Extending Spectral Methods to New Latent Variable Models
(CS 761)

David Merrell **Parikshit Sharma**
Department of Computer Sciences
University of Wisconsin–Madison
Madison, WI
{dmerrell,parikshit}@cs.wisc.edu

Abstract

Latent variable models are widely used in industry and research, though the problem of estimating their parameters has remained challenging; standard techniques (e.g., Expectation-Maximization) offer weak guarantees of optimality. There is a growing body of work reducing latent variable estimation problems to a certain(orthogonal) *spectral decompositions* of symmetric tensors derived from the moments of observed variables. Such decomposition allows a robust and computationally tractable estimation approach for several popular latent variable models; examples include topic mixture models, mixture of gaussians and HMM. In this report, we extend spectral methods to yet another class of latent variable models—topic transition models and mixture language models.

1 Introduction

Bayesian networks are a widely studied family of probabilistic models and find application in many areas, from topic modelling to speech recognition. We refer to Bayesian networks with unobserved variables as *latent variable models*. Estimating the parameters of latent variable models can be challenging, and commonly-used algorithms have weak guarantees of optimality. For example, the popular Expectation-Maximization algorithm only guarantees convergence to a *local* maximum in likelihood.

In response to these challenges, *spectral methods* have arisen as an alternative technique for estimating latent variable models. For some classes of latent variable model, parameter estimation can be formulated as *spectral decompositions* of matrices and tensors derived from empirical cross-moments. Spectral methods have been developed for a variety of latent variable models, including document topic models, Latent Dirichlet Allocation (LDA) [1], hidden Markov models (HMMs) [4], and Gaussian mixture models [3].

In this paper, we add to the existing work by extending spectral methods to some additional latent variable models. Specifically, we approach *topic transition models* and *mixture language models* from the spectral methods perspective. We chose these models, as they represent modest departures from models treated in past works.

2 Background

In this section we give a minimal background of tensors and their decompositions; describe spectral methods for latent variable models; and summarize results from earlier works.

Using NIPS 2017 style file

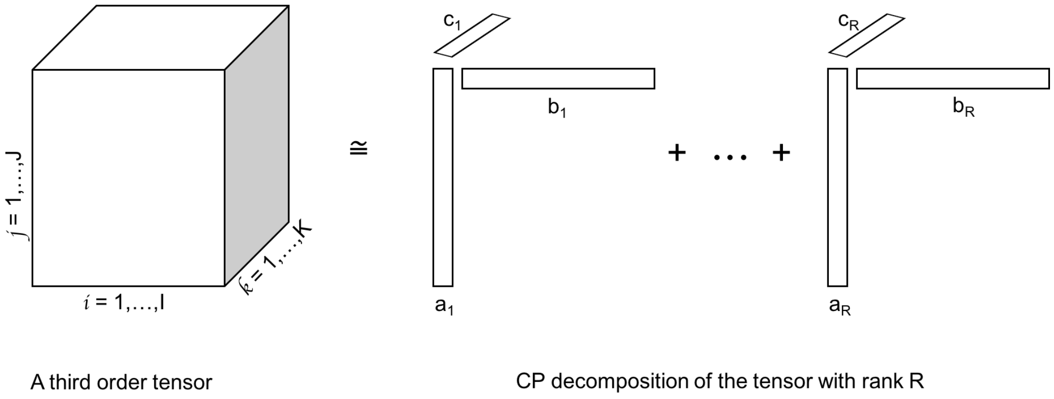


Figure 1: A third order tensor and its CP decomposition

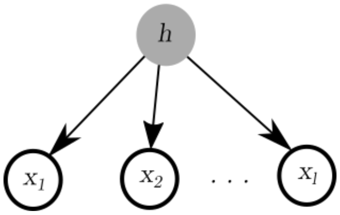


Figure 2: A document topic model; variables $x_1, \dots x_l$ are independent given the topic h.

2.1 Tensor Decompositions

Here we introduce the basics of tensors and their decompositions. Tensors are multidimensional extensions of matrices. An N-th order tensor is an element of the tensor product of N vector spaces, each of which has its own coordinate system. A vector is a 1st order tensor and a matrix is a 2nd order tensor. Figure 1 shows an example of a third order tensor.

Tensor decomposition can be viewed as a higher order generalization of the singular value decomposition (SVD) for matrices. One of the most widely used tensor decompositions is the CANDECOMP/PARAFAC (CP) decomposition [7]. The idea of CP decomposition is to express a tensor as a finite sum of rank-one tensors. For example, given a third order tensor $X \in \mathbb{R}^{I \times J \times K}$, CP decomposition would yield a tensor form like the following:

$$X \approx \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r,$$

where R is a positive integer (the *rank* of the tensor) and $a_r \in \mathbb{R}^I$, $b_r \in \mathbb{R}^J$, and $c_r \in \mathbb{R}^K$ are vectors.

We will show that CP decomposition provides a way to estimate the parameters of latent variable models from observable moments.

2.2 Overview of Spectral Methods

In the context of latent variable models, the purpose of a spectral method is to recover unknown model parameters by performing spectral decompositions (e.g. CP) on moments of observed variables. In this setting, such moments have the forms of tensors or matrices.

We adopt the following notation: let $u \otimes v$ denote the tensor product of vectors u and v; i.e. $u \otimes v$ is a second-order tensor ($uv^\top$), and $u \otimes v \otimes w$ is a third-order tensor.

We guide our discussion of spectral methods with the example of a *document-topic model*; see figure 2 for illustration. Let $h \in \mathbb{R}^k$ be a hidden *topic* variable. Let $x_1, \dots, x_l \in \mathbb{R}^d$ be the *view* variables; these correspond to the l words that appear in the document, chosen from a dictionary of d words. We encode the topic variable $h \in \mathbb{R}^k$ in the following way: $h = e_t$ (the t^{th} standard basis vector) iff

h is the t^{th} topic. Similarly, we encode the word variables $x_i \in \mathbb{R}^d$ as $x_i = e_j$ iff x_i is the j^{th} word in the dictionary. The practicality of this vector encoding will become apparent.

Assume the topic h is distributed according to a probability vector $w \in \mathbb{R}^k$; specifically, $P(h = e_t) = w^\top e_t$. Additionally, let the conditional probability of x_i on h be encoded as a matrix $S^{(x_i)} \in \mathbb{R}^{d\times k}$; given topic $h = e_t$, the distribution of x_i is represented by $S^{(x_i)} e_t$, the t^{th} column of $S^{(x_i)}$. Since h and x_i are encoded as standard basis vectors, their expected values correspond to their distributions; $\mathbb{E}[h] = w$, and $\mathbb{E}[x_i|h] = S^{(x_i)} h$.

Our goal is to recover the unknown parameters of this model—the matrices $S^{(x_i)}$ and the topic probability vector w. To do this, we will *(i)* compute moments of the view variables x_i, *(ii)* show that the resulting expressions contain the unknown parameters, and *(iii)* show that the unknown parameters can be recovered from these expressions.

We begin by computing $M_{1,2}$, a *second-order moment* over variables x_1 and x_2; i.e., pairs of words:

$$\begin{aligned} M_{1,2} &= \mathbb{E}[x_1 \otimes x_2] \\ &= \sum_t \Pr(h = t) \cdot \mathbb{E}[x_1 \otimes x_2 | h = t] \\ &= \sum_t w_t \cdot \mathbb{E}[x_1 | h = t] \otimes \mathbb{E}[x_2 | h = t] \\ &= \sum_t w_t \cdot \left(S^{(x_1)} e_t\right) \otimes \left(S^{(x_2)} e_t\right). \end{aligned}$$

When $S^{(x_1)} = S^{(x_2)} = S$, we say the model is *exchangeable*. Under this assumption, $M_{1,2}$ can be reduced to a *symmetric tensor form*:

$$\begin{aligned} &= \sum_t w_t \cdot (Se_t) \otimes (Se_t) \\ &= \sum_t w_t \cdot \mu_t \otimes \mu_t, \end{aligned} \tag{1}$$

where μ_t is the t^{th} column of the matrix S. This sum is a *spectral decomposition* of the moment matrix $M_{1,2}$. It seems that if we can estimate $M_{1,2}$ and compute its spectral decomposition, then we will recover the model parameters w and S.

Unfortunately, an additional matter of *identifiability* comes into play in the case of $M_{1,2}$: the decomposition in (1) is not generally unique. For example, any of the vectors μ_t can be acted on by a rotation matrix, and the resulting sum will still yield $M_{1,2}$. So the parameters are not identifiable from this decomposition.

This identifiability issue motivates the use of higher-order tensors, which have unique decompositions. For example, the third-order moment over variables x_1, x_2 and x_3 yields a similar expression:

$$\begin{aligned} M_{1,2,3} &= \mathbb{E}[x_1 \otimes x_2 \otimes x_3] \\ &= \sum_t w_t \cdot \mu_t \otimes \mu_t \otimes \mu_t, \end{aligned}$$

under the same exchangeability assumption. Unlike equation (1), however, this decomposition is unique and hence the parameters w and S can be recovered from it. Furthermore, such a symmetric decomposition can be computed using efficient *iterative power methods*, as described in [3, 2]. Older works, e.g. [4, 1] instead recover the parameters by contracting $M_{1,2,3}$ with randomly generated vectors and solving a sequence of eigenproblems on the resulting matrices; this older method is less efficient than the iterative power method, from the standpoints of computational and sample complexity [2].

As described in [2], spectral methods can be extended to models failing the exchangeability assumption. This is done by *(i)* applying linear transformations to the observed variables (in effect, symmetrizing them); *(ii)* performing the previously-discussed technique for symmetric tensors; and *(iii)* recovering the original (asymmetric) parameters by applying inverses of the linear transformations in step *(i)*.

For illustration we apply this process to the previously-discussed document topic model, removing the assumption that $S^{(x_i)} = S^{(x_j)}$ for all i, j. First, define linear transformations $C_{2\to1}$ and $C_{3\to1}$ as

$$C_{2\to1} = M_{1,3}M_{2,3}^{+} \qquad\qquad C_{3\to1} = M_{1,2}M_{3,2}^{+}$$

where $+$ denotes the Moore-Penrose pseudoinverse. Then define variables $\tilde{x}_2, \tilde{x}_3$ as

$$\tilde{x}_2 = C_{2\to1}x_2 \qquad\qquad \tilde{x}_3 = C_{3\to1}x_3.$$

As shown in [2], it follows that

$$\mathbb{E}\left[x_1 \otimes \tilde{x}_2 \otimes \tilde{x}_3\right] = \sum_t w_t \cdot \left(S^{(x_1)}e_t\right) \otimes \left(S^{(x_1)}e_t\right) \otimes \left(S^{(x_1)}e_t\right).$$

Hence, the matrix $S^{(x_1)}$ and vector w can be recovered by decomposing this moment. Furthermore, [2] shows that $S^{(x_2)}$ and $S^{(x_3)}$ can be recovered via

$$S^{(x_2)} = (C_{2\to1})^{-1}S^{(x_1)} \qquad\qquad S^{(x_3)} = (C_{3\to1})^{-1}S^{(x_1)}.$$

This method extends the reach of spectral methods to a host of asymmetric/nonexchangeable models; e.g. LDA and HMMs [2], and Gaussian mixtures [3].

2.3 Related Works

Spectral methods for estimating parameters of multi-view mixture models and HMMs are examined in [4]. Estimation is reduced to a sequence of eigenproblems for matrices derived from second and third moments.

In [3] the connection between latent variable model estimation and tensor decompositions is elucidated. The results of several previous works are couched in terms of low-rank symmetric tensor decompositions, and an efficient iterative power method is proposed for computing the decomposition. This iterative power method is shown to be more efficient than the eigenproblem-based methods from [4].

Estimation for Latent Dirichlet Allocation is formulated as a tensor decomposition in [1], [2], and [3], with [1] giving the most detailed exposition. [1] also establishes the symmetrizing technique for multi-view models demonstrated in section 2.2, allowing a broader set of models (e.g., HMMs) to be estimated via efficient symmetric tensor decompositions.

In [6], the tensor decomposition methods of Anandkumar et. al. are extended to shallow neural networks, under the assumption that the distribution over inputs is known. The resulting procedure trains shallow neural networks to global optimality, avoiding the well known local-optimality disadvantages of gradient-based techniques.

3 Extensions to New Models

In this section, we describe our progress in applying spectral methods to models beyond those addressed in previous works. Each new model differs from previously-treated models in some aspect; they include *topic transition models* and *mixture language models*. By applying spectral methods to new kinds of models, we hope to further characterize the set of models that can be learned effectively this way.

3.1 Topic Transition Models

Past works ([4, 3]) have applied spectral methods to HMMs and document topic models. In this section, we present a model that contains aspects of both HMMs and document topic models; we call it a *topic transition model*. We apply spectral methods to it, and find that it can be learnt in a restricted manner using methods equivalent to those for HMMs.

The topic transition model is illustrated in Figure 3 where each sentence is drawn from a different topic and a transition matrix governs the topic of the next sentence given the topic of the previous sentence. The observables are the words in a sentence and the words are independently drawn from the topic of the sentence. So given the topic $h \in \mathbb{R}^k$, the words are conditionally independent.

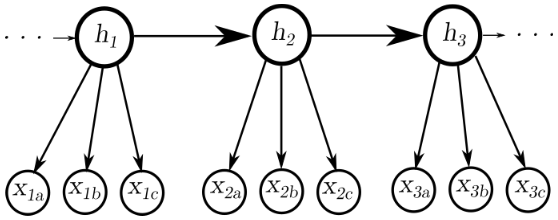
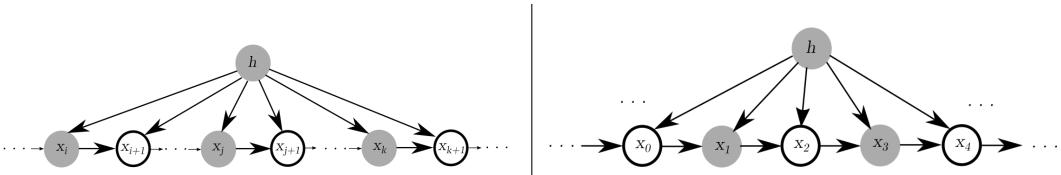


Figure 3: Topic transition model.



(a) First attempt

(b) Second attempt

Figure 4: Mixture language models we attempted to solve. (a) In the first attempt, we *fix* three of the view variables and compute moments over their immediate successors; this is then repeated for every combination of three view variables. (b) In our second attempt, views are assumed to be part of a topic-specific transition system, with x_0 distributed according to the transition system's stationary distribution.

Another application of these models is in activity detection where the observables are different sensors and a transition matrix governs what activity follows next. Observables are grouped into three categories $x_{1a}, x_{2a}, x_{3a} \in \mathbb{R}^p$; $x_{1b}, x_{2b}, x_{3b} \in \mathbb{R}^q$ and $x_{1c}, x_{2c}, x_{3c} \in \mathbb{R}^r$. Think of them as three different types of sensors. Each group follows the same conditional distribution given the topic/activity, as encoded in the matrix $S^{(1)} \in \mathbb{R}^{p\times k}$, $S^{(2)} \in \mathbb{R}^{q\times k}$ and $S^{(3)} \in \mathbb{R}^{r\times k}$, for e.g. given topic $h = e_t$, the distribution of x_{1a} is represented by $S^{(1)}e_t$, the t-th column of $S^{(1)}$. Let $T \in \mathbb{R}^{k\times k}$ be the topic transition matrix i.e. $T_{pq} = \Pr(h_i = p|h_{i-1} = q)$.

We consider one group of random variables, say x_{1a}, x_{2a}, x_{3a}. Suppose we condition the distribution of the three observed variables on h_2. We can see that given h_2, the three random variables are conditionally independent. Let π denote the distribution of the initial state h_1. Then,

- The distribution of h_2 is given by $w := T\pi$
- for all $j \in [k]$:

$$\mathbb{E}[x_{1a}|h_2 = j] = S^{(1)} \operatorname{diag}(\pi)T^\top \operatorname{diag}(w)^{-1}e_j$$
$$\mathbb{E}[x_{2a}|h_2 = j] = S^{(1)}e_j$$
$$\mathbb{E}[x_{3a}|h_2 = j] = S^{(1)}Te_j$$

This now corresponds to the case we had discussed in section 2.2 where we have one topic h and three variables that are distributed differently conditioned on h. We can use the same algorithm to first estimate $\mathbb{E}[x_{2a}|h_2]$, which gives us $S^{(1)}$ and then estimate $\mathbb{E}[x_{2a}|h_2]$ to get T. Similar procedure can be followed for the other two group of variables to estimate $S^{(2)}$ and $S^{(3)}$. However, these matrices are subject to permutation and we could not find a way to align them together unless the $P[h]$ values are distinct and come out to be very close in the three iterations or unless the matrices are identical, which usually happens in topic models.

3.2 Mixture Language Models

The models considered in past works share an important property: their observed variables are all independent, given some latent variable. We explored beyond that property by attempting to apply spectral methods to a bigram *mixture-language* model, as described in [5] and [8]. This model is similar to the document-topic model considered in section 2.2, but assumes additional dependencies between words. See Figure 4 for illustration. Intuitively, it represents the topic-dependence of word *sequences* in a document.

We made multiple attempts at applying spectral methods to this model. Two particular attempts are outlined in the subsequent subsections. In the following subsection, we try to separate the original problem into many smaller problems by holding certain variables fixed, and solving a problem for every combination of those fixed variables. In the subsection after that, we make a direct attempt at computing moments for the full model and expressing them as useful spectral decompositions.

3.2.1 Attempt 1 – Condition on Fixed Observed Variables

Let k be the number of distinct topics in the corpus, d be the number of distinct words in the vocabulary represented by $V := \{v_1, v_2, \ldots, v_d\}$, and $l \geq 5$ be the number of words in each document represented by $\{x_1, x_2, \ldots, x_l\}$. The generative process for a document is as follows: the document's topic h is drawn according to the discrete distribution specified by the probability vector $w \in \mathbb{R}^k$ and then each word x_i is drawn conditioned on the topic of the document and the previous word $x_i \sim \mu_{t,j}$ where $\mu_{t,j} := \Pr(x_i | h = t, x_{i-1} = w_j) \in \mathbb{R}^d$. The parameter estimation task is to estimate w and $\mu_{t,j}$ for $t \in [k]$ and $j \in [d]$. Each document is assumed to begin with a fixed start word $x_0 = v_0$.

Given, the size of the parameter space, it seems natural to design an iterative algorithm to estimate parameters progressively. So we begin with a restricted model. Referring to figure 4a, we fix $x_i = v_a$, $x_j = v_b$ and $x_k = v_c$ from the vocabulary. Under this restricted model, x_{i+1}, x_{j+1} and x_{k+1} are conditionally independent of each other given the fixed words and the topic. $\Pr(x_{i+1}) = \mu_{h,a}$, $\Pr(x_{j+1}) = \mu_{h,b}$ and $\Pr(x_{k+1}) = \mu_{h,c}$. It will be convenient to represent these pairs of words by a single d-dimensional random variable $\in \mathbb{R}^d$.

Let $(x_i, x_{i+1}) := y_1$ and $y_1 = e_k$ if and only if the (i)-th word in the document is w_a and (i+1)-th word is w_k (e_k is the k-th basis vector in $\mathbb{R}^d$). Similarly $(x_j, x_{j+1}) := y_2$ and $(x_k, x_{k+1}) := y_3$.

We now calculate the moment of y_1, y_2 and y_3 under this restricted model.

$$
\begin{aligned}
&= E[y_1 \otimes y_2 \otimes y_3] \\
&= \sum_{1 \leq p,q,r \leq d} \Pr[y_1 = e_p, y_2 = e_q, y_3 = e_r] \cdot e_p \otimes e_q \otimes e_r \\
&= \sum_{1 \leq p,q,r \leq d} \Pr[x_i = v_a, x_{i+1} = v_p, x_j = v_b, x_{j+1} = v_q, x_k = v_c, x_{k+1} = v_r] \cdot e_p \otimes e_q \otimes e_r \\
&= \sum_{1 \leq p,q,r \leq d} \sum_h \Pr[x_i = v_a, x_{i+1} = v_p, x_j = v_b, x_{j+1} = v_q, x_k = v_c, x_{k+1} = v_r, h] \cdot e_p \otimes e_q \otimes e_r \\
&= \sum_{1 \leq p,q,r \leq d} \sum_h \Pr[h] \Pr[x_i = v_a, x_{i+1} = v_p | h] \\
&\qquad\qquad \cdot \Pr[x_j = v_b, x_{j+1} = v_q | h, x_{i+1}] \\
&\qquad\qquad \cdot \Pr[x_k = v_c, x_{k+1} = v_r | h, x_{j+1}] \cdot e_p \otimes e_q \otimes e_r \\
&= \sum_{1 \leq p,q,r \leq d} \sum_h \Pr[h] \Pr[x_{i+1} = v_p | h, x_i = v_a] \Pr[x_i = v_a | h] \\
&\qquad\qquad \cdot \Pr[x_{j+1} = v_q | h, x_j = v_b] \Pr[x_j = v_b | h, x_{i+1}] \\
&\qquad\qquad \cdot \Pr[x_{k+1} = v_r | h, x_k = v_c] \Pr[x_k = v_c | h, x_{j+1}] \cdot e_p \otimes e_q \otimes e_r
\end{aligned}
$$

Note that $\Pr[x_i = v_a | h] = 1$; $\Pr[x_j = v_b | h, x_{i+1}] = 1$; $\Pr[x_k = v_c | h, x_{j+1}] = 1$ since we have fixed these nodes to always have those values i.e. we only consider those documents which have the chosen words v_a, v_b and v_c in the same (i, j, k)-th positions, where the value of i, j and k is variable.

$$
\begin{aligned}
&= \sum_{1\le p,q,r\le d} \sum_h \Pr[h] \Pr[x_{i+1} = v_p | h, x_i = v_a] \\
&\qquad\qquad \cdot \Pr[x_{j+1} = v_q | h, x_j = v_b] \\
&\qquad\qquad \cdot \Pr[x_{k+1} = v_r | h, x_k = v_c] \cdot e_p \otimes e_q \otimes e_r \\
&= \sum_h \Pr[h] \sum_{1\le p\le d} \Pr[x_{i+1} = v_p | h, x_i = v_a] \cdot e_p \\
&\qquad \otimes \sum_{1\le q\le d} \Pr[x_{j+1} = v_q | h, x_j = v_b] \cdot e_q \\
&\qquad \otimes \sum_{1\le r\le d} \Pr[x_{k+1} = v_r | h, x_k = v_c] \cdot e_r \\
&= \sum_h \Pr[h] E[x_{i+1} | h, x_i = v_a] \otimes E[x_{j+1} | h, x_j = v_b] \otimes E[x_{k+1} | h, x_k = v_c] \\
&= \sum_h \Pr[h] \mu_{h,a} \otimes \mu_{h,b} \otimes \mu_{h,c}
\end{aligned}
$$

So we have successfully decomposed the third order tensor into a sum of rank one tensors. The problem now is to estimate $\mathbb{E}[y_1 \otimes y_2 \otimes y_3]$. Note that the (p, q, r)-th entry of $\mathbb{E}[y_1 \otimes y_2 \otimes y_3]$ is

$$
\begin{aligned}
\Pr\,(i^{th} \text{ word} = v_a,\ \ &(i+1)^{th} \text{ word} = v_p, \\
j^{th} \text{ word} = v_b,\ \ &(j+1)^{th} \text{ word} = v_q, \\
k^{th} \text{ word} = v_c,\ \ &(k+1)^{th} \text{ word} = v_r)
\end{aligned}
$$

To calculate this, first we fix the three words v_a, v_b , and v_c. Next we find the index i, j, k such that $x_i = v_a, x_j = v_b, x_k = v_c$ in each document. Documents are clustered based on the value of i, j, k and the set with the highest number of documents is picked for statistical accuracy. Now the task is estimating joint probability of the $i + 1, j + 1, k + 1$ words in this document set, which can be done by counting the occurence of each triplet of words from the vocabulary in these positions.

Design choices for v_a, v_b, v_c.

- $v_a = v_b = v_c$ In this case the decomposition and parameter estimation problem reduces to a symmetric exchangeable model similar to topic model.
- $v_a = w_0 \neq v_b, v_c$ In this case (i + 1) becomes the 1st word and v_b and v_c can be set to be equal or unequal. The model is no longer symmetric and exchangeable. However, it now resembles the asymmetric case introduced in section 2.2 and the same techniques can be used to make the model symmetric for parameter estimation.

Algorithm: We iteratively select v_a, v_b, v_c and use tensor decompsition to learn $\mu_{h,a}, \mu_{h,b}, \mu_{h,c}$ for $h \in [k]$. So, in each iteration we can possibly learn $3 \times k$ parameters of the total $d \times t$ parameters. However, note that the solutions in each iteration are subject to permutation so unless the values of $P[h]$ are distinct and come out reasonably close in each iteration we will not be able to align the $\mu_{t,i}$ values. This is one significant limitation of this approach which made us explore if a direct joint spectral decomposition was possible.

3.2.2 Attempt 2 – Apply spectral method directly

In this section, we describe another attempt at applying spectral methods to a mixture language model—see figure 4b for illustration. In this attempt, we consider views that occur at some point within a sequence of observations.

We again assume the hidden topic variable h is distributed according to a probability vector w, and that the distribution of word x_i depends on the previous word x_{i-1} according to a topic-specific transition matrix, T_t; i.e., $\Pr(x_i | h = t, x_{i-1} = e_j) = T_t e_j$, the j^{th} column of T_t. Furthermore, we assume that the word x_0 occurs sufficiently late in the sequence that its distribution is π_t, the *stationary distribution* of T_t.

We make a straightforward effort to compute second- and third-order moments of the variables x_i. However, it soon becomes apparent that the methods described in section 2.2 are not readily applicable to this model. We describe our work and observations.

In order to obtain a tensor form as was done in section 2.2, we must condition on view variables as well as the topic variable when we compute moments. For example, we condition on variables h, x_1, and x_3 when we compute the following third-order moment:

$$\begin{aligned}\mathbb{E}\left[x_0 \otimes x_2 \otimes x_4\right] &= \sum_{t,i,j} \Pr(h = e_t, x_1 = e_i, x_3 = e_j) \cdot \mathbb{E}\left[x_0 \otimes x_2 \otimes x_4 | h = e_t, x_1 = e_i, x_3 = e_j\right] \\ &= \sum_{t,i,j} \Pr(h = e_t, x_1 = e_i, x_3 = e_j) \cdot \mathbb{E}\left[x_0 | h = e_t, x_1 = e_i, x_3 = e_j\right] \\ &\qquad \otimes \mathbb{E}\left[x_2 | h = e_t, x_1 = e_i, x_3 = e_j\right] \\ &\qquad \otimes \mathbb{E}\left[x_4 | h = e_t, x_1 = e_i, x_3 = e_j\right].\end{aligned}$$

We find the following expressions in terms of the model parameters:

$$\Pr(h, x_1, x_3) = w_t \left(x_3^\top T_h^2 x_1\right) \left(x_1^\top \pi_t\right),$$

$$\begin{aligned}\mathbb{E}\left[x_0 | h = e_t, x_1 = e_i, x_3 = e_j\right] &= \frac{1}{x_1^\top \pi_t} \operatorname{diag}(\pi_h)(T_h^\top) x_1, \\ \mathbb{E}\left[x_2 | h = e_t, x_1 = e_i, x_3 = e_j\right] &= \frac{1}{x_3^\top T_h^2 x_1} \operatorname{diag}(x_3^\top T_h) T_h x_1, \\ \mathbb{E}\left[x_4 | h = e_t, x_1 = e_i, x_3 = e_j\right] &= T_h x_3.\end{aligned}$$

Hence, our third-order moment can be written as

$$\mathbb{E}\left[x_0 \otimes x_2 \otimes x_4\right] = \sum_{t,i,j} w_t \cdot \left(\operatorname{diag}(\pi_t) T_t^\top e_i\right) \otimes \left(\operatorname{diag}(e_j^\top T_t) T_t e_i\right) \otimes \left(T_t e_j\right).$$

It is difficult to apply existing spectral methods to this tensor form for several reasons. We immediately see that it is not symmetric. Additionally, it is a sum over three indices (rather than one), and it is unclear whether any of the extra indices can be eliminated. Without resolving the question about indices, it is unclear whether the form satisfies necessary conditions for applying the symmetrization technique described in section 2.2—specifically, it may not be possible to express each of the conditional expectations in a manner that allows recovery of the transition matrices T. Clearly, applying spectral methods to this model will require the invention of new techniques. We have been unable to develop any such technique in the duration of this project.

4 Conclusions

In this report we summarized the current status of spectral methods for learning latent variable models, and described our efforts to extend those techniques to new classes of models. We found that topic transition models can be treated with the pre-existing spectral methods for HMMs. We also made multiple attempts at applying spectral methods to mixture language models; one attempt yielded an algorithm with high computational cost, while another attempt yielded a tensor form that does not fit into currently existing techniques.

In the future, it would be worthwhile to develop the theory characterizing the correspondence between latent variable models and tensor forms, rather than pursuing the ad-hoc, model-by-model approach taken by existing research.

References

[1] A. Anandkumar, D. Foster, D. Hsu, S. Kakade, and Y. Liu. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012.

[2] A. Anandkumar, D. Foster, D. Hsu, S. Kakade, and Y. Liu. A spectral algorithm for latent dirichlet allocation. In *Algorithmica*, 2015.

[3] A. Anandkumar, R. Ge, D. Hsu, S. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, (15):2773–2832, 2014.

[4] A. Anandkumar, D. Hsu, and S. Kakade. A method of moments for mixture models and hidden markov models. In *25th Annual Conference on Learning Theory*, 2012.

[5] R. Iyer and M. Ostendorf. Modeling long distance dependence in language: Topic mixtures versus dynamic cache models. *IEEE Transactions on Speech and Audio Processing*, 7(1):30–39, 1999.

[6] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. ArXiv:1506.08473, 2015.

[7] T. Kolda and B. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[8] Y. Lobacheva. Discourse mixture language modeling. M.S. thesis, Boston University College of Engineering, Jan 2000.