Efficient Distributed Topic Modeling with Provable Guarantees

Weicong Ding  Mohammad H. Rohban  Prakash Ishwar  Venkatesh Saligrama
Boston University  Boston University  Boston University  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 trade-offs between communication and computation without actually transmitting the documents. Our scheme is based on exploiting the geometry of normalized word-word co-occurrence 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.

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 \( \beta \) 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 \( \langle \beta \rangle \) from the empirical word-frequencies of all documents \( (X) \) 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 cen-
trized approach (i.e. even ignoring communication
costs) requires higher computation cost \(- O(MN^2/\epsilon^2+
WK/\epsilon^2)\) for Arora et al. (2013) and \(O(MN^2 + W^2)\)
for Ding et al. (2013). Next for a lower computational
cost namely, \(O(HN^2/\epsilon^2 + WK/\epsilon^2)\) for Arora et al.
(2013) or \(O(HN^2 + W^2)\) for Ding et al. (2013) per
server which are comparable to our scheme, their sta-
tistical 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 con-
 vex 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 identi-
cally 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 count-
ing 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 at-
tempt 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 pro-
posed 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, \(\beta\), and simplicial assumption
on the normalized second order moments of topic prior
\(^1\). Nevertheless, it is unclear how to directly paral-
lelize 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 mo-
ments 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 \(O(W^2)\) irres-
tive of number of documents. For our distributed con-
text this step has either a computational cost/server
or a communication cost/server that scales as \(O(W^2)\)
irrespective of the other model parameters.

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

2 Topic Geometry and Solid Angle

Let \(\beta\) (of size \(W \times K\)) and \(X (W \times M)\) denote the topic
matrix and the empirical word-by-document matrix.
We obtain \(\tilde{X}\) and \(\tilde{X}'\) 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 \(X^{(l)}\) be empirical words counts
of the documents in the \(l\)-th node, which is a slice of \(X\).
A fusion center controls the whole process and outputs
the estimated topics. We assume that the topic matrix
\(\beta\) is separable and let \(C_k\) be the set of novel words
of topic \(k\) and \(C_0\) be the set of non-novel words. \(A_i\)
denotes the \(i\)-th row vector of a matrix \(A\).

\(^1\)A topic matrix \(\beta \in \mathbb{R}^{W \times K}\) is separable
if for each topic \(k\), there is some word \(i\), called novel word, such that
\(\beta_{i,k} > 0\) and \(\beta_{i,l} = 0, \forall l \neq k\). Let \(a\) and \(R\) be the
mean and correlation matrix of the topic prior. \(R' =
\text{diag}(a)^{-1}R\text{diag}(a)^{-1}\). A topic model is \(\gamma\)-simplicial
if every row vector of \(R'\) is at least \(\gamma > 0\) distant from
the convex hull of other rows of \(R'\). To be precise, Arora et al.
(2013) and Ding et al. (2013) each requires stronger as-
sumption 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 $a$ and $K \times K$ matrix $R$ respectively. We define the normalized second order moment as $R' \overset{\text{def}}{=} \text{diag}^{-1}(a)R\text{diag}^{-1}(a)$ and assume $R'$ to be $\gamma$-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 $E \overset{\text{def}}{=} \beta'R'\beta'^\top$ (of size $W \times W$) where $\beta' = \text{diag}^{-1}(\beta a)\text{diag}(a)$ :

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

**Proof.** It is straightforward to check that $Y = R'\beta'^\top$ is simplicial. Note that $\beta'$ is row stochastic. For $i \in C_k$, $\beta'_{ik} = 1$, so the $i$-th row $E_i = Y_k$. For $i \in C_0$, $E_i$ is convex combination of at least two rows of $Y$. \hfill \Box

Fig. 1 illustrates the geometric property formalized in Lemma 1, i.e., the novel words correspond to extreme points of all rows of $E$. 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 non-zero iff it is an extreme point. Formally,

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

$$q_i \overset{\text{def}}{=} \Pr(\forall j, E_i \neq E_j : \langle E_i, d \rangle > \langle E_j, d \rangle)$$

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

**Lemma 2.** Suppose that $R'$ is $\gamma$-simplicial and $\beta$ 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 $q$’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,

$$q_i = E\left(\mathbb{I}\{\forall j, \|E_i - E_j\| \geq \zeta : E_id \geq E_jd\}\right)$$

where $\zeta$ 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 $E$ along a few iid isotropic directions $d$’s and then (ii) calculating the frequency that word $i$ maximizes the projection values. On the other hand, it has been shown that $MXX'^\top \overset{\text{d}}{\rightarrow} E$ as the number of document $M \rightarrow \infty$ 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 $Ed$ and gathering the indices of the maximizers for a few iid $d$’s. A simple implementation of this idea is to have each server first generate one independent random direction, then calculate $X^{(i)}X^{(i)\top}d$, 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 $Ed$ and gathering the indices of the maximizers for as many documents as possible. We observe that $Ed \approx MX'X'^\top d = M\sum_{l=1}^L X^{(l)}X^{(l)\top}d$, 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 ($O(WP)$) 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
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. $\bar{X}^{(l)}$, $\hat{X}^{(l)}$ on each node; Number of topics $K$; Tolerances $\zeta, \epsilon > 0$.

**Output:** Topic matrix $\hat{E}$.

(Node l) $M_l \leftarrow \text{NovelWordDetect-Nodes}(\bar{X}^{(l)}, \hat{X}^{(l)}, \zeta)$.  
(Center) Set of Novel Words $J \leftarrow \text{NovelWordDetect-Center}(M_l, K, \zeta)$.  
$\hat{E} \leftarrow \text{EstimateTopics}(J, X, \epsilon)$.

### 3.1 Index passing scheme (Alg.-Index)

**Algorithm 2 (Index Passing) NovelWordDetect-Nodes**

**Input:** $\bar{X}^{(l)}$, $\hat{X}^{(l)}$, $\zeta$  
**Output:** $I_l$, indexes of words with max. projs.

$E^{(l)} \leftarrow \emptyset$  
$E^{(l)} \leftarrow H^{(l)} \hat{X}^{(l)}$  
$d^{(l)} \leftarrow$ a sample from an isotropic distribution  
$i^* \leftarrow \arg \max_{1 \leq i \leq W} \langle E^{(l)}_i, d^{(l)} \rangle$  
$I_l \leftarrow I_l \cup \{i^*\}$  
$J_l \leftarrow \{j : E^{(l)}_{i^*, j} + \hat{E}^{(l)}_{i^*, j} - 2\hat{E}^{(l)}_{i^*, j} \geq \zeta/2\}$

for all $k \in J_l^c$:

$J_k \leftarrow \{j : E^{(l)}_{i^*, k} + \hat{E}^{(l)}_{i^*, k} - 2\hat{E}^{(l)}_{i^*, k} \geq \zeta/2\}$

if $\forall j \in J_k : \langle E^{(l)}_j, d^{(l)} \rangle > \langle E^{(l)}_j, d^{(l)} \rangle$  
$I_l \leftarrow I_l \cup \{k\}$

end if

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

$$q_i = \frac{1}{L} \sum_{l=1}^{L} \Pi(l) = 1 + \sum_{l=1}^{L} \Pi(l)$$

where $W \times W$ matrices $\bar{E}^{(l)}_i = H^{(l)} \hat{X}^{(l)}$ and $d^{(l)}$'s are iid directions. Each indicator tests if word $i$ ($\bar{E}^{(l)}_i$) has the maximum projection along $d^{(l)}$. 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 $J_i = \{j : \bar{E}^{(l)}_{i,j} + \hat{E}^{(l)}_{j,j} - 2\hat{E}^{(l)}_{i,j} \geq \zeta/2\}$ which converge to $J_i = \{j : \bar{E}_{i,j} \neq E_j\}$ 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 $d^{(l)}$, most of the indicators in Eq. (3) would be zero. In fact, given $d^{(l)}$, we can first find the maximizer index $i^*$ and corresponding set $J_{i^*}$. Note that $J_{i^*}$ consists of novel words for topics different from $i^*$ and the non-novel words. Hence, all the words in $J_{i^*}$ should make the indicator functions zero for the given $l$. As a consequence, we only evaluate the indicators for words in $J_{i^*} = \{1, \ldots, W\} \setminus J_{i^*}$, which have up to $|C_k|$ elements asymptotically, where $k$ is the topic associated with word $i^*$. Moreover, in a typical dataset, max$_k |C_k|$ is of the order $O(1)$.

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

**Algorithm 3 (Index Passing) NovelWordDetect-Center**

**Input:** Message $I_1, \ldots, I_L$ sent from nodes; $\hat{X}^{(L-1)}, \hat{X}^{(L-1)}$, $\zeta$, $H$, $L$, $K$  
**Output:** $J$ : Indices of $K$ distinct novel words $\hat{q}_1, \ldots, \hat{q}_W \leftarrow 0$.  
for all $1 \leq i \leq L$ and $j \in I_i$  
$\hat{q}_j \leftarrow \hat{q}_j + 1/L$  
$\hat{E}^{(L-1)} \leftarrow H^{(L-1)} \hat{X}^{(L-1)}$  
$J \leftarrow \text{FindNovelWords}(\hat{E}^{(L-1)}, \{\hat{q}_1, \ldots, \hat{q}_W\}, K, \zeta)$

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: $\hat{E}, \{\hat{q}_1, \ldots, \hat{q}_W\}, K, \zeta$
Output: $J$ : Indices of $K$ distinct novel words
$J \leftarrow \emptyset$, $k \leftarrow 0$, $j \leftarrow 1$
while $k < K$ do
    $i \leftarrow$ index of the $j$-th largest value of $\{\hat{q}_1, \ldots, \hat{q}_W\}$
    if $\forall p \in J : \hat{E}_{p,p} + \hat{E}_{i,i} - 2\hat{E}_{i,p} \geq \zeta/2$ then
        $J \leftarrow J \cup \{i\}$, $k \leftarrow k + 1$
    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 $O(HN+W)$ per server. The total communication cost is $O(\log(W))$ 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 $d_r, r = 1, \ldots, P$ as follows:

$$\hat{q}_i = \frac{1}{P} \sum_{r=1}^{P} I(\forall j, \hat{E}_{i,i} + \hat{E}_{j,j} - 2\hat{E}_{i,j} \geq \zeta/2) : \hat{E}_{i,d_r} > \hat{E}_{j,d_r}$$

(4)

Since

$$\hat{E}_{d_r} = M\hat{X}^T \hat{X} d_r = M \sum_{l=1}^{L} \hat{X}^{(l)} \hat{X}^{(l)T} d_r$$

(5)

the partial projection values $\hat{X}^{(l)} \hat{X}^{(l)T} d_r$ of size $W \times 1$ can be calculated locally and transmitted to the fusion center. To synchronize $d_r$’s across servers and to transmit the partial projection values, each server need to communicates $O(WP)$ floating-point numbers.

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

**Proposition 2.** The running time of Alg.-Value is $O(HNP+W)$ per server and $O(WPL+K^2)$ for the fusion center. The communication cost is $O(WP)$ 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: $\hat{X}^{(l)}, \hat{X}^{(l)}, d^{(l)}$, Directions: $d^{(1)}, \ldots, d^{(P)}$
Output: $V^{(l)}$: a $W \times P$ matrix containing the projection values of $\hat{E}^{(l)}$ rows along the given $P$ directions

$$V^{(l)} \leftarrow M\hat{X}^{(l)} \hat{X}^{(l)T} [d^{(1)}, \ldots, d^{(P)}]$$

Algorithm 6 (Value Passing) NovelWordDetect-center
Input: Message $V^{(1)}, \ldots, V^{(l)}$ sent from nodes; $\zeta, K$
Output: $J$ : Indices of $K$ distinct novel words
Generate $P$ iid directions $d^{(1)}, \ldots, d^{(P)}$; Send to each node. Call NovelWordDetect-Node.
$$V \leftarrow \sum_{l=1}^{L} V^{(l)} \text{ and } \forall i : \hat{q}_i \leftarrow 0$$
for all $1 \leq r \leq P$
    $$i^* \leftarrow \arg \max_{1 \leq j \leq W} V_{i,r}$$
    $$\hat{q}_{i^*} \leftarrow \hat{q}_{i^*} + 1/P$$
    $$J \leftarrow \{j : \hat{E}_{i^*,j} + \hat{E}_{j,j} - 2\hat{E}_{i^*,j} \geq \zeta/2\}$$
for all $k \in J$
    $$\hat{J}_k \leftarrow \{j : \hat{E}_{k,k} + \hat{E}_{j,j} - 2\hat{E}_{k,j} \geq \zeta/2\}$$
    $\forall j \in \hat{J}_k : V_{k,j} > V_{j,r}$
    $$\hat{q}_k \leftarrow \hat{q}_k + 1/P$$
end if
$J \leftarrow$ FindNovelWords($\hat{E}, \{\hat{q}_1, \ldots, \hat{q}_W\}, K, \zeta$)

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. $a \in \mathbb{R}^K$ and $R \in \mathbb{R}^{K \times K}$ are the expectation and correlation matrix of topic prior. Let $R' = \text{diag}^{-1}(a)R \text{diag}^{-1}(a)$. We present the results when $d$’s are spherical Gaussian. We also provide sample complexity bounds for other isotopic priors in the supplementary. For Alg.-Index, we have,

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

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

and for

$$L \geq c_3 \frac{\log(W/\delta)}{q_\lambda^2}$$
Alg.-Index fails with probability at most $\delta$, where $c_1$ to $c_3$ are some absolute constants, $\phi = \min_{i,j} \frac{a_{i,j}}{R_{i,j}}$, $\eta = \min_{i,j} \beta_{i,j}$, $\rho = \gamma \min_{i,j} a_{i,j} \neq 1 - \beta_{i,j}$, and $\zeta = \rho^2 / \lambda_V$ where $\lambda_V = \max \{ \text{eig}(R') \}$. $q_\eta = \min_{i,j} q_i > 0$, denotes the minimum solid angle that is strictly positive.

Similarly, for Alg.-Value, we have,

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

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

and for

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

Alg.-Value fails with probability at most $\delta$, where $c_1$ to $c_3$ are some absolute constants and other terms $\zeta$, $\rho$, $\eta$, $\phi$, $q_\eta$ 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 world 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 $\beta_0$. We then generate a semi-synthetic corpus using $\beta_0$ and Dirichlet priors. We refer to it as Synthetic. Note that $\beta_0$ is not guaranteed to be separable. To ensure separability, we add one synthetic novel word 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 $\beta_{sep}$ so that it is column stochastic. The dataset generated by $\beta_{sep}$ 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

$M = 300,000; W = 14,943$; 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(TM) i7-3820M CPU and 16GB RAM, and simulate distributed servers on Intel Core(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 $\ell_1$ error between the ground truth topics and the estimates. We use bipartite matching based on $\ell_1$ 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.

---

Figure 5: Performance of Alg.-Index vs. Alg.-Value as function of $L$ when $P = 150 \times K$, $M = 300,000$. Left: $\ell_1$ 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 $L_1$ 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 $\ell_1$ 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) 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.

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 time reported in Fig. 7.

5.4 Real world text corpus

We apply Alg.-Value on the real world NYT dataset ($W = 14,943; M = 300k$). 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 RecoverL2.

\footnote{http://www.cs.nyu.edu/~halpern/files/anchor-word-recovery.zip}
(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 RecoverL2 and Alg.-Value.

Table 1: Held-out Log probability on NYT dataset for various $K$

<table>
<thead>
<tr>
<th>$K$</th>
<th>RecoverL2</th>
<th>Gibbs</th>
<th>Alg.-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>-8.22</td>
<td>-7.42</td>
<td>-8.54</td>
</tr>
<tr>
<td>100</td>
<td>-7.63</td>
<td>-7.50</td>
<td>-7.35</td>
</tr>
<tr>
<td>150</td>
<td>-8.03</td>
<td>-7.31</td>
<td>-7.94</td>
</tr>
</tbody>
</table>

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, $\beta$, satisfies the separability assumption with $K = 16$ “ground truth” topics that correspond to 16 single limb positions. In this dataset, the second order moments $R'$ 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 $\ell_1$ metric.

![Figure 8](image)

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

![Figure 9](image)

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 $R'$.

The computing time vs $\ell_1$ 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 rank$(R') = 13 < K = 16$, which forces Gram-Schmidt process to stop at a early stage.
References


