate that it is possible to transform this representation so that it only depends on observed variables.
### 7.2 Notation

Before proceeding to present the spectral algorithm, we briefly discuss notation. A latent tree model defines a joint probability distribution over a set of \(O\) observed variables \(\mathcal{O} = \{X_1, \ldots, X_O\}\) and a set of \(H\) hidden variables \(\mathcal{H} = \{X_{O+1}, \ldots, X_{O+H}\}\). The complete set of variables is denoted by \(\mathcal{X} = \mathcal{O} \cup \mathcal{H}\). For simplicity, we assume that all hidden variables have \(S_H\) states and observed variables have \(S_O\) states. The joint distribution of \(\mathcal{X}\) in a latent tree model is fully characterized by the following equation:

\[
P(x_1, \ldots, x_{O+H}) = \prod_{i=1}^{O+H} P(x_i|x_{\pi(i)}). \tag{7.4}
\]

where \(X_{\pi(i)}\) denotes the parent of \(X_i\). For simplicity, assume that all leaves are observed variables and all internal nodes are latent. For further notation, let \(X_r\) be some observed node in the subtree rooted at \(X_i\). Let \(c_i(j)\) denote the \(j^{th}\) child of node \(X_i\), where \(X_i\) has \(\alpha_i\) children. For ease of exposition we will assume that all internal nodes have exactly 3 neighbors.

Furthermore, let \(X^*_i\) denote some observed node in the subtree rooted at \(X_i\) and \(T_i\) denote the set of all observed nodes in the subtree rooted at \(X_i\). Let \(\lambda(i)\) be the left sibling of \(X_i\) and \(\rho(i)\) be the right sibling of \(X_i\) where the order is cyclic. If \(X_i\) only has two children than it can pretend its "uncle" is its left sibling (e.g. set \(\lambda(i)\) to a sibling of \(X_{\pi(i)}\)). A summary of notation is provided in Table 7.2.

### 7.3 Derivation of Alternate Spectral Algorithm

Our spectral derivation has three main components. First we show how the marginal probability tensor \(P(X_1, \ldots, X_O)\) can be factorized into a collection of lower order tensors where the maximum

\[\text{if } X_i \text{ only has one sibling then set to sibling of } X_{\pi(i)}\]
The messages from the children gives the probability estimate: $P(\bar{e}, \bar{f}, \bar{g}, \bar{h}, \bar{i}, \bar{j}) = r^T (M_{c_1(r)} M_{c_2(r)} M_{c_3(r)} 1_r)$. Example: $P(\bar{e}, \bar{f}, \bar{g}, \bar{h}, \bar{i}, \bar{j}) = r^T M_B M_C M_D 1$. 

7.3.1 Factorizing the Marginal Probability Tensor

In this representation, we associate the root node $X_r$ with the marginal probability (labeled) vector $P(X_r)$ and non root nodes with third order labeled tensors $P(X_i| \varnothing X_{\pi(i)})$. Consider Figure 7.3. We have that

$$P(E, F, G, H, I, J) = P(A) \times_A (P(E, F| \varnothing 2 A) \times_A P(G, H| \varnothing 2 A) \times_A P(I, J| \varnothing 2 A) \times_A 1)$$

Decomposing recursively gives

$$P(E, F| \varnothing 2 A) = P(B| \varnothing 2 A) \times_B (P(E| \varnothing 2 B) \times_B P(F| \varnothing 2 B) \times_B 1)$$

$$P(G, H| \varnothing 2 C) = P(C| \varnothing 2 A) \times_C (P(G| \varnothing 2 C) \times_C P(H| \varnothing 2 C) \times_C 1)$$

$$P(I, J| \varnothing 2 D) = P(D| \varnothing 2 A) \times_D (P(I| \varnothing 2 D) \times_B P(J| \varnothing 2 D) \times_D 1)$$

(7.5)

7.3.2 Tensor Message Passing

As in Chapter 4, the recursive factorization procedure described above can be expressed as message passing, a form that will more be notationally convenient for our subsequent derivation. Instead of attempting to reconstruct the entire marginal probability tensor let us simply focus on a single element of this tensor (e.g. $P(\bar{e}, \bar{f}, \bar{g}, \bar{h}, \bar{i}, \bar{j})$). We will associate each leaf with the labeled tensor $L_i = P(X_i| \varnothing 2 X_{\pi(i)})$. Then, the message that the leaf passes to its parent is:

$$M_i = L_i \times_i \delta_{\bar{x}_i}$$

(7.6)

Example: $M_E = P(\bar{e}| \varnothing 2 B) = P(E| \varnothing 2 B) \times_E \delta_{\bar{e}}$

The internal node, associated with the labeled tensor $T_i = P(X_i| \varnothing 2 X_{\pi(i)})$ agglomerates the messages from its children and passes the resulting (diagonal) matrix to its parent (where $1_i$ is a $S_H$ dimensional vector of ones):

$$M_i = T_i \times_i \left( M_{c_1(i)} M_{c_2(i)} 1_i \right)$$

(7.7)

Example: $M_B = P(\bar{e}, \bar{f}| \varnothing 2 A) = P(B| \varnothing 2 A) \times_B (M_E M_F 1_B)$

Finally the root parameter is represented as a labeled vector: $r := P(X_r)$. Combining this with the messages from the children gives the probability estimate:

$$P(\bar{x}_1, ..., \bar{x}_O) = r^T \left( M_{c_1(r)} M_{c_2(r)} M_{c_3(r)} 1_r \right)$$

(7.8)

Example: $P(\bar{e}, \bar{f}, \bar{g}, \bar{h}, \bar{i}, \bar{j}) = r^T M_B M_C M_D 1$. 

101
Note that unlike the tensor representation in Chapter 4, the message now takes the form of a matrix \( M_i \) instead of a vector \( m_i \). Furthermore, the \( 1_i \) is needed unlike in the original tensor representation.

### 7.3.3 Transformed Representation

Next, as in Chapter 4, note we do not need to recover the tensor representation explicitly if our focus is to perform inference using the message passing algorithm as in (7.6)–(7.8). As long as we can recover the tensor representation up to some invertible transformation, we can still obtain the correct marginal probability \( P(\bar{x}_1, \ldots, \bar{x}_O) \).

More specifically, we can insert a identity matrix \( I \) into the message update equation in (7.8) without changing the final probability.

\[
P(\bar{e}, \bar{f}, \bar{g}, h, i, j) = r^\top M_B M_C M_D 1_A
\]

where \( I \) is the \( S_H \times S_H \) identity matrix.

Subsequently, we can then replace this matrix with a pair of matrices \( F \) and \( F^{-1} \), that are inverses of each other, and then regroup the terms:

\[
P(\bar{e}, \bar{f}, \bar{g}, h, i, j) = r^\top F_{A,E} \times \omega_A \left( (F_{A,B}^{-1} \times M_B \times F_{A,C}) \times (F_{A,D}^{-1} M_C F_{A,D}) \times (F_{A,B}^{-1} 1_A) \right) \quad (7.9)
\]

where \( \omega_A \) is a mode label that will be defined in the next section since it depends on the definition of \( F \). These process proceeds recursively e.g.

\[
M_B = \mathcal{P}(B | \emptyset A) \times_{B} F_{B,E} \times_{\omega_B} \left( (F_{B,E}^{-1} M_E F_{B,E}) \times (F_{B,E}^{-1} 1_B) \right)
\]

In general, we can define the following transformed tensor representation:

- **root**: \( F_{i \in \{\} }^T \)
- **internal**: \( \tilde{T}_i = T_i \times_i F_{i \in \{\} } \times_{\pi(i)} F_{\pi(i),i}^{-1} \times_{\pi(i),\rho(i)} F_{\pi(i),\rho(i)} \)
- **leaf**: \( \tilde{L}_i = L_i \times_{\pi(i)} F_{\pi(i),i}^{-1} \times_{\pi(i)} F_{\pi(i),\rho(i)} \)
- **one**: \( \tilde{1}_i = F_{i \in \{\} }^{-1} 1_i \)

### 7.4 Observable Representation

We now derive the observable representation by choosing \( F \) and \( F^{-1} \) systematically, so that we can recover each transformed parameter using the marginal probability of a small set of observed variables. This is summarized by the lemma below:

**Lemma 16.** Define \( F_{ij} = \mathcal{P}(X_j|X_i) \) with mode labels \( \{\omega_i := X_{j'}, X_i\} \).
• root: \( \hat{r} = \mathcal{P}(X_{c_1(o)}) \)

• internal: \( \tilde{T}_i = \mathcal{P}(X_{c_1(o)}', X_{\lambda(o)'}, X_{\rho(o)'}) \times_{\lambda(o)'} \mathcal{P}(X_r, X_{\lambda(o)'})^{-1} \)

• leaf: \( \tilde{L}_i = \mathcal{P}(X_i, X_{\lambda(o)'}, X_{\rho(o)'}) \times_{\lambda(o)'} \mathcal{P}(X_i, X_{\lambda(o)'})^{-1} \)

• one: \( \tilde{1}_i = \mathcal{P}(X_{c_1(o)}', X_{\lambda(c_1(o)')})^{-1} \mathcal{P}(X_{\lambda(c_1(o)')}) \)

**Proof.** The derivation for the root is straightforward:

\[ \hat{r} = \mathcal{P}(X_i) \times_i F_{i,c_1(o)} = \mathcal{P}(X_i) \times_i \mathcal{P}(X_{c_1(o)'}) = \mathcal{P}(X_{c_1(o)'}). \]

**Example (Figure 7.3):** \( X_{c_1(o)}' \in \{E, F\} \) so one possibility is \( \hat{r} = \mathcal{P}(E). \)

For the internal node, first consider the quantity, \( \mathcal{P}(X_r, X_{\lambda(o)'}) \). This can be expanded into,

\[ \mathcal{P}(X_r, X_{\lambda(o)'}) = \mathcal{P}(X_{c_1(o)'}) \mathcal{P}(X_{\lambda(o)'}) \mathcal{P}(X_{\lambda(o)'})^\top \]

(7.10)

Combining this with the definition of \( \tilde{T}_i \) gives

\[ \tilde{T}_i \times_{\lambda(o)'} \mathcal{P}(X_r, X_{\lambda(o)'}) = \tilde{T}_i \times_{\lambda(o)'} \left( \mathcal{P}(X_i | X_{\pi(o)}) \mathcal{P}(X_{\lambda(o)'}) \mathcal{P}(X_{\lambda(o)'})^\top \right) \]

\[ = \mathcal{P}(X_i) \mathcal{P}(X_{\lambda(o)'}) \times_{\lambda(o)'} \left( \mathcal{P}(X_{c_1(o)'}) \mathcal{P}(X_{\lambda(o)'}) \mathcal{P}(X_{\lambda(o)'})^\top \right) \]

\[ = \mathcal{P}(X_{c_1(o)'}, X_{\rho(o)'}, X_{\lambda(o)'}) \]

(7.11)

We can thus conclude that

\[ \tilde{T}_i = \mathcal{P}(X_{c_1(o)'}, X_{\lambda(o)'}, X_{\rho(o)'}) \times_{\lambda(o)'} \mathcal{P}(X_r, X_{\lambda(o)'})^{-1} \]

(7.12)

**Example (Figure 7.3):** For the internal node \( B \), we can set \( X_{c_1(o)'} = X_{c_1(o)'} \in \{E, F\} \), \( X_{\rho(o)'} = \{G, H\} \), \( X_{\lambda(o)'} = \{I, J\} \). This would give \( \tilde{T}_i = \mathcal{P}(E, G, I) \times_{\lambda(o)'} \mathcal{P}(E, I)^{-1}. \)

The leaf is just a special case of the internal node. For the \( \tilde{1}_i \) term, multiply by \( \mathcal{P}(X_{c_1(o)'}, X_{\lambda(c_1(o)')}) \):

\[ \tilde{1}_i^\top \mathcal{P}(X_{c_1(o)'}, X_{\lambda(c_1(o)')}) = \tilde{1}_i^\top \left( \mathcal{P}(X_{c_1(o)'}) \mathcal{P}(X_{\lambda(c_1(o)')}) \right) \]

\[ = \tilde{1}_i^\top \mathcal{P}(X_{c_1(o)'}) \mathcal{P}(X_{\lambda(c_1(o)')}) \]

\[ = \mathcal{P}(X_{\lambda(c_1(o)')})^{-1} \mathcal{P}(X_{\lambda(c_1(o)')}) \]

(7.13)

leading us to conclude that

\[ \tilde{1}_i = \mathcal{P}(X_{c_1(o)'}, X_{\lambda(c_1(o)')})^{-1} \mathcal{P}(X_{\lambda(c_1(o)')}) \]

(7.14)

**Example (Figure 7.3):** For \( \tilde{1}_A, X_{c_1(o)'} \in \{E, F\}, X_{\lambda(c_1(o)')}, \{I, J\} \), so \( \tilde{1}_i = \mathcal{P}(E, I)^{-1} \mathcal{P}(I). \)

akin to the algorithm in Chapter 4 if \( S_O > S_H \) then \( F_{i,j} = U_i^\top \mathcal{P}(X_i | X_i) \) instead of \( F_{i,j} = \mathcal{P}(X_i | X_i) \) where \( U_i \) is set to the top \( s_H \) left singular vectors of \( \mathcal{P}(X_r, X_{\lambda(o)'}) \). The algorithms for training and testing are given in Algorithms 9 and 10.
Algorithm 9 Alternate spectral learning algorithm for latent tree graphical model

In: Tree topology and N i.i.d. samples \(\{x_1, \ldots, x_N\}\)

Out: Estimated observable parameters \(\hat{r}\) for root, \(\hat{T}_i\) for each non-root internal node, \(\hat{L}_i\) for each leaf, \(\hat{I}_i\) for each internal node

1: For each node \(x_i\), perform a “thin” singular value decomposition of \(\hat{P}(X_i, X_{\lambda(i)}) = U \Sigma V^T\); let \(\hat{U}_i = U(:, 1 : S_H)\) be the first \(S_H\) principal left singular vectors.
2: Estimate \(\hat{r}\), and each \(\hat{T}_i, \hat{L}_i, \hat{I}_i\) via

\[
\hat{r} = \hat{U}_i \hat{P}(x_i) \\
\hat{T}_i = \hat{P}(x_i, X_{\lambda(i)} \lambda(i)) \hat{U}_i \hat{P}(X_i, X_{\lambda(i)})^T \hat{U}_i \hat{U}_i \hat{U}_i \hat{U}_i \hat{U}_i \hat{U}_i \hat{U}_i \\
\hat{L}_i = \hat{P}(x_i, X_{\lambda(i)}) \hat{U}_i \hat{P}(X_i, X_{\lambda(i)})^T \hat{U}_i \hat{U}_i \\
\hat{I}_i = (\hat{P}(x_i, X_{\lambda(i)}) \hat{U}_i \hat{P}(X_i, X_{\lambda(i)})^T \hat{U}_i)
\]

7.5 Sample Complexity

We analyze the sample complexity of Algorithm 9 and find that it depends on the tree topology and the spectral properties of the true model. See the Appendix for a proof.

**Theorem 3.** For any \(\varepsilon > 0, 0 < \delta < 1\), let

\[
N \geq \Omega\left(\frac{(\alpha_{\text{max}} S_H)^{2\ell+1}}{\gamma \rho^2} \right) \log \frac{1}{\delta}
\]

where \(\alpha_{\text{max}} = \max \alpha_i, \sigma_{S_H}(\cdot)\) returns the \(S_H\) largest singular value and

\[
\gamma = \min_{i \neq j, i, j \in \Theta} \sigma_{S_H}(P(X_i, X_j))^4 \\
\beta = \min_{i \in \Theta} \sigma_{S_H}(P(X_i, X_{\pi(i)}))^2
\]

Then \(\sum_{x_1, \ldots, x_N} |\hat{P}(x_1, \ldots, x_N) - P(x_1, \ldots, x_N)| \leq \varepsilon\) with probability \(1 - \delta\).

This result implies that the estimation problem gets harder as the maximum degree \(\alpha_{\text{max}}\) of the hidden nodes, the number \(S_H\) of the hidden states, and the length \(\ell\) of the chain of hidden variables

Algorithm 10 Inference with Alternate Spectral Parameters

In: Tree topology, set of spectral parameters \(\hat{r}, \hat{T}_i\) for each node, \(\hat{L}_i\) for each leaf, \(\hat{I}_i\) for each internal node, and set of evidence \(\mathcal{E} = \{\hat{x}_{\epsilon_1}, \ldots, \hat{x}_{\epsilon_K}\}\)

Out: Estimated probability \(P(\hat{x}_{\epsilon_1}, \ldots, \hat{x}_{\epsilon_K})\)

1: In reverse topological order, each node accumulates at message at leaf and sends to parent

- **Evidence Leaf:** \(M_i = \hat{L}_i \times \delta_i\)
- **Non-Evidence Leaf:** \(M_i = \hat{L}_i \times 1\)
- **Internal Node:** \(M_i = \hat{T}_i \times (\hat{M}_{i_{\lambda(i)}} \hat{M}_{i_{\lambda(i)}} \hat{I}_i)\)
- **Root:** \(\hat{P}(\hat{x}_{\epsilon_1}, \ldots, \hat{x}_{\epsilon_K}) = \hat{r} (\hat{M}_{i_{\lambda(1)}} \hat{M}_{i_{\lambda(1)}} \hat{I}_0)\)
7.6 Experiments

In this section, we empirically evaluate this new algorithm. We first empirically compare to the traditional spectral algorithm described in Chapter 4. For both methods we use a linear system size of 1. Since both methods have similar runtime, we only focus on accuracy, particularly as the topology of the tree and nature of the parameters are changed. We measure the performance of joint estimation using $\epsilon = \frac{|\hat{P}(x_1,...,x_O) - P(x_1,...,x_O)|}{P(x_1,...,x_O)}$ averaged over 1000 test points randomly drawn from the underlying model. Spec-Orig or original spectral algorithm denotes the spectral method from Chapter 4 while Spec-Alt or alternate spectral algorithm denotes the spectral method presented in this chapter.

The first set of experiments (Figure 7.4) shows the performance of the methods as the topology of the tree is varied. $S_O$ and $S_H$ are set to 6 and 2 respectively. In all cases the tree is balanced and NumChildren refers to the number of children each internal node has, and Depth indicates the depth of the tree. As one can see for lower sample sizes, the original spectral algorithm is more robust, while as the sample size increases, the alternate algorithm consistently performs better.

The next two sets of experiments measure performance as positive values are added to the diagonals of the conditional probability tables (before normalization) thus increasing the singular values of the resulting conditional probability tables after normalization. For these experiments $S_O$ is set to 4 and $S_H$ is set to 3. Figure 7.5 shows the effect of this for a tree with NumChildren set to 3 (every internal node has 3 children) and depth also fixed to 3 while Figure 7.6 shows the effect of increasing diagonals for NumChildren=4 and Depth=3 (every internal node has 4 children). In both cases, it appears both methods benefit similarly from the increased singular values.
7.6.1 Comparison with Mossel and Roch Algorithm

We now make a separate comparison with the spectral algorithm by [Mossel and Roch, 2006], since it only applies to case where the number of observed states $S_O$ is the same as the number of hidden states $S_H$. We set $S_O = S_H = 2$ and use a variety of topologies shown in Figure 7.7. Although this method is theoretically interesting, it can perform poorly in practice.

The results are shown in Figure 7.7 (the runtime of both methods are similar, and thus not
reported). Our spectral algorithm significantly outperforms the MR algorithm on all trees for practically all sample sizes. This is because our method does not explicitly recover the CPTs, and is thus more robust. We also note that our approach is more general: it can allow for the observation state space to be larger than the hidden state space, which may be preferable in many applications where the observation space can be large (e.g., quantization of a continuous variable), but the hidden factors are simple and have lower dimensions.

7.6.2 Stock Trend Prediction

Finally, we evaluate our algorithm on a stock trend prediction problem. Our goal is to predict whether a stock $X_i$ will go up or down on a particular day given the trends of a set $E$ of other stocks. We acquired closing prices of 59 stocks from 1984 to 2011, which provides us 6800 samples. We randomly partition these samples to 6300 training points and 500 test points. Since we are only predicting whether a stock goes up or down, the data are binarized. From the training data, we learn the latent tree topology using an algorithm by (Choi et al., 2010), and a fully observable Chow-Liu tree (Chow and Liu, 1968b).

We compare our spectral algorithm to $EM^+$ (high precision EM) and $EM^-$ (low precision EM) using the latent tree, and with inference over the Chow-Liu tree (CL). For the prediction task, we need to estimate the conditional, i.e., $P(X_i|x_{j_1}, \ldots, x_{j_{|E|}})$ and $j_1, \ldots, j_{|E|} \in E$. This can be achieved by estimating $P(x_i, x_{j_1}, \ldots, x_{j_{|E|}})$ for each instantiation $x_i$. Then we make prediction by $\hat{x}_i = \text{argmax}_{x_i} P(x_i, x_{j_1}, \ldots, x_{j_{|E|}})$. We measure the prediction error using $\epsilon = |\hat{x}_i - x_i^*|$ where $x_i^*$ is the true label.

We experiment with a varying number of query sizes. For each query size $Q$, we randomly pick $Q$ stocks and predict the value of one stock conditioned on the other $Q-1$, (and repeat for 50 trials). The results are shown in Figure 7.8. In general, all methods get better as we increase the number of evidence. Over the entire range of query sizes, the advantage of latent tree approaches (Spectral, $EM^+/-$) is clear over the Chow-Liu tree. Thus, the latent factors help better model the stock data in this case. Due to the small training sample size, the distinction between our method

---

2www.finance.yahoo.com
and EM is less clear.

7.7 Discussion

In this chapter an alternate observable representation for latent tree graphical models. This representation is considerably different than the one in Chapter 4 and doesn't reduce to it even when the tree is binary. The distinct characteristic of this representation is that the internal/leaf nodes are associated with a 3rd order tensor where the parent variable is on the diagonal. This feature makes it challenging to generalize this factorization to the kernel case where the analogous operator is challenging to define. Similarly, since different children of a clique in a junction tree may be associated with different variables in the parent clique, it also makes it difficult to generalize this approach to the junction tree scenario.

Empirically we show that while this new representation performs slightly worse than the representation in Chapter 4 for small sample sizes, it does considerably better in mid/large sample sizes, especially for trees with larger degree.

7.8 Appendix

7.8.1 Notation for Proof of Theorem 1

We now proceed to prove Theorem 1. \(|\|\|_2\) refers to spectral norm for matrices and tensors (but normal euclidean norm for vectors). \(|\|\|_1\) refers to induced 1 norm for matrices and tensors (max column sum), (but normal l1 norm for vectors). \(|\|\|_F\) refers to Frobenius norm.

The tensor spectral norm (for 3 dimensions) is defined in [Nguyen et al., 2010]:

\[
\|\mathcal{T}\|_2 = \sup_{\|v\|_2 \leq 1} \mathcal{T} \times_3 v_3 \times_2 v_2 \times_1 v_1 \quad (7.19)
\]

We will define the induced 1-norm of a tensor as

\[
\|\mathcal{T}\|_{1,1} = \sup_{\|v\|_1 \leq 1} \|\mathcal{T} \times_1 v\|_1 \quad (7.20)
\]

using the \(\ell_1\) norm of a matrix (i.e., \(\|A\|_1 = \sup_{\|v\|_1 \leq 1} \|Av\|_1\)).

For more information about matrix norms see [Horn and Johnson, 1990].

In general, we suppress the actual subscripts/superscripts on \(U\) and \(O\). It is implied that \(U\) and \(O\) can often be different depending on the transform being considered. However, this makes the notation very messy. It will generally be clear from context which \(U\) and \(O\) are being referred to. When it is not we will arbitrarily index them 1, 2, ..., so that it is clear which corresponds to which.
In general, for simplicity of exposition, we assume that all internal nodes in the tree are unobserved, and all leaves are observed (since this is the hardest case).

The proof generally follows the technique of HKZ \cite{Hsu2009}, but has key differences due to the tree topology instead of the HMM.

We define $\tilde{M}_i = (O^\top \tilde{U})^{-1} M_i (O^\top \tilde{U})$. Then as long as $(O^\top \tilde{U})$ is invertible, $(O^\top \tilde{U})^{-1} M_i (O^\top \tilde{U}) = M_i$. (We admit this is a slight abuse of notation, since $\tilde{M}_i$ was previously defined to be $(U^\top O)^{-1} M_i (U^\top O)$, but as long as $(O^\top \tilde{U})$ is invertible it doesn’t really matter whether it equals $(U^\top O)$ or not for the purposes of this proof). The other quantities are defined similarly.

We seek to prove the following theorem:

**Theorem 4.** Pick any $\epsilon > 0, \delta < 1$. Let

$$N \geq O \left( \frac{1}{\epsilon^2} \left( \frac{(d_{\text{max}} S_H)^{2\ell+1} S_O}{\min_i \sigma_{S_H}(O_i)^2 \min_{i \neq j} \sigma_{S_H}(P_{ij})^4} \right) \right) \log \frac{|S|}{\delta}$$  \hspace{1cm} (7.21)

Then with probability $1 - \delta$

$$\sum_{x_1, \ldots, x_O} \left| \tilde{P}[x_1, \ldots, x_O] - P[x_1, \ldots, x_O] \right| \leq \epsilon$$  \hspace{1cm} (7.22)

In many cases, if the frequency of the observation symbols follow certain distributions, than the dependence on $S_O$ can be removed as showed in HKZ \cite{Hsu2009}. That observation can easily be incorporated into our theorem if desired.

### 7.8.2 Concentration Bounds

\[
\begin{align*}
\epsilon_i &= \left\| \tilde{P}_i - P_i \right\|_F \quad (7.23) \\
\epsilon_{i,j} &= \left\| \tilde{P}_{i,j} - P_{i,j} \right\|_F \quad (7.24) \\
\epsilon_{x,i,j} &= \left\| \tilde{P}_{x,i,j} - P_{x,i,j} \right\|_F \quad (7.25) \\
\epsilon_{i,j,k} &= \left\| \tilde{P}_{i,j,k} - P_{i,j,k} \right\|_F \quad (7.26)
\end{align*}
\]

($x$ denotes a fixed element while $i, j, k$ are over indices).

As the number of samples $N$ gets large, we expect these quantities to be small.

**Lemma 17** (variant of HKZ \cite{Hsu2009}). *If the algorithm independently samples $N$ observation*
triples from the tree, then with probability at least $1 - \delta$. 

\begin{align*}
\epsilon_i & \leq \sqrt{\frac{C}{N} \ln \frac{|\theta|}{\delta}} + \sqrt{\frac{1}{N}} \quad (7.27) \\
\epsilon_{i,j} & \leq \sqrt{\frac{C}{N} \ln \frac{|\theta|}{\delta}} + \sqrt{\frac{1}{N}} \quad (7.28) \\
\epsilon_{i,j,k} & \leq \sqrt{\frac{C}{N} \ln \frac{|\theta|}{\delta}} + \sqrt{\frac{1}{N}} \quad (7.29) \\
\max_x \epsilon_{i,x,j} & \leq \sqrt{\frac{C}{N} \ln \frac{|\theta|}{\delta}} + \sqrt{\frac{1}{N}} \quad (7.30) \\
\max_x \epsilon_{x,i,j} & \leq \sqrt{\frac{C}{N} \ln \frac{|\theta|}{\delta}} + \sqrt{\frac{1}{N}} \quad (7.31)
\end{align*}

where $C$ is some constant (from the union bound over $O(V^3)$). ($V$ is the total number of observed variables in the tree). The proof is the same as that of HKZ (Hsu et al., 2009) except the union bound is larger. The last bound can be made tighter, identical to HKZ, but for simplicity we do not pursue that approach here.

### 7.9 Eigenvalue Bounds

Basically this is Lemma 9 in HKZ (Hsu et al., 2009), which is stated below for completeness:

**Lemma 18.** Suppose $\epsilon_{i,j} \leq \epsilon \times \sigma_{Sh}(P_{i,j})$ for some $\epsilon < 1/2$. Let $\epsilon_0 = \epsilon_{i,j}^2 / ((1 - \epsilon)\sigma_{Sh}(P_{i,j}))^2$. Then:

1. $\epsilon_0 < 1$
2. $\sigma_{Sh}(\tilde{U}^T \tilde{P}_{i,j}) \geq (1 - \epsilon)\sigma_{Sh}(P_{i,j})$
3. $\sigma_{Sh}(\tilde{U}^T P_{i,j}) \geq \sqrt{1 - \epsilon_0}\sigma_{Sh}(P_{i,j})$
4. $\sigma_{Sh}(O^T \tilde{U}) \geq \sqrt{1 - \epsilon}\sigma_{Sh}(O)$

The proof is in HKZ (Hsu et al., 2009).

#### 7.9.1 Bounding the Transformed Quantities

If Lemma 2 holds then $(O^T \tilde{U})$ is invertible. Thus, if we define $\tilde{M}_i = (O^T \tilde{U})^{-1}M_i(O^T \tilde{U})$. Then clearly, $(\tilde{U}^T O)^{-1}\tilde{M}_i(O^T \tilde{U}) = M_i$. (We admit this is a slight abuse of notation, since $\tilde{M}_i$ is previously defined to be $(U^T O)^{-1}M_i(U^T O)$, but as long as $(O^T \tilde{U})$ is invertible it doesn’t really matter whether it equals $(U^T O)$ or not for the purposes of this proof). The other quantities are defined similarly.
We seek to bound the following four quantities:

\[
\begin{align*}
\delta_{\text{one}} &= \| (O_1^T \hat{U})(\hat{u}_i - \hat{1}) \|_1 \\
\gamma_i &= \| (\hat{r}_i - \tilde{r}_i) \times_1 (O_1^T \hat{U})^{-1} \times_2 (O_2^T \hat{U}) \times_3 (O_3^T \hat{U})^{-1} \|_2 \\
\delta_{\text{root}} &= \| (\hat{r} - \tilde{r}) \times_1 (O_{j_1}^T \hat{U})^{-1} \|_\infty \\
\Delta_i &= \sum_{x_i} \| (O_1^T \hat{U})(\hat{M}_i - \tilde{M}_i)(O_2^T \hat{U})^{-1} \|_1
\end{align*}
\] (7.32, 7.33, 7.34, 7.35)

Here \(x_i\) denotes all observations that are in the subtree of node \(i\) (since \(i\) may be hidden or observed). Sometimes we like to distinguish between when \(i\) is observed and when \(i\) is hidden. Thus, we sometimes refer to the quantity \(\Delta_i^{\text{obs}}\) and \(\Delta_i^{\text{hidden}}\) for when \(i\) is observed or hidden respectively.

Again note that the numbering in \((O_1^T \hat{U})\) and \((O_2^T \hat{U})\) is just there to avoid confusion in the same equation (In reality there are many \(U\)’s and \(O\)’s).

**Lemma 19.** Assume \(e_{i,j} \leq \sigma_{SH}(P_{i,j})/3\) for all \(i \neq j\). Then

\[
\begin{align*}
\delta_{\text{root}} &\leq \frac{2 \epsilon_r}{\sqrt{3}\sigma_{SH}(O_{j_1}^T)} \\
\delta_{\text{one}}^{i} &\leq 4 \sqrt{SH} \left( \frac{\epsilon_{i,j}}{\sigma_{SH}(P_{i,j})^2} + \frac{\epsilon_i}{\sqrt{3}\sigma_{SH}(P_{i,j})} \right) \\
\gamma_i &\leq 4 \sqrt{SH} \left( \frac{\epsilon_{m,j}}{\sigma_{SH}(O)} + \frac{\epsilon_{m,jk}}{\sqrt{3}\sigma_{SH}(P_{i,j})} \right) \\
\Delta_i^{\text{hidden}} &\leq \left( 1 + \gamma_i \right) \prod_{k=1}^j (1 + \Delta_{j_k}) \delta_{\text{one}}^{i} + (1 + \gamma_i) \prod_{k=1}^j (1 + \Delta_{j_k}) - m \\
\Delta_i^{\text{obs}} &\leq 4 \sqrt{SH} \left( \frac{\epsilon_{i,j}}{\sigma_{SH}(O)^2} + \frac{\sum_x \epsilon_{m,x,i}}{\sqrt{3}\sigma_{SH}(P_{i,j})} \right)
\end{align*}
\] (7.36, 7.37, 7.38, 7.39, 7.40)

The main challenge in this part is \(\Delta_v\) and \(\gamma_v^{\text{hidden}}\). The rest are similar to HKZ. However, we go through the other bounds to be more explicit about some of the properties used, since sometimes we have used different norms etc.

\(\delta_{\text{root}}\)

We note that \(\hat{r} = P_{j_1}^T \hat{U}\) and similarly \(\tilde{r} = P_{j_1}^T \tilde{U}\).  

111
\[
\delta_{\text{root}} = \| (\hat{r} - r)^T (O_{j_1}^T \hat{U}_{j_1})^{-1} \|_\infty \leq \| P_{j_1}^T - P_{j_1}' \|_2 \| \hat{U}_{j_1} \|_2 \| (O_{j_1}^T \hat{U}_{j_1})^{-1} \|_2 
\leq \| P_{j_1}^T - P_{j_1}' \|_2 \| (O_{j_1}^T \hat{U}_{j_1})^{-1} \|_2 \leq \frac{\epsilon_r}{\sigma_{\text{S}_H}(O_{j_1}^T \hat{U}_{j_1})} 
\] (7.41)

The first inequality follows from the relationship between \(\ell_\infty\) and \(\ell_2\) norm and submultiplicativity. The second follows from a matrix perturbation bound given in Lemma 10.6. We also use the fact that since \(\hat{U}\) is orthonormal it has spectral norm 1.

Assuming that \(\epsilon_{i,j} \leq \sigma_{\text{S}_H}(P_{i,j})/3\) gives \(\delta_{\text{root}} \leq \frac{2\epsilon_r}{\sqrt{3}\sigma_{\text{S}_H}(O_{j_1}^T \hat{U}_{j_1})}\) by Lemma 18.

\[
\delta_{\text{one}}^i = \| (O^T \hat{U})(\hat{1}_i - \tilde{1}_i) \|_1 \leq \sqrt{S_H} \|O\|_2 \| \hat{U} \|_2 \| \hat{1}_i - \tilde{1}_i \|_2 
\leq \sqrt{S_H} \| \hat{1}_i - \tilde{1}_i \|_2 = \sqrt{S_H} \| \hat{1}_i - \tilde{1}_i \|_2 
\] (7.43)

Here we have converted \(\ell_1\) norm to \(\ell_2\) norm, used submultiplicativity, the fact that \(\hat{U}\) is orthonormal so has spectral norm 1, and that \(O\) is a conditional probability matrix and therefore also has spectral norm 1.

We note that \(\tilde{1}_i = \hat{P}_{i,j}^T \hat{U}^T + \hat{P}_j\) and similarly \(\hat{1}_i = (P_{i,j}^T \hat{U}^T + P_j\) where \(i\) and \(j\) are a particular pair of observations described in the main paper.

\[
\| \tilde{1}_i - \hat{1}_i \|_2 = \| (\hat{P}_{m,j}^T \hat{U})^+ \hat{P}_j - (P_{m,j}^T \hat{U})^+ P_j \|_2 \leq \| (\hat{P}_{m,j}^T \hat{U})^+ \hat{P}_j - (P_{m,j}^T \hat{U})^+ \hat{P}_j \|_2 + \| (P_{m,j}^T \hat{U})^+ \hat{P}_j - (P_{m,j}^T \hat{U})^+ P_j \|_2 
\leq \| (\hat{P}_{m,j}^T \hat{U})^+ \hat{P}_j - (P_{m,j}^T \hat{U})^+ \hat{P}_j \|_2 + \| (\hat{P}_{m,j}^T \hat{U})^+ \hat{P}_j - (P_{m,j}^T \hat{U})^+ \hat{P}_j \|_2 
\leq \| (\hat{P}_{m,j}^T \hat{U})^+ - (P_{m,j}^T \hat{U})^+ \|_2 \| \hat{P}_j \|_1 + \| (\hat{P}_{m,j}^T \hat{U})^+ - (P_{m,j}^T \hat{U})^+ \|_2 \| \hat{P}_j - P_j \|_2 
\leq \frac{1 + \sqrt{5}}{2} \frac{\epsilon_{m,j}}{\min(\sigma_{\text{S}_H}(\hat{P}_{m,j}^T \hat{U})), \sigma_{\text{S}_H}(P_{m,j}^T \hat{U}))^2} + \frac{\epsilon_j}{\sigma_{\text{S}_H}(P_{m,j}^T \hat{U}))} 
\] (7.49)

where we have used the triangle inequality in the first inequality and the submultiplicative property of matrix norms in the second. The last inequality follows by matrix perturbation bounds. Thus using the assumption that \(\epsilon_{i,j} \leq \sigma_{\text{S}_H}(P_{i,j})/3\), we get that

\[
\delta_{\text{one}} \leq 4 \sqrt{S_H} \left( \frac{\epsilon_{i,j}}{\sigma_{\text{S}_H}(P_{i,j})^2} + \frac{\epsilon_j}{\sqrt{3}\sigma_{\text{S}_H}(P_{i,j})} \right) 
\] (7.50)

112
Tensor

Recall that \( \tilde{T}_i = T_i \times_1 (O_1^T \hat{U}_1) \times_2 (\hat{U}_2^T O_2) \times_3 (O_3^T \hat{U}_3) = P_{m,j,k} \times_1 \hat{U}_1^T \times_2 (P_{l,j} \hat{U}_2)^+ \times_3 \hat{U}_3^T \). Similarly, \( \tilde{T}_i = P_{m,j,k} \times_1 \hat{U}_1^T \times_2 (P_{l,j} \hat{U}_2)^+ \times_3 \hat{U}_3^T \).

\[
\| (\tilde{T}_i - \tilde{T}_i) \times_1 (O_1^T \hat{U}_1) \times_2 (O_2^T \hat{U}_2) \times_3 (O_3^T \hat{U}_3) \|_{1,1} \leq \frac{\sqrt{S_H}}{\sigma_{SH}(O)} \| \tilde{T}_i - \tilde{T}_i \|_2 \tag{7.51}
\]

This is because both \( \hat{U} \) and \( O \) have spectral norm one and the \( \sqrt{S_H} \) factor is the cost of converting from 1 norm to spectral norm.

\[
\| \tilde{T}_i - \tilde{T}_i \|_2 = \| P_{m,j,k} \times_1 \hat{U}_1^T \times_2 (P_{l,j} \hat{U}_2)^+ \times_3 \hat{U}_3^T - P_{m,j,k} \times_1 \hat{U}_1^T \times_2 (P_{l,j} \hat{U}_2)^+ \times_3 \hat{U}_3^T \|_2 \tag{7.52}
\]

\[
= \| P_{m,j,k} \times_1 \hat{U}_1^T \times_2 (P_{l,j} \hat{U}_2)^+ \times_3 \hat{U}_3^T - P_{m,j,k} \times_1 \hat{U}_1^T \times_2 (P_{l,j} \hat{U}_2)^+ \times_3 \hat{U}_3^T \|_2 \tag{7.53}
\]

\[
+ \| P_{m,j,k} \times_1 \hat{U}_1^T \times_2 ((P_{l,j} \hat{U}_2)^+ - (P_{l,j} \hat{U}_2)^+) \times_3 \hat{U}_3^T \|_2 \tag{7.54}
\]

\[
= \| P_{m,j,k} \times_1 \hat{U}_1^T \times_2 ((P_{l,j} \hat{U}_2)^+ - (P_{l,j} \hat{U}_2)^+) \times_3 \hat{U}_3^T \|_2 \tag{7.55}
\]

\[
+ \| (P_{l,j,k} \times_1 \hat{U}_1^T \times_3 \hat{U}_3^T - P_{l,j,k} \times_1 \hat{U}_1^T \times_3 \hat{U}_3^T) \times_2 (P_{l,j} \hat{U}_2)^+ \|_2 \tag{7.56}
\]

\[
= \| P_{m,j,k} \|_2 \frac{1 + \sqrt{5}}{2} \frac{\epsilon_{i,j}}{\min(\sigma_{SH}(P_{l,j}), \sigma_{SH}(P_{l,j} \hat{U}))^2} + \frac{\epsilon_{m,j,k}}{\sigma_{SH}(P_{l,j} \hat{U})} \tag{7.57}
\]

It is clear that \( \| P_{m,j,k} \|_2 \leq \| P_{m,j,k} \|_F \leq 1 \).

Using the fact that \( \epsilon_{i,j} \leq \sigma_{SH}(P_{l,j})/3 \) gives us the following bound:

\[
\gamma \leq 4 \frac{\sqrt{S_H}}{\sigma_{SH}(O)} \left( \frac{\epsilon_{i,j}}{\sigma_{SH}(P_{l,j})} + \frac{\epsilon_{i,j,k}}{\sqrt{5} \sigma_{SH}(P_{l,j})} \right) \tag{7.58}
\]

**Bounding \( \Delta_i \)**

We now seek to bound \( \Delta_i = \sum_x \| (O_1^T \hat{U}_1)(\tilde{M}_i - \hat{M}_i)(\hat{U}_2^T O_2)^{-1} \|_1 \). There are two cases: either \( i \) is a leaf or it is not.

**\( i \) is leaf node** In this case our proof simply follows from HKZ \(^{(5)}\) and is repeated here for convenience.
\[
\| (O_1^T \tilde{U}_1)(\tilde{M} - \tilde{M})(O_2^T \tilde{U}_2)^{-1} \|_1 \leq \sqrt{S_H} \| O_1 \|_1 \| (\tilde{M}_1 - \tilde{M}_1)(O_2^T \tilde{U}_2)^{-1} \|_2
\]
\[
\leq \sqrt{S_H} \| (\tilde{M}_1 - \tilde{M}_1) \|_2
\]

(7.59)

(7.60)

Note that \( \tilde{M}_1 = (P_{j_i} \tilde{U}_1)^{-1} P_{m,x_{i,j}} \tilde{U}_2 \) and \( \tilde{M}_1 = (P_{j_i} \tilde{U}_1)^{-1} P_{m,x_{i,j}} \tilde{U}_2 \).

\[
\| \tilde{M}_1 - \tilde{M}_1 \|_2 = \| (P_{j_i} \tilde{U}_1)^{-1} P_{m,x_{i,j}} \tilde{U}_2 - (P_{j_i} \tilde{U}_1)^{-1} P_{m,x_{i,j}} \tilde{U}_2 \|_2
\]

(7.61)

\[
= \| (P_{j_i} \tilde{U}_1)^{-1} P_{m,x_{i,j}} \tilde{U}_2 - (P_{j_i} \tilde{U}_1)^{-1} P_{m,x_{i,j}} \tilde{U}_2 + (P_{j_i} \tilde{U}_1)^{-1} P_{m,x_{i,j}} \tilde{U}_2 - (P_{j_i} \tilde{U}_1)^{-1} P_{m,x_{i,j}} \tilde{U}_2 \|_2
\]

\[
\leq \| (P_{j_i} \tilde{U}_1)^{-1} - (P_{j_i} \tilde{U}_1)^{-1} P_{m,x_{i,j}} \tilde{U}_2 \|_2 + \| (P_{j_i} \tilde{U}_1)^{-1} (P_{m,x_{i,j}} \tilde{U}_2 - P_{m,x_{i,j}} \tilde{U}_2) \|_2
\]

(7.62)

\[
\leq \| P_{m,x_{i,j}} \|_2 \frac{1 + \sqrt{5}}{2} \frac{\epsilon_{j_i}}{\min (\sigma_{S_H}(P_{j_i}), \sigma_{S_H}(P_{j_i}))} + \frac{\epsilon_{m,x_{i,j}}}{\sigma_{S_H}(P_{j_i})} \frac{1 + \sqrt{5}}{2} \frac{\epsilon_{j_i}}{\min (\sigma_{S_H}(P_{j_i}), \sigma_{S_H}(P_{j_i}))^2} + \frac{\epsilon_{m,x_{i,j}}}{\sigma_{S_H}(P_{j_i})}. \]

(7.63)

(7.64)

(7.65)

where the first inequality follows from the triangle inequality, and the second uses matrix perturbation bounds (and the fact that spectral norm of \( \tilde{U}_1 \) is 1).

The final inequality follows from the fact that spectral norm is less than frobenius norm which is less than \( \| O \|_1 \) norm:

\[
\| P_{m,x_{i,j}} \|_2 \leq \sqrt{\sum_{m,j} [P_{m,x_{i,j}}]_m^2} \leq \sum_{m,j} [P_{m,x_{i,j}}]_{m,j} \leq P[x_i = x]
\]

(7.66)

The first inequality follows from relation between 1 operator norm and 2 operator norm. Because \( O \) is a conditional probability matrix \( \| O \|_1 = 1 \) (i.e. the max column sum is 1).

Using the fact that \( \epsilon_{j_i} \leq \sigma_{S_H}(P_{j_i})/3 \) gives us the following bound:

\[
\Delta_{i,x} \leq 4 \frac{\sqrt{S_H}}{\sigma_{S_H}(O)} \left( P[x_i = x] \frac{\epsilon_{m,x_{i,j}}}{(\sigma_{S_H}(P_{j_i}))^2} + \frac{\epsilon_{m,x_{i,j}}}{\sqrt{3}\sigma_{S_H}(P_{j_i})} \right)
\]

(7.67)

Summing over \( v \) would give

\[
\Delta_i \leq 4 \frac{\sqrt{S_H}}{\sigma_{S_H}(O)} \left( \frac{\epsilon_{j_i}}{(\sigma_{S_H}(P_{j_i}))^2} + \frac{\sum x_i \epsilon_{m,x_{i,j}}}{\sqrt{3}\sigma_{S_H}(P_{j_i})} \right)
\]

(7.68)
$i$ is not a leaf node Let $\tilde{m}_{j;1} = \tilde{M}_j \ldots \tilde{M}_1\tilde{1}_i$ and $\hat{m}_{j;1} = \hat{M}_j \ldots \hat{M}_1\hat{1}_i$

\[
\sum_{x_i} \left\| (O_2^T \tilde{U}_2)(\tilde{M}_i - \hat{M}_i)(O_3^T \hat{U}_3)^{-1} \right\|_1 
= \sum_{x_i} \left\| (O_2^T \tilde{U}_2)(\tilde{T}_i \times_1 \tilde{M}_u \ldots \tilde{M}_1\tilde{1}_i - \tilde{T}_i \times_1 \hat{M}_u \ldots \hat{M}_1\hat{1}_i)(O_3^T \hat{U}_3)^{-1} \right\|_1 
= \sum_{x_i} \left\| (O_2^T \tilde{U}_2)(\tilde{T}_i \times_1 \hat{m}_{j;1} - \tilde{T}_i \times_1 \hat{m}_{j;1})(O_3^T \hat{U}_3)^{-1} \right\|_1 
= \sum_{x_i} \left\| (O_1^T \tilde{U}_1)(\tilde{T}_i - \tilde{T}_i) \times_1 (O_1^T \tilde{U}_1)^{-1} \times_2 (\tilde{U}_2^T O_2) \times_3 (O_3^T \hat{U}_3)^{-1} \right\|_{1,1} \left\| (O_1^T \tilde{U}_1)\hat{m}_{j;1} \right\|_1 
\leq \sum_{x_i} \left\| (\tilde{T}_i - \tilde{T}_i) \times_1 (O_1^T \tilde{U}_1)^{-1} \times_2 (\tilde{U}_2^T O_2) \times_3 (O_3^T \hat{U}_3)^{-1} \right\|_{1,1} \left\| (O_1^T \tilde{U}_1)\hat{m}_{j;1} \right\|_1 
+ \sum_{x_i} \left\| (O_1^T \tilde{U}_1)(\hat{m}_{j;1} - \hat{m}_{j;1}) \right\|_1 \left\| (\tilde{T}_i - \tilde{T}_i) \times_1 (O_1^T \tilde{U}_1)^{-1} \times_2 (\tilde{U}_2^T O_2) \times_3 (O_3^T \hat{U}_3)^{-1} \right\|_{1,1} 
+ \sum_{x_i} \left\| \tilde{T}_i \times_1 (O_1^T \tilde{U}_1)^{-1} \times_2 (\tilde{U}_2^T O_2) \times_3 (O_3^T \hat{U}_3)^{-1} \right\|_{1,1} \left\| (O_1^T \tilde{U}_1)(\hat{m}_{j;1} - \hat{m}_{j;1}) \right\|_1 

\text{First term is bounded by:} 
\left\| (\tilde{T}_i - \tilde{T}_i) \times_1 (O_1^T \tilde{U}_1)^{-1} \times_2 (\tilde{U}_2^T O_2) \times_3 (O_3^T \hat{U}_3)^{-1} \right\|_{1,1} \leq S_H \leq S_H \gamma_i 

\text{Second term is bounded by:} 
\sum_{x_i} \left\| (O_1^T \tilde{U}_1)(\hat{m}_{j;1} - \hat{m}_{j;1}) \right\|_1 \left\| (\tilde{T}_i - \tilde{T}_i) \times_1 (O_1^T \tilde{U}_1)^{-1} \times_2 (\tilde{U}_2^T O_2) \times_3 (O_3^T \hat{U}_3)^{-1} \right\|_{1,1} 
\leq \gamma_i \sum_{x_i} \left\| (O_1^T \tilde{U}_1)(\hat{m}_{j;1} - \hat{m}_{j;1}) \right\|_1 

\text{Third term is bounded by:} 
\left\| \tilde{T}_i \times_1 (O_1^T \tilde{U}_1)^{-1} \times_2 (\tilde{U}_2^T O_2) \times_3 (O_3^T \hat{U}_3)^{-1} \right\|_{1,1} \sum_{x_i} \left\| (O_1^T \tilde{U}_1)(\hat{m}_{j;1} - \hat{m}_{j;1}) \right\|_1 \leq \sum_{x_i} \left\| (O_1^T \tilde{U}_1)(\hat{m}_{j;1} - \hat{m}_{j;1}) \right\|_1 

\text{In the next section, we will see that} \sum_{x_i} \left\| (O_1^T \tilde{U})(\hat{m}_{j;1} - \hat{m}_{j;1}) \right\|_1 \leq \left( \prod_{k=1}^{j} (1 + \Delta_{jk})^{\delta_{\text{one}}} + S_H \prod_{k=1}^{j} (1 + \Delta_{jk}) - S_H \right). 

\text{So the overall bound is} 
\Delta_i \leq \left( (1 + \gamma_i) \prod_{k=1}^{j} (1 + \Delta_{jk})^{\delta_{\text{one}}} + (1 + \gamma_i)S_H \prod_{k=1}^{j} (1 + \Delta_{jk}) - S_H \right). 

(\text{where} j_1, \ldots, j_1 \text{are children of node} i)
Bounding $\sum_{x_i} \left\| (O^\top \hat{U})(\hat{m}_{j:1} - \hat{m}_{j:1}) \right\|_1$

Lemma 20.

$$\sum_{x_i} \left\| (O^\top \hat{U})(\hat{m}_{j:1} - \hat{m}_{j:1}) \right\|_1 \leq \prod_{k=1}^{u} \left( 1 + \Delta_{j_k} \right) \delta_{\text{one}}^i + S_H \prod_{k=1}^{u} \left( 1 + \Delta_{j_k} \right) - S_H \quad (7.81)$$

(where $j_1, ..., j_u$ are children of node $i$).

The proof is by induction. Base case: $\left\| (O^\top \hat{U})(\hat{1}_i - \hat{1}_i) \right\|_1 \leq \delta_{\text{one}}^i$, by definition of $\delta_{\text{one}}^i$.

Inductive step: Let us say claim holds up until $u - 1$. We show it holds for $u$. Thus

$$\sum_{x_i} \left\| (O^\top \hat{U})(\hat{m}_{u-1:1} - \hat{m}_{u-1:1}) \right\|_1 \leq \prod_{k=1}^{u-1} \left( 1 + \Delta_{j_k} \right) \delta_{\text{one}}^i + S_H \prod_{k=1}^{u-1} \left( 1 + \Delta_{j_k} \right) - S_H \quad (7.82)$$

We now decompose the sum over $x$

$$\sum_{x_{u-1}} \left\| (O^\top \hat{U})(\hat{m}_{u-1:1} - \hat{m}_{u-1:1}) \right\|_1 \quad (7.83)$$

$$= \sum_{x_{u-1}} \left\| (O^\top \hat{U}) \left( \hat{M}_u - \hat{M}_u \right) \hat{m}_{u-1:1} + (\hat{M}_u - \hat{M}_u) (\hat{m}_{u-1:1} - \hat{m}_{u-1:1}) + \hat{m}_{u-1:1} - \hat{m}_{u-1:1} \right\|_1$$

Using the triangle inequality, we get

$$\sum_{x_{u-1}} \left\| (O^\top \hat{U}) (\hat{M}_u - \hat{M}_u) (O^\top \hat{U})^{-1} \right\|_1 \left\| (O^\top \hat{U}) \hat{m}_{u-1:1} \right\|_1 \quad (7.84)$$

$$+ \sum_{x_{u-1}} \left\| (O^\top \hat{U}) (\hat{M}_u - \hat{M}_u) (O^\top \hat{U})^{-1} \right\|_1 \left\| (O^\top \hat{U}) \hat{m}_{u-1:1} - \hat{m}_{u-1:1} \right\|_1 \quad (7.85)$$

$$+ \sum_{x_{u-1}} \left\| (O^\top \hat{U}) \hat{M}_u (O^\top \hat{U})^{-1} \right\|_1 \left\| (O^\top \hat{U}) \hat{m}_{u-1:1} - \hat{m}_{u-1:1} \right\|_1 \quad (7.86)$$

Again we are just numbering the $U$'s and $O$'s for clarity to see which corresponds with which. They are omitted in the actual theorem statements since we will take minimums etc. at the end.

We now must bound these terms. First term:

$$\sum_{x_u} \left\| (O^\top \hat{U}_2) (\hat{M}_u - \hat{M}_u) (O^\top \hat{U}_2)^{-1} \right\|_1 \leq \Delta_u \sum_{x_{u-1}} \left\| (O^\top \hat{U}_1) \hat{m}_{u-1:1} \right\|_1 \leq S_H \Delta_{u} \quad (7.87)$$

since $\Delta_u = \left\| (O^\top \hat{U}_2) (\hat{M}_u - \hat{M}_u) (O^\top \hat{U}_1)^{-1} \right\|_1$. Second term can be bounded by inductive hypothesis:
The third term is bounded by observing that \((O^\top \tilde{U}) \tilde{M}_u (O^\top \tilde{U})^{-1} = \text{diag}(Pr[x|\text{Parent}])\). Thus it is diagonal, and \(Pr[x|\text{Parent}]\) has max row or column sum as 1. This means that the third term is bounded by the inductive hypothesis as well:

\[
\sum_{x_u:1} \left\| (O^\top \tilde{U}) \tilde{M}_u (O^\top \tilde{U})^{-1} \right\|_1 \left\| (O^\top \tilde{U})(\tilde{m}_{(u-1):1} - \hat{m}_{(u-1):1}) \right\|_1 \leq \Delta_u \left( \prod_{k=1}^{u-1} (1 + \Delta_{j_k} \eta^i_{\text{one}} + S_H \prod_{k=1}^{u-1} (1 + \Delta_{j_k}) - S_H \right)^{7.89}
\]

### 7.9.2 Bounding the propagation of error in tree

We now wrap up the proof based on the approach of HKZ [Hsu et al. 2009].

**Lemma 21.**

\[
\sum_{x_1, \ldots, x_O} \left| \tilde{P}[x_1, \ldots, x_O] - P[x_1, \ldots, x_O] \right| \leq S_H \delta_{\text{root}} + (1 + \delta_{\text{root}}) \left( \prod_{k=1}^{I} (1 + \Delta_{j_k} \eta^i_{\text{one}} + S_H \prod_{k=1}^{I} (1 + \Delta_{j_k}) - S_H \right)^{7.90}
\]

\[
\sum_{x_1, \ldots, x_O} \left| \tilde{P}[x_1, \ldots, x_O] - P[x_1, \ldots, x_O] \right| = \sum_{x_1, \ldots, x_O} \left| \tilde{r}^\top \tilde{M}_{j_1} \cdots \tilde{M}_{j_I} \tilde{1}_r - \tilde{r}^\top \tilde{M}_{j_1} \cdots \tilde{M}_{j_I} \tilde{1}_r \right| = \sum_{x_1, \ldots, x_O} \left| (\tilde{r} - \tilde{r})^\top (O^\top \tilde{U})^{-1} (O^\top \tilde{U})(\tilde{M}_{j_1} \tilde{1}) \right| + \sum_{x_1, \ldots, x_O} \left| (\tilde{r} - \tilde{r})^\top (O^\top \tilde{U})^{-1} (O^\top \tilde{U})(\tilde{M}_{j_1} \tilde{1} - \tilde{M}_{j_1} \tilde{1}) \right| + \sum_{x_1, \ldots, x_O} \left| \tilde{r}^\top (O^\top \tilde{U})^{-1} (O^\top \tilde{U})(\tilde{M}_{j_1} \tilde{1} - \tilde{M}_{j_1} \tilde{1}) \right|
\]

The first sum is bounded using Holder inequality and noting that the first term is a conditional probability (of all observed variables conditioned on the root)

\[
\sum_{x_1, \ldots, x_O} \left| (\tilde{r} - \tilde{r})^\top (O^\top \tilde{U})^{-1} (O^\top \tilde{U})(\tilde{M}_{j_1} \tilde{1}) \right| \leq \sum_{x_1, \ldots, x_O} \left\| (\tilde{r} - \tilde{r})^\top (O^\top \tilde{U})^{-1} \right\|_\infty \left\| (O^\top \tilde{U})(\tilde{M}_{j_1} \tilde{1}) \right\|_1 \leq S_H \delta_{\text{root}}
\]
lemma):
\[
\sum_{x_1, \ldots, x_O} \left| \left( \hat{\mathbf{r}} - \bar{\mathbf{r}} \right)^{\top} (O^{\top} \hat{U})^{-1} (O^{\top} \hat{U}) (\hat{M}_{f;1} \mathbf{1}_r - \bar{M}_{f;1} \mathbf{1}_r) \right| \tag{7.97}
\]
\[
\leq \sum_{x_1, \ldots, x_O} \left\| \left( \hat{\mathbf{r}} - \bar{\mathbf{r}} \right)^{\top} (O^{\top} \hat{U})^{-1} \right\|_1 \left\| (O^{\top} \hat{U}) (\hat{M}_{f;1} \mathbf{1}_r - \bar{M}_{f;1} \mathbf{1}_r) \right\|_1 \tag{7.98}
\]
\[
\leq \delta_{\text{root}} \left( \prod_{k=1}^j (1 + \Delta_{jk}) \delta_{\text{one}} + S_H \prod_{k=1}^j (1 + \Delta_{jk}) - S_H \right) \tag{7.99}
\]

The third sum is also bounded by Holder’s Inequality and previous lemmas and noting that \( \hat{\mathbf{r}}^{\top} (U^{\top} O)^{-1} = \mathbb{P}[R = r] \):
\[
\sum_{x_1, \ldots, x_O} \left| \hat{\mathbf{r}}^{\top} (O^{\top} \hat{U})^{-1} (O^{\top} \hat{U}) (\hat{M}_{f;1} \mathbf{1}_r - \bar{M}_{f;1} \mathbf{1}_r) \right| \tag{7.100}
\]
\[
\leq \sum_{x_1, \ldots, x_O} \left\| \hat{\mathbf{r}}^{\top} (O^{\top} \hat{U})^{-1} \right\|_1 \left\| (O^{\top} \hat{U}) (\hat{M}_{f;1} \mathbf{1}_r - \bar{M}_{f;1} \mathbf{1}_r) \right\|_1 \tag{7.101}
\]
\[
\leq \left( \prod_{k=1}^j (1 + \Delta_{jk}) \delta_{\text{one}} + S_H \prod_{k=1}^j (1 + \Delta_{jk}) - S_H \right) \tag{7.102}
\]

Combining these bounds gives us the desired solution.

### 7.10 Putting it all together

We seek for
\[
\sum_{x_1, \ldots, x_O} \left| \mathbb{P}[x_1, \ldots, x_O] - \mathbb{P}[x_1, \ldots, x_O] \right| \leq \epsilon \tag{7.103}
\]

Using the fact that for \( a < .5 \), \( (1 + a/t)^{\top} \leq 1 + 2a \), we get that \( \Delta_{jk} \leq O(\epsilon/(S_H)) \). However, \( \Delta_j \) is defined recursively, and thus the error accumulates exponential in the longest path of hidden nodes. For example, \( \Delta_{\text{obs}}^j \leq O\left( \frac{\epsilon}{(d_{\text{max}} S_H)^{2l+1}} \right) \) where \( l \) is the longest path of hidden nodes. Tracing this back through will gives the result:

Pick any \( \epsilon > 0, \delta < 1 \). Let
\[
N \geq \frac{1}{\epsilon^2} \left( \frac{(d_{\text{max}} S_H)^{2l+1} S_O}{\min_i \sigma_{S_H}(O_i)^2 \min_{i \neq j} \sigma_{S_H}(P_{ij})^4} \right) \log \frac{\delta}{\epsilon} \tag{7.104}
\]

Then with probability \( 1 - \delta \)
\[
\sum_{x_1, \ldots, x_O} \left| \mathbb{P}[x_1, \ldots, x_O] - \mathbb{P}[x_1, \ldots, x_O] \right| \leq \epsilon \tag{7.105}
\]
In many cases, if the frequency of the observation symbols follow certain distributions, than the dependence on $S_O$ can be removed as showed in HKZ [Hsu et al., 2009].

### 7.10.1 Matrix Perturbation Bounds

This is Theorem 3.8 from pg. 143 in Stewart and Sun, 1990 [Stewart and Sun, 1990]. Let $A \in \mathbb{R}^{m \times n}$, with $m \geq n$ and let $\tilde{A} = A + E$. Then

$$\left\| \tilde{A}^+ - A^+ \right\|_2 \leq \frac{1 + \sqrt{5}}{2} \max(\|A^+\|_2^2, \|\tilde{A}\|_2^2) \|E\|_2 \tag{7.106}$$
Part II

Spectral Models for Natural Language Processing
Chapter 8

Spectral Unsupervised Parsing with Additive Tree Metrics

In this chapter we tackle unsupervised syntactic parsing, an important problem in NLP where existing methods are very sensitive to local optima and therefore require careful initialization. Although traditionally formulated as a parameter learning problem, we take a different approach that revolves around structure learning. Thus, some of the key ideas of this chapter are closely related to the structure learning algorithms discussed in Chapter 6. We discuss this relationship further in §8.5.

Contribution of this chapter: We propose a spectral approach for unsupervised constituent parsing that comes with theoretical guarantees on latent structure recovery. The main algorithm is based on lifting the concept of additive tree metrics for structure learning of latent trees in the phylogenetic and machine learning communities to the case where the tree structure varies across examples. Although finding the “minimal” latent tree is NP-hard in general, for the case of projective trees we find that it can be found using bilexical parsing algorithms. Empirically, our algorithm performs favorably compared to the constituent context model of Klein and Manning (2002) without the need for careful initialization.

Outline: An introduction is first presented, followed by a description of our conditional latent tree model. We then propose a provably consistent learning algorithm and finally present experiments.

Prerequisites: This chapter assumes a general understanding of latent variable models as presented in Chapter 2.2 and the connection between latent variable models and low rank factorization in Chapter 3.

8.1 Introduction

Solutions to the problem of grammar induction have been long sought after since the early days of computational linguistics and are interesting both from cognitive and engineering perspectives.
Cognitively, it is more plausible to assume that children obtain only terminal strings of parse trees and not the actual parse trees. This means the unsupervised setting is a better model for studying language acquisition. From the engineering perspective, training data for unsupervised parsing exists in abundance (i.e., sentences and part-of-speech tags), and it is much cheaper than the syntactically annotated data required for supervised training.

Most existing solutions treat the problem of unsupervised parsing by assuming a generative process over parse trees, e.g., probabilistic context-free grammars (Jelinek et al., 1992), and the constituent context model (Klein and Manning, 2002). Learning then reduces to finding a set of parameters that are estimated by identifying a local maximum of an objective function such as the likelihood (Klein and Manning, 2002) or a variant of it (Smith and Eisner, 2005; Cohen and Smith, 2009; Headden et al., 2009; Spitkovsky et al., 2010b; Gillenwater et al., 2010; Golland and DeNero, 2012). Unfortunately, finding the global maximum for these objective functions is usually intractable (Cohen and Smith, 2012), which often leads to severe local optima problems (but see Gormley and Eisner, 2013). Thus, strong experimental results are often achieved by initialization techniques (Klein and Manning, 2002; Gimpel and Smith, 2012), incremental dataset use (Spitkovsky et al., 2010a) and other specialized techniques to avoid local optima such as count transforms (Spitkovsky et al., 2013). These approaches, while empirically promising, generally lack theoretical justification.

On the other hand, recently proposed spectral methods approach the problem via restriction of the PCFG model (Hsu et al., 2012) or matrix completion (Bailly et al., 2013). These novel perspectives offer strong theoretical guarantees but are not designed to achieve competitive empirical results.

In this chapter, we suggest a different approach, to provide a first step to bridging this theory-experiment gap. More specifically, we approach unsupervised constituent parsing from the perspective of structure learning as opposed to parameter learning. We associate each sentence with an undirected latent tree graphical model, which is a tree consisting of both observed variables (corresponding to the words in the sentence) and an additional set of latent variables that are unobserved in the data. This undirected latent tree is then directed via a direction mapping to give the final constituent parse.

In our framework, parsing reduces to finding the best latent structure for a given sentence. However, due to the presence of latent variables, structure learning of latent trees is substantially more complicated than in observed models. As before, one solution would be local search heuristics.

Intuitively, however, latent tree models encode low rank dependencies among the observed variables permitting the development of “spectral” methods that can lead to provably correct solutions. In particular we leverage the concept of additive tree metrics (Buneman, 1971, 1974) in phylogenetics and machine learning that can create a special distance metric among the observed variables as a function of the underlying spectral dependencies (Choi et al., 2011; Song et al., 2011b; Anandkumar et al., 2011; Ishteva et al., 2012). Additive tree metrics can be leveraged by “meta-algorithms” such as neighbor-joining (Saitou and Nei, 1987) and recursive grouping (Choi et al., 2011) to provide consistent learning algorithms for latent trees.

Moreover, we show that it is desirable to learn the “minimal” latent tree based on the tree metric (“minimum evolution” in phylogenetics). While this criterion is in general NP-hard (Desper and Gascuel, 2005), for projective trees we find that a bilexical parsing algorithm can be used to find...
The bear ate the fish

Figure 8.1: Example for the tag sequence (DT, NN, VBD, DT, NN) showing the overview of our approach. We first learn a undirected latent tree for the sequence (left). We then apply a direction mapping $h_{\text{dir}}$ to direct the latent tree (center). This can then easily be converted into a bracketing (right).

Figure 8.2: Candidate constituent parses for $x = (\text{VBD, DT, NN})$ (left-correct, right -incorrect)

an exact solution efficiently (Eisner and Satta, 1999).

Unlike in phylogenetics and graphical models, where a single latent tree is constructed for all the data, in our case, each part of speech sequence is associated with its own parse tree. This leads to a severe data sparsity problem even for moderately long sentences. To handle this issue, we present a strategy that is inspired by ideas from kernel smoothing in the statistics community (Zhou et al., 2010; Kolar et al., 2010b,a). This allows principled sharing of samples from different but similar underlying distributions.

We provide theoretical guarantees on the recovery of the correct underlying latent tree and characterize the associated sample complexity under our technique. Empirically we evaluate our method on data in English, German and Chinese. Our algorithm performs favorably to Klein and Manning’s (2002) constituent-context model (CCM), without the need for careful initialization. In addition, we also analyze CCM’s sensitivity to initialization, and compare our results to Seginer’s algorithm (Seginer, 2007).

8.2 Learning Setting and Model

In this section, we detail the learning setting and a conditional tree model we learn the structure for.

8.2.1 Learning Setting

Let $w = (w_1, ..., w_\ell)$ be a vector of words corresponding to a sentence of length $\ell$. Each $w_i$ is represented by a vector in $\mathbb{R}^p$ for $p \in \mathbb{N}$. The vector is an embedding of the word in some space,
chosen from a fixed dictionary that maps word types to \( \mathbb{R}^p \). In addition, let \( x = (x_1, ..., x_\ell) \) be the associated vector of part-of-speech (POS) tags (i.e. \( x_i \) is the POS tag of \( w_i \)).

In our learning algorithm, we assume that examples of the form \((w^{(i)}, x^{(i)})\) for \( i \in [N] = \{1, \ldots, N\} \) are given, and the goal is to predict a bracketing parse tree for each of these examples. The word embeddings are used during the learning process, but the final decoder that the learning algorithm outputs maps a POS tag sequence \( x \) to a parse tree. While ideally we would want to use the word information in decoding as well, much of the syntax of a sentence is determined by the POS tags, and relatively high level of accuracy can be achieved by learning, for example, a supervised parser from POS tag sequences.

Just like our decoder, our model assumes that the bracketing of a given sentence is a function of its POS tags. The POS tags are generated from some distribution, followed by a deterministic generation of the bracketing parse tree. Then, latent states are generated for each bracket, and finally, the latent states at the yield of the bracketing parse tree generate the words of the sentence (in the form of embeddings). The latent states are represented by vectors \( z \in \mathbb{R}^m \) where \( m < p \).

### 8.2.2 Intuition

For intuition, consider the simple tag sequence \( x = (\text{VBD}, \text{DT}, \text{NN}) \). Two candidate constituent parse structures are shown in Figure 8.2 and the correct one is boxed in green (the other in red). Recall that our training data contains word phrases that have the tag sequence \( x \) e.g. \( w^{(1)} = (\text{hit}, \text{the}, \text{ball}) \), \( w^{(2)} = (\text{ate}, \text{an}, \text{apple}) \).

Intuitively, the words in the above phrases exhibit dependencies that can reveal the parse structure. The determiner \( w_2 \) and the direct object \( w_3 \) are correlated in that the choice of determiner depends on the plurality of \( w_3 \). However, the choice of verb \( w_1 \) is mostly independent of the determiner. We could thus conclude that \( w_2 \) and \( w_3 \) should be closer in the parse tree than \( w_1 \) and \( w_2 \), giving us the correct structure. Informally, the latent state \( z \) corresponding to the \((w_2, w_3)\) bracket would store information about the plurality of \( z \), the key to the dependence between \( w_2 \) and \( w_3 \). It would then be reasonable to assume that \( w_2 \) and \( w_3 \) are independent given \( z \).

### 8.2.3 A Conditional Latent Tree Model

Following this intuition, we propose to model the distribution over the latent bracketing states and words for each tag sequence \( x \) as a latent tree graphical model, which encodes conditional independences among the words given the latent states.

Let \( \mathcal{V} := \{w_1, ..., w_\ell, z_1, ..., z_H\} \), with \( w_i \) representing the word embeddings, and \( z_i \) representing the latent states of the bracketings. Then, according to our base model it holds that:
\[ P(\mathbf{w}, \mathbf{z}|\mathbf{x}) = \prod_{i=1}^{H} P(z_i|\pi_x(z_i), \theta(\mathbf{x})) \times \prod_{i=1}^{\ell(\mathbf{x})} P(w_i|\pi_x(w_i), \theta(\mathbf{x})) \]

(8.1)

where \( \pi_x() \) returns the parent node index of the argument in the latent tree corresponding to tag sequence \( x \). If \( z \) is the root, then \( \pi_x(z) = \emptyset \). All the \( w_i \) are assumed to be leaves while all the \( z_i \) are internal (i.e. non-leaf) nodes. The parameters \( \theta(\mathbf{x}) \) control the conditional probability tables. We do not commit to a certain parametric family, but see more about the assumptions we make about \( \theta \) in §8.3.2. The parameter space is denoted \( \Theta \). The model assumes a factorization according to a latent-variable tree. The latent variables can incorporate various linguistic properties, such as head information, valence of dependency being generated, and so on. This information is expected to be learned automatically from data.

Our generative model deterministically maps a POS sequence to a bracketing via an undirected latent-variable tree. The orientation of the tree is determined by a direction mapping \( h_{\text{dir}}(u) \), which is fixed during learning and decoding. This means our decoder first identifies (given a POS sequence) an undirected tree, and then orients it by applying \( h_{\text{dir}} \) on the resulting tree (see below).

Define \( \mathcal{U} \) to be the set of undirected latent trees where all internal nodes have degree exactly 3 (i.e. they correspond to binary bracketing), and in addition \( h_{\text{dir}}(u) \) for any \( u \in \mathcal{U} \) is projective (explained in the \( h_{\text{dir}} \) section). In addition, let \( \mathcal{T} \) be the set of binary bracketings. The complete generative model that we follow is then:

- Generate a tag sequence \( x = (x_1, \ldots, x_\ell) \)
- Decide on \( u(x) \in \mathcal{U} \), the undirected latent tree that \( x \) maps to.
- Set \( t \in \mathcal{T} \) by computing \( t = h_{\text{dir}}(u) \).
- Set \( \theta \in \Theta \) by computing \( \theta = \theta(\mathbf{x}) \).
- Generate a tuple \( \mathbf{v} = (w_1, \ldots, w_\ell, z_1, \ldots, z_H) \) where \( w_i \in \mathbb{R}^p, z_j \in \mathbb{R}^m \) according to Eq. \(8.1\)

See Figure 1 (left) for an example.

**The Direction Mapping** \( h_{\text{dir}} \). Generating a bracketing via an undirected tree enables us to build on existing methods for structure learning of latent-tree graphical models [Choi et al., 2011; Anandkumar et al., 2011]. Our learning algorithm focuses on recovering the undirected tree based for the generative model that was described above. This undirected tree is converted into a directed tree by applying \( h_{\text{dir}} \). The mapping \( h_{\text{dir}} \) works in three steps:

- It first chooses a top bracket \( ([1, R - 1], [R, \ell]) \) where \( R \) is the mid-point of the bracket and \( \ell \) is the length of the sentence.
• It marks the edge $e_{i,j}$ that splits the tree according to the top bracket as the “root edge” (marked in red in Figure 8.1(center))

• It then creates $t$ from $u$ by directing the tree outward from $e_{i,j}$ as shown in Figure 8.1(center)

The resulting $t$ is a binary bracketing parse tree. As implied by the above definition of $h_{\text{dir}}$, selecting which edge is the root can be interpreted as determining the top bracket of the constituent parse. For example, in Figure 8.1, the top bracket is $[(1, 2), [3, 5)] = ([\text{DT}, \text{NN}], [\text{VBD}, \text{DT}, \text{NN}])$. Note that the “root” edge $e_{z_1,z_2}$ partitions the leaves into precisely this bracketing. As indicated in the above section, we restrict the set of undirected trees to be those such that after applying $h_{\text{dir}}$ the resulting $t$ is projective i.e. there are no crossing brackets. In §8.4.1, we discuss an effective heuristic to find the top bracket without supervision.

8.3 Spectral Learning Algorithm based on Additive Tree Metrics

Our goal is to recover $t \in \mathcal{T}$ for tag sequence $x$ using the data $\mathcal{D} = \{(w^{(i)}, x^{(i)})\}_{i=1}^N$. To get an intuition about the algorithm, consider a partition of the set of examples $\mathcal{D}$ into $\mathcal{D}(x) = \{(w^{(i)}, x^{(i)}) \in \mathcal{D} | x^{(i)} = x\}$, i.e. each section in the partition has an identical sequence of part of speech tags. Assume for this section $|\mathcal{D}(x)|$ is large (we address the data sparsity issue in §8.3.4).

We can then proceed by learning how to map a POS sequence $x$ to a tree $t \in \mathcal{T}$ (through $u \in \mathcal{U}$) by focusing only on examples in $\mathcal{D}(x)$.

Directly attempting to maximize the likelihood unfortunately results in an intractable optimization problem and greedy heuristics are often employed (Harmeling and Williams, 2011). Instead we propose a method that is provably consistent and returns a tree that can be mapped to a bracketing using $h_{\text{dir}}$.

If all the variables were observed, then the Chow-Liu algorithm (Chow and Liu, 1968c) could be used to find the most likely tree structure $u \in \mathcal{U}$. The Chow-Liu algorithm essentially computes the distances among all pairs of variables (the negative of the mutual information) and then finds the minimum cost tree. However, the fact that the $z_i$ are latent variables makes this strategy substantially more complicated. In particular, it becomes challenging to compute the distances among pairs of latent variables. What is needed is a “special” distance function that allows us to reverse engineer the distances among the latent variables given the distances among the observed variables. This is the key idea behind additive tree metrics that are the basis of our approach.

In the following sections, we describe the key steps to our method. §3.1 and §3.2 largely describe existing background on additive tree metrics and latent tree structure learning, while §3.3 and §3.4 discuss novel aspects that are unique to our problem.

8.3.1 Additive Tree Metrics

Let $u(x)$ be the true undirected tree of sentence $x$ and assume the nodes $\mathcal{V}$ to be indexed by $[M] = \{1, \ldots, M\}$ such that $M = |\mathcal{V}| = H + \ell$. Furthermore, let $v \in \mathcal{V}$ refer to a node in the
undirected tree (either observed or latent). We assume the existence of a distance function that allows us to compute distances between pairs of nodes. For example, as we see in §8.3.2 we will define the distance \(d(i, j)\) to be a function of the covariance matrix \(\mathbb{E}[v_i v_j^\top | u(x), \theta(x)]\). Thus if \(v_i\) and \(v_j\) are both observed variables, the distance can be directly computed from the data.

Moreover, the metrics we construct are such that they are tree additive, defined below:

**Definition 2.** A function \(d_{u(x)} : [M] \times [M] \to \mathbb{R}\) is an additive tree metric [Erdos et al., 1999] for the undirected tree \(u(x)\) if it is a distance metric, and furthermore, \(\forall i, j \in [M]\) the following relation holds:

\[
d_{u(x)}(i, j) = \sum_{(a, b) \in \text{path}_{u(x)}(i, j)} d_{u(x)}(a, b)
\]

where \(\text{path}_{u(x)}(i, j)\) is the set of all the edges in the (undirected) path from \(i\) to \(j\) in the tree \(u(x)\).

As we describe below, given the tree structure, the additive tree metric property allows us to compute “backwards” the distances among the latent variables as a function of the distances among the observed variables.

Define \(D\) to be the \(M \times M\) distance matrix among the \(M\) variables, i.e. \(D_{ij} = d_{u(x)}(i, j)\). Let \(D_{WW}\), \(D_{ZW}\) (equal to \(D_{WZ}^\top\)), and \(D_{ZZ}\) indicate the word-word, latent-word and latent-latent sub-blocks of \(D\) respectively. In addition, since \(u(x)\) is assumed to be known from context, we denote \(d_{u(x)}(i, j)\) just by \(d(i, j)\).

Given the fact that the distance between a pair of nodes is a function of the random variables they represent (according to the true model), only \(D_{WW}\) can be empirically estimated from data. However, if the underlying tree structure is known, then Definition 2 can be leveraged to compute \(D_{ZZ}\) and \(D_{ZW}\) as we show below.

We first show how to compute \(d(i, j)\) for all \(i, j\) such that \(i\) and \(j\) are adjacent to each other in \(u(x)\), based only on observed nodes. It then follows that the other elements of the distance matrix can be computed based on Definition 2. To show how to compute distances between adjacent nodes, consider the two cases: (1) \((i, j)\) is a leaf edge; (2) \((i, j)\) is an internal edge.

---

2This means that it satisfies \(d(i, j) = 0\) if and only if \(i = j\), the triangle inequality and is also symmetric.
Case 1 (leaf edge, figure [8.3(a)]). Assume without loss of generality that \( j \) is the leaf and \( i \) is an internal latent node. Then \( i \) must have exactly two other neighbors \( a \in [M] \) and \( b \in [M] \). Let \( A \) denote the set of nodes that are closer to \( a \) than \( i \) and similarly let \( B \) denote the set of nodes that are closer to \( b \) than \( i \). Let \( A^* \) and \( B^* \) denote all the leaves (word nodes) in \( A \) and \( B \) respectively. Then using path additivity (Definition 2), it can be shown that for any \( a^* \in A^*, b^* \in B^* \) it holds that:

\[
d(i, j) = \frac{1}{2} (d(j, a^*) + d(j, b^*) - d(a^*, b^*)) \tag{8.3}
\]

Note that the right-hand side only depends on distances between observed random variables.

Case 2 (internal edge, figure [8.3(b)]). Both \( i \) and \( j \) are internal nodes. In this case, \( i \) has exactly two other neighbors \( a \in [M] \) and \( b \in [M] \), and similarly, \( j \) has exactly other two neighbors \( g \in [M] \) and \( h \in [M] \). Let \( A \) denote the set of nodes closer to \( a \) than \( i \), and analogously for \( B, G, \) and \( H \). Let \( A^*, B^*, G^*, \) and \( H^* \) refer to the leaves in \( A, B, G, \) and \( H \) respectively. Then for any \( a^* \in A^*, b^* \in B^*, g^* \in G^*, \) and \( h^* \in H^* \) it can be shown that:

\[
d(i, j) = \frac{1}{4} \left( d(a^*, g^*) + d(a^*, h^*) + d(b^*, g^*) + d(b^*, h^*) - 2d(a^*, b^*) - 2d(g^*, h^*) \right) \tag{8.4}
\]

Empirically, one can obtain a more robust empirical estimate \( \tilde{d}(i, j) \) by averaging over all valid choices of \( a^*, b^* \) in Eq. 8.35 and all valid choices of \( a^*, b^*, g^*, h^* \) in Eq. 8.36 (Desper and Gascuel, 2005).

### 8.3.2 Constructing a Spectral Additive Metric

In constructing our distance metric, we begin with the following assumption on the distribution in Eq. 8.1 (analogous to the assumptions made in Anandkumar et al., 2011).

**Assumption 1 (Linear, Rank \( m \), Means).**

\[
\mathbb{E}[z_i | \pi_x(z_i), x] = A_{[z_i | \pi_x(z_i), x]} \pi_x(z_i) \quad \forall i \in [H]
\]

where \( A_{[z_i | \pi_x(z_i), x]} \in \mathbb{R}^{n \times m} \) has rank \( m \).

\[
\mathbb{E}[w_i | \pi_x(w_i), x] = C_{[w_i | \pi_x(w_i), x]} \pi_x(w_i) \quad \forall i \in [\ell(x)]
\]

where \( C_{[w_i | \pi_x(w_i), x]} \in \mathbb{R}^{p \times m} \) has rank \( m \).

Also assume that \( \mathbb{E}[z_i z_i^T | x] \) has rank \( m \) \( \forall i \in [H] \).

Note that the matrices \( A \) and \( C \) are a direct function of \( \theta(x) \), but we do not specify a model family for \( \theta(x) \). The only restriction is in the form of the above assumption. If \( w_i \) and \( z_i \) were discrete, represented as binary vectors, the above assumption would correspond to requiring all conditional probability tables in the latent tree to have rank \( m \). Assumption [1] allows for the \( w_i \) to be high dimensional features, as long as the expectation requirement above is satisfied. Similar
assumptions are made with spectral parameter learning methods e.g. (Hsu et al., 2009), (Bailly et al. 2009), (Parikh et al. 2011), and (Cohen et al. 2012).

Furthermore, Assumption 1 makes it explicit that regardless of the size of \( p \), the relationships among the variables in the latent tree are restricted to be of rank \( m \), and are thus low rank since \( p > m \). To leverage this low rank structure, we propose using the following additive metric, a normalized variant of that in (Anandkumar et al., 2011):

\[
d_{\text{spectral}}(i, j) = -\log \Lambda_m(\Sigma_x(i, j)) + \frac{1}{2} \log \Lambda_m(\Sigma_x(i, i)) + \frac{1}{2} \log \Lambda_m(\Sigma_x(j, j)) \tag{8.5}
\]

where \( \Lambda_m(A) \) denotes the product of the top \( m \) singular values of \( A \) and \( \Sigma_x(i, j) := \mathbb{E}[v_i v_j^\top | x] \), i.e. the uncentered cross-covariance matrix.

We can then show that this metric is additive:

**Lemma 22.** If Assumption 1 holds then, \( d_{\text{spectral}} \) is an additive tree metric (Definition 2).

A proof is in the appendix for completeness. From here, we use \( d \) to denote \( d_{\text{spectral}} \), since that is the metric we use for our learning algorithm.

### 8.3.3 Recovering the Minimal Projective Latent Tree

It has been shown (Rzhetsky and Nei, 1993) that for any additive tree metric, \( u(\alpha) \) can be recovered by solving \( \arg \min_{u \in \mathcal{U}} c(u) \) for \( c(u) \):

\[
c(u) = \sum_{(i, j) \in \mathcal{E}_u} d(i, j). \tag{8.6}
\]

where \( \mathcal{E}_u \) is the set of pairs of nodes which are adjacent to each other in \( u \) and \( d(i, j) \) is computed using Eq. 8.35 and Eq. 8.36.

Note that the metric \( d \) we use in defining \( c(u) \) is based on the expectations from the true distribution. In practice, the true distribution is unknown, and therefore we use an approximation for the distance metric \( \hat{d} \). As we discussed in §8.3.1 all elements of the distance matrix are functions of observable quantities if the underlying tree \( u \) is known. However, only the word-word sub-block \( D_{\text{WW}} \) can be directly estimated from the data without knowledge of the tree structure.

This subtlety makes solving the minimization problem in Eq. 8.6 NP-hard (Desper and Gascuel 2005) if \( u \) is allowed to be an arbitrary undirected tree. However, if we restrict \( u \) to be in \( \mathcal{U} \), as we do in the above, then maximizing \( \hat{c}(u) \) over \( \mathcal{U} \) can be solved using the bilexical parsing algorithm from (Eisner and Satta 1999). This is because the computation of the other sub-blocks of the distance matrix only depend on the partitions of the nodes shown in Figure 8.3 into \( A, B, G, \) and \( H \), and not on the entire tree structure.

Therefore, the procedure to find a bracketing for a given POS tag \( x \) is to first estimate the distance matrix sub-block \( \hat{D}_{\text{WW}} \) from raw text data (see §8.3.4), and then solve the optimization problem \( \arg \min_{u \in \mathcal{U}} \hat{c}(u) \) using a variant of the Eisner-Satta algorithm where \( \hat{c}(u) \) is identical to \( c(u) \) in Eq. 8.6, with \( d \) replaced with \( \hat{d} \).
We now address the data sparsity problem, in particular that estimating \(d\) must be estimated from data, as discussed below.

If the true distance metric \(d(x)\) can be computed based only on the text, without needing to identify the latent information \(h_{\text{dir}}\), then \(d\) can be fully recovered by optimizing the cost function \(c(u)\). However, in practice the distance metric is known, with respect to the true distribution that generates the words in a sentence, then \(u\) can be fully recovered by optimizing the cost function \(c(u)\). However, in practice the distance metric must be estimated from data, as discussed above.

### 8.3.4 Estimation of \(d\) from Sparse Data

We now address the data sparsity problem, in particular that \(D(x)\) can be very small, and therefore estimating \(d\) for each POS sequence separately can be problematic.\(^3\)

In order to estimate \(d\) from data, we need to estimate the covariance matrices \(\Sigma_x(i,j)\) (for \(i,j \in \{1, \ldots, \ell(x)\}\)) from Eq. 8.5.

To give some motivation to our solution, consider estimating the covariance matrix \(\Sigma_x(1,2)\) for the tag sequence \(x = (\text{DT}_1, \text{NN}_2, \text{VBD}_3, \text{DT}_4, \text{NN}_5)\). \(D(x)\) may be insufficient for an accurate empirical

\(^3\)This data sparsity problem is quite severe – for example, the Penn treebank [Marcus et al., 1993] has a total number of 43,498 sentences, with 42,246 unique POS tag sequences, averaging \(|D(x)|\) to be 1.04.
estimate. However, consider another sequence \(x' = (\text{RB}_1, \text{DT}_2, \text{NN}_3, \text{VBD}_4, \text{DT}_5, \text{ADJ}_6, \text{NN}_7)\). Although \(x\) and \(x'\) are not identical, it is likely that \(\Sigma_{x'}(2,3)\) is similar to \(\Sigma_x(1,2)\) because the determiner and the noun appear in similar syntactic context. \(\Sigma_{x'}(5,7)\) also may be somewhat similar, but \(\Sigma_{x'}(2,7)\) should not be very similar to \(\Sigma_x(1,2)\) because the noun and the determiner appear in a different syntactic context.

The full learning algorithm is given in Figure ???. The first step in the algorithm is to estimate the covariance matrix block \(\tilde{\Sigma}_{x(i)}(j,k)\) for each training example \(x^{(i)}\) and each pair of preterminal positions \((j,k)\) in \(x^{(i)}\). Instead of computing this block by computing the empirical covariance matrix for positions \((j,k)\) in the data \(D(x)\), the algorithm uses all of the pairs \((j',k')\) from all of \(N\) training examples. It averages the empirical covariance matrices from these contexts using a kernel weight, which gives a similarity measure for the position \((j,k)\) in \(x^{(i)}\) and \((j',k')\) in another example \(x^{(i')}\). \(\gamma\) is the kernel “bandwidth”, a user-specified parameter that controls how inclusive the kernel will be with respect to examples in \(D\) (see §8.4.1 for a concrete example). Note that the learning algorithm is such that it ensures that \(\tilde{\Sigma}_{x(i)}(j,k) = \Sigma_{x(i')}((j',k')\text{ if }G(j,x^{(i)}) = G(j',x^{(i')}))\) and \(G(k,x^{(i)}) = G(k',x^{(i')}\).

Once the empirical estimates for the covariance matrices are obtained, a variant of the Eisner-Satta algorithm is used, as mentioned in §8.3.3.

### 8.3.5 Theoretical Guarantees

Our main theoretical guarantee is that Algorithm 11 will recover the correct tree \(u \in \mathcal{U}\) with high probability, if the given top bracket is correct and if we obtain enough examples \((u^{(i)}, x^{(i)})\) from the model in §2. We give the theorem statement below. The constants lurking in the \(O\)-notation and the full proof are in the Appendix.

Denote \(\sigma_x(j,k)^{(r)}\) as the \(r\text{th}\) singular value of \(\Sigma_x(j,k)\). Let \(\sigma^*(x) := \min_{j,k \in \ell(x)} \min \left(\sigma_x(j,k)^{(m)}\right)\).

**Theorem 5.** Define \(\hat{u}\) as the tree for tag sequence \(x\) being output by Algorithm 11 and \(u(x)\) as the correct tree. Let 

\[\Delta(x) := \min_{u' \in \mathcal{U} \setminus u(x)} (c(u(x)) - c(u'))/(8|\ell(x)|)\]

Let \(0 < \delta < 1\). Assume that

\[N \geq O\left(\frac{m^2 \log \left(\frac{\sigma^*(x)^2}{\delta}\right)}{\min(\sigma^*(x)^2 \Delta(x)^2, \sigma^*(x)^2 \nu_x(\gamma)^2)}\right)\]
Then with probability $1 - \delta$, $\hat{u} = u(x)$.

where $\nu_x(\gamma)$, defined in the Appendix, is a function of the underlying distribution over the tag sequences $x$ and the kernel bandwidth $\gamma$.

Thus, the sample complexity of our approach depends on the dimensionality of the latent and observed states ($m$ and $p$), the underlying singular values of the cross-covariance matrices ($\sigma^*(x)$) and the difference in the cost of the true tree compared to the cost of the incorrect trees ($\triangle(x)$).

8.4 Experiments

We report results on three different languages: English, German, and Chinese. For English we use the Penn treebank \cite{Marcus1993}, with sections 2–21 for training and section 23 for final testing. For German and Chinese we use the Negra treebank and the Chinese treebank respectively and the first 80% of the sentences are used for training and the last 20% for testing. All punctuation from the data is removed.

We primarily compare our method to the constituent-context model (CCM) of \cite{Klein2002}. We also compare our method to the algorithm of \cite{Seginer2007}.

8.4.1 Experimental Settings

**Top bracket heuristic** Our algorithm requires the top bracket in order to direct the latent tree. In practice, we employ the following heuristic to find the bracket using the following three steps:

- If there exists a comma/semicolon/colon at index $i$ that has at least a verb before $i$ and both a noun followed by a verb after $i$, then return $([0, i - 1], [i, \ell(x)])$ as the top bracket. (Pick the rightmost comma/semicolon/colon if multiple satisfy the criterion).
- Otherwise find the first non-participle verb (say at index $j$) and return $([0, j - 1], [j, \ell(x)])$.
- If no verb exists, return $([0, 1], [1, \ell(x)])$.

**Word embeddings** As mentioned earlier, each $w_i$ can be an arbitrary feature vector. For all languages we use Brown clustering \cite{Brown1992} to construct a $\log(C) + C$ feature vector where the first $\log(C)$ elements indicate which mergable cluster the word belongs to, and the last $C$ elements indicate the cluster identity. For English, more sophisticated word embeddings are easily obtainable, and we experiment with neural word embeddings \cite{Turian2010} of length 50. We also explored two types of CCA embeddings: OSCCA and TSCCA, given in \cite{Dhillon2012b}. The OSCCA embeddings behaved better on the English the dataset, so we only report their results.

\footnote{We make brief use of punctuation for our top bracket heuristic detailed below before removing it.}
### Table 8.1: Comparison of different CCM variants on English (training). U stands for universal POS tagset, OB stands for conjoining original POS tags with Brown clusters and UB stands for conjoining universal POS tags with Brown clusters. The best setting is just the vanilla setting, CCM.

<table>
<thead>
<tr>
<th>$\ell$</th>
<th>CCM</th>
<th>CCM-U</th>
<th>CCM-OB</th>
<th>CCM-UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq 10$</td>
<td>72.5</td>
<td>57.1</td>
<td>58.2</td>
<td>62.9</td>
</tr>
<tr>
<td>$\leq 15$</td>
<td>54.1</td>
<td>36</td>
<td>24</td>
<td>23.7</td>
</tr>
<tr>
<td>$\leq 20$</td>
<td>50</td>
<td>34.7</td>
<td>19.3</td>
<td>19.1</td>
</tr>
<tr>
<td>$\leq 25$</td>
<td>47.2</td>
<td>30.7</td>
<td>16.8</td>
<td>16.6</td>
</tr>
<tr>
<td>$\leq 30$</td>
<td>44.8</td>
<td>29.6</td>
<td>15.3</td>
<td>15.2</td>
</tr>
<tr>
<td>$\leq 40$</td>
<td>26.3</td>
<td>13.5</td>
<td>13.9</td>
<td>13.8</td>
</tr>
</tbody>
</table>

### Table 8.2: $F_1$ bracketing measure for the test sets and train sets in three languages. NN, CC, and BC indicate the performance of our method for neural embeddings, CCA embeddings, and Brown clustering respectively, using the heuristic for $h_{\text{dir}}$ described in §8.4.1. NN-O, CC-O, and BC-O indicate that the oracle (i.e. true top bracket) was used for $h_{\text{dir}}$.

<table>
<thead>
<tr>
<th>$\ell$</th>
<th>English</th>
<th>German</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell$</td>
<td>$\text{NN-O}$</td>
<td>$\text{NN}$</td>
<td>$\text{CC-O}$</td>
</tr>
<tr>
<td>$\leq 10$</td>
<td>70.9</td>
<td>69.2</td>
<td>70.4</td>
</tr>
<tr>
<td>$\leq 15$</td>
<td>60.3</td>
<td>58.3</td>
<td>58.6</td>
</tr>
<tr>
<td>$\leq 20$</td>
<td>54.1</td>
<td>52.3</td>
<td>52.3</td>
</tr>
<tr>
<td>$\leq 25$</td>
<td>50.8</td>
<td>49.0</td>
<td>48.6</td>
</tr>
<tr>
<td>$\leq 30$</td>
<td>48.1</td>
<td>46.3</td>
<td>45.6</td>
</tr>
<tr>
<td>$\leq 40$</td>
<td>45.5</td>
<td>43.8</td>
<td>43.0</td>
</tr>
</tbody>
</table>

**Choice of kernel** For our experiments, we use the kernel

$$K_{\gamma}(j,k,j',k'|x,x') = \max \left\{ 0, 1 - \frac{\kappa(j,k,j',k'|x,x')}{\gamma} \right\}$$

where $\gamma$ denotes the user-specified bandwidth, and $\kappa(j,k,j',k'|x,x') = \frac{|j-k| - |j'-k'|}{|j-k| + |j'-k'|}$ if $x(j) = x(j')$ and $x(k') = x(k)$, and $\text{sign}(j-k) = \text{sign}(j'-k')$ (and $\infty$ otherwise).

The kernel is non-zero if and only if the tags at position $j$ and $k$ in $x$ are identical to the ones in position $j'$ and $k'$ in $x'$, and if the direction between $j$ and $k$ is identical to the one between $j'$ and $k'$. Note that the kernel is not binary, as opposed to the theoretical kernel in the Appendix. Our experiments show that using a non-zero value different than 1 that is a function of the distance between $j$ and $k$ compared to the distance between $j'$ and $k'$ does better in practice.
Figure 8.4: Histogram showing performance of CCM across 100 random restarts for sentences of length $\leq 10$.

**Choice of data** For CCM, we found that if the full dataset (all sentence lengths) is used in training, then performance degrades when evaluating on sentences of length $\leq 10$. We therefore restrict the data used with CCM to sentences of length $\leq \ell$, where $\ell$ is the maximal sentence length being evaluated. This does not happen with our algorithm, which manages to leverage lexical information whenever more data is available. We therefore use the full data for our method for all lengths.

We also experimented with the original POS tags and the universal POS tags of Petrov et al. (2011). Here, we found out that our method does better with the universal part of speech tags. For CCM, we also experimented with the original parts of speech, universal tags (CCM-U), the cross-product of the original parts of speech with the Brown clusters (CCM-OB), and the cross-product of the universal tags with the Brown clusters (CCM-UB). The results in Table 8.1 indicate that the vanilla setting is the best for CCM.

Thus, for all results, we use universal tags for our method and the original POS tags for CCM. We believe that our approach substitutes the need for fine-grained POS tags with the lexical information. CCM, on the other hand, is fully unlexicalized.

**Parameter Selection** Our method requires two parameters, the latent dimension $m$ and the bandwidth $\gamma$. CCM also has two parameters, the number of extra constituent/distituent counts used for smoothing. For both methods we chose the best parameters for sentences of length $\ell \leq 10$ on the English Penn Treebank (training) and used this set for all other experiments. This resulted in $m = 7$, $\gamma = 0.4$ for our method and 2, 8 for CCM’s extra constituent/distituent counts respectively. We also tried letting CCM choose different hyperparameters for different sentence lengths based on dev-set likelihood, but this gave worse results than holding them fixed.
8.4.2 Results

**Test I: Accuracy**  Table 8.2 summarizes our results. CCM is used with the initializer proposed in (Klein and Manning, 2002)\(^5\) NN, CC, and BC indicate the performance of our method for neural embeddings, CCA embeddings, and Brown clustering respectively, using the heuristic for \(h_{dir}\) described in § 8.4.1. NN-O, CC-O, and BC-O indicate that the oracle (i.e. true top bracket) was used for \(h_{dir}\). For our method, test set results can be obtained by using Algorithm ?? (except the distances are computed using the training data).

For English, while CCM behaves better for short sentences (\(\ell \leq 10\)), our algorithm is more robust with longer sentences. This is especially noticeable for length \(\leq 40\), where CCM breaks down and our algorithm is more stable. We find that the neural embeddings modestly outperform the CCA and Brown cluster embeddings.

The results for German are similar, except CCM breaks down earlier at sentences of \(\ell \leq 30\). For Chinese, our method substantially outperforms CCM for all lengths. Note that CCM performs very poorly, obtaining only around 20% accuracy even for sentences of \(\ell \leq 20\). We didn’t have neural embeddings for German and Chinese (which worked best for English) and thus only used Brown cluster embeddings.

For English, the disparity between NN-O (oracle top bracket) and NN (heuristic top bracket) is rather low suggesting that our top bracket heuristic is rather effective. However, for German and Chinese note that the “BC-O” performs substantially better, suggesting that if we had a better top bracket heuristic our performance would increase.

**Test II: Sensitivity to initialization**  The EM algorithm with the CCM requires very careful initialization, which is described in (Klein and Manning, 2002). If, on the other hand, random initialization is used, the variance of the performance of the CCM varies greatly. Figure 8.4 shows a histogram of the performance level for sentences of length \(\leq 10\) for different random initializers. As one can see, for some restarts, CCM obtains accuracies lower than 30% due to local optima. Our method does not suffer from local optima and thus does not require careful initialization.

**Test III: Comparison to Seginer’s algorithm**  Our approach is not directly comparable to Seginer’s because he uses punctuation, while we use POS tags. Using Seginer’s parser we were able to get results on the training sets. On English: 75.2\% (\(\ell \leq 10\)), 64.2\% (\(\ell \leq 20\)), 56.7\% (\(\ell \leq 40\)). On German: 57.8\% (\(\ell \leq 10\)), 45.0\% (\(\ell \leq 20\)), and 39.9\% (\(\ell \leq 40\)). On Chinese: 56.6\% (\(\ell \leq 10\)), 45.1\% (\(\ell \leq 20\)), and 38.9\% (\(\ell \leq 40\)).

Thus, while Seginer’s method performs better on English, our approach performs 2-3 points better on German, and both methods give similar performance on Chinese.

\(^5\)We used the implementation available at http://tinyurl.com/lhwk5n6.
8.5 Conclusion

We described a spectral approach for unsupervised constituent parsing that comes with theoretical guarantees on latent structure recovery. Empirically, our algorithm performs favorably to the CCM of Klein and Manning (2002) without the need for careful initialization.

The tree metric used in this work is closely related to that used in Chapter 6 except it is finite dimensional. However, the meta-algorithm used to learn the tree structure is rather different. Because of the domain, we are able to restrict the space of tree structures to those that are projective, thus enabling us to find the minimum cost tree. In Chapter 6 however, we did not have any restrictions on the tree structure, and thus solving for the minimum cost tree is NP-hard. Therefore in Chapter 6 we resorted to using neighbor joining (Saitou and Nei, 1987), which is consistent but does not have any guarantees when the model assumptions do not hold (since it is greedy). On the other hand, our meta-algorithm will find the minimum cost tree regardless of whether the data was generated by the assumed model or not.

8.6 Appendix

8.6.1 Path Additivity

We first prove that our proposed tree metric is path additive based on the proof technique of Lemma 15 Chapter 6.

**Lemma 23.** If Assumption 1 holds then, $d_{\text{spectral}}$ is an additive metric.

**Proof.** For conciseness, we simply prove the property for paths of length 2. The proof for more general cases follows similarly (e.g. see (Anandkumar et al., 2011)).

First note that the relationship between eigenvalues and singular values allows us to rewrite the distance metric as

$$d_{\text{spectral}}(i, j) = -\frac{1}{2} \log \Lambda_m(\Sigma_x(i, j)\Sigma_x(i, j)^\top) + \frac{1}{4} \log \Lambda_m(\Sigma_x(i, i)\Sigma_x(i, i)^\top) + \frac{1}{4} \log \Lambda_m(\Sigma_x(j, j)\Sigma_x(j, j)^\top)$$

(8.7)

Furthermore, $\Sigma_x(i, j)\Sigma_x(i, j)^\top$ is rank $m$ by Assumption 1 and the conditional independence statements implied by the latent tree model. Thus $\Lambda_m(\Sigma_x(i, j)\Sigma_x(i, j)^\top)$ is equivalent to taking the pseudo-determinant $|\cdot|_+$ of $(\Sigma_x(i, j)\Sigma_x(i, j)^\top)$, which is the product of the non-zero eigenvalues. The pseudo-determinant can be alternatively defined in the following limit form:

$$|B|_+ = \lim_{\alpha \to 0} \frac{|B + \alpha I|}{\alpha^{p-m}}$$

(8.8)

where $B$ is a $p \times p$ matrix of rank $m$ and $I$ is the $p \times p$ identity and $|\cdot|$ equals the standard determinant. Note that if $m = p$ then $|B|_+ = |B|$.
Thus, our distance metric can be further rewritten as:

\[
d^\text{spectral}(i, j) = -\frac{1}{2} \log |\Sigma_x(i, j)\Sigma_x(i, j)^T|_+ + \frac{1}{4} \log |\Sigma_x(i, i)\Sigma_x(i, i)^T|_+ + \frac{1}{4} \log |\Sigma_x(j, j)\Sigma_x(j, j)^T|_+ \tag{8.9}
\]

We now proceed with the proof. For paths of length 2, \(i - j - k\), there are three cases:

- \(i \leftarrow j \rightarrow k\)
- \(i \leftarrow j \leftarrow k\)
- \(i \rightarrow j \rightarrow k\)

**Case 1: \((i \leftarrow j \rightarrow k)\)**     Note that in this case \(j\) must be latent but \(i\) and \(k\) can be either observed or latent. We assume below that both \(i\) and \(k\) are observed. The same proof strategy works for the other cases.

Due to Assumption 1,

\[
\Sigma_x(i, k) = \mathbb{E}[v_i v_k^T | x] = C_{i|x} \Sigma_x(j, j) C_{j|x}^T \tag{8.10}
\]

Thus,

\[
\Sigma_x(i, k)\Sigma_x(i, k)^T = C_{i|x} \Sigma_x(j, j) C_{j|x}^T C_{k|x} \Sigma_x(j, j) C_{k|x}^T \tag{8.11}
\]

Combining this definition with Sylvester’s Determinant Theorem (Akritas et al., 1996), gives us that:

\[
|\Sigma_x(i, k)\Sigma_x(i, k)^T|_+ = |C_{i|x} \Sigma_x(j, j) C_{j|x}^T C_{k|x} \Sigma_x(j, j) C_{k|x}^T|_+ = |C_{i|x}^T C_{j|x} \Sigma_x(j, j) C_{k|x}^T C_{k|x} \Sigma_x(j, j)|_+ \tag{8.12}
\]

(i.e. we can move \(C_{k|x}^T\) the to the front).

Now \(C_{i|x}^T C_{j|x} \Sigma_x(j, j) C_{k|x}^T C_{k|x} \Sigma_x(j, j)\) is \(m \times m\) and has rank \(m\). Thus, the pseudo-determinant equals the normal determinant in this case. Using the fact that \(|AB| = |A||B|\) if \(A\) and \(B\) are square, we get

\[\text{although the additive distance metric is undirected, } A \text{ and } C \text{ in Assumption 1 are defined with respect to parent-child relationships, so we must consider direction}\]
we get

$$|\Sigma_x(i, k)\Sigma_x(i, k)^\top|_+ = |C_{i|j, x}^TC_{i|j, x}^\top\Sigma_x(j, j)C_{k|j, x}^TC_{k|j, x}^\top\Sigma_x(j, j)|$$

$$= |C_{i|j, x}^TC_{i|j, x}^\top\Sigma_x(j, j)||C_{k|j, x}^TC_{k|j, x}^\top\Sigma_x(j, j)|$$

$$= \frac{|\Sigma_x(j, j)||C_{i|j, x}^TC_{i|j, x}^\top\Sigma_x(j, j)|}{|\Sigma_x(j, j)|} \times \frac{|\Sigma_x(j, j)||C_{k|j, x}^TC_{k|j, x}^\top\Sigma_x(j, j)|}{|\Sigma_x(j, j)|}$$

(8.13)

Furthermore, note that

$$|\Sigma_x(j, j)C_{i|j, x}^TC_{i|j, x}^\top\Sigma_x(j, j)| = |C_{i|j, x}C_{i|j, x}(j, j)C_{i|j, x}^\top|_+ = |\Sigma_x(i, j)\Sigma_x(i, j)^\top|_+$$

(8.14)

This gives,

$$|\Sigma_x(i, k)\Sigma_x(i, k)^\top|_+ = \frac{|\Sigma_x(i, j)\Sigma_x(i, j)^\top|_+}{|\Sigma_x(j, j)|} \times \frac{|\Sigma_x(k, j)\Sigma_x(k, j)^\top|_+}{|\Sigma_x(j, j)|}$$

(8.15)

Substituting back into Eq. (8.9) proves that

$$d^{\text{spectral}}(i, k) = -\frac{1}{2} \log |\Sigma_x(i, j)\Sigma_x(i, j)^\top|_+ - \frac{1}{2} \log |\Sigma_x(j, k)\Sigma_x(j, k)^\top|_+ + \frac{1}{2} \log |\Sigma_x(j, j)\Sigma_x(j, j)^\top|_+$$

$$+ \frac{1}{2} \log |\Sigma_x(i, i)\Sigma_x(i, i)^\top|_+ + \frac{1}{2} \log |\Sigma_x(k, k)\Sigma_x(k, k)^\top|_+$$

$$= d^{\text{spectral}}(i, j) + d^{\text{spectral}}(j, k)$$

(8.16)

**Case 2: i ← j ← k** The proof is similar to above. Here since only leaf nodes can be observed, j and k must be latent but i can be either observed or latent. We assume it is observed, the latent case follows similarly.

Due to Assumption 1,

$$\Sigma_x(i, k) = \mathbb{E}[\nu_i\nu_k^\top|x] = C_{i|k, x}A_{j|k, x}\Sigma_x(k, k)$$

(8.17)

Thus, via Sylvester’s Determinant Theorem as before,

$$|\Sigma_x(i, k)\Sigma_x(i, k)^\top|_+ = |C_{i|j, x}^TC_{i|j, x}^\top\Sigma_x(k, k)\Sigma_x(k, k)\Sigma_x(k, k)^\topC_{i|k, x}^TC_{i|k, x}^\top|_+$$

$$= |C_{i|j, x}^TC_{i|j, x}^\top\Sigma_x(k, k)\Sigma_x(k, k)^\topC_{i|k, x}^TC_{i|k, x}^\top|_+$$

(8.18)

Now $C_{i|j, x}^TC_{i|j, x}A_{j|k, x}\Sigma_x(k, k)\Sigma_x(k, k)^\topA_{j|k, x}^\top$ is $m \times m$ and has rank $m$. Thus, the pseudo-determinant equals the normal determinant in this case. Using the fact that $|AB| = |A||B|$ if $A$ and $B$ are square, we get.
\[
|\Sigma_x(i, k)\Sigma_x(i, k)^\top|_+ = \left| C_{ij, x}^{\top} C_{ij, x} A_{jk, x} \Sigma_x(k, k) \Sigma_x(k, k) A_{jk, x}^{\top} \right|
= \left| C_{ij, x}^{\top} C_{ij, x} \Sigma_x(k, k) \Sigma_x(k, k) A_{jk, x}^{\top} \right|
= \left| C_{ij, x}^{\top} C_{ij, x} \Sigma_x(j, k) \Sigma_x(j, k)^\top \right|
= \frac{\left| \Sigma_x(j, j) \right||C_{ij, x}^{\top} C_{ij, x}| |\Sigma_x(j, j)|}{|\Sigma_x(j, j)|||\Sigma_x(j, j)|} \times |\Sigma_x(j, k)\Sigma_x(j, k)^\top|
\]

Furthermore, note that
\[
|\Sigma_x(j, j)C_{ij, x}^{\top} C_{ij, x} \Sigma_x(j, j)| = |C_{ij, x}^{\top} \Sigma_x(j, j) \Sigma_x(j, j) C_{ij, x}|_+
= |\Sigma_x(i, j)\Sigma_x(i, j)^\top|_+
\]

This gives,
\[
|\Sigma_x(i, k)\Sigma_x(i, k)^\top|_+ = \frac{\left| \Sigma_x(i, j)\Sigma_x(i, j)^\top \right|_+}{|\Sigma_x(j, j)||\Sigma_x(j, j)|} \times |\Sigma_x(j, k)\Sigma_x(j, k)^\top|_+
\]

Substituting back into Eq. (8.9) proves that
\[
d_{\text{spectral}}(i, k) = -\frac{1}{2} \log |\Sigma_x(i, j)\Sigma_x(i, j)^\top|_+ - \frac{1}{2} \log |\Sigma_x(j, k)\Sigma_x(j, k)^\top|_+ + \frac{1}{2} \log |\Sigma_x(j, j)\Sigma_x(j, j)^\top|_+ (8.23)
\]
\[
+ \frac{1}{4} \log |\Sigma_x(i, i)\Sigma_x(i, i)^\top|_+ + \frac{1}{4} \log |\Sigma_x(k, k)\Sigma_x(k, k)^\top|_+ (8.24)
\]
\[
= d_{\text{spectral}}(i, j) + d_{\text{spectral}}(j, k) (8.25)
\]

Case 3: \(i \rightarrow j \rightarrow k\) The same argument as case 2 holds here.

\[
\square
\]

8.6.2 Theoretical Guarantees For Algorithm 11

Our main theoretical guarantee is that the learning Algorithm 11 will recover the correct undirected tree \(u \in \mathcal{U}\) with high probability, if the given top bracket is correct and if we obtain a sufficient number of examples \((u^{(i)}, x^{(i)})\) being generated from the model in §2.

Our kernel is controlled by its “bandwidth” \(\gamma\), a typical kernel parameter that appears when using kernel smoothing. The larger this positive number is, the more inclusive the kernel will be with respect to examples in \(\mathcal{D}\) in order to estimate a given covariance matrix.

In order for the learning algorithm to be consistent the kernel must be able to effectively manage
the bias-variance trade-off with the bandwidth parameter. Intuitively, if the sample size is small, then the bandwidth should be relatively large to control the variance. As the sample size increases, the bandwidth should be decreased at a certain rate to reduce the bias.

In kernel density estimation in $\mathbb{R}^d$, one can put smoothness assumptions on the density being estimated, such as bounded second derivatives. With these conditions, it can be shown that setting $\gamma = O(N^{-1/5})$ will optimally trade-off the bias and the variance in a way that leads to a consistent estimator.

However, our space of possible sequences is discrete and thus it is much more difficult to define these analogous smoothness conditions. Therefore, giving an asymptotic rate to set the bandwidth as a function of the sample size is difficult. To solve this issue, we consider a specific theoretical kernel where the bandwidth directly relates to the bias of the estimator. Define the following $\gamma$-ball

$$B_{j,k,x}(\gamma) = \{ (j', k', x') : \| \Sigma_x(j', k') - \Sigma_x(j, k) \|_F \leq \gamma \}$$

where $\|A\|_F$ for a matrix $A$ is the Frobenius norm of that matrix, i.e.: $\|A\|_F = \sqrt{\sum_{jk} A_{jk}^2}$.

We then define the following kernel:

$$K_\gamma(j, k, j', k'|x, x') = \begin{cases} 
1 & (j', k', x') \in B_{j,k,x}(\gamma) \\
0 & (j', k', x') \not\in B_{j,k,x}(\gamma) 
\end{cases}$$

Furthermore, to quantify the expected effective sample size using kernel smoothing we define the following quantity:

$$\nu_{j,k,x}(\gamma) = \left( \sum_{x' \in \mathcal{X}} \frac{p(x')}{l(x')^2} \sum_{j', k' \in \ell(x')} K_\gamma(j, k, j', k'|x, x') \right)$$

where $p(x)$ the prior distribution over tag sequences. Let $\nu_x(\gamma) = \min_{j,k} \nu_{j,k,x}(\gamma)$. Intuitively, $\nu_{j,k,x}(\gamma)$ represents the probability of finding contexts that are “similar” to $(j, k, x)$ as defined by the kernel. Consequently, $N\nu_x(\gamma)$ represents a lower bound on the expected effective sample size for tag sequence $x$.

Denote $\sigma_x(j, k)^{(m)}$ as the $r^{th}$ singular value of $\Sigma_x(j, k)$. Let $\sigma^*(x) := \min_{j,k \in \ell(x)} \min \left( \sigma_x(j, k)^{(m)} \right)$. Finally, define $\phi$ as the difference of the maximum possible entry in $w$ and the minimum possible entry in $w$ (i.e. we assume that the embeddings are bounded vectors).

Using the above definitions and leveraging the proof technique in (Zhou et al., 2010), we establish that our strategy will result in a consistent estimation of the word-word distance subblock $D_{WW}$.
Lemma 24. Assume the kernel used is that in Eq. (8.27) with bandwidth \( \gamma = \frac{C_1 \epsilon \sigma^*(x)}{m} \) and that

\[
N \geq \frac{m^2 \phi^2}{C_3 \epsilon^2 \sigma^*(x)^2 \nu_x(\gamma)^2} \log \left( \frac{C_2 p^2 \ell(x)^2}{\delta} \right) \tag{8.30}
\]

Then, for a fixed tag sequence \( x \), and any \( \epsilon < \frac{\sigma^*(x)}{2} \):

\[
|d^{\text{spectral}}(j, k) - d^{\text{spectral}}(j, k)| \leq \epsilon \quad \forall j \neq k \in \ell(x) \tag{8.31}
\]

with probability \( 1 - \delta \).

Then we have the following theorem.

Theorem 6. Define

\[
\Delta(x) := \frac{\min_{u' \in U: u' \neq u(x)} (c(u(x)) - c(u'))}{8|\ell(x)|} \tag{8.32}
\]

where \( u(x) \) is the correct tree and \( u' \) is any other tree.

Let \( \hat{u} \) be the estimated tree for tag sequence \( x \). Assume that the kernel in Eq. (8.27) is used with bandwidth \( \gamma = \frac{\Delta(x) \sigma^*(x)}{C_4 m} \) and that

\[
N \geq \frac{C_5 m^2 \phi^2}{\min(\sigma^*(x)^2 \Delta(x)^2, \sigma^*(x)^2 \nu_x(\gamma)^2)} \log \left( \frac{C_2 p^2 \ell(x)^2}{\delta} \right) \tag{8.33}
\]

Then with probability \( 1 - \delta \), \( \hat{u} = u(x) \).

Proof. By Lemma 1, and this choice of \( N \) and \( \gamma \),

\[
|d^{\text{spectral}}(j, k) - d^{\text{spectral}}(j, k)| \leq \Delta(x) \quad \forall j \neq k \in \ell(x) \tag{8.34}
\]

Now for any tree \( u \in U \), we can compute \( c(u) \) by summing over the distances \( d(i, j) \) where \( i \) and \( j \) are adjacent in \( u \).

Recall the formulas for \( d(i, j) \) (from the main chapter):

- leaf edge:

\[
d(i, j) = \frac{1}{2} (d(j, a^*) + d(j, b^*) - d(a^*, b^*)) \tag{8.35}
\]

- internal edge:

\[
d(i, j) = \frac{1}{4} (d(a^*, g^*) + d(a^*, h^*) + d(b^*, g^*) + d(b^*, h^*) - 2d(a^*, b^*) - 2d(g^*, h^*)) \tag{8.36}
\]
where \(a^*, b^*, g^*, h^*\) are leaves (i.e. word nodes). Thus, Eq. (8.34) implies that

\[
|\hat{d}(i, j) - d(i, j)| \leq 2\Delta(x) \quad \forall j \neq k \in [M] \tag{8.37}
\]

Since there are \(\leq 2|\ell(x)|\) edges in the tree, this is sufficient to guarantee that \(|\hat{c}(u) - c(u)| \leq 4|\ell(x)|\Delta(x) \forall u \in \mathcal{U}\). Thus,

\[
\hat{c}(u(x)) - \hat{c}(u') < 0 \quad \forall u' \in \mathcal{U} : u' \neq u(x) \tag{8.38}
\]

Thus, the correct tree is the one with minimum estimated cost and Algorithm 1 will return the correct tree. \(\square\)

### 8.7 Lemma Proofs

Instead of proving Lemma 1 directly, we divide it into two stages. First, we show that our strategy yields consistent estimation of the covariance matrices (Lemma 25). We then show that a concentration bound on the covariance matrices implies that the empirical distance is close to the true distance (Lemma 26). Putting the results together proves Lemma 24.

**Lemma 25.** Assume the kernel used is that in Eq. (8.27) with bandwidth \(\gamma = \epsilon/2\) and that

\[
N \geq \frac{\log(C\nu^2\ell(x)^2\delta^2)}{C_2\epsilon^2\nu_x(\epsilon/2)^2} \tag{8.39}
\]

Then for a fixed tag sequence \(x\) and \(\forall j, k \in \ell(x)\),

\[
\|\hat{\Sigma}_x(j, k) - \tilde{\Sigma}_x(j, k)\|_F \leq \epsilon \tag{8.40}
\]

with probability \(1 - \delta\).

**Proof.** Define the quantity,

\[
\tilde{\Sigma}_x(j, k) = \frac{\sum_{i=1}^{M} \sum_{j', k' \in [\ell(x)_i]} K_\gamma(j, k, j', k'|x, x_i) \Sigma_x(j', k')}{\sum_{i=1}^{N} \sum_{j', k' \in [\ell(x)_i]} K_\gamma(j, k, j', k'|x, x_i)} \tag{8.41}
\]

Note that via triangle inequality,

\[
\|\hat{\Sigma}_x(j, k) - \tilde{\Sigma}_x(j, k)\|_F \leq \|\hat{\Sigma}_x(j, k) - \tilde{\Sigma}_x(j, k)\|_F + \|\tilde{\Sigma}_x(j, k) - \tilde{\Sigma}_x(j, k)\|_F \tag{8.42}
\]

The first term (the bias) is bounded by \(\epsilon/2\) using the definition of the kernel in Eq. (8.27) with bandwidth \(\epsilon/2\). The proof for bound for the second term (the variance) can be derived using the technique of (Zhou et al., 2010) and is in Utility Lemma. \(\square\)

Lemma 25 can then be used to prove that with high probability the estimated mutual information is close to the true mutual information.
Lemma 26. Assume that

\[ \|\hat{\Sigma}_x(j, k) - \Sigma_x(j, k)\|_F \leq \epsilon \leq \frac{\sigma'(x)}{2} \quad \forall j, k \in [\ell(x)] \]  

(8.43)

Then,

\[ |d^{\text{spectral}}(j, k) - \hat{d}^{\text{spectral}}(j, k)| \leq \frac{C_3 \epsilon}{\sigma'(x)} \quad \forall j \neq k \in [\ell(x)] \]  

(8.44)

Proof. By triangle inequality,

\[
|d^{\text{spectral}}(j, k) - \hat{d}^{\text{spectral}}(j, k)| \\
\leq |\log \Lambda_m(\Sigma_x(j, k)) - \log \Lambda_m(\hat{\Sigma}_x(j, k))| \\
+ \frac{1}{2}|\log \Lambda_m(\Sigma_x(j, j)) - \log \Lambda_m(\hat{\Sigma}_x(j, j))| \\
+ \frac{1}{2}|\log \Lambda_m(\hat{\Sigma}_x(k, k)) - \log \Lambda_m(\hat{\Sigma}_x(k, k))| 
\]

(8.45)

We show the bound on \( |\log \Lambda_m(\Sigma_x(j, j)) - \log \Lambda_m(\hat{\Sigma}_x(j, j))| \), the others follow similarly. By definition of \( \Lambda_m(\cdot) \), and triangle inequality we have that,

\[
|\log \Lambda_m(\Sigma_x(j, j)) - \log \Lambda_m(\hat{\Sigma}_x(j, j))| \leq \sum_{r=1}^{m} |\log(\sigma_x(j, j)^{(r)}) - \log(\hat{\sigma}_x(j, j)^{(r)})| 
\]

To translate our concentration bounds on Frobenius norm to concentration of eigenvalues, we apply Weyl’s Theorem. Assume first that \( \hat{\sigma}_x(j, j)^{(r)} \geq \sigma_x(j, j)^{(r)} \). Then,

\[
|\log(\sigma_x(j, j)^{(r)}) - \log(\hat{\sigma}_x(j, j)^{(r)})| \\
\leq |\log(\sigma_x(j, j)^{(r)}) - \log(\sigma_x(j, j)^{(r)} + \epsilon)| \\
\leq |\log(\sigma_x(j, j)^{(r)}) - \log(\sigma_x(j, j)^{(r)}(1 + \frac{\epsilon}{\sigma_x(j, j)^{(r)}}))| \\
\leq \log(1 + \frac{\epsilon}{\sigma_x(j, j)^{(r)}}) \leq \frac{\epsilon}{\sigma_x(j, j)^{(r)}} 
\]

(8.46)

Similarly, when \( \hat{\sigma}_x(j, j)^{(r)} < \sigma_x(j, j)^{(r)} \), then

\[
|\log(\sigma_x(j, j)^{(r)}) - \log(\hat{\sigma}_x(j, j)^{(r)})| \leq \frac{\epsilon}{\sigma_x(j, j)^{(r)}} = \frac{\epsilon}{\sigma_x(j, j)^{(r)} - \epsilon} 
\]

(8.47)

Using the fact that \( \epsilon \leq \frac{\sigma'(x)}{2} \) we obtain that,

\[
|\log(\sigma_x(j, j)^{(r)}) - \log(\hat{\sigma}_x(j, j)^{(r)})| \leq \frac{2\epsilon}{\sigma_x(j, j)^{(r)}} 
\]

(8.48)

As a result,

\[
|\log \Lambda_m(\Sigma_x(j, j)) - \log \Lambda_m(\hat{\Sigma}_x(j, j))| \leq m \frac{2\epsilon}{\sigma_x(j, j)^{(m)}} 
\]

(8.49)

143
Repeating this for the other terms in Eq. (8.45) gives the bound.

Utility Lemma 1.

\[ P(||\tilde{\Sigma}_x(j,k) - \Sigma_x(j,k)||_F \geq \xi) \leq C_1 p^2 \ell(x)^2 \exp\left(-\frac{C N \nu_x(\gamma)^2 \xi^2}{\phi^2}\right) \quad \forall j, k \in [\ell(x)] \]  

Proof. The proof uses the technique of Zhou et al. [2010].

Define, \( N_{x,j;k}(\gamma) = \sum_{i=1}^{N} \sum_{j^i,j^i \in [\ell(x)]} K(\gamma, j, j^i, k, x, x_i) \) which is the empirical effective sample size under the smoothing kernel.

We first show that the \( N_{x,j;k}(\gamma) \) is close to the expected effective sample size \( N \nu_x(\gamma) \).

Using Hoeffding’s Inequality, we obtain that

\[ P\left(|N_{x,j;k}(\gamma) - N \nu_x(\gamma)| \geq N \nu_x(\gamma)/2\right) \leq C \exp(-N \nu_x(\gamma)^2/2) \]  

(8.51)

Note that

\[ ||\tilde{\Sigma}_x(j,k) - \Sigma_x(j,k)||_F \leq p^2 \max_{a,b} |\tilde{\zeta}_x(j,k;a,b) - \zeta_x(j,k;a,b)| \]  

(8.52)

where \( \tilde{\zeta}_x(j,k;a,b) \) is the element on the \( a^th \) row and \( b^th \) column of \( \tilde{\Sigma}_x(j,k) \). Thus, it suffices to bound \( |\tilde{\zeta}_x(j,k;a,b) - \zeta_x(j,k;a,b)| \).

Define the boolean variable \( E = \mathbb{I}[N_{x,j;k}(\gamma) > N \nu_x(\gamma)/2] \). Then,

\[ P(||\tilde{\zeta}_x(j,k;a,b) - \zeta_x(j,k;a,b)||_F \geq \xi) = P(||\tilde{\zeta}_x(j,k;a,b) - \zeta_x(j,k;a,b)||_F \geq \xi | E = 1) P(E = 1) \]

\[ + P(||\tilde{\zeta}_x(j,k;a,b) - \zeta_x(j,k;a,b)||_F \geq \xi | E = 0) P(E = 0) \]

\[ \leq P(||\tilde{\zeta}_x(j,k;a,b) - \zeta_x(j,k;a,b)||_F \geq \xi | E = 1) + P(E = 0) \]  

(8.53)

The second term is bounded by Eq. (8.51). We prove the first term below.

For shorthand, refer to \( P(||\tilde{\zeta}_x(j,k;a,b) - \zeta_x(j,k;a,b)||_F \geq \xi | E = 1) \) as \( P_E(||\tilde{\zeta}_x(j,k;a,b) - \zeta_x(j,k;a,b)||_F \geq \xi | E = 1) \). Define the following quantities:

For convenience define the following quantity, where \( w_{j,a}^{(i)} \) is the \( a^th \) element of the vector \( w_j^{(i)} \)

\[ \delta^{(i)}_{x}(j,k;a,b) = w_{j,a}^{(i)}(w_{k,b}^{(i)})^\top - \zeta_x^{(i)}(j,k;a,b) \]  

(8.54)
For every $\kappa > 0$, by Markov’s inequality,

$$
P_E \left( \sum_{i \in [N]} \sum_{j', k' \in \{\chi^{(i)}\}} K_\gamma(j, k, j', k' | x, x^{(i)}) \delta_{\chi^{(i)}}(j, k; a, b) > N_{x; j;k}(\gamma) \xi \right)
$$

$$
= P_E \left( \exp \left( \kappa \sum_{i \in [N]} \sum_{j', k' \in \{\chi^{(i)}\}} K_\gamma(j, k, j', k' | x, x^{(i)}) \delta_{\chi^{(i)}}(j, k; a, b) \right) > \exp \left( \kappa N_{x; j;k}(\gamma) \xi \right) \right)
$$

$$
\leq \mathbb{E} \left[ \exp \left( \kappa \sum_{i \in [N]} \sum_{j', k' \in \{\chi^{(i)}\}} K_\gamma(j, k, j', k' | x, x^{(i)}) \delta_{\chi^{(i)}}(j, k; a, b) \right) \right] \exp \left( \kappa N_{x; j;k}(\gamma) \xi \right) \quad (8.55)
$$

We now bound the numerator. Using the fact that the samples are independent, we have that

$$
\mathbb{E} \left[ \exp \left( \kappa \sum_{i \in [N]} \sum_{j', k' \in \{\chi^{(i)}\}} K_\gamma(j, k, j', k' | x, x^{(i)}) \delta_{\chi^{(i)}}(j, k; a, b) \right) \right] \quad (8.56)
$$

$$
= \prod_{i \in [N]} \prod_{j', k' \in \{\chi^{(i)}\}} \mathbb{E} \left[ \exp \left( \kappa K_\gamma(j, k, j', k' | x, x^{(i)}) \delta_{\chi^{(i)}}(j, k; a, b) \right) \right] \quad (8.57)
$$

$$
= \prod_{i \in [N]} \prod_{j', k' \in \{\chi^{(i)}\}} \mathbb{I}(j', k', x^{(i)}) \in B_\gamma(j, k, x) \mathbb{E} [\exp(\kappa \delta_{\chi^{(i)}}(j, k; a, b))] \quad (8.58)
$$

$$
= \prod_{i \in [N]} \prod_{j', k' \in \{\chi^{(i)}\}} \mathbb{I}(j', k', x^{(i)}) \in B_\gamma(j, k, x) \exp(\kappa^2 \phi^2 / 8) \quad (8.59)
$$

From Utility Lemma\(^2\) we can conclude that

$$
\prod_{i \in [N]} \prod_{j', k' \in \{\chi^{(i)}\}} \mathbb{I}(j', k', x^{(i)}) \in B_\gamma(j, k, x) \mathbb{E} [\exp(\kappa \delta_{\chi^{(i)}}(j, k; a, b))] \quad (8.60)
$$

$$
\leq \prod_{i \in [N]} \prod_{j', k' \in \{\chi^{(i)}\}} \mathbb{I}(j', k', x^{(i)}) \in B_\gamma(j, k, x) \exp(\kappa^2 \phi^2 / 8) \quad (8.61)
$$

$$
\leq \exp(N_{x; j;k}(\gamma) \kappa^2 \phi^2 / 8) \quad (8.62)
$$

$$
P_E \left( \sum_{i \in [N]} \sum_{j', k' \in \{\chi^{(i)}\}} K_\gamma(j, k, j', k' | x, x^{(i)}) \delta_{\chi^{(i)}}(j, k; a, b) > N_{x; j;k}(\gamma) \xi \right) \quad (8.63)
$$

$$
\leq \exp(-\kappa N_{x; j;k}(\gamma) \xi) \exp(N_{x; j;k}(\gamma) \kappa^2 \phi^2 / 8) \quad (8.64)
$$
Setting $\kappa = \frac{4\xi}{\phi^2}$, gives us that

$$P_E \left( \sum_{i \in [N]} \sum_{j', k' \in \ell(x)} K_y(j, k, j', k' | x, x') \delta_x(j, k; a, b) > N_{x, j, k}(\gamma) \xi \right)$$

(8.65)

$$\leq \exp(-2N_{x, j, k}(\gamma) \xi^2 / \phi^2)$$

(8.66)

$$\leq \exp(-N_{x}(\gamma) \xi^2 / \phi^2)$$

(8.67)

Combining the above with Eq. (8.51) and taking some union bounds over $a, b, j, k$ proves the lemma. $\square$

**Helper Lemmas**

We make use of the following helper lemma that is standard in the proof of Hoeffding’s Inequality, e.g. [Casella and Berger, 1990].

**Utility Lemma 2.** Suppose that $\mathbb{E}(X) = 0$ and that $a \leq X \leq b$. Then

$$\mathbb{E}[e^{tX}] \leq e^{t^2(b-a)^2/8}$$

(8.68)
Chapter 9

Language Modeling via Power Low Rank Ensembles

Finally, we propose to use the linear algebra point of view of probabilistic modeling to tackle language modeling, a classic challenge in NLP with many applications such as machine translation and speech recognition.

Contribution of this chapter: We present power low rank ensembles (PLRE), a flexible framework for n-gram language modeling consisting of ensembles of low rank matrices and tensors. Our method can be understood as a generalization of n-gram modeling to non-integer n, and includes standard techniques such as absolute discounting and Kneser-Ney smoothing as special cases. PLRE training is efficient and our approach outperforms state-of-the-art modified Kneser Ney baselines in terms of perplexity on large corpora as well as on BLEU score in a downstream machine translation task.

Outline: We first present an introduction (§9.1) and then review existing n-gram smoothing methods (§9.2). We then present the intuition behind the key components of our technique: rank (§9.3.1) and power (§9.3.2). We then show how these can be interpolated into an ensemble (§9.4). An experimental evaluation on English and Russian corpora (§9.5) is then presented.

Prerequisites: This chapter assumes a general understanding of the connection between latent variable models and low rank factorization in Chapter 3.

9.1 Introduction

Language modeling is the task of estimating the probability of sequences of words in a language and is an important component in, among other applications, automatic speech recognition (Rabiner and Juang, 1993) and machine translation (Koehn, 2010). The predominant approach to language modeling is the n-gram model, wherein the probability of a word sequence $P(w_1, ..., w_\ell)$ is decomposed using the chain rule, and then a Markov assumption is made: $P(w_1, ..., w_\ell) \approx \prod_{i=1}^\ell P(w_i|w_{i-n+1})$. While this assumption substantially reduces the modeling com-
plexity, parameter estimation remains a major challenge. Due to the power-law nature of language (Zipf, 1949), the maximum likelihood estimator massively overestimates the probability of rare events and assigns zero probability to legitimate word sequences that happen not to have been observed in the training data (Manning and Schütze, 1999).

Many smoothing techniques have been proposed to address the estimation challenge. These reassign probability mass (generally from over-estimated events) to unseen word sequences, whose probabilities are estimated by interpolating with or backing off to lower order n-gram models (Chen and Goodman, 1999).

Somewhat surprisingly, these widely used smoothing techniques differ substantially from techniques for coping with data sparsity in other domains, such as collaborative filtering (Koren et al., 2009; Su and Khoshgoftaar, 2009) or matrix completion (Candes and Recht, 2009; Cai et al., 2010). In these areas, low rank approaches based on matrix factorization play a central role (Lee and Seung, 2001; Salakhutdinov and Mnih, 2008; Mackey et al., 2011). For example, in recommender systems, a key challenge is dealing with the sparsity of ratings from a single user, since typical users will have rated only a few items. By projecting the low rank representation of a user’s (sparse) preferences into the original space, an estimate of ratings for new items is obtained. These methods are attractive due to their computational efficiency and mathematical well-foundedness.

In this chapter, we introduce **power low rank ensembles** (PLRE), in which low rank tensors are used to produce smoothed estimates for n-gram probabilities. Ideally, we would like the low rank structures to discover semantic and syntactic relatedness among words and n-grams, which are used to produce smoothed estimates for word sequence probabilities.

In contrast to the few previous low rank language modeling approaches, PLRE is not orthogonal to n-gram models, but rather a general framework where existing n-gram smoothing methods such as Kneser-Ney smoothing are special cases. A key insight is that PLRE does not compute low rank approximations of the original joint count matrices (in the case of bigrams) or tensors i.e. multi-way arrays (in the case of 3-grams and above), but instead altered quantities of these counts based on an element-wise power operation, similar to how some smoothing methods modify their lower order distributions.

Moreover, PLRE has two key aspects that lead to easy scalability for large corpora and vocabularies. First, since it utilizes the original n-grams, the ranks required for the low rank matrices and tensors tend to be remain tractable (e.g. around 100 for a vocabulary size $V \approx 1 \times 10^6$) leading to fast training times. This differentiates our approach over other methods that leverage an underlying latent space such as neural networks (Bengio et al., 2003; Mnih and Hinton, 2007; Mikolov et al., 2010) or soft-class models (Saul and Pereira, 1997) where the underlying dimension is required to be quite large to obtain good performance. Moreover, at test time, the probability of a sequence can be queried in time $O(\kappa_{\text{max}})$ where $\kappa_{\text{max}}$ is the maximum rank of the low rank matrices/tensors used. While this is larger than Kneser Ney’s virtually constant query time, it is substantially faster than conditional exponential family models (Chen and Rosenfeld, 2000; Chen, 2009; Nelakanti et al., 2013) and neural networks which require $O(V)$ for exact computation of the normalization constant. See Section 9.7 for a more detailed discussion of related work.
9.2 Discount-based Smoothing

We first provide background on absolute discounting (Ney et al., 1994) and Kneser-Ney smoothing (Kneser and Ney, 1995), two common n-gram smoothing methods. Both methods can be formulated as back-off or interpolated models; we describe the latter here since that is the basis of our low rank approach.

9.2.1 Notation

Let \( c(w) \) be the count of word \( w \), and similarly \( c(w, w_{i-1}) \) for the joint count of words \( w \) and \( w_{i-1} \). For shorthand we will define \( w_i^j \) to denote the word sequence \( \{w_i, w_{i+1}, \ldots, w_{j-1}, w_j\} \). Let \( \hat{P}(w) \) refer to the maximum likelihood estimate (MLE) of the probability of word \( w_i \), and similarly \( \hat{P}(w_i | w_{i-1}) \) for the probability conditioned on a history, or more generally, \( \hat{P}(w_i | w_{i-n+1}) \).

Let \( N_-(w_i) := |\{w : c(w_i, w) > 0\}| \) be the number of distinct words that appear before \( w_i \). More generally, let \( N_-(w_i^j_{i-n+1}) = |\{w : c(w_i^j_{i-n+1}, w) > 0\}| \). Similarly, let \( N_+(w_{i-n+1}^{i-1}) = |\{w : c(w, w_{i-n+1}^{i-1}) > 0\}| \).

\( V \) denotes the vocabulary size.

9.2.2 Absolute Discounting

Absolute discounting works on the idea of interpolating higher order n-gram models with lower-order n-gram models. However, first some probability mass must be “subtracted” from the higher order n-grams so that the leftover probability can be allocated to the lower order n-grams. More specifically, define the following discounted conditional probability:

\[
\hat{P}_D(w_i | w_{i-n+1}^{i-1}) = \max\{c(w_i, w_{i-n+1}^{i-1}) - D, 0\} \left/ \frac{c(w_{i-n+1}^{i-1})}{c(w_{i-n+1}^{i-1})} \right.
\]

Then absolute discounting \( P_{abs}(\cdot) \) uses the following (recursive) equation:

\[
P_{abs}(w_i | w_{i-n+1}^{i-1}) = \hat{P}_D(w_i | w_{i-n+1}^{i-1}) + \gamma(w_{i-n+1}^{i-1}) P_{abs}(w_i | w_{i-n+2}^{i-1})
\]

where \( \gamma(w_{i-n+1}^{i-1}) \) is the leftover weight (due to the discounting) that is chosen so that the conditional distribution sums to one: \( \gamma(w_{i-n+1}^{i-1}) = \frac{D}{c(w_{i-n+1}^{i-1})} N_+(w_{i-n+1}^{i-1}) \). For the base case, we set \( P_{abs}(w_i) = \hat{P}(w_i) \).

Discontinuity: Note that if \( c(w_{i-n+1}^{i-1}) = 0 \), then \( \gamma(w_{i-n+1}^{i-1}) = 0 \), in which case \( \gamma(w_{i-n+1}^{i-1}) \) is set to 1. We will see that this discontinuity appears in PLRE as well.
9.2.3 Kneser Ney Smoothing

Ideally, the smoothed probability should preserve the observed unigram distribution:

\[ \hat{P}(w_i) = \sum_{w_{i-n+1}} P_{sm}(w_i|w_{i-n+1}) \hat{P}(w_{i-n+1}) \]  

(9.3)

where \( P_{sm}(w_i|w_{i-n+1}) \) is the smoothed conditional probability that a model outputs. Unfortunately, absolute discounting does not satisfy this property, since it exclusively uses the unaltered MLE unigram model as its lower order model. In practice, the lower order distribution is only utilized when we are unsure about the higher order distribution (i.e., when \( \gamma() \) is large). Therefore, the unigram model should be altered to condition on this fact.

This is the inspiration behind Kneser-Ney (KN) smoothing, an elegant algorithm with robust performance in \( n \)-gram language modeling. KN smoothing defines alternate probabilities \( P_{alt}(\cdot) \):

\[
P_{alt}(w_i|w_{i-n+1}) = \begin{cases} \hat{P}(w_i|w_{i-n+1}), & \text{if } n' = n \\ \frac{\max\{N_i(w_{i-n+1})\} - D, 0\} }{\sum_{i} N_i(w_{i-n+1})}, & \text{if } n' < n \\ \end{cases}
\]

(9.4)

The base case for unigrams reduces to \( P_{alt}(w_i) = \frac{N_i(w_i)}{\sum_{i} N_i(w_i)} \). Intuitively \( P_{alt}(w_i) \) is proportional to the number of unique words that precede \( w_i \). Thus, words that appear in many different contexts will be given higher weight than words that consistently appear after only a few contexts. These alternate distributions are then used with absolute discounting:

\[
P_{kn}(w_i|w_{i-n+1}) = P_{alt}(w_i|w_{i-n+1}) + \gamma(w_{i-n+1})P_{kn}(w_i|w_{i-n+2})
\]

(9.5)

where we set \( P_{kn}(w_i) = P_{alt}(w_i) \). By definition, KN smoothing satisfies the marginal constraint in Eq. 9.3 (Kneser and Ney, 1995).

9.3 Power Low Rank Ensembles

In \( n \)-gram smoothing methods, if a bigram count \( c(w_i, w_{i-1}) \) is zero, the unigram probabilities are used, which is equivalent to assuming that \( w_i \) and \( w_{i-1} \) are independent (and similarly for general \( n \)). However, in this situation, instead of backing off to a 1-gram, we may like to back off to a “1.5-gram” or more generally an order between 1 and 2 that captures a coarser level of dependence between \( w_i \) and \( w_{i-1} \) and does not assume full independence.

Inspired by this intuition, our strategy is to construct an ensemble of matrices and tensors that not only consists of MLE-based count information, but also contains quantities that represent levels of dependence in-between the various orders in the model. We call these combinations power low rank ensembles (PLRE), and they can be thought of as \( n \)-gram models with non-integer \( n \). Our approach can be recursively formulated as:
\[ P_{\text{plre}}(w_i|w_{i-n+1}) = P_{\text{alt}}^D(w_i|w_{i-n+1}) \]
\[ + \gamma_0(w_{i-n+1}^{-1}) \left( Z_{D_1}(w_i|w_{i-n+1}^{-1}) + \ldots + \gamma_{\eta-1}(w_{i-n+1}^{-1}) \left( Z_{D_\eta}(w_i|w_{i-n+1}) + \gamma_{\eta}(w_{i-n+1}) \left( P_{\text{plre}}(w_i|w_{i-n+2}) \right) \right) \right) \]
\[(9.6)\]

where \( Z_1, \ldots, Z_\eta \) are conditional probability matrices that represent the intermediate \( n \)-gram orders and \( D \) is a discount function (specified in §9.4).

This formulation begs answers to a few critical questions. How to construct matrices that represent conditional probabilities for intermediate \( n \)? How to transform them in a way that generalizes the altered lower order distributions in KN smoothing? How to combine these matrices such that the marginal constraint in Eq. 9.3 still holds? The following propose solutions to these three queries:

1. **Rank** (Section 9.3.1): Rank gives us a concrete measurement of the dependence between \( w_i \) and \( w_{i-1} \). By constructing low rank approximations of the bigram count matrix and higher-order count tensors, we obtain matrices that represent coarser dependencies, with a rank one approximation implying that the variables are independent.

2. **Power** (Section 9.3.2): In KN smoothing, the lower order distributions are not the original counts but rather altered estimates. We propose a continuous generalization of this alteration by taking the element-wise power of the counts.

3. **Creating the Ensemble** (Section 9.4): Lastly, PLRE also defines a way to interpolate the specifically constructed intermediate \( n \)-gram matrices. Unfortunately a constant discount, as presented in Section 9.2, will not in general preserve the lower order marginal constraint (Eq. 9.3). We propose a generalized discounting scheme to ensure the constraint holds.

### 9.3.1 Rank

We first show how rank can be utilized to construct quantities between an \( n \)-gram and an \( n-1 \)-gram. In general, we think of an \( n \)-gram as an \( n \)-th order tensor i.e. a multi-way array with \( n \) indices \([i_1, \ldots, i_n]\). (A vector is a tensor of order 1, a matrix is a tensor of order 2 etc.) Computing a special rank one approximation of slices of this tensor produces the \( n-1 \)-gram. Thus, taking rank \( \kappa \) approximations in this fashion allows us to represent dependencies between an \( n \)-gram and \( n-1 \)-gram.

Consider the bigram count matrix \( B \) with \( N \) counts which has rank \( V \). Note that \( \hat{P}(w_i|w_{i-1}) = \frac{B(w_i,w_{i-1})}{\sum_w B(w,w_{i-1})} \). Additionally, \( B \) can be considered a random variable that is the result of sampling \( N \) tuples of \((w_i, w_{i-1})\) and agglomerating them into a count matrix. Assuming \( w_i \) and \( w_{i-1} \) are independent, the expected value (with respect to the empirical distribution) \( E[B] = NP(w_i)P(w_{i-1}) \), which can be rewritten as being proportional to the outer product of the unigram probability vector with itself, and is thus rank one.

---

1 with a slight abuse of notation, let \( Z_{D_j} \) be shorthand for \( Z_{j,D_j} \)
This observation extends to higher order n-grans as well. Let $C^n$ be the n-th order tensor where $C^n(w_{ij}, ..., w_{i-n+1}) = c(w_{ij}, ..., w_{i-n+1})$. Furthermore denote $C^n(\cdot, \tilde{w}_{i-n+2}, \cdot)$ to be the $V \times V$ matrix slice of $C^n$ where $\tilde{w}_{i-n+2}, ..., w_{i-n+1}$ are held fixed to a particular sequence $\tilde{w}_{i-n+2}, ..., \tilde{w}_{i-1}$. Then if $w_i$ is conditionally independent of $w_{i-n+1}$ given $w_{i-1}$, then $\mathbb{E}[C^n(\cdot, \tilde{w}_{i-n+2}, \cdot)]$ is rank one $\forall \tilde{w}_{i-n+2}$.

However, it is rare that these matrices are actually rank one, either due to sampling variance or the fact that $w_i$ and $w_{i-1}$ are not independent. What we would really like to say is that the best rank one approximation $B^{(1)}$ (under some norm) of $B$ is $\hat{P}(w_i)\hat{P}(w_{i-1})$. While this statement is not true under the $\ell_2$ norm, it is true under generalized KL divergence (Lee and Seung, 2001): $gKL(A\|B) = \sum_{ij} \left( A_{ij} \log(\frac{A_{ij}}{B_{ij}}) - A_{ij} + B_{ij} \right)$.

In particular, generalized KL divergence preserves row and column sums: if $M^{(\kappa)}$ is the best rank $\kappa$ approximation of $M$ under gKL then the row sums and column sums of $M^{(\kappa)}$ and $M$ are equal (Ho and Van Dooren, 2008). Leveraging this property, it is straightforward to prove the following lemma:

**Lemma 27.** Let $B^{(\kappa)}$ be the best rank $\kappa$ approximation of $B$ under gKL. Then $B^{(1)} \propto \hat{P}(w_i)\hat{P}(w_{i-1})$ and $\forall w_{i-1}$ s.t. $c(w_{i-1}) \neq 0$:

$$\hat{P}(w_i) = \frac{B^{(1)}(w_i, w_{i-1})}{\sum_{w} B^{(1)}(w, w_{i-1})} \quad (9.7)$$

For more general $n$, let $C^{n,(\kappa)}(\cdot, \tilde{w}_{i-n+2}, \cdot)$ be the best rank $\kappa$ approximation of $C^n(\cdot, \tilde{w}_{i-n+2}, \cdot)$ under gKL. Then similarly, $\forall w_{i-n+1}$ s.t. $c(w_{i-n+1}) > 0$:

$$\hat{P}(w_i|w_{i-1}, ..., w_{i-2}) = \frac{C^{n,(1)}(w_i, w_{i-1})}{\sum_{w} C^{n,(1)}(w, w_{i-1})} \quad (9.8)$$

Thus, by selecting $1 < \kappa < V$, we obtain count matrices and tensors between $n$ and $n - 1$-grams. The condition that $c(w_{i-n+1}) > 0$ corresponds to the discontinuity discussed in §9.2.2.

### 9.3.2 Power

Since KN smoothing alters the lower order distributions instead of simply using the MLE, varying the rank is not sufficient in order to generalize this suite of techniques. Thus, PLRE computes low rank approximations of altered count matrices. Consider taking the elementwise power $\rho$ of the bigram count matrix, which is denoted by $B^\rho$. For example, the observed bigram count matrix and associated row sum:

$$B^1 = \begin{pmatrix} 1.0 & 2.0 & 1.0 \\ 0 & 5.0 & 0 \\ 2.0 & 0 & 0 \end{pmatrix} \quad \text{row sum} \rightarrow \begin{pmatrix} 4.0 \\ 5.0 \\ 2.0 \end{pmatrix}$$

As expected the row sum is equal to the unigram counts (which we denote as $u$). Now consider
\[ B^{0.5} : \]

\[
B^{0.5} = \begin{pmatrix}
1.0 & 1.4 & 1.0 \\
0 & 2.2 & 0 \\
1.4 & 0 & 0
\end{pmatrix}
\rightarrow
\begin{pmatrix}
3.4 \\
2.2 \\
1.4
\end{pmatrix}
\]

Note how the row sum vector has been altered. In particular since \( w_1 \) (corresponding to the first row) has a more diverse history than \( w_2 \), it has a higher row sum (compared to in \( u \) where \( w_2 \) has the higher row sum). Lastly, consider the case when \( p = 0 \):

\[
B^0 = \begin{pmatrix}
1.0 & 1.0 & 1.0 \\
0 & 1.0 & 0 \\
1.0 & 0 & 0
\end{pmatrix}
\rightarrow
\begin{pmatrix}
3.0 \\
1.0 \\
1.0
\end{pmatrix}
\]

The row sum is now the number of unique words that precede \( w_i \) (since \( B^0 \) is binary) and is thus equal to the (unnormalized) Kneser Ney unigram. This idea also generalizes to higher order \( n \)-grams and leads us to the following lemma:

**Lemma 28.** Let \( B^{(\rho,\kappa)} \) be the best rank \( \kappa \) approximation of \( B^p \) under \( g_{KL} \). Then \( \forall w_{i-1} \) s.t. \( c(w_{i-1}) \neq 0 \):

\[
P^{alt}(w_i) = \frac{B^{(0,1)}(w_i, w_{i-1})}{\sum_w B^{(0,1)}(w, w_{i-1})}
\]

For more general \( n \), let \( C^{n,(\rho,\kappa)}_{i-1,...,i-n+2} \) be the best rank \( \kappa \) approximation of \( C^{n,(\rho)}_{i-1,...,i-n+2} \) under \( g_{KL} \). Similarly, \( \forall w_{i-1} \) s.t. \( c(w_{i-1}) > 0 \):

\[
P^{alt}(w_i|w_{i-1},...,w_{i-n+2}) = \frac{C^{n,(0,1)}_{i-1,...,i-n+2}(w_i, w_{i-n+1})}{\sum_w C^{n,(0,1)}_{i-1,...,i-n+2}(w, w_{i-n+1})}
\]

### 9.4 Creating the Ensemble

Recall our overall formulation in Eq. 9.6; a naive solution would be to set \( Z_1, ..., Z_\eta \) to low rank approximations of the count matrices/tensors under varying powers, and then interpolate through constant absolute discounting. Unfortunately, the marginal constraint in Eq. 9.3 will generally not hold if this strategy is used. Therefore, we propose a generalized discounting scheme where each non-zero \( n \)-gram count is associated with a different discount \( D_j(w_i, w_{i-n+1}) \). The low rank approximations are then computed on the discounted matrices, leaving the marginal constraint intact.

For clarity of exposition, we focus on the special case where \( n = 2 \) with only one low rank matrix before stating our general algorithm:

\[
P_{pre}(w_i|w_{i-1}) = \hat{P}_{D_\kappa}(w_i|w_{i-1})
\]

\[
+ \gamma_0(w_{i-1}) \left( Z_{D_\kappa}(w_i|w_{i-1}) + \gamma_1(w_{i-1}) P^{alt}(w_i) \right)
\]

(9.10)
Our goal is to compute $D_0, D_1$ and $Z_1$ so that the following lower order marginal constraint holds:

$$\hat{P}(w_i) = \sum_{w_{i-1}} P_{\text{pre}}(w_i|w_{i-1})\hat{P}(w_{i-1})$$  \hspace{1cm} (9.11)

Our solution can be thought of as a two-step procedure where we compute the discounts $D_0, D_1$ (and the $\gamma(w_{i-1})$ weights as a by-product), followed by the low rank quantity $Z_1$. First, we construct the following intermediate ensemble of powered, but full rank terms. Let $Y^{(p)}$ be the matrix such that $Y^{(p)}(w_i, w_{i-1}) := c(w_i, w_{i-1})^{p_j}$. Then define

$$P_{\text{pwr}}(w_i|w_{i-1}) := Y^{(p_0=1)}(w_i|w_{i-1}) + \gamma_0(w_{i-1})\left(Y^{(p_1)}(w_i|w_{i-1}) + \gamma_1(w_{i-1})Y^{(p_2=0)}(w_i|w_{i-1})\right)$$ \hspace{1cm} (9.12)

where with a little abuse of notation:

$$Y^{(p_j)}(w_i|w_{i-1}) = \frac{c(w_i, w_{i-1})^{p_j} - D_j(w_i, w_{i-1})}{\sum_{w_i} c(w_i, w_{i-1})^{p_j}}$$

Note that $P_{\text{alt}}(w_i)$ has been replaced with $Y^{(p_2=0)}(w_i|w_{i-1})$, based on Lemma 28, and will equal $P_{\text{alt}}(w_i)$ once the low rank approximation is taken as discussed in § 9.4.2.

Since we have only combined terms of different power (but all full rank), it is natural choose the discounts so that the result remains unchanged i.e., $P_{\text{pwr}}(w_i|w_{i-1}) = \hat{P}(w_i|w_{i-1})$, since the low rank approximation (not the power) will implement smoothing. Enforcing this constraint gives rise to a set of linear equations that can be solved (in closed form) to obtain the discounts as we now show below.

### 9.4.1 Step 1: Computing the Discounts

To ensure the constraint that $P_{\text{pwr}}(w_i|w_{i-1}) = \hat{P}(w_i|w_{i-1})$, it is sufficient to enforce the following two local constraints:

$$Y^{(p_j)}(w_i|w_{i-1}) = Y^{(p_j)}(w_i|w_{i-1}) + \gamma_j(w_{i-1})Y^{(p_{j+1})}(w_i|w_{i-1}) \text{ for } j = 0, 1$$ \hspace{1cm} (9.13)

This allows each $D_j$ to be solved for independently of the other $\{D_j\}_{j \neq j}$. Let $c_{i,i-1} = c(w_i, w_{i-1})$, $c^j_{i,i-1} = c(w_i, w_{i-1})^{p_j}$, and $d^j_{i,i-1} = D_j(w_i, w_{i-1})$. Expanding Eq. 9.13 yields that $\forall w_i, w_{i-1}$:

$$\frac{c^j_{i,i-1}}{\sum_i c^j_{i,i-1}} = \frac{c^j_{i,i-1} - d^j_{i,i-1}}{\sum_i c^j_{i,i-1}} + \left(\frac{\sum_i d^j_{i,i-1}}{\sum_i c^j_{i,i-1}}\right) \frac{c^{j+1}_{i,i-1}}{\sum_i c^{j+1}_{i,i-1}}$$ \hspace{1cm} (9.14)

which can be rewritten as:

$$-d^j_{i,i-1} + \left(\sum_i d^j_{i,i-1}\right) \frac{c^{j+1}_{i,i-1}}{\sum_i c^{j+1}_{i,i-1}} = 0$$ \hspace{1cm} (9.15)
Algorithm 12 Compute $D$

**In:** Count tensor $C^n_i$, powers $\rho_j, \rho_{j+1}$ such that $\rho_j \geq \rho_{j+1}$, and parameter $d_*$.

**Out:** Discount $D_j$ for powered counts $C^{n,i_j}$ and associated leftover weight $\gamma_j$

1: Set $D_j(w_i, w_{i-n+1}^{-1}) = d_*, c(w_i, w_{i-n+1}^{-1})^{\rho_{j+1}}$.

2:

$$\gamma_j(w_i, w_{i-n+1}^{-1}) = \frac{d_* \sum c(w_i, w_{i-n+1}^{-1})^{\rho_{j+1}}}{\sum w_i c(w_i, w_{i-n+1}^{-1})^{\rho_j}}$$

Note that Eq. 9.15 decouples across $w_{i-1}$ since the only $d_{i,i-1}^j$ terms that are dependent are the ones that share the preceding context $w_{i-1}$.

It is straightforward to see that setting $d_{i,i-1}^j$ proportional to $c_{i,i-1}^{j+1}$ satisfies Eq. 9.15. Furthermore it can be shown that all solutions are of this form (i.e., the linear system has a null space of exactly one). Moreover, we are interested in a particular subset of solutions where a single parameter $d_*$ (independent of $w_{i-1}$) controls the scaling as indicated by the following lemma:

**Lemma 29.** Assume that $\rho_j \geq \rho_{j+1}$. Choose any $0 \leq d_* \leq 1$. Set $d_{i,i-1}^j = d_* c_{i,i-1}^{j+1} \forall i, j$. The resulting discounts satisfy Eq. 9.15 as well as the inequality constraints $0 \leq d_{i,i-1}^j \leq c_{i,i-1}^j$. Furthermore, the leftover weight $\gamma_j$ takes the form:

$$\gamma_j(w_{i-1}) = \frac{\sum d_{i,i-1}^j}{\sum c_{i,i-1}^j} = \frac{d_* \sum c_{i,i-1}^{j+1}}{\sum c_{i,i-1}^j}$$

(9.16)

**Proof.** Clearly this choice of $d_{i,i-1}^j$ satisfies Eq. 9.15. The largest possible value of $d_{i,i-1}^j$ is $c_{i,i-1}^{j+1}$, $\rho_j \geq \rho_{j+1}$ implies $c_{i,i-1}^{j+1} \geq c_{i,i-1}^j$. Thus the inequality constraints are met. It is then easy to verify that $\gamma$ takes the above form. □

The above lemma generalizes to longer contexts (i.e. $n > 2$) as shown in Algorithm 12. Note that if $\rho_j = \rho_{j+1}$ then Algorithm 12 is equivalent to scaling the counts e.g. deleted-interpolation/Jelinek Mercer smoothing (Jelinek and Mercer, 1980). On the other hand, when $\rho_{j+1} = 0$, Algorithm 12 is equal to the absolute discounting that is used in Kneser-Ney. Thus, depending on $\rho_{j+1}$, our method generalizes different types of interpolation schemes to construct an ensemble so that the marginal constraint is satisfied.

### 9.4.2 Step 2: Computing Low Rank Quantities

The next step is to compute low rank approximations of $Y_{D_j}^{(\rho_j)}$ to obtain $Z_{D_j}$, such that the intermediate marginal constraint in Eq. 9.11 is preserved. This constraint trivially holds for the intermediate ensemble $P_{\text{pwr}}(w_i|w_{i-1})$ due to how the discounts were derived in § 9.4.1. For our running bigram example, define $Z_{D_j}^{(\rho_j,k_j)}$ to be the best rank $k_j$ approximation to $Y_{D_j}^{(\rho_j,k_j)}$ according
Algorithm 13 Compute $Z$

**In:** Count tensor $C^*$, power $\rho$, discounts $D$, rank $\kappa$

**Out:** Discounted low rank conditional probability table $Z^{(\rho,\kappa)}_D(w_i|w_{i-n+1}^j)$ (represented implicitly)

1: Compute powered counts $C^{\rho,\kappa}_D$.
2: Compute denominators $\sum w_c(w_i, w_{i-n+1}^j)^\rho$ s.t. $c(w_{i-n+1}^j) > 0$.
3: Compute discounted powered counts $C^{\rho,\kappa}_D(p) = C^{\rho,\kappa}_D - D$.
4: For each slice $M^{(\rho,\kappa)}_{i,n+2} := C^{\rho,\kappa}_D(:,\tilde{w}_{i-n+2}^j)$ compute
   
   $M^{(\kappa)} := \min_{A \in \mathbb{R}^{n \times n}} \| M^{(\rho,\kappa)}_{i,n+2} - A \|_{KL}$
   
   (stored implicitly as $M^{(k)} = LR$)

5: Note that

\[
Z^{(\rho,\kappa)}_D(w_i|w_{i-n+1}^j) = \frac{Z^{(\rho,\kappa)}_D(w_i, w_{i-n+1}^j)}{\sum w_c(w_i, w_{i-n+1}^j)^\rho}
\]  

Note that $Z^{(\rho,\kappa)}_D(w_i|w_{i-1})$ is a valid (discounted) conditional probability since $gKL$ preserves row/column sums so the denominator remains unchanged under the low rank approximation. Then using the fact that $Z^{(0,1)}(w_i|w_{i-1}) = P^{alt}(w_i)$ (Lemma 28) we can embellish Eq. 9.10 as

\[
P_{plre}(w_i|w_{i-1}) = P_{D_{\kappa}}(w_i|w_{i-1}) + \gamma_0(w_{i-1}) \left( Z^{(\rho,\kappa)}_D(w_i|w_{i-1}) + \gamma_1(w_{i-1}) P^{alt}(w_i) \right)
\]  

Leveraging the form of the discounts and row/column sum preserving property of $gKL$, we then have the following lemma (the proof is in the Appendix):

**Lemma 30.** Let $P_{plre}(w_i|w_{i-1})$ indicate the PLRE smoothed conditional probability as computed by Eq. 9.10 and Algorithms 12 and 13. Then, the marginal constraint in Eq. 9.11 holds i.e.:

\[
\hat{P}(w) = \sum_{w_{i-1}} P_{plre}(w_i|w_{i-1}) \hat{P}(w_{i-1})
\]  

9.4.3 More general algorithm

In general, the principles outlined in the previous sections hold for higher order $n$-grams. Assume that the discounts are computed according to Algorithm 12 with parameter $d$, and $Z^{(\rho,\kappa)}_{D_{\kappa}}$ is computed according to Algorithm 13. Note that, as shown in Algorithm 13, for higher order $n$-grams, the $Z^{(\rho,\kappa)}_{D_{\kappa}}$ are created by taking low rank approximations of slices of the (powered) count
We can leverage the form of the discount, and using the fact that $\rho$, where $\rho$ is chosen to equal $(\eta + 1)$, we can leverage the form of Eq. 9.30 in the following form:

$$P_{\text{plre}}(w_i|w_{i-n+1}^{i-1}) = P_{\text{alt}}^{\text{terms}}(w_i|w_{i-n+1}^{i-1}) + \gamma_0(w_{i-n+1}^{i-1})P_{\text{plre}}(w_i|w_{i-n+1}^{i-2})$$

(9.21)

where $P_{\text{alt}}^{\text{terms}}(w_i|w_{i-n+1}^{i-1})$ contains the terms in Eq. 9.30 except the last, and $\gamma_0(w_{i-n+1}^{i-1}) = \prod_{h=0}^n \gamma_h(w_{i-n+1}^{i-1})$, we can leverage the form of the discount, and using the fact that $\rho_{\eta+1} = \gamma_0(w_{i-n+1}^{i-1})$.

With this form of $\gamma()$, Eq. 9.21 is remarkably similar to KN smoothing (Eq. 9.5) if KN’s discount parameter $D$ is chosen to equal $(d_*)^{(\eta + 1)}$.

The difference is that $P_{\text{alt}}^{\text{terms}}()$ has been replaced with the alternate estimate $P_{\text{alt}}^{\text{terms}}(w_i|w_{i-n+1}^{i-1})$, which have been enriched via the low rank structure. Since these alternate estimates were constructed via our ensemble strategy they contain both very fine-grained dependencies (the original $n$-grams) as well as coarser dependencies (the lower rank $n$-grams) and is thus fundamentally different than simply taking a single matrix/tensor decomposition of the trigram/bigram matrices.

Moreover, it provides a natural way of setting $d_*$ based on the Good-Turing (GT) estimates employed by KN smoothing. In particular, we can set $d_*$ to be the $(\eta + 1)^{\text{th}}$ root of the KN discount $D$ that can be estimated via the GT estimates.

9.4.5 Computational Considerations

PLRE scales well even as the order $n$ increases. To compute a low rank bigram, one low rank approximation of a $V \times V$ matrix is required. For the low rank trigram, we need to compute a low rank approximation of each slice $C_{D}^{w_i,p}(\cdot, w_{i-2}, :)$ $\forall w_{i-2}$. While this may seem daunting at first, in practice the size of each slice (number of non-zero rows/columns) is usually much, much smaller than $V$, keeping the computation tractable.

Similarly, PLRE also evaluates conditional probabilities at evaluation time efficiently. As shown

\cite{2} for derivation see proof of Lemma 30 in the Appendix.
Table 9.1: Perplexity results on small corpora for all methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>class-1024(3)</th>
<th>BO-KN(3)</th>
<th>int-KN(3)</th>
<th>BO-MKN(3)</th>
<th>int-MKN(3)</th>
<th>PLRE(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small-English Dev</td>
<td>115.64</td>
<td>99.20</td>
<td>99.73</td>
<td>99.95</td>
<td>95.63</td>
<td>91.18</td>
</tr>
<tr>
<td>Small-English Test</td>
<td>119.70</td>
<td>103.86</td>
<td>104.56</td>
<td>104.55</td>
<td>100.07</td>
<td>95.15</td>
</tr>
<tr>
<td>Small-Russian Dev</td>
<td>286.38</td>
<td>281.29</td>
<td>265.71</td>
<td>287.19</td>
<td>263.25</td>
<td>241.66</td>
</tr>
<tr>
<td>Small-Russian Test</td>
<td>284.09</td>
<td>277.74</td>
<td>262.02</td>
<td>283.70</td>
<td>260.19</td>
<td>238.96</td>
</tr>
</tbody>
</table>

in Algorithm 13, the normalizer can be precomputed on the sparse powered matrix/tensor. As a result our test complexity is $O(\sum \eta_{total} \kappa_i)$ where $\eta_{total}$ is the total number of matrices/tensors in the ensemble. While this is larger than Kneser Ney’s practically constant complexity of $O(n)$, it is much faster than other recent methods for language modeling such as neural networks and conditional exponential family models where exact computation of the normalizing constant costs $O(V)$.

9.5 Experiments

To evaluate PLRE, we compared its performance on English and Russian corpora with several variants of KN smoothing, class-based models, and the log-bilinear neural language model (Mnih and Hinton, 2007). We evaluated with perplexity in most of our experiments, but also provide results evaluated with BLEU (Papineni et al., 2002) on a downstream machine translation (MT) task. We have made the code for our approach publicly available.3

To build the hard class-based LMs, we utilized mkcls4, a tool to train word classes that uses the maximum likelihood criterion (Och, 1995) for classing. We subsequently trained trigram class language models on these classes (corresponding to 2nd-order HMMs) using SRILM (Stolcke, 2002), with KN-smoothing for the class transition probabilities. SRILM was also used for the baseline KN-smoothed models.

For our MT evaluation, we built a hierarchical phrase translation (Chiang, 2007) system using cdec (Dyer et al., 2010). The KN-smoothed models in the MT experiments were compiled using KenLM (Heafield, 2011).

9.5.1 Datasets

For the perplexity experiments, we evaluated our proposed approach on 4 datasets, 2 in English and 2 in Russian. In all cases, the singletons were replaced with “<unk>” tokens in the training corpus, and any word not in the vocabulary was replaced with this token during evaluation. There is a general dearth of evaluation on large-scale corpora in morphologically rich languages such as Russian, and thus we have made the processed Large-Russian corpus available for comparison.4

- **Small-English**: APNews corpus (Bengio et al., 2003): Train - 14 million words, Dev - 963,000, Test - 963,000. Vocabulary- 18,000 types.
For the MT evaluation, we used the parallel data from the WMT 2013 shared task, excluding the Common Crawl corpus data. The newstest2012 and newstest2013 evaluation sets were used as the development and test sets respectively.

9.5.2 Small Corpora

For the class-based baseline LMs, the number of classes was selected from \{32, 64, 128, 256, 512, 1024\} (Small-English) and \{512, 1024\} (Small-Russian). We could not go higher due to the computation-ally laborious process of hard clustering. For Kneser-Ney, we explore four different variants: back-off (BO-KN) interpolated (int-KN), modified back-off (BO-MKN), and modified interpolated (int-MKN). Good-Turing estimates were used for discounts. All models trained on the small corpora are of order 3 (trigrams).

For PLRE, we used one low rank bigram and one low rank trigram in addition to the MLE n-gram estimates. The powers of the intermediate matrices/tensors were fixed to be 0.5 and the discounts were set to be square roots of the Good Turing estimates (as explained in §9.4.4). The ranks were tuned on the development set. For Small-English, the ranges were \{1e−3, 5e−3\} (as a fraction of the vocabulary size) for both the low rank bigram and low rank trigram models. For Small-Russian the ranges were \{5e−4, 1e−3\} for both the low rank bigram and the low rank trigram models.

The results are shown in Table 9.1. The best class-based LM is reported, but is not competitive with the KN baselines. PLRE outperforms all of the baselines comfortably. Moreover, PLRE’s performance over the baselines is highlighted in Russian. With larger vocabulary sizes, the low rank approach is more effective as it can capture linguistic similarities between rare and common words.

Next we discuss how the maximum n-gram order affects performance. Figure 9.1 shows the relative percentage improvement of our approach over int-MKN as the order is increased from 2 to 4 for both methods. The Small-English dataset has a rather small vocabulary compared to the number of tokens, leading to lower data sparsity in the bigram. Thus the PLRE improvement is small for order = 2, but more substantial for order = 3. On the other hand, for the Small-Russian dataset, the vocabulary size is much larger and consequently the bigram counts are sparser. This leads to similar improvements for all orders (which are larger than that for Small-English).

On both these datasets, we also experimented with tuning the discounts for int-MKN to see if the baseline could be improved with more careful choices of discounts. However, this achieved
only marginal gains (reducing the perplexity to 98.94 on the Small-English test set and 259.0 on the Small-Russian test set).

**Comparison to neural language models:** Although not the focus of this work, we also briefly compare our method to neural language models. Mnih and Hinton (2007) evaluate on the Small-English dataset (but remove end markers and concatenate the sentences). They obtain perplexities 117.0 and 107.8 using contexts of size 5 and 10 respectively. With this preprocessing, a 4-gram (context 3) PLRE achieves 108.4 perplexity. We also obtained results for a recurrent neural network (RNN) model (Mikolov et al., 2010) on the Small-English dataset using the RNNLM toolkit (Mikolov, 2012), and the RNN-ME variant obtains a perplexity of 82.1 (infinite context). Note that neural networks generally have a test complexity of $O(k\sqrt{V})$ using word-classing (Goodman, 2001; Mikolov et al., 2011) which is considerably larger than the test complexity $O(kn)$ of our approach (The test complexity of Kneser-Ney is $O(n)$).

### 9.5.3 Large Corpora

Results on the larger corpora for the top 2 performing methods “PLRE” and “int-MKN” are presented in Table 9.2. Due to the larger training size, we use 4-gram models in these experiments. However, including the low rank 4-gram tensor provided little gain and therefore, the 4-gram PLRE only has additional low rank bigram and low rank trigram matrices/tensors. As above, ranks were tuned on the development set. For Large-English, the ranges were $\{1e-4, 5e-4, 1e-3\}$ (as a fraction of the vocabulary size) for both the low rank bigram and low rank trigram models. For Small-Russian the ranges were $\{1e-5, 5e-5, 1e-4\}$ for both the low rank bigram and the low rank trigram models. For statistical validity, 10 test sets of size equal to the original test set were generated by randomly sampling sentences with replacement from the original test set. Our method outperforms “int-MKN” with gains similar to that on the smaller datasets. As shown in Table 9.3, our method obtains fast training times even for large datasets.
Table 9.2: Mean perplexity results on large corpora, with standard deviation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>int-MKN(4)</th>
<th>PLRE(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large-English Dev</td>
<td>73.21</td>
<td>71.21</td>
</tr>
<tr>
<td>Large-English Test</td>
<td>77.90 ± 0.203</td>
<td>75.66 ± 0.189</td>
</tr>
<tr>
<td>Large-Russian Dev</td>
<td>326.9</td>
<td>297.11</td>
</tr>
<tr>
<td>Large-Russian Test</td>
<td>289.63 ± 6.82</td>
<td>264.59 ± 5.839</td>
</tr>
</tbody>
</table>

Table 9.3: PLRE training times for a fixed parameter setting. 8 Intel Xeon CPUs were used.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PLRE Training Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small-English</td>
<td>3.96 min (order 3) / 8.3 min (order 4)</td>
</tr>
<tr>
<td>Small-Russian</td>
<td>4.0 min (order 3) / 4.75 min (order 4)</td>
</tr>
<tr>
<td>Large-English</td>
<td>3.2 hrs (order 4)</td>
</tr>
<tr>
<td>Large-Russian</td>
<td>8.3 hrs (order 4)</td>
</tr>
</tbody>
</table>

Table 9.4: Results on English-Russian translation task (mean ± stdev). See text for details.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>int-MKN(4)</td>
<td>17.63 ± 0.11</td>
</tr>
<tr>
<td>PLRE(4)</td>
<td>17.79 ± 0.07</td>
</tr>
<tr>
<td>Smallest Diff</td>
<td>PLRE+0.05</td>
</tr>
<tr>
<td>Largest Diff</td>
<td>PLRE+0.29</td>
</tr>
</tbody>
</table>

9.6 Machine Translation Task

Table 9.4 presents results for the MT task, translating from English to Russian. We used MIRA (Chiang et al., 2008) to learn the feature weights. To control for the randomness in MIRA, we avoid retuning when switching LMs - the set of feature weights obtained using int-MKN is the same, only the language model changes. The procedure is repeated 10 times to control for optimizer instability (Clark et al., 2011). Unlike other recent approaches where an additional feature weight is tuned for the proposed model and used in conjunction with KN smoothing (Vaswani et al., 2013), our aim is to show the improvements that PLRE provides as a substitute for KN. On average, PLRE outperforms the KN baseline by 0.16 BLEU, and this improvement is consistent in that PLRE never gets a worse BLEU score.

9.7 Related Work

Recent attempts to revisit the language modeling problem have largely come from two directions: Bayesian nonparametrics and neural networks. Teh (2006) and Goldwater et al. (2006) discovered the connection between interpolated Kneser Ney and the hierarchical Pitman-Yor process. These have led to generalizations that account for domain effects (Wood and Teh, 2009) and unbounded contexts (Wood et al., 2009).
The idea of using neural networks for language modeling is not new (Miikkulainen and Dyer, 1991), but recent efforts (Mnih and Hinton, 2007; Mikolov et al., 2010) have achieved impressive performance. These methods can be quite expensive to train and query (especially as the vocabulary size increases). Techniques such as noise contrastive estimation (Gutmann and Hyvärinen, 2012; Mnih and Teh, 2012; Vaswani et al., 2013), subsampling (Xu et al., 2011), or careful engineering approaches for maximum entropy LMs (which can also be applied to neural networks) (Wu and Khudanpur, 2000) have improved training of these models, but querying the probability of the next word given still requires explicitly normalizing over the vocabulary, which is expensive for big corpora or in languages with a large number of word types. Mnih and Teh (2012) and Vaswani et al. (2013) propose setting the normalization constant to 1, but this is approximate and thus can only be used for downstream evaluation, not for perplexity computation. An alternate technique is to use word-classing (Goodman, 2001; Mikolov et al., 2011), which can reduce the cost of exact normalization to $O(\sqrt{V})$. In contrast, our approach is much more scalable, since it is trivially parallelized in training and does not require explicit normalization during evaluation.

There are a few low rank approaches (Saul and Pereira, 1997; Bellegarda, 2000; Hutchinson et al., 2011), but they are only effective in restricted settings (e.g. small training sets, or corpora divided into documents) and do not generally perform comparably to state-of-the-art models. Roark et al. (2013) also use the idea of marginal constraints for re-estimating back-off parameters for heavily-pruned language models, whereas we use this concept to estimate $n$-gram specific discounts.

### 9.8 Conclusion

We presented power low rank ensembles, a technique that generalizes existing $n$-gram smoothing techniques to non-integer $n$. By using ensembles of sparse as well as low rank matrices and tensors, our method captures both the fine-grained and coarse structures in word sequences. Our discounting strategy preserves the marginal constraint and thus generalizes Kneser Ney, and under slight changes can also extend other smoothing methods such as deleted-interpolation/Jelinek-Mercer smoothing. Experimentally, PLRE convincingly outperforms Kneser-Ney smoothing as well as class-based baselines.

### 9.9 Appendix

The primary purpose of this section is to provide a proof of Lemma 30. We also show that Lemma 30 extends to $n > 2$.

#### 9.9.1 Proof of Lemma 30

Proof. Assume the following more general form where multiple low rank matrices can be used i.e.
$$P_{\text{pre}}(w_i|w_{i-1}) = \left. P_{\text{alt}}(w_i|w_{i-1}) + \right. \left. \gamma_0(w_{i-1}) \left( Z_{D_i}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) + \ldots + \gamma_{\eta-1}(w_{i-1}) \left( Z_{D_i}^{(\rho_{\eta-1},\kappa_{\eta-1})}(w_i|w_{i-1}) + \gamma_{\eta}(w_{i-1}) \left( Z^{(\rho_{\eta+1}=0,\kappa_{\eta+1}=1)}(w_i|w_{i-1}) \right) \right) \right) \right) \tag{9.23}$$

where we note that $Z^{(\rho_{\eta+1}=0,\kappa_{\eta+1}=1)}(w_i|w_{i-1})$ is equivalent to $P_{\text{alt}}(w_i)$. It is assumed that $1 \geq \rho_0 \geq \ldots \rho_{\eta+1} = 0$.

First unroll the recursion and rewrite $P_{\text{pre}}(w_i|w_{i-1})$ as:

$$P_{\text{pre}}(w_i|w_{i-1}) = \sum_{j=0}^{\eta+1} \gamma_{0,j-1}(w_{i-1}) Z_{D_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1})$$

where $\gamma_{0,j-1}(w_{i-1}) = \prod_{h=0}^{j-1} \gamma_h(w_{i-1})$ and $\gamma_{0,-1}(w_{i-1}) = 1$. Note that $P_{\text{wtr}}(w_i|w_{i-1})$ can be written in the same way.

$$P_{\text{wtr}}(w_i|w_{i-1}) = \sum_{j=0}^{\eta} \gamma_{0,j}(w_{i-1}) Y_{D_j}^{(\rho_j)}(w_i|w_{i-1}) \tag{9.24}$$

Note that $P_{\text{wtr}}(w_i|w_{i-1})$ already satisfies the marginal constraint i.e.

$$\hat{P}(w) = \sum_{w_i} P_{\text{wtr}}(w_i|w_{i-1}) \hat{P}(w_{i-1}) \tag{9.25}$$

because the discounts were chosen such that $P_{\text{wtr}}(w_i|w_{i-1}) = \hat{P}(w_i|w_{i-1})$.

Thus it suffices to show that for all $j = 0, \ldots, \eta + 1$:

$$\sum_{w_{i-1}} \hat{P}(w_{i-1}) \gamma_{0,j-1}(w_{i-1}) Y_{D_j}^{(\rho_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}} \hat{P}(w_{i-1}) \gamma_{0,j-1}(w_{i-1}) Z_{D_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) \tag{9.26}$$

The statement above is trivially true when $j = 0$. For all other cases, note that due to the way we have set the discounts, $\gamma_{0,j-1}$ takes a special form:

$$\prod_{h=0}^{j-1} \gamma_h(w_{i-1}) = \frac{d_s \sum_i c_{i,j-1}^{(\rho_1)} d_s \sum_i c_{i,j-1}^{(\rho_2)} \ldots d_s \sum_i c_{i,j-1}^{(\rho_{j-1})}}{\sum_i c_{i,j-1}^{(\rho_0)} \sum_i c_{i,j-1}^{(\rho_1)} \ldots \sum_i c_{i,j-1}^{(\rho_{j-1})}} \tag{9.27}$$

Using this form in Eq. 5 and simplifying yields:

$$\sum_{w_{i-1} \left( \sum_i c_{i,j-1}^{(\rho_j)} \right)} Y_{D_j}^{(\rho_j)}(w_i|w_{i-1}) = \sum_{w_{i-1} \left( \sum_i c_{i,j-1}^{(\rho_j)} \right)} Z_{D_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1})$$
which is equivalent to requiring that
\[
\sum_{w_{i-1}} \mathbf{Y}^{(\rho_j)}_{D_j}(w_i, w_{i-1}) = \sum_{w_{i-1}} \mathbf{Z}^{(\rho_j, \kappa_j)}_{D_j}(w_i, w_{i-1})
\]  
(9.28)

which holds because rank minimization under gKL preserves row and column sums. \(\square\)

### 9.9.2 Generalization to \(n > 2\)

**Theorem 7.** Let \(P_{\text{plre}}(w_i|w_{i-1}^{j-1})\) indicate the PLRE smoothed conditional probability and \(\hat{P}(w)\) indicate the MLE probability of \(w\). Then,
\[
\hat{P}(w) = \sum_{w_{i-1}^{j-1}} P_{\text{plre}}(w_i|w_{i-1}^{j-1}) \hat{P}(w_{i-1}^{j-1})
\]  
(9.29)

**Proof.** Recall that,
\[
P_{\text{plre}}(w_i|w_{i-1}^{j-1}) = P_{\text{alt}}^{D_0}(w_i|w_{i-1}^{j-1})
+ \gamma_0(w_{i-1}^{j-1}) \left( Z^{(\rho_1, \kappa_1)}_{D_1}(w_i|w_{i-1}^{j-1}) + \ldots \right)
+ \gamma_{\eta-1}(w_{i-1}^{j-1}) \left( Z^{(\rho_\eta, \kappa_\eta)}_{D_\eta}(w_i|w_{i-1}^{j-1}) + \gamma_{\eta}(w_{i-1}^{j-1}) \left( P_{\text{plre}}(w_i|w_{i-1}^{j-1}) \right) \right)\ldots
\]  
(9.30)

Define,
\[
P_{\text{pwr}}(w_i|w_{i-1}^{j-1}) = P_{\text{alt}}^{D_0}(w_i|w_{i-1}^{j-1})
+ \gamma_0(w_{i-1}^{j-1}) \left( Y^{(\rho_1, \kappa_1)}_{D_1}(w_i|w_{i-1}^{j-1}) + \ldots \right)
+ \gamma_{\eta-1}(w_{i-1}^{j-1}) \left( Y^{(\rho_\eta, \kappa_\eta)}_{D_\eta}(w_i|w_{i-1}^{j-1}) + \gamma_{\eta}(w_{i-1}^{j-1}) \left( P_{\text{pwr}}(w_i|w_{i-1}^{j-1}) \right) \right)\ldots
\]  
(9.31)

where with a little abuse of notation
\[
Y^{(\rho_j)}_{D_j}(w_i|w_{i-1}^{j-1}) = \frac{\tilde{c}(w_i, w_{i-1}^{j-1})^{(\rho_j)} - D_j(w_i, w_{i-1}^{j-1})}{\sum_{w_j} \tilde{c}(w_i, w_{i-1}^{j-1})^{(\rho_j)}}
\]  
(9.32)
and

\[ \hat{c}(w_i, w_{i-n'+1}^{i-1}) = \begin{cases} 
  c(w_i, w_{i-n'+1}^{i-1}), & \text{if } n' = n \\
  N_-(w_{i-n'+1}^{i-1}) & \text{if } n' < n 
\end{cases} \]

Furthermore, define

\[ P_{\text{pwr}}(w_i | w_{i-n'+1}^{i-1}) = P_{\text{alt}}(w_i | w_{i-n'+1}^{i-1}) + \gamma_0(w_{i-n'+1}^{i-1}) \left( \gamma_{0\rightarrow\eta}^{(p_{i,n})} Y_{D_i}^{(p_{i,n})}(w_i | w_{i-n'+1}^{i-1}) + \ldots \right) \]

(9.33)

Note that because of the way the discounts are computed in Algorithm 12,

\[ P_{\text{pwr}}(w_i | w_{i-n'+1}^{i-1}) = P_{\text{alt}}(w_i | w_{i-n'+1}^{i-1}) \]

for all \( n' \leq n \).

As a result, (for some choice of discount)

\[ P_{\text{pwr}}(w_i | w_{i-n'+1}^{i-1}) = P_{\text{kn}}(w_i | w_{i-n'+1}^{i-1}) \]  

(9.35)

Since, we know that Kneser Ney satisfies the marginal constraint (Chen and Goodman, 1999), this implies that,

\[ \hat{P}(w) = \sum_{w_{i-n'+1}^{i-1}} P_{\text{pwr}}(w_i | w_{i-n'+1}^{i-1}) \hat{P}(w_{i-n'+1}^{i-1}) \]

(9.36)

Thus, all we have to do is prove that

\[ \sum_{w_{i-n'+1}^{i-1}} P_{\text{pwr}}(w_i | w_{i-n'+1}^{i-1}) \hat{P}(w_{i-n'+1}^{i-1}) = \sum_{w_{i-n'+1}^{i-1}} P_{\text{pfr}}(w_i | w_{i-n'+1}^{i-1}) \hat{P}(w_{i-n'+1}^{i-1}) \]

(9.37)

Now, we follow the same argument as with \( n = 2 \) (i.e. unrolling the recursion and applying the fact that \( gKL \) preserves row/column sums).

For notational simplicity assume that \( n = 3 \). Then, we can write \( P_{\text{pwr}}(w_i | w_{i-1}^{i-1}) \) as:

\[ P_{\text{pwr}}(w_i | w_{i-2}^{i-1}) = \sum_{j=0}^{\eta-1} \gamma_{j-1}(w_{i-2}^{i-1}) Y_{D_j}^{(p_{i,n})}(w_i | w_{i-2}^{i-1}) + \sum_{j=0}^{\eta+1} \gamma_{j-1}(w_{i-2}^{i-1}) Y_{D_j}^{(p_{i,n})}(w_i | w_{i-1}) \]

(9.38)
where $\gamma_{0:-1}(w_{i-2}^{j-1}) = 1$ and

$$
\gamma_{0:-1}(w_{i-2}^{j-1}) = \prod_{h=0}^{j-1} \gamma_h(w_{i-2}^{j-1}) = \frac{d_s \sum_{i} c_{i,j-1,i-2}^{\theta_1} d_s \sum_{i} c_{i,j-1,i-2}^{\theta_2} ... d_s \sum_{i} c_{i,j-1,i-2}^{\theta_k}}{\sum_{i} \hat{c}_{i,j-1,i-2}} = \frac{(d_s) \sum_{i} \hat{c}_{i,j-1,i-2}}{\sum_{i} \hat{c}_{i,j-1,i-2}}
$$

(9.39)

Here $\hat{c}_{i,j-1,i-2}$ is shorthand for $\hat{c}(w_i, w_{i-2}^{j-1})$.

Similarly, $\gamma_{0:-1}(w_{i-1}) = 1$ and

$$
\gamma_{0:-1}(w_{i-1}) = \prod_{h=0}^{j-1} \gamma_h(w_{i-1}) = \frac{d_s \sum_{i} c_{i,j,i-1}^{\theta_1} d_s \sum_{i} c_{i,j,i-1}^{\theta_2} ... d_s \sum_{i} c_{i,j,i-1}^{\theta_k}}{\sum_{i} \hat{c}_{i,j,i-1}} = \frac{(d_s) \sum_{i} \hat{c}_{i,j,i-1}}{\sum_{i} \hat{c}_{i,j,i-1}}
$$

(9.40)

(Again, it is assumed that $1 \geq \rho_0 \geq ..., \rho_{\eta+1} = 0$.)

Analogously,

$$
P_{\text{pre}}(w_i|w_{i-2}^{j-1}) = \sum_{j=0}^{\eta} \gamma_{0:j-1}(w_{i-2}^{j-1}) Z_{D_j}^{(\rho_{i},\chi_j)}(w_i|w_{i-2}^{j-1}) + \sum_{j=0}^{\eta+1} \gamma_{0:j}(w_{i-2}^{j-1}) \gamma_{0:j-1}(w_{i-1}) Z_{D_j}^{(\rho_{i},\chi_j)}(w_i|w_{i-1})
$$

(9.41)

Now for any trigram term we prove that

$$
\sum_{w_{i-2}^{j-1}} \gamma_{0:j-1}(w_{i-2}^{j-1}) Y_{D_j}^{(\rho_{i},\chi_j)}(w_i|w_{i-2}^{j-1}) \hat{p}(w_{i-2}^{j-1}) = \sum_{w_{i-2}^{j-1}} \gamma_{0:j-1}(w_{i-2}^{j-1}) Z_{D_j}^{(\rho_{i},\chi_j)}(w_i|w_{i-2}^{j-1}) \hat{p}(w_{i-2}^{j-1})
$$

(9.42)

Plugging in the definition of $\gamma_{0:j-1}$ and simplifying gives

$$
\sum_{w_{i-2}^{j-1}} \left(\sum_{i} \hat{c}_{i,j-1,i-2}^{\theta_j}\right) Y_{D_j}^{(\rho_{i},\chi_j)}(w_i|w_{i-2}^{j-1}) = \sum_{w_{i-2}^{j-1}} \left(\sum_{i} \hat{c}_{i,j-1,i-2}^{\theta_j}\right) Z_{D_j}^{(\rho_{i},\chi_j)}(w_i|w_{i-2}^{j-1})
$$

(9.43)

which is equivalent to

$$
\sum_{w_{i-2}^{j-1}} Y_{D_j}^{(\rho_{i},\chi_j)}(w_i, w_{i-2}^{j-1}) = \sum_{w_{i-2}^{j-1}} Z_{D_j}^{(\rho_{i},\chi_j)}(w_i, w_{i-2}^{j-1})
$$

(9.44)

which holds because of the definition of $Z$ and the fact that rank minimization under $gKL$ preserves row/column sums.
Now consider any bigram term. We seek to show that:

$$\sum_{w_{i-2}^{j-1}} \gamma_{0,j}(w_{i-2}^{j-1}) \gamma_{0:j-1}(w_{i-1}) Y_{D_j}^{(p_j,\kappa_j)}(w_i|w_{i-1}) \hat{P}(w_{i-1}^{j-1}) = \sum_{w_{i-2}^{j-1}} \gamma_{0,j}(w_{i-2}^{j-1}) \gamma_{0:j-1}(w_{i-1}) Z_{D_j}^{(p_j,\kappa_j)}(w_i|w_{i-1}) \hat{P}(w_{i-1}^{j-1})$$ \hspace{1cm} (9.45)

Substituting definition of $\gamma_{0,j}(w_{i-2}^{j-1})$ gives

$$\sum_{w_{i-2}^{j-1}} \frac{(d_j)^{\eta+1}}{\sum_i \tilde{c}_{i,j-1,i-2}} \sum_i \bar{c}_{i,j-1,i-2} \gamma_{0:j-1}(w_{i-1}) Y_{D_j}^{(p_j,\kappa_j)}(w_i|w_{i-1}) \hat{P}(w_{i-1}^{j-1})$$

$$= \sum_{w_{i-2}^{j-1}} \frac{(d_j)^{\eta+1}}{\sum_i \tilde{c}_{i,j-1,i-2}} \sum_i \bar{c}_{i,j-1,i-2} \gamma_{0:j-1}(w_{i-1}) Z_{D_j}^{(p_j,\kappa_j)}(w_i|w_{i-1}) \hat{P}(w_{i-1}^{j-1})$$ \hspace{1cm} (9.46)

Simplifying and pushing in the sum over $w_{i-2}$ gives,

$$\sum_{w_{i-1}} (\sum_{i,j-2} \bar{c}_{i,j-1,i-2}) \gamma_{0:j-1}(w_{i-1}) Y_{D_j}^{(p_j,\kappa_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}} (\sum_{i,j-2} \bar{c}_{i,j-1,i-2}) \gamma_{0:j-1}(w_{i-1}) Z_{D_j}^{(p_j,\kappa_j)}(w_i|w_{i-1})$$ \hspace{1cm} (9.47)

Note that since $p_{\eta+1} = 0$, $\sum_{i,j-2} \bar{c}_{i,j-1,i-2} = \sum_i \tilde{c}_{i,j-1}$ (by definition of $\tilde{c}$).

Using this fact and substituting definition of $\gamma_{0:j-1}(w_{i-1})$ gives

$$\sum_{w_{i-1}} (\sum_i \tilde{c}_{i,j-1}) \frac{(d_j)^{\eta+1}}{\sum_i \tilde{c}_{i,j-1}} \sum_i \bar{c}_{i,j-1,i-2} Y_{D_j}^{(p_j,\kappa_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}} (\sum_i \tilde{c}_{i,j-1}) \frac{(d_j)^{\eta+1}}{\sum_i \tilde{c}_{i,j-1}} \sum_i \bar{c}_{i,j-1,i-2} Z_{D_j}^{(p_j,\kappa_j)}(w_i|w_{i-1})$$ \hspace{1cm} (9.48)

Simplifying gives,

$$\sum_{w_{i-1}} Y_{D_j}^{(p_j)}(w_i, w_{i-1}) = \sum_{w_{i-1}} Z_{D_j}^{(p_j,\kappa_j)}(w_i, w_{i-1})$$ \hspace{1cm} (9.49)

which holds because rank minimization under KL divergence preserves row and column sums. \hspace{1cm} □
Chapter 10

Conclusion

In this thesis, we have approached probabilistic modeling through the lens of linear and tensor algebra. Viewing latent variable models from this perspective allowed us to leverage tools such as matrix/tensor decomposition, inversion, and additive metrics to propose many novel solutions to parameter and structure learning as well as modeling with latent variables.

In particular, Part I proposed spectral learning algorithms for latent tree graphical models (Chapter 4 and 7) and latent junction trees (Chapter 5). Unlike traditional methods, our algorithms are provably consistent, local-optima-free, and 1-2 orders of magnitude faster than EM for large sample sizes. They also easily generalize to the continuous, nonparametric setting (Chapter 6).

In Part II, we used these insights to approach latent variable modeling for NLP from a linear algebra perspective. Our method for unsupervised parsing (Chapter 8) is the first algorithm that has both theoretical guarantees and is also practical, performing favorably to the CCM method of Klein and Manning. We also developed power low rank ensembles (Chapter 9), a framework for language modeling that generalizes existing \( n \)-gram techniques to non-integer \( n \). It consistently outperforms state-of-the-art Kneser Ney baselines and can train on billion-word datasets in a few hours.

Finally, this thesis also leaves open many directions for future work, two of which are described below.

**Discriminative spectral learning methods:** Spectral learning methods have been developed primarily for generative latent variable modeling. However, in many cases, discriminative techniques lead to better predictive power. A natural question to ask is can spectral learning be integrated with discriminative training?

This is an exciting problem that has recently been of interest in the spectral learning community ([Chaganty and Liang, 2013](#)) and is challenging since most discriminative methods (even for fully observed models) rely on optimization that can be difficult to integrate with spectral estimation. One promising idea is to discriminatively learn the projection matrix \( U \), that would learn a subspace that is more optimal for the prediction problem (an idea suggested by [Quattoni et al., 2014](#) for sequence models). There has been work along these lines in context of oriented PCA ([Platt et al., 2010](#)) for translingual document representations.
We are currently exploring spectral methods for max-margin parsing although the project is still in its preliminary stages (joint work with Mrinmaya Sachan, Shay Cohen, and Eric Xing).

Connecting Neural Networks and Latent Variable Models: This thesis has sought to develop connections between low rank decomposition and latent variable modeling, two different formulations of a similar intuition: that there exists an underlying “simpler” explanation of the data that is unobserved. However, neural networks are another popular technique that exploit a similar insight. Can these approaches be related to either linear algebra methods and/or latent variable models?

For example, let us consider recurrent neural networks for language modeling (Mikolov et al., 2010) that model sequences of words. Let \( w_t \) denote the \( V \times 1 \) indicator vector representation for \( w_t \) (the \( t \)th word in the sequence) and \( w_t(j) \) indicate its \( j \)th index. Let \( s_t \) denote the \( K \times 1 \) dimensional state vector at position \( t \) and \( y_t \) be the \( V \times 1 \) dimensional output vector that will represent the estimated probability \( \hat{P}(w_{t+1}|w_1, \ldots, w_t) \). Then, a recurrent neural network is defined by the following equations:

\[
\begin{align*}
  s_t(j) &= f \left( \sum_{i=1}^{V} w_t(i) \mu_{ji} + \sum_{l=1}^{K} s_{t-1}(l) \kappa_{jl} \right) \\
  y_t(m) &= g \left( \sum_{j=1}^{K} s_t(j) \nu_{mj} \right)
\end{align*}
\]

where \( \mu_{ji}, \kappa_{jl}, \nu_{mj} \) are weights to be learned. Typically \( f \) and \( g \) are set to:

\[
\begin{align*}
  f(z) &= \frac{1}{1 + \exp(-z)} \\
  g(z_m) &= \frac{\exp(z_m)}{\sum_k \exp(z_k)}
\end{align*}
\]

Therefore in a recurrent neural net, \( y_t \) only depends on the state vector \( s_t \). \( s_{t+1} \) depends on \( w_t \) and \( s_t \). As a result, the RNN requires \( O(K^2 + KV) \) parameters where \( K \) is the size of the latent dimension and \( V \) is the vocabulary size.

Note that the dependencies in an RNN are different from a traditional hidden Markov model (HMM) in one aspect, namely that \( s_{t+1} \) depends on both \( s_t \) and \( w_t \). This is different from a traditional HMM where the state at time \( t \) only depends on the state \( t - 1 \) via the transition matrix.

However, it is easy to alter the HMM so that it can also have the same dependences as the RNN with the same number of parameters. In particular we can have define the following model parameters for the prior, transition, and observation probabilities respectively:

\[
\begin{align*}
  \pi(i) &:= P(h_t = i) \\
  T(i, j, m) &:= P(h_{t+1} = i|h_t = j, w_t = m) := \frac{\exp(\alpha_{ij} + \beta_{i,m})}{\sum_n \exp(\alpha_{nj} + \beta_{n,m})} \\
  O(m, i) &:= P(w_t = m|h_t = i)
\end{align*}
\]
Note that here we have denoted the state by $h$ instead of $s$ as with the RNN since both models don’t have the same definition of state. A state in an HMM is a discrete valued variable (although the probability distribution over it is a real-valued vector) while the RNN state $s$ is itself a real-valued (non-probabilistic) vector. Moreover the transition tensor $T$ is now parameterized using log-linear models so that the number of parameters is still $O(K^2 + KV)$.

However, despite their similarity, there have been no theoretical or empirical comparisons between these two models. Such a comparison could shed light into the differences in representation power between these two classes of approaches. This is a line of work I’m currently exploring.
Bibliography


Hitchcock, F. L. (1927). The expression of a tensor or a polyadic as a sum of products. sn.


175


