Efficient Distributed Topic Modeling with Provable Guarantees

Weicong Ding Boston University
Mohammad H. Rohban Boston University
Prakash Ishwar Boston University
Venkatesh Saligrama Boston University

Abstract

Topic modeling for large-scale distributed web-collections requires distributed techniques that account for both computational and communication costs. We consider topic modeling under the separability assumption and develop novel computationally efficient methods that provably achieve the statistical performance of the state-of-the-art *centralized* approaches while requiring insignificant communication between the distributed document collections. We achieve tradeoffs between communication and computation without actually transmitting the documents. Our scheme is based on exploiting the geometry of normalized word-word cooccurrence matrix and viewing each row of this matrix as a vector in a high-dimensional space. We relate the solid angle subtended by extreme points of the convex hull of these vectors to topic identities and construct distributed schemes to identify topics.

1 Introduction

Large and web-scale document collections are ubiquitous as evidenced by Google online libraries, Twitter streaming, and Flickr image hosting databases. Such large-scale collections are generally archived in *distributed* servers worldwide. The goal of this paper is topic discovery, i.e., extract the common dominant themes among a corpus. Due to the distributed nature of the corpora and limited communication bandwidth between servers, new techniques that account for both computation and communication costs are required. In this paper, we develop novel computationally efficient methods that provably achieve the statistical performance of the state-of-the-art *centralized* approaches while requiring insignificant communication between distributed servers which contain documents.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

We consider a corpus of M documents composed of words drawn from a vocabulary of size W among L servers. As is now common (Blei, 2012) we view each document as a collection of N i.i.d word drawings from an unknown $W \times 1$ document word-distribution vector. Each document word-distribution vector is modeled as an unknown probabilistic mixture of $K \ll W$ unknown $W \times 1$ latent *topic* word-distribution vectors. These are grouped into a topic matrix $\boldsymbol{\beta}$ and are shared among the M documents. Each document-specific mixture over topics is assumed to be sampled i.i.d from a prior, e.g., Dirichlet in Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Log-normal in Correlated Topic Model (Blei and Lafferty, 2007). The goal is to estimate the latent topic-word distribution matrix ($\boldsymbol{\beta}$) from the empirical word-frequencies of all documents ($\mathbf{X}$) while keeping communication and computation costs low.

Topic modeling has been extensively studied and has resulted in a number of different approaches. Our approach here is based on Nonnegative Matrix Factorization (NMF) with additional separability assumption on the topic matrix. The separability assumption amounts to existence of "novel" words unique to each topic. While the general NMF problem, even without the distributed aspect, is known to be $\mathcal{NP}$-hard (Vavasis, 2009), it has recently been shown that under the separability assumption NMF has a polynomial time solution (Arora et al., 2012). A key step in the solution strategy in this setting is the identification of novel words corresponding to each topic (e.g., Arora et al., 2013; Ding et al., 2013). Once the novel words are identified, the topic matrix can be estimated by means of linear regression.

Our focus here is on distributed methods for novel word detection since the linear regression can be efficiently parallelized. For $H = W/L$ documents/server one of our schemes has a $\mathcal{O}(HNK + W)$ computation cost/server and communication cost that scales as $\mathcal{O}(WK)$ bits/server. The previously proposed state-of-the-arts (e.g., Arora et al., 2013; Ding et al., 2013), however, are not particularly amenable to parallelization. Furthermore, our parallelization has other ben-

efits namely better computational efficiency at the cost of modest communication. Indeed, to achieve similar statistical efficiency as our method, their centralized approach (i.e. even ignoring communication costs) requires higher computation cost - $\mathcal{O}(MN^2/\epsilon^2 + WK/\epsilon^2)$ for Arora et al. (2013) and $\mathcal{O}(MN^2 + W^2)$ for Ding et al. (2013). Next for a lower computational cost namely, $\mathcal{O}(HN^2/\epsilon^2 + WK/\epsilon^2)$ for Arora et al. (2013) or $\mathcal{O}(HN^2 + W^2)$ for Ding et al. (2013) per server which are comparable to our scheme, their statistical accuracy degrades significantly.

Our scheme hinges on exploiting the geometry of the normalized word-word co-occurrence matrix. Words are associated with row vectors of this matrix and the novel words correspond to extreme points of the convex hull spanned by these row vectors. It follows that the solid angle subtended by each novel word (i.e., the associated row vectors) is strictly positive but identically zero for non-novel words. The essence of our scheme then boils down to detecting whether or not the solid angle for a word is non-zero. We can do this through random projections. Specifically, we project each word along random directions and the number of times a word achieves a maximum value along a random direction is proportional to the solid angle. This process of random projections followed by counting the number of times a word is a maximizer can be efficiently parallelized leading to our results.

1.1 Related Work

Topic models and their distributed variations have been studied before. Bayesian based approaches attempt to fit a MAP/ML estimate to the data using heuristics such as variational Bayes (Blei et al., 2003) and Gibbs Sampling (Griffiths and Steyvers, 2004). To deal with a distributed corpus techniques such as collapsed Gibbs sampling and distributed MAP inference have been proposed (Asuncion et al., 2009; Newman et al., 2009). These distributed approaches appear to empirically achieve the performance of their centralized counterparts. An alternative approach is based on NMF with appropriate regularization (Cichocki et al., 2009). Distributed variants of NMF have been proposed (Liu et al., 2010; Gemulla et al., 2011) based on stochastic gradient descent to account for communication costs. Nevertheless, to our best knowledge, none of the distributed approaches proposed so far come with statistical, computational, and communication guarantees.

One possible direction is to attempt to parallelize existing topic modeling algorithms that do come with computational and statistical guarantees. While the general problem is known to be $\mathcal{NP}$-hard (Arora et al., 2012), a recent trend shows that the topic discovery problem lends itself to polynomial time solutions with additional structural assumptions (e.g., Anandkumar et al., 2012, 2013; Arora et al., 2012, 2013; Ding et al., 2013; Recht et al., 2012; Kumar et al., 2013).

The setup of our paper is closely related to Arora et al. (2012, 2013); Ding et al. (2013) that consider topic models under the so called separability assumption on the topic matrix, $\boldsymbol{\beta}$, and simplicial assumption on the normalized second order moments of topic prior [1]. Nevertheless, it is unclear how to directly parallelize these approaches. For instance, the key step of Arora et al. (2013) is based on Gram-Schmidt type of process over rows of the (normalized) empirical word-word co-occurrence matrix. Gram-Schmidt process being inherently sequential, it is hard to parallelize this key step. Similarly, the key step in Ding et al. (2013) scans each row of the normalized second order moments and declares a word i as novel if in the i-th row, the diagonal term is the maximum by some margin. In the worst case, this step scales as $\mathcal{O}(W^2)$ irrespective of number of documents. For our distributed context this step has either a computational cost/server or a communication cost/server that scales as $\mathcal{O}(W^2)$ irrespective of the other model parameters.

The rest of this paper is organized as follows. In section 2, we describe the main intuition which is geometric. In section 3, we provide the algorithms along with their computational and communication costs. In section 4, we summarize its statistical guarantees. In section 5, we present a set of experiments to demonstrate the superiority of the algorithms in various aspects.

2 Topic Geometry and Solid Angle

Let $\boldsymbol{\beta}$ (of size $W \times K$) and $\mathbf{X}$($W \times M$) denote the topic matrix and the empirical word-by-document matrix. We obtain $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{X}}'$ by first splitting each document into two independent copies and then scaling the rows to make them row-stochastic. Let L be the number of distributed servers and $\mathbf{X}^{(l)}$ be empirical words counts of the documents in the l-th node, which is a slice of $\mathbf{X}$. A fusion center controls the whole process and outputs the estimated topics. We assume that the topic matrix $\boldsymbol{\beta}$ is separable and let $\mathcal{C}_k$ be the set of novel words of topic k and $\mathcal{C}_0$ be the set of non-novel words. $\mathbf{A}_i$ denotes the i-th row vector of a matrix $\mathbf{A}$.

[1] A topic matrix $\boldsymbol{\beta} \in \mathbb{R}^{W \times K}$ is **separable** if for each topic k, there is some word i, called novel word, such that $\boldsymbol{\beta}_{i,k} > 0$ and $\boldsymbol{\beta}_{i,l} = 0$, $\forall\ l \neq k$. Let $\mathbf{a}$ and $\mathbf{R}$ be the mean and correlation matrix of the topic prior. $\mathbf{R}' = \mathrm{diag}(\mathbf{a})^{-1}\mathbf{R}\,\mathrm{diag}(\mathbf{a})^{-1}$. A topic model is γ-**simplicial** if every row vector of $\mathbf{R}'$ is at least $\gamma > 0$ distant from the convex hull of other rows of $\mathbf{R}'$. To be precise, Arora et al. (2013) and Ding et al. (2013) each requires stronger assumption on the topic priors that each implies simplicial.

We denote the mean and correlation matrix of the topic prior by the $K \times 1$ vector $\mathbf{a}$ and $K \times K$ matrix $\mathbf{R}$ respectively. We define the normalized second order moment as $\mathbf{R}' \triangleq \text{diag}^{-1}(\mathbf{a})\mathbf{R}\,\text{diag}^{-1}(\mathbf{a})$ and assume $\mathbf{R}'$ to be γ-simplicial, i.e., its row vectors are extreme points of the convex hull they themselves constitute. This induces a simple geometric picture of the normalized word co-occurrence matrix $\mathbf{E} \triangleq \boldsymbol{\beta}'\mathbf{R}'\boldsymbol{\beta}'^{\top}$ (of size $W \times W$) where $\boldsymbol{\beta}' = \text{diag}^{-1}(\boldsymbol{\beta}\mathbf{a})\boldsymbol{\beta}\,\text{diag}(\mathbf{a})$:

Lemma 1. *Suppose that $\mathbf{R}'$ is γ-simplicial and $\boldsymbol{\beta}$ is separable. Then, word i is a novel word if and only if the i-th row of $\mathbf{E}$ is an extreme point of the convex hull spanned by rows of $\mathbf{E}$.*

Proof. It is straightforward to check that $\mathbf{Y} = \mathbf{R}'\boldsymbol{\beta}'^{\top}$ is simplicial. Note that $\boldsymbol{\beta}'$ is row stochastic. For $i \in \mathcal{C}_k$, $\boldsymbol{\beta}'_{ik} = 1$, so the i-th row $\mathbf{E}_i = \mathbf{Y}_k$. For $i \in \mathcal{C}_0$, $\mathbf{E}_i$ is convex combination of at least two rows of $\mathbf{Y}$. □

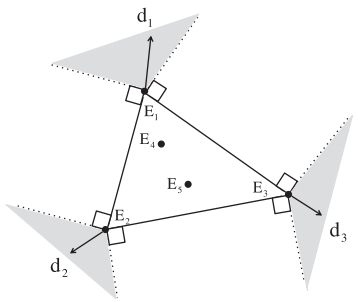

Figure 1: Schematic view of convex hull spanned by rows of $\mathbf{E}$. Word $1, 2, 3$ are novel, word $4, 5$ are non-novel. The shaded regions depict the set of directions in which each of the extreme points has the maximum projection along. Solid Angle q_i, for a novel word i, is the probability of the i-th shaded region considering $\mathbf{E}_i$ as the origin.

Fig. 1 illustrates the geometric property formalized in Lemma 1, i.e., the novel words correspond to extreme points of all rows of $\mathbf{E}$. Thus novel words can be efficiently detected through an extreme point finding algorithm. To solve this geometric problem, we exploit a key quantity, the normalized **Solid Angle** of an extreme point, as indicated by the shaded angles in Fig. 1. From a statistical viewpoint, it can be defined as the probability that a certain point has the maximum projection value along an isotropically distributed random direction. This value is strictly nonzero iff it is an extreme point. Formally,

Definition 1. *The normalized solid angle of word i is*

$$q_i \triangleq \Pr(\forall j, \mathbf{E}_i \neq \mathbf{E}_j : \langle \mathbf{E}_i, \mathbf{d} \rangle > \langle \mathbf{E}_j, \mathbf{d} \rangle) \qquad (1)$$

where $\mathbf{d}$ is drawn from an isotropic distribution (e.g., spherical Gaussian).

Lemma 2. *Suppose that $\mathbf{R}'$ is γ-simplicial and $\boldsymbol{\beta}$ is separable. Then, $q_i > 0$ if and only if i is a novel word.*

Lemma 2 depicts our high level approach : (1) Estimate the solid angles q_i's. (2) Select K words with largest q_i's whose word co-occurrence patterns are distinct as novel words. (3) Estimate topic matrix using constrained linear regression as in Arora et al. (2013) and Ding et al. (2013).

Estimate Solid Angle through projections:

We can rewrite q_i defined in (1) as,

$$q_i = \mathbb{E}\left(\mathbb{I}\{\forall j, \|\mathbf{E}_i - \mathbf{E}_j\| \geq \zeta : \ \mathbf{E}_i\mathbf{d} \geq \mathbf{E}_j\mathbf{d}\}\right) \qquad (2)$$

where ζ is some proper threshold. We first note that the expectation in (2) can be well approximated by an empirical mean by (i) first Projecting all the rows of $\mathbf{E}$ along a few iid isotropic directions $\mathbf{d}$'s and then (ii) calculating the frequency that word i maximizes the projection values. On the other hand, it has been shown that $M\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}^{\top} \xrightarrow{p} \mathbf{E}$ as the number of document $M \to \infty$ while N is fixed (c.f., Arora et al., 2013). In sum we can estimate q_i consistently and efficiently in a *centralized* setting.

Estimate Solid Angle from distributed servers:

Now we consider the distributed settings where the M documents are evenly distributed over L servers. Recall that our approach boils down to calculating the projection values $\mathbf{Ed}$ and gathering the indices of the maximizers for a few iid $\mathbf{d}$'s. A simple implementation of this idea is to have each server first generate one independent random direction, then calculate $\mathbf{X}^{(l)}\mathbf{X}^{(l)\top}\mathbf{d}$, determine the the word which maximizes the projection and transmit only its index to the fusion center for novel word detection. (Fig. 2 (left))

The above scheme has low communication cost. However, as we shall see later, to obtain desired statistical accuracy, it is crucial to calculate the projection values $\mathbf{Ed}$ with as many documents as possible. We observe that $\mathbf{Ed} \approx M\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}^{\top}\mathbf{d} = M\sum_{l=1}^{L}\widetilde{\mathbf{X}}'^{(l)}\widetilde{\mathbf{X}}^{(l)\top}\mathbf{d}$, which is a summation of L $(W \times 1)$ partial projection values that can be calculated locally. Thus if a common set of P random directions has been pre-distributed to the servers (or they have access to the same random number generator with a common seed), they can compute and transmit the partial projections to the fusion center instead of sending the entire set of their local documents. The solid angle estimates based on the number of times the projection of the row-vectors corresponding to words is maximum can be calculated at the fusion center. This scheme (Fig. 2, right) has a higher but moderate communication cost $(\mathcal{O}(WP))$ while its statistical accuracy can match that of a centralized algorithm which has access to the documents from all the servers.

We refer to the scheme which is based on transmitting

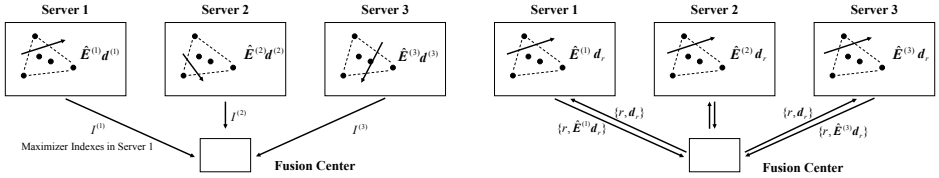

Figure 2: Two schemes to estimate solid angle : index passing (Left) and projection value passing (Right).

the maximizer indices as Alg.Index and the scheme based on transmitting partial projections as Alg.Value. While Alg.-Index offers some intuitions in analysis, Alg.-Value achieves all the desired performance.

3 Distributed Novel Words Detection

Algorithm 1 below sketches the key steps that are applicable to both Alg.-Index and Alg.-Value. The two main steps are 1) Novel words detection; 2) Topic Estimation. Recall that M documents are stored on L distributed servers (thus $H = M/L$ docs/server) which are connected to the Fusion Center.

Algorithm 1 Algorithm in High Level

Input: Text docs. $\widetilde{\mathbf{X}}^{(l)}$, $\widetilde{\mathbf{X}}'^{(l)}$ on each node; Number of topics K; Tolerances $\zeta, \epsilon > 0$.
Output: Topic matrix $\boldsymbol{\beta}$.
(Node l) $\mathcal{M}_l \leftarrow$NovelWordDetect-Nodes($\widetilde{\mathbf{X}}^{(l)}, \widetilde{\mathbf{X}}'^{(l)}, \zeta$).
(Center) Set of Novel Words $\mathcal{J} \leftarrow$NovelWordDetect-Center($\mathcal{M}_l$, K, ζ).
$\hat{\boldsymbol{\beta}} \leftarrow$EstimateTopics($\mathcal{I}, \mathbf{X}, \epsilon$)

3.1 Index passing scheme (Alg.-Index)

Algorithm 2 (Index Passing) NovelWordDetect-Nodes

Input: $\widetilde{\mathbf{X}}_{(l)}$, $\widetilde{\mathbf{X}}_{\prime(l)}$, ζ
Output: $\mathcal{I}_l$, indexes of words with max. projs..
$\mathcal{I}_l \leftarrow \varnothing$
$\widehat{\mathbf{E}}^{(l)} \leftarrow H\widetilde{\mathbf{X}}'^{(l)}\widetilde{\mathbf{X}}^{(l)\top}$
$\mathbf{d}^{(l)} \leftarrow$ a sample from an isotropic distribution
$i^* \leftarrow \arg\max_{1\le i\le W}\langle\widehat{\mathbf{E}}_i^{(l)}, \mathbf{d}^{(l)}\rangle$
$\mathcal{I}_l \leftarrow \mathcal{I}_l \cup \{i^*\}$
$\hat{J}_{i^*} \leftarrow \{j : \widehat{E}^{(l)}_{i^*,i^*} + \widehat{E}^{(l)}_{j,j} - 2\widehat{E}^{(l)}_{i^*,j} \ge \zeta/2\}$
for all $k \in \hat{J}^c_{i^*}$
  $\hat{J}_k \leftarrow \{j : \widehat{E}^{(l)}_{k,k} + \widehat{E}^{(l)}_{j,j} - 2\widehat{E}^{(l)}_{k,j} \ge \zeta/2\}$
  if $\forall j \in \hat{J}_k : \langle\widehat{\mathbf{E}}_k^{(l)}, \mathbf{d}^{(l)}\rangle > \langle\widehat{\mathbf{E}}_j^{(l)}, \mathbf{d}^{(l)}\rangle$
    $\mathcal{I}_l \leftarrow \mathcal{I}_l \cup \{k\}$
  end if

Alg.-Index approximates the Solid Angles q_i defined in Eqn. (2) as follows,

$$\hat{q}_i = \frac{1}{L}\sum_{l=1}^{L}\mathbb{I}(\forall j, \widehat{E}^{(l)}_{i,i} + \widehat{E}^{(l)}_{j,j} - 2\widehat{E}^{(l)}_{i,j} \ge \zeta/2 : \widehat{\mathbf{E}}_i^{(l)}\mathbf{d}^{(l)} > \widehat{\mathbf{E}}_j^{(l)}\mathbf{d}^{(l)}) \quad (3)$$

where $W \times W$ matrices $\widehat{\mathbf{E}}^{(l)} = H\widetilde{\mathbf{X}}'^{(l)}\widetilde{\mathbf{X}}^{(l)\top}$ and $\mathbf{d}^{(l)}$'s are iid directions. Each indicator tests if word i ($\widehat{\mathbf{E}}_i^{(l)}$) has the maximum projection along $\mathbf{d}^{(l)}$. This is carried out within each server. To be more exact, for each word i, we only compare its projection value against that of the words in the set $\hat{J}_i = \{j : \widehat{E}^{(l)}_{i,i} + \widehat{E}^{(l)}_{j,j} - 2\widehat{E}^{(l)}_{i,j} \ge \zeta/2\}$ which converge to $J_i = \{j : \mathbf{E}_i \ne \mathbf{E}_j\}$ as $H \to \infty$ (Ding et al., 2013). For a novel word i, J_i excludes all the other novel words of the same topic as i to avoid technical difficulties.

We can simplify the calculation of Eq. (3), as sketched in Alg. 2. First we note that for each $\mathbf{d}^{(l)}$, most of the indicators in Eq. (3) would be zero. In fact, given $\mathbf{d}^{(l)}$, we can first find the maximizer index i^* and corresponding set $\hat{J}_{i^*}$. Note that $\hat{J}_{i^*}$ consists of novel words for topics different from i^* and the non-novel words. Hence, all the words in $\hat{J}_{i^*}$ should make the indicator functions zero for the given l. As a consequence, we only evaluate the indicators for words in $\hat{J}^c_{i^*} = \{1, \ldots, W\} \setminus \hat{J}_{i^*}$, which have up to $|\mathcal{C}_k|$ elements asymptotically, where k is the topic associated with word i^*. Moreover, in a typical dataset, $\max_k |\mathcal{C}_k|$ is of the order $\mathcal{O}(1)$.

Secondly, we note that the matrices $\widehat{\mathbf{E}}^{(l)}$'s do not need to be calculated explicitly. Recall that the projection values $\widehat{\mathbf{E}}^{(l)}\mathbf{d}^{(l)} = H\widetilde{\mathbf{X}}'^{(l)}\widetilde{\mathbf{X}}^{(l)\top}\mathbf{d}^{(l)}$ and matrices $\widetilde{\mathbf{X}}'^{(l)}$ are sparse. Each $W \times 1$ projection vector requires $\mathcal{O}(HN)$ computation time. Similarly, for i^*, all the corresponding $\widehat{E}^{(l)}_{i^*,j}$'s can be viewed as projecting along $\mathbf{d} = \mathbf{e}_{i^*}$ whose only non-zero element is i^*.

Algorithm 3 (Index Passing) NovelWordDetect-Center

Input: Message $\mathcal{I}_1, \ldots, \mathcal{I}_L$ sent from nodes; $\widetilde{\mathbf{X}}^{(L+1)}$, $\widetilde{\mathbf{X}}'^{(L+1)}$, ζ, H, L, K
Output: $\mathcal{J}$: Indices of K distinct novel words
$\hat{q}_1, \ldots, \hat{q}_W \leftarrow 0$.
for all $1 \le i \le L$ and $j \in \mathcal{I}_i$
  $\hat{q}_j \leftarrow \hat{q}_j + 1/L$
$\widehat{\mathbf{E}}^{(L+1)} \leftarrow H\widetilde{\mathbf{X}}'^{(L+1)}\widetilde{\mathbf{X}}^{(L+1)\top}$
$\mathcal{J} \leftarrow$ FindNovelWords($\widehat{\mathbf{E}}^{(L+1)}, \{\hat{q}_1, \ldots, \hat{q}_W\}, K, \zeta$)

Alg. 3 summarizes steps for the fusion center. After gathering the maximizer indices, the fusion center estimates the $\hat{q}_i$'s, sort them in a descending order, and

Algorithm 4 FindNovelWords

Input: $\widehat{\mathbf{E}}, \{\hat{q}_1, \ldots, \hat{q}_W\}, K, \zeta$
Output: $\mathcal{J}$: Indices of K distinct novel words
  $\mathcal{J} \leftarrow \varnothing$, $k \leftarrow 0$, $j \leftarrow 1$
  while $k < K$ **do**
    $i \leftarrow$ index of the j-th largest value of $\{\hat{q}_1, \ldots, \hat{q}_W\}$
    if $\forall p \in \mathcal{J} : \widehat{E}_{p,p} + \widehat{E}_{i,i} - 2\widehat{E}_{i,p} \geq \zeta/2$ **then**
      $\mathcal{J} \leftarrow \mathcal{J} \cup \{i\}$, $k \leftarrow k+1$
    end if
    $j \leftarrow j+1$
  end while

select distinct novel words from the top as in Alg. 4. We defer the consistency result in Sec.4. Its running time and communication cost are summarized as follows and proved in supplementary section:

Proposition 1. *The running time of Alg.-Index is $\mathcal{O}(HN+W)$ per server. The total communication cost is $\mathcal{O}(\log(W))$ bits/server.*

3.2 Projection value passing (Alg.-Value)

Algorithm 5 and 6 outline the Alg.-Value. Similar in structure to Alg.-Index, the main idea here is to use *all* the M documents in the corpus to calculate the projections and then estimate solid angles q_i's using P *globally synchronized* directions $\mathbf{d}_r, r = 1, \ldots, P$ as follows:

$$\hat{q}_i = \frac{1}{P}\sum_{r=1}^{P} \mathbb{I}(\forall j, \widehat{E}_{i,i} + \widehat{E}_{j,j} - 2\widehat{E}_{i,j} \geq \zeta/2 : \widehat{\mathbf{E}}_i \mathbf{d}_r > \widehat{\mathbf{E}}_j \mathbf{d}_r) \quad (4)$$

Since

$$\widehat{\mathbf{E}}\mathbf{d}_r = M\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}^\top \mathbf{d}_r = M\sum_{l=1}^{L} \widetilde{\mathbf{X}}'^{(l)}\widetilde{\mathbf{X}}^{(l)\top}\mathbf{d}_r \quad (5)$$

the partial projection values $\widetilde{\mathbf{X}}'^{(l)}\widetilde{\mathbf{X}}^{(l)\top}\mathbf{d}_r$ of size $W \times 1$ can be calculated locally and transmitted to the fusion center. To synchronize $\mathbf{d}_r$'s across servers and to transmit the partial projection values, each server need to communicates $\mathcal{O}(WP)$ floating-point numbers.

The fusion center executes all the remaining procedures. We point out to calculate Eqn. (5), rows of $\widetilde{\mathbf{X}}$ are normalized globally. Using similar tricks as in Alg.-Index, we have:

Proposition 2. *The running time of Alg.-Value is $\mathcal{O}(HNP+W)$ per server and $\mathcal{O}(WPL+K^2)$ for the fusion center. The communication cost is $\mathcal{O}(WP)$ floating numbers/server.*

We refer to supplementary for the computational and communication cost of each step. We further prove its sample complexity in Sec. 4. As demonstrated in Sec. 5, it attains desirable performance on empirical corpora.

Algorithm 5 (Value Passing) NovelWordDetect-node

Input: $\widetilde{\mathbf{X}}^{(l)}$, $\widetilde{\mathbf{X}}'^{(l)}$, P Directions: $\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(P)}$
Output: $\mathbf{V}^{(l)}$: a $W \times P$ matrix containing the projection values of $\widehat{\mathbf{E}}^{(l)}$ rows along the given P directions

$\mathbf{V}^{(l)} \leftarrow M\widetilde{\mathbf{X}}'^{(l)}\widetilde{\mathbf{X}}^{(l)\top}\left[\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(P)}\right]$

Algorithm 6 (Value Passing) NovelWordDetect-center

Input: Message $\mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(l)}$ sent from nodes; ζ, K
Output: $\mathcal{J}$: Indices of K distinct novel words
  Generate P iid directions $\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(P)}$; Send to each node, Call NovelWordDetect-Node.
  $\mathbf{V} \leftarrow \sum_{l=1}^{L} \mathbf{V}^{(l)}$, and $\forall i : \hat{q}_i \leftarrow 0$
  for all $1 \leq r \leq P$
    $i^* \leftarrow \arg\max_{1 \leq j \leq W} V_{j,r}$
    $\hat{q}_{i^*} \leftarrow \hat{q}_{i^*} + 1/P$
    $\hat{J}_{i^*} \leftarrow \{j : \widehat{E}_{i^*,i^*} + \widehat{E}_{j,j} - 2\widehat{E}_{i^*,j} \geq \zeta/2\}$
    for all $k \in \hat{J}_{i^*}^c$
      $\hat{J}_k \leftarrow \{j : \widehat{E}_{k,k} + \widehat{E}_{j,j} - 2\widehat{E}_{k,j} \geq \zeta/2\}$
      if $\forall j \in \hat{J}_k : V_{k,r} > V_{j,r}$
        $\hat{q}_k \leftarrow \hat{q}_k + 1/P$
      end if
  $\mathcal{J} \leftarrow$ **FindNovelWords**$(\widehat{\mathbf{E}}, \{\hat{q}_1, \ldots, \hat{q}_W\}, K, \zeta)$

We note that the topic estimation step is similar to Ding et al. (2013); Arora et al. (2013). In this step, the W number of regressions for estimation are decoupled and distributable. We defer the operational details of this step to the supplementary.

4 Theoretical analysis

We present the sample complexity bound for the proposed algorithms in this section. Recall that H, L, and P denote the number of documents/server, the number of servers, and the number of projections (in Alg.-Value), respectively. $\mathbf{a} \in \mathbb{R}^K$ and $\mathbf{R} \in \mathbb{R}^{K \times K}$ are the expectation and correlation matrix of topic prior. Let $\mathbf{R}' = \mathrm{diag}^{-1}(\mathbf{a})\mathbf{R}\,\mathrm{diag}^{-1}(\mathbf{a})$. We present the results when $\mathbf{d}$'s are spherical Guassian. We also provide sample complexity bounds for other isotopic priors in the supplementary. For Alg.-Index, we have,

Theorem 1. *Let $\mathbf{R}'$ be γ-simplicial and the topic matrix $\boldsymbol{\beta}$ be separable. Let $\mathbf{d}^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$. Then, Alg.-Index outputs all the novel words of K topics consistently as both $H, L \to \infty$. Furthermore, $\forall \delta > 0$, for*

$$H \geq \max\left\{ c_1 \frac{\log(8W/q_\wedge)}{\zeta^2\phi^2\eta^4},\ c_2 \frac{W^2\log^2(8W/q_\wedge)}{\pi\rho^2 q_\wedge^2 \phi^2 \eta^4} \right\}$$

and for

$$L \geq c_3 \frac{\log(W/\delta)}{q_\wedge^2}$$

Alg.-Index fails with probability at most δ, where c_1 to c_3 are some absolute constants, $\phi = \min_{i,j} \frac{a_i a_j}{R_{i,j}}$, $\eta = \min_{1 \le i \le W} \boldsymbol{\beta}_i \mathbf{a}$, $\rho = \gamma \min_{\beta'_{j,k} \neq 1}(1 - \beta'_{j,k})$, and $\zeta = \rho^2/\lambda_\vee$ where $\lambda_\vee = \max \text{eig}(\mathbf{R}')$. $q_\wedge = \min_{q_i > 0} q_i$ denotes the minimum solid angle that is strictly positive.

Similarly, for Alg.-Value, we have,

Theorem 2. *Let $\mathbf{R}'$ be γ-simplicial and the topic matrix $\boldsymbol{\beta}$ be separable. Let $\mathbf{d}_r \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$. Then, Alg.-Value outputs all the novel words of K topics consistently as total number of documents $M = HL \to \infty$ and $P \to \infty$. Furthermore, $\forall \delta > 0$, for*

$$M \ge \max\left\{ c_1 \frac{\log(3W/\delta)}{\zeta^2 \phi^2 \eta^4},\ c_2 \frac{W^2 \log(2W/q_\wedge) \log(3W/\delta)}{\rho^2 q_\wedge^2 \phi^2 \eta^4} \right\}$$

and for

$$P \ge c_3 \frac{\log(3W/\delta)}{q_\wedge^2}$$

Alg.-Value fails with probability at most δ, where c_1 to c_3 are some absolute constants and other terms ζ, ρ, η, ϕ, $q_\wedge$ are defined in Theorem 1.

Detailed proofs can be found in the supplementary. As Theorem 1 and 2 shows, our approach requires only the simplicial assumption, which is a weaker assumption than that required for provable approaches that are also practical such as those of Arora et al. (2013); Ding et al. (2013).

5 Experimental Results

5.1 Dataset and Measures

Following Arora et al. (2013), we generate *semi-synthetic* corpora to ensure that the synthetic documents resemble the dimensionality and sparsity of the real word corpus. To this end, given a real-world dataset, we first train an LDA model using Gibbs Sampling (Griffiths and Steyvers, 2004) and obtain a topic matrix $\boldsymbol{\beta}_0$. We then generate a semi-synthetic corpus using $\boldsymbol{\beta}_0$ and Dirichlet priors. We refer to it as **Synthetic**. Note that $\boldsymbol{\beta}_0$ is not guaranteed to be separable. To ensure separability, we add **one** synthetic novel words to each topic. We assign the probability of novel word to be equal to the most probable word in the topic. We then renormalize columns of the new **separable** $(W + K) \times K$ topic matrix $\boldsymbol{\beta}_{sep}$ so that it is column stochastic. The dataset generated by $\boldsymbol{\beta}_{sep}$ is referred to as **Synthetic+novel**.

For the real-world dataset we use New York Times (NYT) articles data set (Frank and Asuncion, 2010). Following Arora et al. (2013), we prune the vocabulary based on document frequencies and then remove a standard stop-word list. After pruning, we get

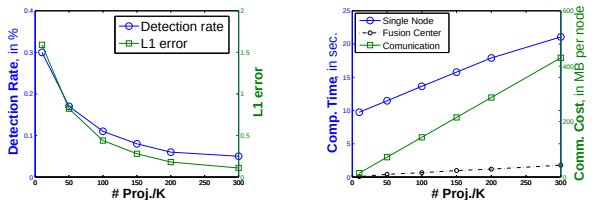

Figure 3: Performance of Alg.-Value as function of P when $L = 300$, $M = 300,000$. Left: ℓ_1 and Detection error; Right: Computation and communication cost.

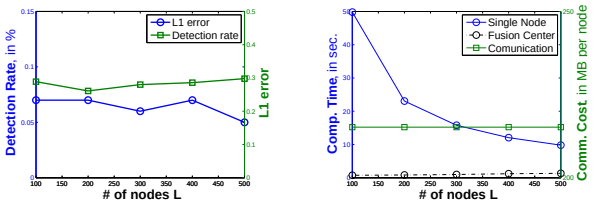

Figure 4: Performance of Alg.-Value as function of L when $P = 150 \times K$, $M = 300,000$. Left: ℓ_1 and Detection error; Right: Computation and communication cost.

$M = 300,000$; $W = 14,943$; average document length $N = 298$. We train a LDA model with $K = 100$ on it using Gibbs Sampling [2]. We then generate synthetic dataset for various M by fixing $N = 300$ and using a Dirichlet prior with symmetric hyper-parameters 0.03.

We simulate the Fusion Center on an Intel Core$^{\text{TM}}$i7-3820M CPU and 16GB RAM, and simulate distributed servers on Intel Core$^{\text{TM}}$i5-3210M CPU and 8GB RAM. We allows only 1 thread/server. For simplicity, we report total number of floats/server to be transmitted as the communication cost.

For semi-synthetic dataset, we compute the ℓ_1 error between the ground truth topics and the estimates. We use bipartite matching based on ℓ_1 distance to match two set of topics. For the synthetic+novel dataset, we further measure the Detection Error, defined as the percentage of topics whose novel words are not detected by the algorithms.

5.2 Properties of Alg.-Index and Alg.-Value

The time and statistical performance of Alg.-Index and Alg.-Value depend on the number of nodes L and the number of projections P. In this section, we show these properties using the synthetic+novel NYT dataset ($W = 15043, K = 100$). This will guide the settings for rest of the experiments. For results in this section, a MATLAB implementation is used to show the relative time cost. All the results are averaged across 5 random runs.

[2] Code available at http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

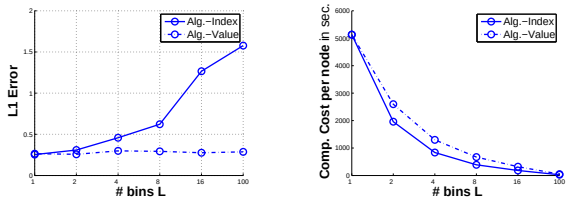

Figure 5: Performance of Alg.-Index vs. Alg.-Value as function of L when $P = 150 \times K$, $M = 300,000$. Left: ℓ_1 Error, (the x-axis is not uniform); Right: Computing time.

Figure 3 depicts the performance as function of P when L is fixed. As P increases, the time cost grows linearly while L1 error and miss detection rate decreases. These suggest that some moderate number of projections suffices to achieve good statistical performance. Fig. 4 on the other hand demonstrates the case when P is fixed. The computation time of each server scales inversely proportional to L, while the accuracy remains the same level. This shows Alg.-Value benefits for the fully distributed case without loss of accuracy. Based on the above observations we choose a large L for Alg.-Value and moderate P.

Next, we compare Alg.-Index vs Alg.-Value. To be fair, for Alg.-Index, we perform P/L projections/server. As the results in Fig. 5 shows, the ℓ_1 error of Alg.-Index is only comparable to Alg.-Value for very small $L \leq 4$, with only marginal improvement in time cost. This shows that the number of documents/server is crucial to the success of Alg.-Index. Therefore, in practical cases, unless L is really small or there exist hard constraints on communication, Alg.-Value is a better choice.

5.3 Semi-synthetic Dataset

In this section, we show performance of our algorithms on synthetic+novel NYT and synthetic NYT for varying M. We fix the number of nodes $L = 200$ and $P = 150 \times K$. We compare against centralized algorithms RecoverL2 (Arora et al., 2013)[3] and DDP (Ding et al., 2013). A Python implementation for our algorithm is used for this section.

As summarized in Fig. 6, our *distributed* Alg.-Value approach can achieve similar estimation accuracy as the *centralized* RecoverL2 on both Synthetic+Novel and Synthetic. This shows that our algorithm does not lose statistical efficiency due to the distributed setup. Moreover, the computation cost vs. error plot (Fig. 7) fully depicts the merits of our approach, i.e., Alg.-Value can achieve the same level of statistical accuracy with lower the computation cost than RecoverL2 or DDP.

[3] http://www.cs.nyu.edu/~halpern/files/anchor-word-recovery.zip

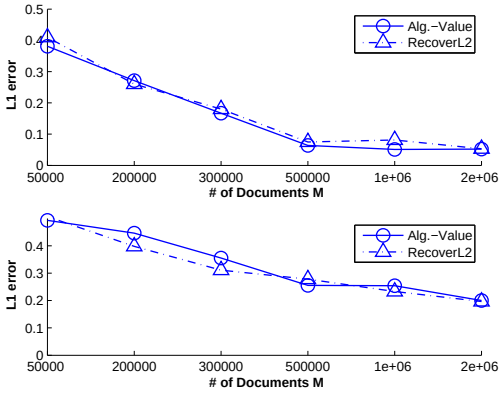

Figure 6: ℓ_1 error of estimated topic matrix for various M on Upper: Synthetic+Novel NYT; Lower: Synthetic NYT. $P = 150K$. The *distributed* Alg.-Value have similar accuracy as the *centralized* RecoverL2.

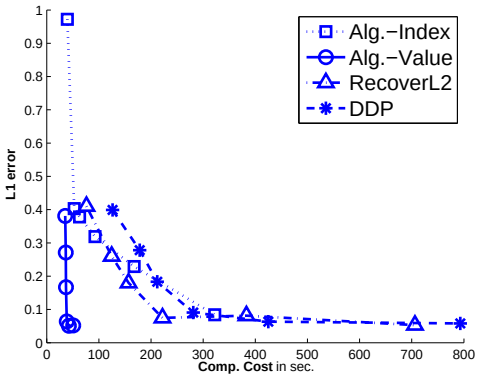

Figure 7: Computation cost vs. L1 error on Synthetic+novel NYT. $P = 150 \times K$ and $M = 50k, 200k, 300k, 500k, 1m, 2m$. $L = 200$ parallel threads are simulated for *centralized* RecoverL2 and DDP in Regression. $L = 200$ for Alg.-Value. $L = 4$ for Alg.-Index. Alg.-Value achieves the best accuracy-time tradeoff.

The C-implementation we used for Gibbs Sampling requires 6918sec. in estimating topics using 100 iterations for a $M = 300,000$ corpus. This is much longer than the the time reported in Fig. 7.

5.4 Real world text corpus

We apply Alg.-Value on the real world NYT dataset ($W = 14,943; M = 300$k). We adopt the held-out probability as the performance metric as is now the accepted standard. Following the same setting as in Arora et al. (2013), we randomly select 60k documents for testing (240k for training). Again, $P = 150 \times K$ and $L = 200$ are used.

We report the held-out (log) probability normalized by the total number of words in test docs. (averaged over 5 runs) in Table 1. We compare against RecovelL2

(Arora et al., 2013) and the baseline Gibbs Sampling approach (Griffiths and Steyvers, 2004). Gibbs produces the best description power. Alg.-Value and RecoverL2 has somewhat worse performance than Gibbs. This could be attributed to the missing novel words that appear only in the test set, which is crucial to the success of RecovelL2 and Alg.-Value.

Table 1: Held-out log probability on NYT dataset for various K.

K	RecoverL2	Gibbs	Alg.-Value
50	-8.22	-7.42	-8.54
100	-7.63	-7.50	-7.35
150	-8.03	-7.31	-7.94

Table 2: Examples of estimated topics on NYT by Alg.-Value

"weather"	weather wind air storm rain cold
"feeling"	feeling sense love character heart emotion
"election"	election zzz_florida ballot vote zzz_al_gore recount
"game"	yard game team season play zzz_nfl

We show example topics extracted by our Alg.-Value on the entire NYT dataset in Table 2. For each topic, its most frequent words are listed. As we can see, the estimated topics do form recognizable themes.

5.5 The Swimmer Image Dataset

Following Ding et al. (2013) , we apply our algorithm on a synthetic *swimmer* image dataset introduced by Donoho and Stodden (2004). There are $M = 256$ binary images, each of $W = 32 \times 32 = 1024$ pixels. Each image represents a swimmer composed of four limbs, each of which can be in one of 4 distinct positions, along with an invariant torso. By interpreting pixel positions as words, each image is viewed as a document composed of non-zero valued pixel positions. Since each position of a single limb features unique pixels in the image, the topic matrix, $\boldsymbol{\beta}$, satisfies the separability assumption with $K = 16$ "ground truth" topics that correspond to 16 single limb positions. In this dataset, the second order moments $\mathbf{R}'$ is rank-deficient (Ding et al., 2013).

Following Ding et al. (2013), body pixel in original binary image is set to 10 and background pixel values 1. We then choose a "clean" images, suitably normalized, as an underlying distribution across pixels and generate a "noisy" document with $N = 200$ "words" sampled from it. Examples are shown in Fig. 8. We prune the vocabulary according to document frequency.

We compare Alg.-Value against centralized RecoverL2-Cen (Arora et al., 2013) and DDP-Cen (Ding et al., 2013). Since $M = 256$ is rather small in the original dataset, we sampled $M = 8192$ images from the original with replacement and simulated $L = 32$ servers. Each server contains 256 images, the same as the original data. In addition, we consider a naive distributed version of RecoverL2 and DDP (denoted by RecoverL2-Dis and DDP-Dis resp.), where each server runs the algorithm independently with its local documents and then averaged the estimated topics across nodes. The topics are aligned by means of ℓ_1 metric.

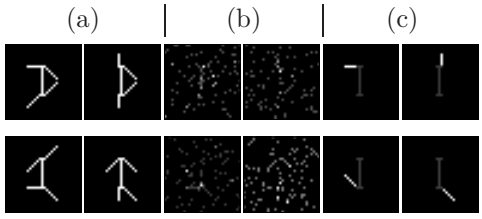

Figure 8: (a) Example "clean" images in Swimmer dataset; (b) Corresponding images with sampling "noise"; (c) Examples of ideal topics.

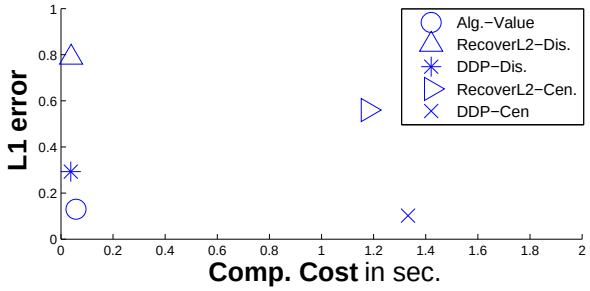

Figure 9: Computation cost vs. L1 error on swimmer dataset. $P = 150 \times K, M = 8192, L = 32$. Alg.-Value achieves the best accuracy-time tradeoff, while handles the rank deficient $\mathbf{R}'$.

The computing time vs ℓ_1 error Fig. 9 reveals the same story as Fig. 7. Again, we can see that our distributed algorithm can achieve minimum error rate with much shorter running time. We note that the DDP-Dis results in higher error rate than reported in (Ding et al., 2013) (where $M = 256$). One reason is that when we sample $M = 8192$ images with replacement from the original 256 images, the $L = 256$ images on each server do not necessarily contains all the $K = 16$ topics. DDP-Cen has an accuracy which is comparable to that of Alg.-Value, but this comes at the price of a much higher computational cost. RecoverL2 perform consistently bad since $\text{rank}(\mathbf{R}') = 13 < K = 16$, which forces Gram-Schmidt process to stop at a early stage.

References

A. Anandkumar, D. Foster, D. Hsu, S. Kakade, and Y. K. Liu. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25*, pages 926–934, Lake Tahoe, NV, Dec. 2012.

A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade. Learning linear bayesian networks with latent variables. In *the 30th Int. Conf. on Machine Learning*, Atlanta, GA, Jun. 2013.

S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond SVD. In *53rd IEEE Annu. Symp. Foundations of Computer Science*, pages 1–10, New Brunswick, NJ, Oct. 2012.

S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *the 30th Int. Conf. on Machine Learning*, Atlanta, GA, Jun. 2013.

A. Asuncion, P. Smyth, and M. Welling. Asynchronous distributed learning of topic models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 81–88, 2009.

D. Blei. Probabilistic topic models. *Commun. of the ACM*, 55(4):77–84, 2012.

D. Blei and J. Lafferty. A correlated topic model of science. *The Ann. of Applied Statistics*, 1(1):17–35, 2007.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama. Topic Discovery through Data Dependent and Random Projection. In *the 30th Int. Conf. on Machine Learning*, Atlanta, GA, Jun. 2013.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems 16*, pages 1141–1148, Cambridge, MA, 2004. MIT press.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

R. Gemulla, E. Nijkamp, P. J. Haas, and y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD*, pages 69–77, 2011.

T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, Apr. 2004.

A. Kumar, V. Sindhwani, and P. Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *the 30th Int. Conf. on Machine Learning*, Atlanta, GA, Jun. 2013.

C. Liu, H. Yang, J. Fan, L. He, and Yi-Min Y. Wang. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce. In *Proceedings of the 19th international conference on World wide web*, pages 681–690, 2010.

D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *JMLR*, 10: 1801–1828, December 2009.

B. Recht, C. Re, J. Tropp, and V. Bittorf. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems 25*, pages 1223–1231, Lake Tahoe, NV, Dec. 2012.

S. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM J. on Optimization*, 20(3): 1364–1377, Oct. 2009.