A provable SVD-based algorithm for learning topics in dominant admixture corpus

Trapit Bansal†, C. Bhattacharyya‡
Department of Computer Science and Automation
Indian Institute of Science
Bangalore -560012, India
†trapitbansal@gmail.com
‡chiru@csa.iisc.ernet.in
Ravindran Kannan
Microsoft Research
India
kannan@microsoft.com

Abstract

Topic models, such as Latent Dirichlet Allocation (LDA), posit that documents are drawn from admixtures of distributions over words, known as topics. The inference problem of recovering topics from such a collection of documents drawn from admixtures, is NP-hard. Making a strong assumption called separability, [4] gave the first provable algorithm for inference. For the widely used LDA model, [6] gave a provable algorithm using clever tensor-methods. But [4, 6] do not learn topic vectors with bounded $l_1$ error (a natural measure for probability vectors).

Our aim is to develop a model which makes intuitive and empirically supported assumptions and to design an algorithm with natural, simple components such as SVD, which provably solves the inference problem for the model with bounded $l_1$ error. A topic in LDA and other models is essentially characterized by a group of co-occurring words. Motivated by this, we introduce topic specific Catchwords, a group of words which occur with strictly greater frequency in a topic than any other topic individually and are required to have high frequency together rather than individually. A major contribution of the paper is to show that under this more realistic assumption, which is empirically verified on real corpora, a singular value decomposition (SVD) based algorithm with a crucial pre-processing step of thresholding, can provably recover the topics from a collection of documents drawn from Dominant admixtures. Dominant admixtures are convex combination of distributions in which one distribution has a significantly higher contribution than the others. Apart from the simplicity of the algorithm, the sample complexity has near optimal dependence on $\omega_0$, the lowest probability that a topic is dominant, and is better than [4]. Empirical evidence shows that on several real world corpora, both Catchwords and Dominant admixture assumptions hold and the proposed algorithm substantially outperforms the state of the art [5].

1 Introduction

Topic models [1] assume that each document in a text corpus is generated from an ad-mixture of topics, where, each topic is a distribution over words in a Vocabulary. An admixture is a convex combination of distributions. Words in the document are then picked in i.i.d. trials, each trial has a multinomial distribution over words given by the weighted combination of topic distributions. The problem of inference, recovering the topic distributions from such a collection of documents, is provably NP-hard. Existing literature pursues techniques such as variational methods [2] or MCMC procedures [3] for approximating the maximum likelihood estimates.

Given the intractability of the problem one needs further assumptions on topics to derive polynomial time algorithms which can provably recover topics. A possible (strong) assumption is that each
A document has only one topic but the collection can have many topics. A document with only one topic is sometimes referred as a pure topic document. [7] proved that a natural algorithm, based on SVD, recovers topics when each document is pure and in addition, for each topic, there is a set of words, called primary words, whose total frequency in that topic is close to 1. More recently, [6] show using tensor methods that if the topic weights have Dirichlet distribution, we can learn the topic matrix. Note that while this allows non-pure documents, the Dirichlet distribution gives essentially uncorrelated topic weights.

In an interesting recent development [4, 5] gave the first provable algorithm which can recover topics from a corpus of documents drawn from admixtures, assuming separability. Topics are said to be separable if in every topic there exists at least one Anchor word. A word in a topic is said to be an Anchor word for that topic if it has a high probability in that topic and zero probability in remaining topics. The requirement of high probability in a topic for a single word is unrealistic.

Our Contributions: Topic distributions, such as those learnt in LDA, try to model the co-occurrence of a group of words which describes a theme. Keeping this in mind we introduce the notion of Catchwords. A group of words are called Catchwords of a topic, if each word occurs strictly more frequently in the topic than other topics and together they have high frequency. This is a much weaker assumption than separability. Furthermore we observe, empirically, that posterior topic weights assigned by LDA to a document often have the property that one of the weights is significantly higher than the rest. Motivated by this observation, which has not been exploited by topic modeling literature, we suggest a new assumption. It is natural to assume that in a text corpus, a document, even if it has multiple themes, will have an overarching dominant theme. In this paper we focus on document collections drawn from dominant admixtures. A document collection is said to be drawn from a dominant admixture if for every document, there is one topic whose weight is significantly higher than other topics and in addition, for every topic, there is a small fraction of documents which are nearly purely on that topic. The main contribution of the paper is to show that from a corpus of documents drawn from admixtures, assuming separability if for every document, there is one topic whose weight is significantly higher than the rest. Our algorithm, which we call TSVD, indeed provably finds a good approximation in total $L_1$ error to the topic matrix. We prove a bound on the error of our approximation which does not grow with dictionary size $d$, unlike [5] where the error grows linearly with $d$.

Empirical evidence shows that on semi-synthetic corpora constructed from several real world datasets, as suggested by [5], TSVD substantially outperforms the state of the art [5]. In particular it is seen that compared to [5] TSVD gives 27% lower error in terms of $L_1$ recovery on 90% of the topics.

Problem Definition: $d, k, s$ will denote respectively, the number of words in the dictionary, number of topics and number of documents. $d, s$ are large, whereas, $k$ is to be thought of as much smaller. Let $S_k = \{x = (x_1, x_2, \ldots, x_k) : x_l \geq 0; \sum_l x_l = 1\}$. For each topic, there is a fixed vector in $S_k$ giving the probability of each word in that topic. Let $M$ be the $d \times k$ matrix with these vectors as its columns.

Documents are picked in i.i.d. trials. To pick document $j$, one first picks a $k-$ vector $W_{ij}, W_{2j}, \ldots, W_{kj}$ of topic weights according to a fixed distribution on $S_k$. Let $P_{ij} = MW_{ij}$ be the weighted combination of the topic vectors. Then the $m$ words of the document are picked in i.i.d. trials; each trial picks a word according to the multinomial distribution with $P_{ij}$ as the probabilities. All that is given as data is the frequency of words in each document, namely, we are given the $d \times s$ matrix $A$, where $A_{ij} = \frac{\text{Number of occurrences of word } i \text{ in Document } j}{m}$. Note that $E(A|W) = P$, where, the expectation is taken entry-wise.

In this paper we consider the problem of finding $M$ given $A$.

2 Previous Results

In this section we review literature related to designing provable algorithms for topic models. For an overview of topic models we refer the reader to the excellent survey[1]. Provably algorithms for recovering topic models was started by [7]. Latent Semantic Indexing(LSI) [8] remains a successful method for retrieving similar documents by using SVD. [7] showed that one can recover $M$ from a collection of documents, with pure topics, by using SVD based procedure under the additional Primary Words Assumption. [6] showed that in the admixture case, if one assumes Dirichlet
distribution for the topic weights, then, indeed, using tensor methods, one can learn $\mathbf{M}$ to $l_2$ error provided some added assumptions on numerical parameters like condition number are satisfied.

The first provably polynomial time algorithm for admixture corpus was given in [4, 5]. For a topic $l$, a word $i$ is an anchor word if $M_{i,l} \geq p_0$; $M_{i,l'} = 0 \forall l' \neq l$.

**Theorem 2.1** [4] If every topic has an anchor word, there is a polynomial time algorithm that returns an $\hat{\mathbf{M}}$ such that with high probability,

$$\sum_{l=1}^{k} \sum_{i=1}^{d} |\hat{M}_{il} - M_{il}| \leq d\varepsilon \text{ provided } s \geq \text{Max} \left\{ O \left( \frac{k^6 \log d}{a^2 \varepsilon^2 p_0^6 \gamma^2 m} \right) , O \left( \frac{k^4}{\gamma^2 a^2} \right) \right\},$$

where, $\gamma$ is the condition number of $E(WW^T)$, $a$ is the minimum expected weight of a topic and $m$ is the number of words in each document.

Note that the error grows linearly in the dictionary size $d$, which is often large. Note also the dependence of $s$ on parameters $p_0$, which is $1/p_0^6$ and on $a$, which is $1/a^4$. If, say, the word “run” is an anchor word for the topic “baseball” and $p_0 = 0.1$, then the requirement is that every 10th word in a document on this topic is “run”. This seems too strong to be realistic. It would be more realistic to ask that a set of words like - “run”, “hit”, “score”, etc. together have frequency at least 0.1 which is what our catchwords assumption does.

## 3 Learning Topics from Dominant Admixtures

Informally, a document is said to be drawn from a Dominant Admixture if the document has one dominant topic. Besides its simplicity, we show empirical evidence from real corpora to demonstrate that topic dominance is a reasonable assumption. The Dominant Topic assumption is weaker than the Pure Topic assumption. More importantly SVD based procedures proposed by [7] will not apply. Inspired by the Primary words assumption we introduce the assumption that each topic has a set of Catchwords which individually occur more frequently in that topic than others. This is again a much weaker assumption than both Primary Words and Anchor Words assumptions and can be verified experimentally. In this section we establish that by applying SVD on a matrix, obtained by thresholding the word-document matrix, and subsequent $k$ means clustering can learn topics having Catchwords from a Dominant Admixture corpus.

### 3.1 Assumptions: Catchwords and Dominant admixtures

Let $\alpha, \beta, \rho, \delta, \varepsilon_0$ be non-negative reals satisfying:

\begin{align*}
\beta + \rho & \leq (1 - \delta)\alpha. \quad (1) \\
\alpha + 2\delta & \leq 0.5; \quad \delta \leq 0.08. \quad (2)
\end{align*}

**Dominant topic Assumption** (a) For $j = 1, 2, \ldots, s$, document $j$ has a dominant topic $l(j)$ such that $W_{l(j),j} \geq \alpha$ and $W_{l',j} \leq \beta$, $\forall l' \neq l(j)$.

(b) For each topic $l$, there are at least $\varepsilon_0 w_0 s$ documents in each of which topic $l$ has weight at least $1 - \delta$.

**Catchwords Assumption:** There are $k$ disjoint sets of words - $S_1, S_2, \ldots, S_k$ such that with $\varepsilon$ defined in (9)

\begin{align*}
\forall i \in S_l, \forall l' \neq l, \ M_{il'} & \leq p M_{il} \quad (3) \\
\sum_{i \in S_l} M_{il} & \geq p_0 \quad (4) \\
\forall i \in S_l, m\delta^2 a M_{il} & \geq 8\ln \left( \frac{20}{\varepsilon w_0} \right). \quad (5)
\end{align*}

Part (b.) of the Dominant Topic Assumption is in a sense necessary for “identifiability” - namely for the model to have a set of $k$ document vectors so that every document vector is in the convex
hull of these vectors. The Catchwords assumption is natural to describe a theme as it tries to model a unique group of words which is likely to co-occur when a theme is expressed. This assumption is close to topics discovered by LDA like models, which try to model co-occurrence of words. If \( \alpha, \delta \in \Omega(1) \), then, the assumption (5) says \( M_{il} \in \Omega^*(1/m) \). In fact if \( M_{il} \in o(1/m) \), we do not expect to see word \( i \) (in topic \( l \)), so it cannot be called a catchword at all.

A slightly different (but equivalent) description of the model will be useful to keep in mind. What is fixed (not stochastic) are the matrices \( M \) and the distribution of the weight matrix \( W \). To pick document \( j \), we can first pick the dominant topic \( l \) in document \( j \) and condition the distribution of \( W_{lj} \) on this being the dominant topic. One could instead also think of \( W_{lj} \) being picked from a mixture of \( k \) distributions. Then, we let \( P_{ij} = \sum_{l=1}^{M} M_{il} W_{lj} \) and pick the \( m \) words of the document in i.i.d multinomial trials as before. We will assume that

\[
T_l = \{ j : l \text{ is the dominant topic in document } j \} \text{ satisfies } |T_l| = w_l s,
\]

where, \( w_l \) is the probability of topic \( l \) being dominant. This is only approximately valid, but the error is small enough that we can disregard it.

For \( \zeta \in \{0, 1, 2, \ldots, m\} \), let \( p_i(\zeta, l) \) be the probability that \( j \in T_l \) and \( A_{ij} = \zeta / m \) and \( q_i(\zeta, l) \) the corresponding “empirical probability”:

\[
p_i(\zeta, l) = \int_{W_{lj}} \left( \frac{m}{\zeta} \right) P_{ij}^\zeta (1 - P_{ij})^{m-\zeta} \text{Prob}(W_{lj} \mid j \in T_l) \text{Prob}(j \in T_l), \quad \text{where, } P_{ij} = MW_{lj}. \tag{6}
\]

\[
q_i(\zeta, l) = \frac{1}{s} | \{ j \in T_l : A_{ij} = \zeta / m \} |. \tag{7}
\]

Note that \( p_i(\zeta, l) \) is a real number, whereas, \( q_i(\zeta, l) \) is a random variable with \( E(q_i(\zeta, l)) = p_i(\zeta, l) \). We need a technical assumption on the \( p_i(\zeta, l) \) (which is weaker than unimodality).

**No-Local-Min Assumption:** We assume that \( p_i(\zeta, l) \) does not have a local minimum, in the sense:

\[
p_i(\zeta, l) > \text{Min}(p_i(\zeta - 1, l), p_i(\zeta + 1, l)) \forall \zeta \in \{1, 2, \ldots, m - 1\}. \tag{8}
\]

The justification for the this assumption is two-fold. First, generally, Zipf’s law kind of behaviour where the number of words plotted against relative frequencies declines as a power function has often been observed. Such a plot is monotonically decreasing and indeed satisfies our assumption. But for Catchwords, we do not expect this behaviour - indeed, we expect the curve to go up initially as the relative frequency increases, then reach a maximum and then decline. This is a unimodal function and also satisfies our assumption. Indeed, we have empirically observed, see **EXPTS**, that these are essentially the only two behaviours.

**Relative sizes of parameters** Before we close the section we discuss the values of the parameters are in order. Here, \( s \) is large. For asymptotic analysis, we can think of it as going to infinity. \( 1/w_0 \) is also large and can be thought of as going to infinity. [In fact, if \( 1/w_0 \in O(1) \), then, intuitively, we see that there is no use of a corpus of more than constant size - since our model has i.i.d. documents, intuitively, the number of samples we need should depend mainly on \( 1/w_0 \).] \( m \) is (much) smaller, but need not be constant.

\( c \) refers to a generic constant independent of \( m, s, 1/w_0, \varepsilon, \delta \); its value may be different in different contexts.

### 3.2 The TSVD Algorithm

Existing SVD based procedures for clustering on raw word-document matrices fail because the spread of frequencies of a word within a topic is often more (at least not significantly less) than the gap between the word’s frequencies in two different topics. Hypothetically the frequency for the word *run*, in the topic *Sports*, may range from say 0.01 on up. But in other topics, it may range from 0 to 0.005 say. The success of the algorithm will lie on correctly identifying the dominant topics such as sports by identifying that the word *run* has occurred with high frequency. In this example, the gap (0.01-0.005) between Sports and other topics is less than the spread within Sports (1.0-0.01), so a 2-clustering approach (based on SVD) will split the topic Sports into two. While this is a toy
Finally, we identify nearly pure documents in \( T \) are now (approximately) identified as the most frequently occurring words in documents in \( l \) columns. Also, \( T \) thresholding on \( M \) catchwords occur the most. Then we get an approximation to Threshold SVD based K-means (TSVD) documents. We now describe the precise algorithm.

A of algorithm proposed by \[12\]. As we will show, the partition produced after clustering, \( k \) cluster centers. The clustering so found is not yet satisfactory. We use the classic Lloyd’s \( k \)-means algorithm proposed by [12]. As we will show, the partition produced after clustering, \( \{ R_1, \ldots, R_k \} \) of \( A^{(2)} \) is close to the partition induced by the Dominant Topics, \( \{ T_1, \ldots, T_k \} \). Catchwords of topic \( l \) are now (approximately) identified as the most frequently occurring words in documents in \( R_l \). Finally, we identify nearly pure documents in \( T_l \) (approximately) as the documents in which the catchwords occur the most. Then we get an approximation to \( M_{-l} \) by averaging these nearly pure documents. We now describe the precise algorithm.

3.3 Topic recovery using Thresholded SVD

Threshold SVD based K-means (TSVD)

\[
\varepsilon = \min \left( \frac{1}{900\alpha^2 \delta^2}, \frac{\varepsilon_0 \sqrt{\alpha \rho_0 \delta}}{640m \sqrt{k}} \right).
\]  

(9)

1. Randomly partition the columns of \( A \) into two matrices \( (1) \) and \( (2) \) of \( s \) columns each.

2. Thresholding

(a) Compute Thresholds on \( (1) \): For each \( i \), let \( \zeta_i \) be the highest value of \( \zeta \in \{0, 1, 2, \ldots, m\} \) such that \(|\{ j : A^{(1)}_{ij} > \zeta_i \}| \geq \frac{m \delta}{2s}; \{ j : A^{(1)}_{ij} = \zeta_i \} \leq 3\varepsilon w_0 s.

(b) Do the thresholding on \( (2) \): \( B_{ij} = \begin{cases} \sqrt{\zeta_i} & \text{if } A_{ij}^{(2)} > \zeta_i/m \text{ and } \zeta_i \geq 8 \ln(20/\varepsilon w_0) \\ 0 & \text{otherwise} \end{cases} \)

3. SVD Find the best rank \( k \) approximation \( B^{(k)} \) to \( B \).

4. Identify Dominant Topics

(a) Project and Cluster Find (approximately) optimal \( k \)– means clustering of the columns of \( B^{(k)} \).

(b) Lloyd’s Algorithm Using the clustering found in Step 4(a) as the starting clustering, apply Lloyd’s algorithm \( k \) means algorithm to the columns of \( B \) (not \( B^{(k)} \)).

(c) Let \( R_1, R_2, \ldots, R_k \) be the \( k \)–partition of \([s]\) corresponding to the clustering after Lloyd’s. //Will prove that \( R_l \approx T_l^{*l/} //

5. Identify Catchwords

(a) For each \( i, l \), compute \( g(i, l) = \left( \varepsilon_0 w_0 s/2 \right) \) th highest element of \( \{ A_{ij}^{(2)} : j \in R_l \} \).

(b) Let \( J_l = \{ i : g(i, l) > \frac{4}{m \delta \gamma} \ln(20/\varepsilon w_0), \max_{i \neq i'} g(i, l') \} \), where, \( \gamma = \frac{1 - 2\delta}{(1 + s) (\delta + \rho)} \).

6. Find Topic Vectors Find the \( \left( \varepsilon_0 w_0 s/2 \right) \) th highest \( \sum_{i \in J_l} A_{ij}^{(2)} \) among all \( j \in [s] \) and return the average of these \( A_{-j} \) as our approximation \( M_{-l} \) to \( M_{-l} \).
Theorem 3.1 Main Theorem Under the Dominant Topic, Catchwords and No-Local-Min assumptions, the algorithm succeeds with high probability in finding an $M$ so that

$$\sum_{i,l} |M_{il} - \hat{M}_{il}| \in O(k \delta), \text{ provided } s \in \Omega^*(\frac{1}{w_0} \left( \frac{k^6 m^2}{\alpha^2 p_0^2} + \frac{m^2 k}{\varepsilon_0^2 \delta^2 \alpha p_0} + \frac{d}{\varepsilon_0 \delta^2} \right)) .$$

A note on the sample complexity is in order. Notably, the dependence of $s$ on $w_0$ is best possible (namely $s \in \Omega^*(1/w_0)$) within logarithmic factors, since, if we had fewer than $1/w_0$ documents, a topic which is dominant with probability only $w_0$ may have none of the documents in the collection. The dependence of $s$ on $d$ needs to be at least $d/\varepsilon_0 w_0 \delta^2$: to see this, note that we only assume that there are $r = O(\varepsilon_0 w_0 s)$ nearly pure documents on each topic. Assuming we can find this set (the algorithm approximately does), their average has standard deviation of about $\sqrt{M_0}/\sqrt{r}$ in coordinate $i$. If topic vector $M_{ij}$ has $O(d)$ entries, each of size $O(1/d)$, to get an approximation of $M_{ij}$ to $l_1$ error $\delta$, we need the per coordinate error $1/\sqrt{d \gamma}$ to be at most $\delta/d$ which implies $s \geq \frac{d}{\varepsilon_0 w_0 \delta^2}$. Note that to get comparable error in [4], we need a quadratic dependence on $d$.

There is a long sequence of Lemmas to prove the theorem. The Lemmas and the proofs are given in Appendix. The essence of the proof lies in proving that the clustering step correctly identifies the partition induced by the dominant topics. For this, we take advantage of a recent development on the $k$-means algorithm from [9] [see also [10]], where, it is shown that under a condition called the Proximity Condition, Lloyd’s $k$ means algorithm starting with the centers provided by the SVD-based algorithm, correctly identifies almost all the documents’ dominant topics. We prove that indeed the Proximity Condition holds. This calls for machinery from Random Matrix theory (in particular bounds on singular values). We prove that the singular values of the thresholded word-document matrix are nicely bounded. Once the dominant topic of each document is identified, we are able to find the Catchwords for each topic. Now, we rely upon part (b.) of the Dominant Topic assumption: that is there is a small fraction of nearly Pure Topic-documents for each topic. The Catchwords help isolate the nearly pure-topic documents and hence find the topic vectors. The proofs are complicated by the fact that each step of the algorithm induces conditioning on the data-for example, after clustering, the document vectors in one cluster are not anymore independent.

4 Experimental Results

We compare the thresholded SVD based k-means (TSVD\textsuperscript{2}) algorithm 3.2 with the algorithms of [5], Recover-KL and Recover-L2, using the code made available by the authors\textsuperscript{3}. We first provide empirical support for the algorithm assumptions in Section 3.1, namely the dominant topic and the catchwords assumption. Then we show on 4 different semi-synthetic data that TSVD provides as good or better recovery of topics than the Recover algorithms. Finally on real-life datasets, we show that the algorithm performs as well as [5] in terms of perplexity and topic coherence.

Implementation Details: TSVD parameters $(w_0, \varepsilon, \varepsilon_0, \gamma)$ are not known in advance for real corpus. We tested empirically for multiple settings and the following values gave the best performance. Thresholding parameters used were: $w_0 = \frac{1}{4}, \varepsilon = \frac{1}{6}$. For finding the catchwords, $\gamma = 1.1, \varepsilon_0 = \frac{1}{3}$ in step 5. For finding the topic vectors (step 6), taking the top 50% $(w_0 \varepsilon_0 = \frac{1}{2})$ gave empirically better results. The same values were used on all the datasets tested. The new algorithm is sensitive to the initialization of the first k-means step in the projected SVD space. To remedy this, we run 10 independent random initializations of the algorithm with K-Means++ [13] and report the best result.

Datasets: We use four real word datasets in the experiments. As pre-processing steps we removed standard stop-words, selected the vocabulary size by term-frequency and removed documents with less than 20 words. Datasets used are: (1) NIPS\textsuperscript{2}: Consists of 1,500 NIPS full papers, vocabulary of 2,000 words and mean document length 1023. (2) NYT\textsuperscript{4}: Consists of a random subset of 30,000

\textsuperscript{1}The superscript $^*$ hides a logarithmic factor in $dsk/\delta_{\text{fail}}$, where, $\delta_{\text{fail}} > 0$ is the desired upper bound on the probability of failure.

\textsuperscript{2}Resources available at: http://mllab.csa.iisc.ernet.in/tsvd

\textsuperscript{3}http://www.cs.nyu.edu/~halpern/files/anchor-word-recovery.zip

\textsuperscript{4}http://archive.ics.uci.edu/ml/datasets/Bag+of+Words
we use step 5 of TSVD (Algorithm 3.2) to find the set of catchwords for each topic-cluster, i.e. all documents which have highest posterior probability for the same topic into one cluster. Then

Catchwords assumption: is also a substantial fraction of documents satisfying almost pure topic assumption.

Dominant topic assumption: Table 1 shows the fraction of the documents in each corpus which satisfy this assumption with (minimum probability of dominant topic) and (maximum probability of non-dominant topics). The fraction of documents which have almost pure topics with highest topic weight at least 0.95 ( ) is also shown. The results indicate that the dominant topic assumption is well justified (on average 64% documents satisfy the assumption) and there is also a substantial fraction of documents satisfying almost pure topic assumption.

Catchwords assumption: We first find a -clustering of the documents by assigning all documents which have highest posterior probability for the same topic into one cluster. Then we use step 5 of TSVD (Algorithm 3.2) to find the set of catchwords for each topic-cluster, i.e. , with the parameters: , , (taking into account constraints in Section 3.1). Table 1 reports the fraction of topics with non-empty set of catchwords and the average per topic frequency of the catchwords. Results indicate that most topics on real data contain catchwords (Table 1, second-last column). Moreover, the average per-topic frequency of the word-frequency for that topic is also quite high (Table 1, last column).

No-Local-Min Assumption: To provide support and intuition for the local-min assumption we consider the quantity \( q_i(\zeta, l) \), in (7). Recall that \( \mathbb{E}[q_i(\zeta, l)] = p_i(\zeta, l) \), we will analyze the behavior of \( q_i(\zeta, l) \) as a function of \( \zeta \) for some topics \( l \) and words \( i \). As defined, we need a fixed \( m \) to check this assumption and so we generate semi-synthetic data with a fixed \( m \) from LDA model trained on the real NYT corpus (as explained in Section 4.2.1). We find catchwords and the sets \( \{ T_i \} \) as in the catchwords assumption above and plot \( q_i(\zeta, l) \) separately for some random catchwords and non-catchwords by fixing some random \( l \in [k] \). Figure 1 shows the plots. As explained in 3.1, the plots are monotonically decreasing for non-catchwords and satisfy the assumption. On the other hand, the plots for catchwords are almost unimodal and also satisfy the assumption.

| Corpus  | \( s \) | \( k \) | \% s with Dominant Topics (\( \alpha = 0.4 \)) | \% s with Pure Topics (\( \delta = 0.05 \)) | \% Topics with CW | CW Mean Frequency |
|---------|------|------|-----------------|-----------------|-----------------|-----------------|
| NYT     | 1500 | 50   | 56.6%           | 2.3%            | 96%             | 0.05            |
| NYT     | 30000| 50   | 63.7%           | 8.5%            | 98%             | 0.07            |
| Pubmed  | 30000| 50   | 62.2%           | 5.1%            | 78%             | 0.05            |
| 20NG    | 13389| 20   | 74.1%           | 39.5%           | 85%             | 0.06            |

Table 1: Algorithm Assumptions. For dominant topic assumption, fraction of documents with satisfy the assumption for (\( \alpha, \beta \)) = (0.4, 0.3) are shown. % documents with almost pure topics (\( \delta = 0.05 \), i.e. 95% pure) are also shown. Last two columns show results for catchwords (CW) assumption.

4.1 Algorithm Assumptions

To check the dominant topic and catchwords assumptions, we first run 1000 iterations of Gibbs sampling on the real corpus and learn the posterior document-topic distribution (\( \{ W_{.,j} \} \)) for each document in the corpus (by averaging over 10 saved-states separated by 50 iterations after the 500 burn-in iterations). We will use this posterior document-topic distribution as the document generating distribution to check the two assumptions.

Dominant topic assumption: Table 1 shows the fraction of the documents in each corpus which satisfy this assumption with (minimum probability of dominant topic) and (maximum probability of non-dominant topics). The fraction of documents which have almost pure topics with highest topic weight at least 0.95 ( ) is also shown. The results indicate that the dominant topic assumption is well justified (on average 64% documents satisfy the assumption) and there is also a substantial fraction of documents satisfying almost pure topic assumption.

Catchwords assumption: We first find a -clustering of the documents \( \{ T_1, \ldots, T_k \} \) by assigning all documents which have highest posterior probability for the same topic into one cluster. Then we use step 5 of TSVD (Algorithm 3.2) to find the set of catchwords for each topic-cluster, i.e. \( \{ S_1, \ldots, S_k \} \), with the parameters: , , , , . Table 1 reports the fraction of topics with non-empty set of catchwords and the average per topic frequency of the catchwords. Results indicate that most topics on real data contain catchwords (Table 1, second-last column). Moreover, the average per-topic frequency of the group of catchwords for that topic is also quite high (Table 1, last column).

No-Local-Min Assumption: To provide support and intuition for the local-min assumption we consider the quantity \( q_1(\zeta, l) \), in (7). Recall that \( \mathbb{E}[q_1(\zeta, l)] = p_1(\zeta, l) \), we will analyze the behavior of \( q_1(\zeta, l) \) as a function of \( \zeta \) for some topics \( l \) and words \( i \). As defined, we need a fixed \( m \) to check this assumption and so we generate semi-synthetic data with a fixed \( m \) from LDA model trained on the real NYT corpus (as explained in Section 4.2.1). We find catchwords and the sets \( \{ T_i \} \) as in the catchwords assumption above and plot \( q_1(\zeta, l) \) separately for some random catchwords and non-catchwords by fixing some random \( l \in [k] \). Figure 1 shows the plots. As explained in 3.1, the plots are monotonically decreasing for non-catchwords and satisfy the assumption. On the other hand, the plots for catchwords are almost unimodal and also satisfy the assumption.

4.2 Empirical Results

4.2.1 Topic Recovery on Semi-Synthetic Data

Semi-synthetic Data: Following [5], we generate semi-synthetic corpora from LDA model trained by MCMC, to ensure that the synthetic corpora retain the characteristics of real data. Gibbs sampling

\[ \frac{1}{k} \sum_{l=1}^{k} \frac{1}{|T_l|} \sum_{i \in S_l} \sum_{j \in T_l} A_{ij} \]
Figure 1: Plot of $q_i(\zeta, l)$ for some random catchwords (left) and non-catchwords (right). Each of three plots for catchword is for one topic ($l$) with two random catchwords ($i$) for each topic and each plot on right is for one non-catchword ($i$) with curves for multiple topics ($l$).

| Corpus | Documents | Recover-L2 | Recover-KL | TSVD | % Improvement |
|--------|-----------|------------|------------|------|---------------|
| NIPS   | 40,000    | 0.342      | 0.308      | 0.115| 62.7%         |
|        | 50,000    | 0.346      | 0.308      | 0.145| 52.9%         |
|        | 60,000    | 0.346      | 0.311      | 0.131| 57.9%         |
| Pubmed | 40,000    | 0.588      | 0.332      | 0.288| 13.3%         |
|        | 50,000    | 0.378      | 0.326      | 0.280| 14.1%         |
|        | 60,000    | 0.372      | 0.328      | 0.284| 13.4%         |
| 20NG   | 40,000    | 0.126      | **0.120**  | 0.124| -3.3%         |
|        | 50,000    | 0.118      | 0.114      | 0.113| 0.9%          |
|        | 60,000    | 0.114      | 0.110      | 0.106| 3.6%          |
| NYT    | 40,000    | 0.214      | 0.208      | 0.195| 6.3%          |
|        | 50,000    | 0.211      | 0.206      | 0.188| 10.2%         |
|        | 60,000    | 0.205      | 0.200      | 0.194| 5.0%          |

Table 2: L1 reconstruction error on various semi-synthetic datasets. Last column is percent improvement over Recover-KL (best performing Recover algorithm).

is run for 1000 iterations on all the four datasets and the final word-topic distribution is used to generate varying number of synthetic documents with document-topic distribution drawn from a symmetric Dirichlet with hyper-parameter 0.01. For NIPS, NYT and Pubmed we use $k = 50$ topics, for 20NewsGroup $k = 20$, and mean document lengths of 1000, 300, 100 and 200 respectively. Note that the synthetic data is not guaranteed to satisfy dominant topic assumption for every document (on average about 80% documents satisfy the assumption for value of $(\alpha, \beta)$ tested in Section 4.1).

**Topic Recovery:** We learn the word-topic distributions ($\hat{M}$) for the semi-synthetic corpora using TSVD and the Recover algorithms of [5]. Given these learned topic distributions and the original data-generating distributions ($M$), we align the topics of $M$ and $\hat{M}$ by bipartite matching and rearrange the columns of $\hat{M}$ in accordance to the matching with $M$. Topic recovery is measured by the average of the $l_1$ error across topics (called reconstruction error [5]), $\Delta(M, \hat{M})$, defined as:

$$\Delta(M, \hat{M}) = \frac{1}{k} \sum_{l=1}^{k} \sum_{d=1}^{d} |M_{dl} - \hat{M}_{dl}|.$$

We report reconstruction error in Table 2 for TSVD and the Recover algorithms. Recover-L2 and Recover-KL. TSVD has smaller error on most datasets than the Recover-KL algorithm. We observed performance of TSVD to be always better than Recover-L2. Best performance is observed on NIPS which has largest mean document length, indicating that larger $m$ leads to better recovery. Results on 20NG are slightly worse than Recover-KL for small sample size (though better than Recover-L2), but the difference is small. While the values in Table 2 are averaged values, Figure 2 shows that TSVD algorithm achieves much better topic recovery (27% improvement in $l_1$ error over Recover-KL) for majority of the topics (>90%) on most datasets.
Figure 2: Histogram of $l_1$ error across topics for 40,000 synthetic documents. TSVD (blue, solid border) gets better recovery on most topics (> 90%) for most datasets (leaving small number of outliers) than Recover-KL (green, dashed border).

| Corpus    | Perplexity | Topic Coherence |
|-----------|------------|-----------------|
|           | R-KL       | R-L2            | TSVD    | R-KL       | R-L2            | TSVD    |
| NIPS      | 754        | 749             | 835     | -86.4 ± 24.5 | -88.6 ± 22.7   | -65.2 ± 29.4 |
| NYT       | 1579       | 1685            | 1555    | -105.2 ± 25.0 | -102.1 ± 28.2  | -107.6 ± 25.7 |
| Pubmed    | 1188       | 1203            | 1307    | -94.0 ± 22.5  | -94.4 ± 22.5   | -84.5 ± 28.7  |
| 20NG      | 2431       | 2565            | 2390    | -93.7 ± 13.6  | -89.4 ± 20.7   | -90.4 ± 27.0  |

Table 3: Perplexity and Topic Coherence. R-KL is Recover-KL, R-L2 is Recover-L2. Standard deviation for topic coherence across topics is also shown.

4.2.2 Topic Recovery on Real Data

Perplexity: A standard quantitative measure used to compare topic models and inference algorithms is perplexity [2]. Perplexity of a set of $D$ test documents, where each document $j$ consists of $m_j$ words, denoted by $w_j$, is defined as: $perp(D_{test}) = exp \left\{ -\frac{\sum_{D_j=1}^{D} \log p(w_j)}{\sum_{D_j=1}^{D_j} m_j} \right\}$. To evaluate perplexity on real data, the held-out sets consist of 350 documents for NIPS, 10000 documents for NYT and Pubmed, and 6780 documents for 20NewsGroup. Table 3 shows the results of perplexity on the 4 datasets. TSVD gives comparable perplexity with Recover-KL, results being slightly better on NYT and 20NewsGroup which are larger datasets with moderately high mean document lengths.

Topic Coherence: [11] proposed Topic Coherence as a measure of semantic quality of the learned topics by approximating user experience of topic quality on top $d_0$ words of a topic. Topic coherence is defined as $TC(d_0) = \frac{1}{d_0} \sum_{i=1}^{d_0} \sum_{j<i} \log \frac{D(w_i, w_j) + e}{D(w_j)}$, where $D(w)$ is the document frequency of a word $w$, $D(w_i, w_j)$ is the document frequency of $w_i$ and $w_j$ together, and $e$ is a small constant. We evaluate TC for the top 5 words of the recovered topic distributions and report the average and standard deviation across topics. TSVD gives comparable results on Topic Coherence (see Table 3).

Topics on Real Data: Table 4 shows the top 5 words of all 50 matched pair of topics on NYT dataset for TSVD, Recover-KL and Gibbs sampling. Most of the topics recovered by TSVD are more closer to Gibbs sampling topics. Indeed, the total average $l_1$ error with topics from Gibbs sampling for topics from TSVD is 0.034, whereas for Recover-KL it is 0.047 (on the NYT dataset).

Summary: We evaluated the proposed algorithm, TSVD, rigorously on multiple datasets with respect to the state of the art (Recover), following the evaluation methodology of [5]. In Table 2 we
show that the L1 reconstruction error for the new algorithm is small and on average 19.6% better than the best results of the Recover algorithms [5]. We also demonstrate that on real datasets the algorithm achieves comparable perplexity and topic coherence to Recover (Table 3). Moreover, we show on multiple real datasets that the algorithm assumptions are well justified in practice.

**Conclusion**

Real world corpora often exhibits the property that in every document there is one topic dominantly present. A standard SVD based procedure will not be able to detect these topics, however TSVD, a thresholded SVD based procedure, as suggested in this paper, discovers these topics. While SVD is time-consuming, there have been a host of recent sampling-based approaches which make SVD easier to apply to massive corpora which may be distributed among many servers. We believe that apart from topic recovery, thresholded SVD can be applied even more broadly to similar problems, such as matrix factorization, and will be the basis for future research.

Table 4: Top 5 words of matched topic pairs for TSVD, Recover-KL and Gibbs sampling. Catchwords and anchor words in top 5 words are highlighted for TSVD and Recover-KL.
Table 4: Top 5 words of matched topic pairs for TSVD, Recover-KL and Gibbs sampling. Catchwords and anchor words in top 5 words are highlighted for TSVD and Recover-KL.

| TSVD                  | Recover-KL                                                                 | Gibbs                                                                                      |
|-----------------------|-----------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| actor film play movie | goal play team season game                                                  | film movie award actor *zzz_oscar*                                                      |
| character             |                                                                             |                                                                             |
| school student teacher| school student program million children                                     | school student teacher program children                                                  |
| district program      |                                                                             |                                                                             |
| tax taxes cut billion | *zzz_governor_bush* tax campaign taxes plan                                 | tax plan billion million cut                                                            |
| investor              |                                                                             |                                                                             |
| team player season coach |                                                           |                                                                             |
| *zzz_nfl*             |                                                                             |                                                                             |
| family home friend room|                                                                             |                                                                             |
| school                |                                                                             |                                                                             |
| primary *zzz_mccain* voter *zzz_john_mccain* |                                                                             |                                                                             |
| *zzz_microsoft* court* company case law |                                                                             |                                                                             |
| company million percent |                                                                             |                                                                             |
| percent shares billion |                                                                             |                                                                             |
| site web sites com www |                                                                             |                                                                             |
| scientist human cell study researcher |                                                                             |                                                                             |
| baby mom percent home family |                                                                             |                                                                             |
| point game half shot team |                                                                             |                                                                             |
| *zzz_russia* *zzz_vladimir_putin* *zzz_russian* *zzz_boris_yeltsin* *zzz_moscow* |                                                                             |                                                                             |
| com *zzz_canada* www fax information |                                                                             |                                                                             |
| room restaurant building fish painting |                                                                             |                                                                             |
| loved family show friend play |                                                                             |                                                                             |
| prices percent worker oil price |                                                                             |                                                                             |
| *zzz_clinton* flag official federal *zzz_white_house* |                                                                             |                                                                             |
| million test shares air president |                                                                             |                                                                             |
| *zzz_brady* *zzz_al_gore* campaign *zzz_gore* *zzz_clinton* |                                                                             |                                                                             |
| files article computer art ball |                                                                             |                                                                             |
| con percent *zzz_mexico* federal official |                                                                             |                                                                             |
| involving book film case right |                                                                             |                                                                             |
| test women study student found |                                                                             |                                                                             |
| *zzz_internet* companies company business customer |                                                                             |                                                                             |
| company companies deal *zzz_internet* *zzz_time_warner* |                                                                             |                                                                             |

Continued on next page.
Table 4: Top 5 words of matched topic pairs for TSVD, Recover-KL and Gibbs sampling. Catchwords and anchor words in top 5 words are highlighted for TSVD and Recover-KL.

| TSVD                  | Recover-KL                           | Gibbs                               |
|-----------------------|---------------------------------------|-------------------------------------|
| zzz_internet companies| newspaper zzz_chronicle               | million money worker                |
| company business      | zzz_examiner zzz_hearst million        | company pay                         |
| customer              |                                       |                                     |
| goal play games king  | zzz_tiger_wood shot tournament tour    | zzz_tiger_wood tour                 |
| game                  | player                                | shot player                         |
| zzz_american          |                                       |                                     |
| zzz_united_states     | zzz_israel zzz_lebanon peace          | zzz_israel peace palestinian        |
| zzz_nato              | zzz_syria israeli                     | talk israel                         |
| camp war              |                                       |                                     |
| team season game      | team game point season player         | race won win fight team             |
| play                  |                                       |                                     |
| reporter zzz_earl_caldwell | corp group list oil meeting      | black white zzz_black hispanic reporter |
| zzz_black black look  |                                       |                                     |
| campaign zzz_republican | zzz_bush zzz_mccain campaign republican voter | gun bill law zzz_congress legislation |
| republican primary    |                                       |                                     |
| zzz_bush zzz_mccain   | flag black zzz_confederate right group|                                     |
| campaign primary      |                                       |                                     |
| republican            |                                       |                                     |
| zzz_bush zzz_mccain   | campaign zzz_mccain                   | flag zzz_confederate                |
| campaign primary      | campaign republican voter             | zzz_congress legislation            |
| republican            |                                       |                                     |
| zzz_bush zzz_mccain   |                                       |                                     |
| zzz_bush zzz_mccain   |                                       |                                     |
| campaign               |                                       |                                     |
| zzz_mccain            |                                       |                                     |
| campaign               |                                       |                                     |
| republican             |                                       |                                     |
| zzz_mccain            |                                       |                                     |
| campaign               |                                       |                                     |
| zzz_bush               |                                       |                                     |
| zzz_bush               |                                       |                                     |
| campaign               |                                       |                                     |
| zzz_bush               |                                       |                                     |
| zzz_government case officer security | court law case lawyer right |                                     |

References

[1] Blei, D. Introduction to probabilistic topic models. Communications of the ACM, pp. 77–84, 2012.
[2] Blei, D., Ng, A., and Jordan, M. Latent Dirichlet allocation. Journal of Machine Learning Research, pp. 3:993–1022, 2003. Preliminary version in Neural Information Processing Systems 2001.
[3] Griffiths, T. L. and Steyvers, M. Finding scientific topics. Proceedings of the National Academy of Sciences, 101:5228–5235, 2004.
[4] Arora, S., Ge, R., and Moitra, A. Learning topic models – going beyond SVD. In Foundations of Computer Science, 2012b.
[5] Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu M. A practical algorithm for topic modeling with provable guarantees. In International Conference on Machine Learning, 2013.
[6] Anandkumar, A., Foster, D., Hsu, D., Kakade, S., and Liu, Y. A Spectral Algorithm for Latent Dirichlet Allocation In Neural Information Processing Systems, 2012.
[7] Papadimitriou, C., Raghavan, P., Tamaki H., and Vempala S. Latent semantic indexing: a probabilistic analysis. Journal of Computer and System Sciences, pp. 217–235, 2000. Preliminary version in PODS 1998.
[8] Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and Harshman, R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, pp. 391–407, 1990.
[9] Kumar, A., and Kannan, R. Clustering with spectral norm and the k-means algorithm. In Foundations of Computer Science, 2010.
[10] Awashiti, P., and Sheffet, O. Improved spectral-norm bounds for clustering. In Proceedings of Approx/Random, 2012.
[11] Mimno, D., Wallach, H., Talley, E., Leenders, M. and McCallum, A. Optimizing semantic coherence in topic models. In Empirical Methods in Natural Language Processing, pp. 262–272, 2011.
A Line of Proof

We describe the Lemmas we prove to establish the result. The detailed proofs are in the Section B.

A.1 General Facts

We start with a consequence of the no-local-minimum assumption. We use that assumption solely through this Lemma.

**Lemma A.1** Let $p_i(\zeta, l)$ be as in (6). If for some $\zeta_0 \in \{0, 1, \ldots, m\}$ and $\nu \geq 0$, $\sum_{\zeta \geq \zeta_0} p_i(\zeta, l) \geq \nu$ and also $\sum_{\zeta \leq \zeta_0} p_i(\zeta, l) \geq \nu$ then, $p_i(\zeta_0, l) \geq \frac{\nu}{m}$.

Next, we state a technical Lemma which is used repeatedly. It states that for every $i, \zeta, l$, the empirical probability that $A_{ij} = \zeta/m$ is close to the true probability. Unsurprisingly, we prove it using H-C. But we will state a consequence in the form we need in the sequel.

**Lemma A.2** Let $p_i(\zeta, l)$ and $q_i(\zeta, l)$ be as in (6) and (7). We have

$$\forall i, l, \zeta : \Pr \left( \left| p_i(\zeta, l) - q_i(\zeta, l) \right| \geq \frac{\varepsilon}{2} \sqrt{\frac{w_0}{m}} \sqrt{p_i(\zeta, l)} + \frac{\varepsilon^2 w_0}{2} \right) \leq 2 \exp(-\varepsilon^2 s w_0/8).$$

From this it follows that with probability at least $1 - 2 \exp(-\varepsilon^2 s w_0/8)$,

$$\frac{1}{2} q_i(\zeta, l) - \varepsilon^2 w_0 \leq p_i(\zeta, l) \leq 2q_i(\zeta, l) + 2\varepsilon^2 w_0.$$

A.1.1 Properties of Thresholding

Say that a threshold $\zeta_i$ “splits” $T_i^{(2)}$ if $T_i^{(2)}$ has a significant number of $j$ with $A_{ij} > \zeta_i/m$ and also a significant number of $j$ with $A_{ij} \leq \zeta_i/m$. Intuitively, it would be desirable if no threshold splits any $T_i$, so that, in B, for each $i, l$, either most $j \in T_i^{(2)}$ have $B_{ij} = 0$ or most $j \in T_i^{(2)}$ have $B_{ij} = \sqrt{\zeta_i}$. We now prove that this is indeed the case with proper bounds. We henceforth refer to the conclusion of the Lemma below by the mnemonic “no threshold splits any $T_i$”.

**Lemma A.3** (No Threshold Splits any $T_i$) For a fixed $i, l$, with probability at least $1 - 2 \exp(-\varepsilon^2 w_0 s/8)$, the following holds:

$$\min \left( \Pr(\bar{A}_{ij}^{(2)} \leq \frac{\zeta_i}{m} : j \in T_i^{(2)}), \quad \Pr(A_{ij}^{(2)} > \frac{\zeta_i}{m} : j \in T_i^{(2)}) \right) \leq 4m \varepsilon w_0.$$

Let $\mu$ be a $d \times s$ matrix whose columns are given by

$$\forall j \in T_i^{(2)}, \quad \mu_{., j} = E(B_{., j} | j \in T_i).$$

$\mu$’s columns corresponding to all $j \in T_i$ are the same. The entries of the matrix $\mu$ are fixed (real numbers) once we have $A^{(1)}$ (and the thresholds $\zeta_i$ are determined). Note: We have “integrated out $W$”, i.e.,

$$\mu_{ij} = \int_{W_{., j}} \Pr(W_{., j} | j \in T_i) E(B_{ij} | W_{., j}).$$

(So, think of $W_{., j}$ for $A^{(1)}$’s columns being picked first from which $\zeta_i$ is calculated. $W_{., j}$ for columns of $A^{(2)}$ are not yet picked until the $\zeta_i$ are determined.) But $\mu_{ij}$ are random variables before we fix $A^{(1)}$. The following Lemma is a direct consequence of “no threshold splits any $T_i$”.

---

[12] Lloyd, Stuart P. Least squares quantization in PCM, *IEEE Transactions on Information Theory* 28 (2): 1291–1237, 1982.

[13] Arthur, D., and Vassilvitskii, S. K-means++: The advantages of careful seeding. In *Proceedings of ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, 2007.

[14] McDiarmid, C. On the method of Bounded Differences. *Surveys in Combinatorics: London Math. Soc. Lecture Note Series* 141. Cambridge University Press., 1989.

[15] Vershynin R. Introduction to non-asymptotic analysis of random matrices. In *ArXiv:1011.3027v6 [math.PR] 4 Oct 2011*.
Lemma A.4 Let $\zeta'_l = \text{Max}(\zeta_l, 8 \ln(20/\epsilon w_0))$. With probability at least $1 - 4kd \exp(-\epsilon^2 sw_0/8)$ (over the choice of $A^{(1)}$):

$$\forall l, \forall j \in T_l, \forall i: \mu_{ij} \leq \epsilon_l \sqrt{\zeta'_l} \text{ OR } \mu_{ij} \geq \sqrt{\zeta'_l}(1 - \epsilon_l)$$

$$\forall l, \forall i, \text{Var}(B_{ij}) \leq 2\epsilon_l \zeta'_l,$$  \hspace{0.7cm} \hspace{0.7cm} (10)

where, $\epsilon_l = 4m\epsilon w_0/w_l$.

So far, we have proved that for every $i$, the threshold does not split any $T_l$. But this is not sufficient in itself to be able to cluster (and hence identify the $T_l$), since, for example, this alone does not rule out the extreme cases that for most $j$ in every $T_l$, $A_{ij}^{(2)}$ is above the threshold (whence $\mu_{ij} \geq (1 - \epsilon_l)\sqrt{\zeta'_l}$ for almost all $j$) or for most $j$ in no $T_l$ is $A_{ij}^{(2)}$ above the threshold, whence, $\mu_{ij} \leq \epsilon_l \sqrt{\zeta'_l}$ for almost all $j$. Both these extreme cases would make us lose all the information about $T_l$ due to thresholding; this scenario and milder versions of it have to be proven not to occur. We do this by considering how thresholds handle catchwords. Indeed we will show that for a catchword $i \in S_l$, a $j \in T_l$ has $A_{ij}^{(2)}$ above the threshold and a $j \not\in T_l$ has $A_{ij}^{(2)}$ below the threshold. Both statements will only hold with high probability, of course and using this, we prove that $\mu_{.,j}$ and $\mu_{.,j'}$ are not too close for $j, j'$ in different $T_l$'s. For this, we need the following Lemmas.

Lemma A.5 For $i \in S_l$, and $l' \neq l$, we have with $\eta_i = \lceil M_i(\alpha + \beta + \rho)m/2 \rceil$,

$$\text{Prob}(A_{ij} \leq \eta_i/m \mid j \in T_l) \leq \epsilon w_0/20, \quad \text{Prob}(A_{ij} \geq \eta_i/m \mid j \in T_{l'}) \leq \epsilon w_0/20.$$ 

Lemma A.6 With probability at least $1 - 8mdk \exp(-\epsilon^2 sw_0/8)$, we have

$$\text{for } j \in T_l, j' \not\in T_l, \mid \mu_{.,j} - \mu_{.,j'} \mid \geq \frac{2}{9} \alpha p_0 m.$$ 

A.1.2 Proximity

Next, we wish to show that clustering as in TSVD identifies the dominant topics correctly for most documents, i.e., that $R_l \approx T_l$ for all $l$. For this, we will use a theorem from [9] [see also [10]] which in this context says:

Theorem A.7 If all but a $f$ fraction of the the $B_{.,j}$ satisfy the “proximity condition”, then TSVD identifies the dominant topic in all but $c_1 f$ fraction of the documents correctly after polynomial number of iterations.

To describe the proximity condition, first let $\sigma$ be the maximum over all directions $v$ of the square root of the mean-squared distance of $B_{.,j}$ to $\mu_{.,j}$, i.e.,

$$\sigma^2 = \text{Max}_{\|v\|=1} \frac{1}{8}v^T(B - \mu)^2 = \frac{1}{8}\|B - \mu\|^2.$$ 

The parameter $\sigma$ should remind the reader of standard deviation, which is indeed what this is, since $E(B_{.,j}|T_1, T_2, \ldots, T_l) = \mu$. Our random variables $B_{.,j}$ being $d-$ dimensional vectors, we take the maximum standard deviation in any direction.

Definition: $B$ is said to satisfy the proximity condition with respect to $\mu$, if for each $l$ and each $j \in T_l$ and and each $l' \neq l$ and $j' \in T_{l'}$, the projection of $B_{.,j}$ onto the line joining $\mu_{.,j}$ and $\mu_{.,j'}$ is closer to $\mu_{.,j}$ by at least

$$\Delta = \frac{c_0 k}{\sqrt{w_0}} \sigma,$$

than it is to $\mu_{.,j'}$. [Here, $c_0$ is a constant.]

To prove proximity, we need to bound $\sigma$. This will be the task of the subsection B.1 which relies heavily on Random Matrix Theory.
Proof: (of Lemma A.1) Abbreviate \( p_i(\cdot, l) \) by \( f(\cdot) \). We claim that either (i) \( f(\zeta) \geq f(\zeta - 1)\forall 1 \leq \zeta \leq \zeta_0 \) or (ii) \( f(\zeta + 1) \leq f(\zeta)\forall m - 1 \geq \zeta \geq \zeta_0 \). To see this, note that if both (i) and (ii) fail, we have \( \zeta_1 \leq \zeta_0 \) and \( \zeta_2 \geq \zeta_0 \) with \( f(\zeta_1) - f(\zeta_1 - 1) < 0 < f(\zeta_2 + 1) - f(\zeta_2) \). But then there has to be a local minimum of \( f \) between \( \zeta_1 \) and \( \zeta_2 \). If (i) holds, clearly, \( f(\zeta_0) \geq f(\zeta)\forall \zeta < \zeta_0 \) and so the lemma follows. So, also if (ii) holds.

Proof: (of Lemma A.2) Note that \( q_i(\zeta, l) = \frac{1}{2} \left\{ |\{j \in T_i : A_{ij} = \zeta/m\}| - \frac{1}{s} \sum_{j=1}^{s} X_j \right\} \), where, \( X_j \) is the indicator variable of \( A_{ij} = \zeta/m \) \( \land j \in T_i \). \( \frac{1}{2} \sum_{j} E(X_j) = p_i(\zeta, l) \) and we apply H-C with \( t = \frac{1}{2} \varepsilon \sqrt{w_0} \sqrt{p_i(\zeta, l)} + \frac{1}{2} \varepsilon^2 w_0 \) and \( \mu = p_i(\zeta, l) \). We have \( \frac{t^2}{\mu + t} \geq \varepsilon^2 w_0/4 \), as is easily seen by calculating the roots of the quadratic \( t^2 - \frac{1}{4} t \varepsilon^2 w_0 - \frac{1}{4} \varepsilon^2 w_0 \mu = 0 \). Thus we get the claimed for \( T_i \). Note that the same proof applies for \( T_i^{(1)} \) as well as \( T_i^{(2)} \). To prove the second assertion, let \( a = q_i(\zeta, l) \) and \( b = \sqrt{p_i(\zeta, l)} \), then, \( b \) satisfies the quadratic inequalities:

\[
b^2 - \frac{1}{2} \varepsilon \sqrt{w_0} b - (a + \frac{1}{2} \varepsilon^2 w_0) \leq 0 ; \quad b^2 + \frac{1}{2} \varepsilon \sqrt{w_0} b - (a - \frac{1}{2} \varepsilon^2 w_0) \geq 0.
\]

By bounding the roots of these quadratics, it is easy to see the second assertion after some calculation.

Proof: (of Lemma A.3) Note that \( \zeta_i \) is a random variable which depends only on \( A^{(1)} \). So, for \( j \in T_i^{(2)} \), \( A_{ij} \) are independent of \( \zeta_i \). Now, if

\[
\operatorname{Prob}(A_{ij} \leq \frac{\zeta_i}{m} ; j \in T_i^{(2)}) > 4m \varepsilon w_0 \quad \text{and} \quad \operatorname{Prob}(A_{ij} > \frac{\zeta_i}{m} ; j \in T_i^{(2)}) > 4m \varepsilon w_0 ,
\]

by Lemma (A.1), we have

\[
\operatorname{Prob}(A_{ij} = \frac{\zeta_i}{m} ; j \in T_i^{(2)}) > 4 \varepsilon w_0 .
\]

Since \( \operatorname{Prob}(A_{ij} = \zeta/m ; j \in T_i^{(1)}) = \operatorname{Prob}(A_{ij} = \zeta/m ; j \in T_i^{(2)}) \) for all \( \zeta \), we also have

\[
\operatorname{Prob}(A_{ij} = \frac{\zeta_i}{m} ; j \in T_i^{(1)}) = p_i(\zeta_i, l) > 4 \varepsilon w_0 . \quad (11)
\]

Pay a failure probability of \( 2 \exp(-\varepsilon^2 w_0/8) \) and assume the conclusion of Lemma (A.2) and we have:

\[
\frac{1}{s} \left\{ j \in T_i^{(1)} : A_{ij} = \frac{\zeta_i}{m} \right\} = q_i(\zeta_i, l) \geq p_i(\zeta_i, l) - \frac{\varepsilon}{2} \sqrt{w_0} p_i(\zeta_i, l) - \frac{\varepsilon^2}{2} w_0.
\]

Now, it is easy to see that \( p_i(\zeta_i, l) - \frac{\varepsilon}{2} \sqrt{w_0} p_i(\zeta_i, l) \) increases as \( p_i(\zeta_i, l) \) increases subject to (11). So,

\[
p_i(\zeta_i, l) - \frac{\varepsilon}{2} \sqrt{w_0} p_i(\zeta_i, l) - \frac{\varepsilon^2}{2} w_0 > (4 \varepsilon - \varepsilon^{3/2} - \frac{1}{2} \varepsilon^2) w_0 \geq 3 \varepsilon w_0,
\]

contradicting the definition of \( \zeta_i \) in the algorithm. This completes the proof of the Lemma.

Proof: (of Lemma A.4): After paying a failure probability of \( 4kd \exp(-\varepsilon^2 w_0/8) \), assume no threshold splits any \( T_i \). [The factors of \( k \) and \( d \) come in because we are taking the union bound over
all words and all topics.] Then,

\[
\text{Prob}(A_{ij}^{(2)} \leq \frac{\zeta_i}{m} \mid j \in T_i^{(2)}) = \sum_{\zeta = 0}^{\zeta_i} p_i(\zeta, l)/\text{Prob}(j \in T_l) \leq 4m\varepsilon \frac{w_0}{w_l}
\]

or \(\text{Prob}(A_{ij}^{(2)} > \frac{\zeta_i}{m} \mid j \in T_i^{(2)}) = \sum_{\zeta = \zeta_i+1}^{m} p_i(\zeta, l)/\text{Prob}(j \in T_l) \leq 4m\varepsilon \frac{w_0}{w_l}\).

Wlg, assume that \(\text{Prob}(A_{ij} \leq \zeta_i/m \mid j \in T_i) \leq \varepsilon_1\). Then, with probability, at least \(1 - \varepsilon_1\), \(A_{ij}^{(2)} > \zeta_i/m\). Now, either \(\zeta_i < 8\ln(20/\varepsilon w_0)\) and all \(B_{ij}, j \in T_i\) are zero and then \(\mu_{ij} = 0\), or \(\zeta_i \geq 8\ln(20/\varepsilon w_0)\), whence, \(E(B_{ij} | j \in T_i) = [(1 - \varepsilon_1)\sqrt{\zeta_i'} \sqrt{\zeta_i}]\). So, \(\mu_{ij} \geq (1 - \varepsilon_1)\sqrt{\zeta_i'}\) and \(\text{Prob}(B_{ij} = 0) \leq \varepsilon_1\). So,

\[
\text{Var}(B_{ij} | j \in T_i) \leq (\sqrt{\zeta_i'} - (1 - \varepsilon_1)\sqrt{\zeta_i'})^2 \text{Prob}(B_{ij} = \sqrt{\zeta_i'} | j \in T_i) + (\sqrt{\zeta_i'} - 0)^2 \text{Prob}(B_{ij} = 0 | j \in T_i) \leq 2\varepsilon_1\zeta_i'.
\]

This proves the lemma in this case. The other case is symmetric.

**Proof:** (of Lemma A.5) Recall that \(P_{ij} = \sum_i M_{il}W_{ij}\) is the probability of word \(i\) in document \(j\) conditioned on \(W\). Fix an \(i \in S_i\). From the dominant topic assumption,

\[
\forall j \in T_l, P_{ij} = \sum_{l_1} M_{il_1}W_{i_1,j} \geq M_{il}W_{ij} \geq M_{il}\alpha. \tag{12}
\]

The \(P_{ij}\) are themselves random variables. Note that (12) holds with probability 1. From Catchword assumption and (1), we get that

\[
M_{il}\alpha - (\eta_i/m) \geq M_{il}\alpha - M_{il}((\alpha + \beta + \rho)/2) \geq M_{il}\alpha\delta/2.
\]

Now, we will apply H-C with \(\mu - t = \eta_i/m\) and \(\mu \geq M_{il}\alpha\) for the \(m\) independent words in a document. By Calculus, the probability bound from H-C of \(\exp(-t^2w_1s/2\mu) = \exp(-\mu - (\eta_i/m)/2\mu)\) is highest subject to the constraints \(\mu \geq M_{il}\alpha; \eta_i \leq mM_{il}(\alpha + \beta + \rho)/2\), when \(\mu = M_{il}\alpha\) and \(\eta_i = mM_{il}(\alpha + \beta + \rho)/2\), whence, we get

\[
\text{Prob}(A_{ij} \leq \eta_i/m \mid j \in T_l) \leq \exp(-M_{il}\alpha\delta^2m/8) \leq \varepsilon w_0/20,
\]

using (5). Now, we prove the second assertion of the Lemma.

\[
\forall j \in T_{l'}, l' \neq l, \sum_{l_1} M_{il_1}W_{i_1,j} = M_{il}W_{ij} + \sum_{l_1 \neq l} M_{il_1}W_{i_1,j} \\
\leq M_{il}W_{ij} + (\max_{l_1 \neq l} M_{il_1})(1 - W_{ij}) \\
\leq M_{il}(\beta + \rho). \tag{13}
\]

\[
\eta_i - M_{il}(\beta + \rho) \geq \frac{M_{il}(\alpha + \beta + \rho)}{2} - M_{il}(\beta + \rho) - \frac{1}{m} \geq \frac{3M_{il}\alpha\delta}{8},
\]

using (5) and (1). Applying the first inequality of Lemma (B.1) with \(\mu + t = \eta_i/m\) and \(\mu \leq M_{il}(\beta + \rho)\) and again using (5),

\[
\text{Prob}(A_{ij} \geq \eta_i/m \mid j \in T_{l'} \leq \exp \left(-9M_{il}\alpha\delta^2m/64\right) \leq \varepsilon w_0/20.
\]

**Lemma B.2** For \(i \in S_i\), \(\text{Prob}(\zeta_i < \eta_i) \leq 3mke^{-e^2sw_0/8},\) with \(\eta_i\) as defined in Lemma A.5.

**Proof:** Fix attention on \(i \in S_i\). After paying the failure probability of \(3mke^{-e^2sw_0/8}\), assume the conclusions of Lemma (A.2) hold for all \(\zeta, l\). It suffices to show that

\[
|\{j : A_{ij}^{(1)} > \eta_i/m\}| \geq \frac{w_0^2}{2}, \quad |\{j : A_{ij}^{(1)} = \frac{\eta_i}{m}\}| < 3w_0\varepsilon s,
\]

since, \(\eta_i\) is an integer and \(\zeta_i\) is the largest integer satisfying the inequalities. The first statement follows from first assertion of Lemma A.5. The second statement is slightly more complicated. Using both the first and second assertions of Lemma A.5, we get that for all \(l'\) (including \(l' = l\)), we have

\[
\text{Prob}(A_{ij} = \eta_i/m | j \in T_{l'}^{(1)}) \leq \varepsilon w_0/20.
\]
\[ \left| \{ j \in T_{l'}^{(1)} : A_{ij} = \eta_i/m \} \right| \leq \varepsilon w_0 w_0 s/20 + \frac{\varepsilon}{2} \sqrt{w_0 w_0} \sqrt{\varepsilon w_0} / 20 w_0 s + \frac{\varepsilon^2 w_0 w_0}{2} \]
\[ \leq \frac{\varepsilon w_0 s}{8} (w_0 + \sqrt{\varepsilon w_0}) + \frac{\varepsilon^2 w_0 s}{2}. \]

Now, adding over all \( l' \) and using \( \sum_{l'} \sqrt{w_0} \leq \sqrt{k} \sum_{l'} w_0 = \sqrt{k} \), we get
\[ \left| \{ j : A_{ij}^{(1)} = \eta_i/m \} \right| \leq \varepsilon w_o s, \]

since \( \varepsilon \leq 1/k. \)

**Lemma B.3** Define \( I_i = \{ i \in S_i : \zeta_i \geq \eta_i \} \). With probability at least \( 1 - 8mk \exp(-\varepsilon^2 w_0 s/8) \), we have for all \( l, \)
\[ \sum_{i \in I_i} \zeta_i' \geq m\alpha p_0/2. \]

**Proof:** After paying the failure probability, we assume the conclusion of Lemma A.2 holds for all \( i, \zeta, l. \) Now, by Lemma B.2, we have (with \( 1 \) denoting the indicator function)
\[ E \left( \sum_{i \in S_i} M_{d1}(\zeta_i < \eta_i) \right) \leq 3mk \exp(-\varepsilon^2 w_0 / 8) \sum_{i \in S_i} M_d, \]
which using Markov inequality implies that with probability at least \( 1 - 6mk \exp(-\varepsilon^2 w_0 / 8), \)
\[ \sum_{i \in I_i} M_{d1} \geq \frac{1}{2} \sum_{i \in S_i} M_d \geq p_0/2, \]  
(14)
using (4). Note that by (5), no catchword has \( \zeta_i' \) set to zero. So,
\[ \sum_{i \in I_i} \zeta_i' = \sum_{i \in I_i} \zeta_i \geq \sum_{i \in I_i} \eta_i \geq \sum_{I_i} M_{d1} \alpha / 2 \geq \alpha p_0 m / 2. \]

**Proof:** (of Lemma A.6) For this proof, \( i \) will denote an element of \( I_i \). By Lemma A.5,
\[ \forall i \in I_i, i' \neq i, \text{Prob}(A_{ij} > \frac{\zeta_i}{m}, j \in T_{l'}^{(1)}) \leq \text{Prob}(A_{ij} > \eta_i/m, j \in T_{l'}^{(1)}) \leq \frac{\varepsilon w_0}{20}. \]  
(15)
This implies by Lemma A.2, for \( i' \neq i, \)
\[ \left| \{ j \in T_{l'}^{(1)} : A_{ij} > \frac{\zeta_i}{m} \} \right| \leq w_0 s \left( \frac{\varepsilon w_0}{20} + \varepsilon \sqrt{w_0 \sqrt{w_0} \sqrt{\varepsilon w_0} / 4} \right) + w_0 \varepsilon^2 s / 2. \]  
(16)
Summing over all \( i' \neq i \), we get (using \( \sum_{l'} \sqrt{w_0} \leq \sqrt{\sum_{l'} w_0} = 1 / \sqrt{\varepsilon} \) by (9))
\[ \sum_{i' \neq i} \left| \{ j \in T_{l'}^{(1)} : A_{ij} > \frac{\zeta_i}{m} \} \right| \leq \varepsilon w_0 s. \]

Now the definition of \( \zeta_i \) in the algorithm implies that:
\[ \sum_{\zeta > \zeta_i} q_i(\zeta, l) = \left| \{ j \in T_{l'}^{(1)} : A_{ij} > \frac{\zeta_i}{m} \} \right| \geq \left( \frac{w_0}{2} - \varepsilon w_0 \right) s \geq w_0 s / 4. \]

So, by Lemma A.2,
\[ \text{Prob}(j \in T_i; A_{ij} > \zeta_i/m) = \sum_{\zeta > \zeta_i} p_i(\zeta, l) \geq \frac{1}{2} \sum_{\zeta > \zeta_i} q_i(\zeta, l) - \varepsilon^2 w_0 m \]
\[ \geq \frac{w_0}{8} - \varepsilon^2 w_0 m \geq w_0 / 9, \]
using (9). Next let \( \tilde{\rho} = \text{Prob}(A_{ij} = \zeta_i/m; j \in T_i). \) Since \( \left| \{ j \in T_{l'}^{(1)} : A_{ij} = \zeta_i/m \} \right| \leq 3 \varepsilon w_0 s, \) by the definition of \( \zeta_i \) in the algorithm, we get by a similar argument
\[ \tilde{\rho} \leq 2 q_i(\zeta_i, l) + 2 \varepsilon^2 w_0 \leq 7 \varepsilon w_0. \]  
(17)
Now, by Lemma A.1, we have
\[ \hat{p} \geq \min \left( \frac{w_0}{9m}, \frac{1}{m} \Pr(A_{ij} \leq \zeta_i/m; j \in T_i^{(2)}) \right). \]

By (9), \(7\varepsilon w_0 < w_0/9m\) and so we get:
\[ \Pr(A_{ij} \leq \zeta_i/m; j \in T_i^{(2)}) \leq 7\varepsilon mw_0. \]

Noting that by (5), no catchword has \(\zeta_i\) set to zero, \(\Pr(B_{ij} = 0; j \in T_i^{(2)}) \leq 7\varepsilon mw_0/w_l \leq 1/6\), by (9). This implies
\[ \mu_{ij} \geq \frac{5}{6} \sqrt{\zeta_i}. \]

Now, by (15), we have for \(j' \notin T_i\),
\[ \mu_{ij'} \leq \sqrt{\zeta_i}/6. \]

So, we have
\[ |\mu_{-j} - \mu_{-j'}|^2 \geq \sum_{i \in I_l}(\mu_{ij} - \mu_{ij'})^2 \geq (4/9)\sum_{i \in I_l} \zeta_i. \]

Now Lemma (B.3) implies the current Lemma.

B.1 Bounding the Spectral norm

**Theorem B.4** Fix an \(l\). For \(j \in T_l\), let \(R_{-j} = B_{-j} - \mu_{-j}\). [The \(R_{-j}, j \in T_l\) are vector-valued random variables which are independent, even conditioned on the partition \(T_1, T_2, \ldots, T_k\).] With probability at least \(1 - 10mdk \exp(-\varepsilon^2 w_0 s/8)\), we have \(\|R\|^2 \leq c k w_0 \varepsilon sm^2\). Thus,
\[ \|B - \mu\|^2 \leq c \varepsilon w_0 s m^2 k^2. \]

We will apply Random Matrix Theory, in particular the following theorem, to prove Theorem B.4.

**Theorem B.5** [15, Theorem 5.44] Suppose \(R\) is a \(d \times r\) matrix with columns \(R_{-j}\) which are independent identical vector-valued random variables. Let \(U = E(R_{-j} R_{-j}^T)\) be the inertial matrix of \(R_{-j}\). Suppose \(|R_{-j}| \leq \nu\) always. Then, for any \(t > 0\), with probability at least \(1 - \exp(-ct^2)\), we have\(^4\)
\[ ||R|| \leq ||U||^{1/2} \sqrt{t} + tv. \]

We need the following Lemma first.

**Lemma B.6** With probability at least \(1 - \exp(-s\varepsilon w_0/3)\), we have
\[ \zeta_0 \leq 4m \lambda; \sum_i \zeta_i' \leq 4km \tag{18} \]

**Proof:** The probability of word \(i\) in document \(j\), is given by:
\[ P_{ij} = \sum_i M_{ij} W_{ij} \leq \lambda_i \text{ (where, } \lambda_i = \max_i M_{ij}). \]
If \(\lambda_i < \frac{1}{m} \ln(20/\varepsilon w_0)\), then, \(\Pr(A_{ij} > (8/m) \ln(20/\varepsilon w_0)) \leq \varepsilon w_0\) by H-C (since \(A_{ij}\) is the average of \(m\) i.i.d. trials). Let \(X_j\) be the indicator function of \(A_{ij} > (8/m) \ln(20/\varepsilon w_0)\). \(X_j\) are independent and so using H-C, we see that with probability at least \(1 - \exp(-\varepsilon w_0 s/3)\), less than \(w_0 s/2\) of the \(A_{ij}\) are greater \((8/m) \ln(20/\varepsilon w_0)\), whence, \(\zeta_i' = 0\). So we have (using the union bound over all words):
\[ \Pr\left( \sum_{i: \lambda_i < (1/m) \ln(20/\varepsilon w_0)} \zeta_i' > 0 \right) \leq d \exp(-\varepsilon w_0 s/3). \]

If \(\lambda_i \geq (1/m) \ln(20/\varepsilon w_0)\), then
\[ \Pr(A_{ij} > 4\lambda_i) \leq e^{-\lambda_i m} \leq \varepsilon w_0/2. \]

\(^4||R||\) denotes the spectral norm of \(R\).
which implies by the same $X_j$ kind of argument that with probability at least $1 - \exp(-\varepsilon w_0 s/4)$, for a fixed $i$, $\zeta_i \leq 4\lambda_i m$. Using the union bound over all words and adding all $i$, we get that with probability at least $1 - 2d \exp(-\varepsilon w_0 s/4), \sum_i \zeta_i \leq 4m \sum_i \lambda_i \leq 4m \sum_i \mu_{il} \leq 4km.$

Now we prove the bound on $\zeta_0$. For each fixed $i, j$, we have $\Pr(A_{ij} \geq 4\lambda) \leq e^{-m\lambda} \leq \varepsilon w_0$. Now, let $Y_{ij}$ be the indicator variable of $A_{ij} \geq 4\lambda$. The $Y_{ij}, j = 1, 2, \ldots, s$ are independent (for each fixed $i$). So, $\Pr(\zeta_i \geq 4m\lambda) \leq \Pr(\sum_j Y_{ij} \geq w_0 s/2) \leq e^{-\varepsilon w_0 s/3}$. Using an union bound over all words, we get that $\Pr(\zeta_0 > 4m\lambda) \leq e^{-\varepsilon w_0 s/3}$ by H-C.

**Proof:** (of Theorem B.4) First,

$$||U|| = \max_{|v|=1} E(v^T R_{.,j})^2 \leq E(||R_{.,j}||^2) \leq 2\varepsilon_l \sum_i \zeta_i' \leq 8\varepsilon_l km,$$

by Lemma (B.6) and Lemma (A.4). We can also take $\nu = 2\sqrt{km}$ in Theorem B.5 and with $t = \sqrt{\varepsilon w_0 s}$, the first statement of the current theorem follows (noting $r = w_0 s$). The second statement follows by just paying a factor of $k$ for the $k$ topics.

**B.2 Proving Proximity**

From Theorem (B.4), the $\sigma$ in definition A.1.2 is $\varepsilon \sqrt{w_0 m^2 k^2}$. So, the $\Delta$ in definition A.1.2 is $\varepsilon \sqrt{w_0 m^2 k^2}$. Let $\hat{B}_{.,j}$ be the projection of $B_{.,j}$ onto the line joining $\mu_{.,j}$ and $\mu_{.,j'}$. The probability that $|\hat{B}_{.,j} - \mu_{.,j'}| \leq |\hat{B}_{.,j} - \mu_{.,j}| + c_0 k^2 \sqrt{m}$ is at most $\varepsilon \sqrt{w_0 w_0 \sqrt{\hat{R}}}/\sqrt{\alpha P_0}$. Hence, with probability at least $1 - c_0 k^2 \varepsilon^2 m$, the number of $j$ for which $B_{.,j}$ does not satisfy the proximity condition is at most $\varepsilon \sqrt{w_0 w_0 s}/10\varepsilon_1$.

**Proof:** After paying the failure probability of $c_0 k^2 \varepsilon^2 m$, of Lemmas (B.6) and (A.6), assume that $\zeta_0 \leq 4m\lambda$, $|\mu_{.,j} - \mu_{.,j'}| \geq 2\alpha \varepsilon_0 w_0/9$ and $\sum_i \zeta_i' \leq 4km$.

Let $X = (B_{.,j} - \mu_{.,j}) \cdot (\mu_{.,j'} - \mu_{.,j})$. $X$ is a random variable, whose expectation is 0 conditioned on $j \in T_0^{(2)}$. Since $\Pr(B_{ij} = \sqrt{\zeta_i'}, j \in T) = \mu_{ij}/\sqrt{\zeta_i'}$, we have:

$$E[X] \leq E \left[ \sum_i |B_{ij} - \mu_{ij}| |\mu_{ij'} - \mu_{ij}| \right]$$

$$= \sum_i \left[ \left| \mu_{ij}/\sqrt{\zeta_i} \right| + \left| \mu_{ij}/\sqrt{\zeta_i} \right| \right] |\mu_{ij} - \mu_{ij'}|$$

$$\leq 2\varepsilon_l \sum_i \sqrt{\zeta_i} |\mu_{ij} - \mu_{ij'}| \quad \text{by Lemma A.4}$$

$$\leq 2\varepsilon_l \left( \sum_i \zeta_i' \right) \frac{1}{2} |\mu_{.,j} - \mu_{.,j'}| \leq 4\varepsilon_l \sqrt{\varepsilon km} |\mu_{.,j} - \mu_{.,j'}|.$$

Now apply Markov inequality to get

$$\Pr(|X| \geq \frac{1}{8} |\mu_{.,j} - \mu_{.,j'}|^2) \leq 32\varepsilon_l \sqrt{\varepsilon km}/|\mu_{.,j} - \mu_{.,j'}| \leq 80\varepsilon_l \sqrt{\varepsilon km}.$$

If $|X| \leq |\mu_{.,j} - \mu_{.,j'}|^2/8$, then, $|\hat{B}_{.,j} - \mu_{.,j'}| \geq |\hat{B}_{.,j} - \mu_{.,j}| + 3|\mu_{.,j} - \mu_{.,j'}|/4 \geq |\hat{B}_{.,j} - \mu_{.,j}| + c_0 k^2 \sqrt{m}$, by (9). This proves the first assertion of the Lemma.

The second statement of Lemma follows by applying H-C to the random variable $\sum_j Z_{j}/s$, where, $Z_j$ is the indicator random variable of $B_{.,j}$ not satisfying the proximity condition (and using (9).)
The last Lemma implies that the algorithm TSVD correctly identifies the dominant topic in all but at most $\varepsilon_0 w_0 /10$ fraction of the documents by Theorem (A.7).

**Lemma B.8** With probability at least $1 - \exp(-w_0 s^2 /8)$, TSVD correctly identifies the dominant topic in all but at most $\varepsilon_0 w_0 \delta /10$ fraction of documents in each $T_i$.

### B.3 Identifying Catchwords

Recall the definition of $J_l$ from Step 5a of the algorithm. The two lemmas below are roughly converses of each other which prove roughly that $J_l$ consists of those $i$ for which $M_{il}$ is strictly higher than $M_{il'}$. Using them, Lemma B.11 says that almost all the $\varepsilon_0 w_0 s /2$ documents found in Step 6 of the algorithm are $1 - \delta$ pure for topic $l$.

**Lemma B.9** Let $\nu = (1 - 2\delta) / (1 + \delta)$. If $i \in J_l$, then for all $l' \neq l$, $M_{il} \geq \nu M_{il'}$ and $M_{il} \geq 3 \frac{m \delta}{s} \ln (20 / \varepsilon w_0)$.

**Proof:** It is easy to check that the assumptions (2) and (1) imply $\nu \geq 2$. Let $i \in J_l$. By the definition of $J_l$ in the algorithm, $g(i, l) \geq (4 / \nu \delta^2) \ln (20 / \varepsilon w_0)$. Note that $P_{ij} \leq \text{Max}_l M_{il}$ for all $j$. So,

$$\max_{i,l} M_{il} \geq 3 \frac{m \delta}{s} \ln (20 / \varepsilon w_0). \quad (19)$$

If the Lemma is false, then, for $l'$ attaining $\text{Max}_{i,l} M_{il}$, we have $M_{il} < \nu M_{il'}$. Recall $R_l$ defined in Step 4c of the algorithm. Let

$$\hat{T}_l = R_l \cap \{ \text{the set of } 1 - \delta \text{ pure documents in } T_l \}.$$ 

Since all but $\varepsilon_0 w_0 s /10$ documents in $T_l$ belong to $R_l$, we have $|\hat{T}_l| \geq 0.9\varepsilon_0 w_0 s$. For $j \in \hat{T}_l$, $P_{ij} \geq M_{il} W_{ij} \geq (1 - \delta) M_{il}$. So, $\text{Prob}(A_{ij} < M_{il} (1 - 2\delta)) \leq \exp(-M_{il}^2 /3) \leq \varepsilon w_0 /4$ using (19). Thus the number of documents in $R_l$ for which $A_{ij} \geq M_{il} (1 - 2\delta)$ is at least $0.9\varepsilon_0 w_0 s - 3\varepsilon w_0 s \geq 0.6\varepsilon_0 w_0 s$. This implies that with probability at least $1 - \exp(-c \varepsilon^2 w_0 s)$, $g(i, l') \geq M_{il'} (1 - 2\delta)$.

Now, for $j \in T_l$, $P_{ij} \leq \text{Max}_l (M_{il}, M_{il'}) \leq \nu M_{il'}$. So, $\text{Prob}(A_{ij} > M_{il'} \nu (1 + \delta)) \leq \varepsilon w_0 /4$, again using (19). At most $\varepsilon_0 w_0 s /10$ documents of other $T_{i'}, i' \neq l$ are in $R_l$ (by Lemma B.8). So, whp, $g(i, l) \leq M_{il'} \nu (1 + \delta)$ and so we have

$$g(i, l) \leq \frac{\nu (1 + \delta)}{1 - 2\delta} g(i, l'),$$

contradicting the definition of $J_l$. So, we must have that $M_{il} \geq \nu M_{il'}$ for all $l' \neq l$. The second assertion of the Lemma now follows from (19).

**Lemma B.10** If $M_{il} \geq \text{Max}_{l'} \left( \frac{5 \ln (20 / \varepsilon w_0)}{m \delta}, \frac{1}{\rho} M_{il'} \right)$, then, with probability at least $1 - \exp(-c \varepsilon^2 w_0 s)$, we have that $i \in J_l$. So, $S_l \subseteq J_l$.

**Proof:** Let $\hat{T}_l = R_l \cap \{ \text{set of } 1 - \delta \text{ pure documents in } T_l \}$. For $j \in \hat{T}_l$, $P_{ij} \geq M_{il} (1 - \delta)$ which implies that whp, (since $|\hat{T}_l| \geq 0.9\varepsilon_0 s$, again by Lemma B.8)

$$g(i, l) \geq M_{il} (1 - 2\delta) \quad (20)$$

On the other hand, for $j \in T_l$ and for $l' \neq l$, $i : M_{il'} \leq \rho M_{il}$ (hypothesis of the Lemma), $P_{ij} \leq M_{il} W_{ij} + \rho M_{il} (1 - W_{ij}) \leq M_{il} (\beta + \rho)$. So whp,

$$g(i, l') \leq M_{il} (\beta + \rho) (1 + \delta). \quad (21)$$

From (20) and (21) and hypothesis of the Lemma, it follows that

$$g(i, l) \geq \text{Max}_{l'} \left( \frac{4 \ln (1 / \varepsilon w_0)}{m \delta^2}, \frac{1 - 2\delta}{(1 + \delta)(\beta + \rho)} g(i, l') \right).$$

So, $i \in J_l$ as claimed. It only remains to check that $i$ in $S_l$ satisfies the hypothesis of the Lemma which is obvious.
Lemma B.11 Let $v_i = \sum_{i \in J_i} M_{i,i}$ and let $L$ be the set of $\{(s_0 w_0/2)\}$'s whose average is returned in Step 6 of the TSVD Algorithm as $\hat{M}_{i,j}$. With probability at least $1 - c \exp(-c \varepsilon^2 w_0 s)$, we have:

$$\left| \frac{1}{|L|} \sum_{j \in L} (A_{i,j} - \hat{M}_{i,j}) \right|_1 \leq O(\delta). \quad (22)$$

Proof: The proof needs care since $J_i$ is itself a random set dependent on $A^{(2)}$. To understand the proof intuitively, if we pretend that there is no conditioning of $J_i$ on $A^{(2)}$, then, basically, our arguments in Lemma B.9 would yield this Lemma. However, we have to work harder to avoid conditioning effects. Define

$$K_i = \{ i : M_{i,i} \geq \nu M_{i,i} \forall \ell' \neq \ell; M_{i,i} \geq (3/m\delta^2) \ln(20/\varepsilon w_0) \}.$$ 

Note that $K_i$ is not a random set; it does not depend on $A$, just on $M$ which is fixed. Lemma B.9 proved that $J_i \subseteq K_i$. Since $\sum_i M_{i,i} = 1$, we have $|K_i| \leq m\delta^2/3$. The probability bounds given here will be after conditioning on $W$. [In other words, we prove statements of the form $\Prob(E | W) \leq a$ which is (the usual) shorthand for: for each possible value $w$ of the matrix $W$, $\Prob(E | W = w) \leq a$.] This will be possible, since, even after fixing $W$, the $A_{i,j}$ are independent, though certainly not identically distributed now, since the $W_{i,j}$ may differ.

For $i \in K_i$, we have for all $j$, $P_{ij} = \sum_{i'} M_{i,i'} W_{i'j} \leq M_{i,i}$, since, $M_{i,i'} \leq M_{i,i}/\nu \leq M_{i,i}/2$ for $\ell' \neq \ell$. For any $x \leq M_{i,i}$,

$$\Prob(|A_{ij}^{(2)} - P_{ij}| \geq \delta M_{i,i} | W, P_{ij} = x) \leq 2 \exp \left( -\frac{\delta^2 M_{i,i}^2 m}{2(1 + \delta)x} \right) \leq 2 \exp \left( -\frac{m\delta^2 M_{i,i}}{3} \right).$$

Noting that $m\delta^2 M_{i,i} \geq 3 \ln(20/\varepsilon w_0)$ for $i \in K_i$, we get

$$\Prob(|A_{ij}^{(2)} - P_{ij}| \geq \delta M_{i,i} | W) \leq \varepsilon w_0/20.$$

Using the union bound over all $i \in K_i$ yields (for each $j \in [s]$),

$$\Prob(\exists i \in K_i : |A_{ij}^{(2)} - P_{ij}| \geq \delta M_{i,i} | W) \leq \frac{m\delta^2 \varepsilon w_0}{20} \leq \frac{\varepsilon_0 w_0 \delta^2}{20},$$

by (9). Let

$$BAD = \{ j : \exists i \in K_i : |A_{ij}^{(2)} - P_{ij}| \geq \delta M_{i,i} \}. $$

Using the independence of $A_{i,j}$, (even conditioned on $W$), apply H-C to get that for the event $\mathcal{E} : |BAD| \geq \frac{8\varepsilon_0 w_0 \delta}{10}$

$$\Prob(\mathcal{E} | W) \leq 2 \exp(-c\varepsilon w_0 s). \quad (23)$$

After paying the failure probability, for the rest of the proof, assume that $\neg \mathcal{E}$ holds. Let $U_i = \{ j : W_{ij} \geq 1 - \delta \}$. By the dominant topic assumption, we know that $|U_i| \geq \varepsilon_0 w_0 s$. So, $|U_i \setminus BAD| \geq 4\varepsilon_0 w_0 s/5$ and we get (using (9)):

$$\forall N_i \subseteq K_i, \left| \{ j : W_{ij} \geq 1 - \delta ; \sum_{i \in N_i} A_{ij}^{(2)} \geq (1 - 2\delta) \sum_{i \in N_i} M_{i,i} \} \right| \geq 4\varepsilon_0 w_0 s/5. \quad (24)$$

Now consider $j : W_{ij} \leq (1 - 6\delta)$ and $i \in K_i$.

$$P_{ij} \leq M_{i,i} W_{ij} + \sum_{\ell' \neq \ell} M_{i,i'} W_{i'j} \leq M_{i,i}(1 - 6\delta) + \frac{M_{i,i}}{\nu} 6\delta \leq M_{i,i}(1 - 3\delta),$$

since by (2) and (1), we have that $\nu \geq 2$. So, for a $j$ with $W_{ij} \leq 1 - 6\delta$ to have $\sum_{i \in J_i} A_{ij}^{(2)} \geq (1 - 2\delta)\nu_{i,j}$, $j$ must be in $BAD$. This gives us

$$\forall N_i \subseteq K_i, \left| \{ j : W_{ij} \leq (1 - 6\delta) ; \sum_{i \in N_i} A_{ij}^{(2)} \geq (1 - 2\delta) \sum_{i \in N_i} M_{i,i} \} \right| \leq \varepsilon_0 w_0 \delta s/10. \quad (25)$$
Let $L$ be the set of $\lfloor \varepsilon_0 w_0 s/2 \rfloor$ words picked to be in the document. But then we would have $|\delta_i| \leq \varepsilon_0 w_0 s/10$ on the number of $\ell$’s and gives us $|X - EX| \leq O(\varepsilon)s/|L|$. We still have to bound $EX$. By Jensen’s inequality,

$$EX \leq \frac{1}{|L|} \sum_i \left( E \left( \sum_{j \in L} (A_{ij} - P_{ij})^2 \right)^{1/2} \right) \leq \frac{1}{|L|} \sum_i \sqrt{\sum_{j \in L} P_{ij}} \leq \sqrt{d/|L|},$$

where, we have used the independence of $A_{ij}$ and the fact that $E(A_{ij} - P_{ij})^2 = \text{Var}(A_{ij})$. This proves the claim.

We now bound $\frac{1}{|L|} \sum_{j \in L} (P_{ij} - M_{ij})$. Note that by (24) and (25), all but at most $\varepsilon_0 w_0 s/10$ of the $j$’s in $L$ have $W_{ij} \geq 1 - 6\delta$, whence, we get $|P_{ij} - M_{ij}| \leq 6\delta$ for these $j$. For the $j$ with $W_{ij} < 1 - 6\delta$, we just use $|P_{ij} - M_{ij}| \leq 2$. So

$$\frac{1}{|L|} \sum_{j \in L} (P_{ij} - M_{ij}) \leq 6\delta + \frac{0.2\varepsilon_0 w_0 s}{10|L|} \in O(\varepsilon).$$

This finishes the proof of (22).