Supplementary Material

We first list two of our key assumptions

Assumption 1. *(Separable condition) A topic matrix $\boldsymbol{\beta} \in \mathbb{R}^{W \times K}$ is* **separable** *if for each topic k, there is some word i such that $\boldsymbol{\beta}_{i,k} > 0$ and $\boldsymbol{\beta}_{i,l} = 0, \forall\, l \neq k$.*

Assumption 2. *(Simplicial condition) Let $\mathbf{a}$ and $\mathbf{R}$ denote the expectation and correlation matrix of the priors on columns of $\boldsymbol{\theta}$. Define the normalized second order moments as $\mathbf{R}' \triangleq \operatorname{diag}^{-1}(\mathbf{a})\mathbf{R}\operatorname{diag}^{-1}(\mathbf{a})$. $\mathbf{R}'$ is* γ**-simplicial** *if every row vector of $\mathbf{R}'$ is at an Euclidean distance of at least $\gamma > 0$ distant from the convex hull of the remaining rows of $\mathbf{R}'$.*

We also denote $\boldsymbol{\beta}' = \operatorname{diag}^{-1}(\boldsymbol{\beta}\mathbf{a})\boldsymbol{\beta}\operatorname{diag}(\mathbf{a})$. Denote $\mathbf{E} = \boldsymbol{\beta}'\mathbf{R}'\boldsymbol{\beta}'^{\top}$ and $\mathbf{E}_i$ as i-th row of it. For convenient, for a novel word $i \in \mathcal{C}_k$, define $\bar{\mathcal{N}}(i)^* = {\mathcal{C}_k}^c$. Similarly for a non-novel word $i \in \mathcal{C}_0$, define $\bar{\mathcal{N}}(i)^* = {\mathcal{C}_0}^c$. H, L, P denote the number of document in each node (computing unit/server), number of nodes, and number of projections respectively.

The following lemma depict the key topic geometry. It is **Lemma 1** in Sec. 2 of the main paper.

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 $\boldsymbol{\beta}_i'\mathbf{R}'\boldsymbol{\beta}'^{\top} = \mathbf{E}_i$, i.e., the i-th row of $\mathbf{E}$, is an extreme point of the convex hull spanned by the rows of $\mathbf{E}$.*

Proof. By the following Prop. 1, $\mathbf{Y} = \mathbf{R}'\boldsymbol{\beta}'^{\top}$ is γ-simplicial. Therefore by definition, no row of $\mathbf{Y}$ is in the convex hull of the remaining rows.

For a novel word $i \in \mathcal{C}_k$,$\boldsymbol{\beta}'_{i,k} = 1$ and $\boldsymbol{\beta}'_{i,l} = 0, l \neq k$, therefore, $\mathbf{E}_i = \mathbf{Y}_k$. For an non-novel word $j \in \mathcal{C}_0$, the row vector $\boldsymbol{\beta}'_j$ has at least two non-zero components and $\mathbf{E}_i = \sum\limits_{k=1}^{K} \beta'_{i,k}\mathbf{Y}_k$. Therefore, it is a convex combination at least two rows of $\mathbf{Y}$, i.e., the rows $\mathbf{E}_i$ that correspond to novel words $i \in \mathcal{C}_k$. □

Proposition 1. *Let $\mathbf{R}'$ be γ-simplicial and $\boldsymbol{\beta}$ be separable. Then matrix $\mathbf{R}'\boldsymbol{\beta}'^{\top}$ is at least γ-simplicial.*

Proof. Since $\mathbf{R}'$ is γ-simplicial, $\forall c_j \geq 0,\ \sum_{j\neq i} c_j = 1$, $\|\mathbf{R}'_i - \sum_{j\neq i} c_j\mathbf{R}'_j\| \geq \gamma$. Note that $\boldsymbol{\beta}$ is separable, by reordering the words, $\boldsymbol{\beta}^{\top} = \left[\mathbf{I}_{K\times K},\ \mathbf{J}_{K\times(W-K)}\right]$ and we obtain $\mathbf{R}'\boldsymbol{\beta}'^{\top} = [\mathbf{R}',\ \mathbf{R}'\mathbf{J}]$. Now we can check the simplicity of $\mathbf{R}'\boldsymbol{\beta}$ by definition. $\forall c_j \geq 0,\ \sum_{j\neq i} c_j = 1$, we can write $\|(\mathbf{R}'\boldsymbol{\beta}'^{\top})_i - \sum_{j\neq i} c_j(\mathbf{R}'\boldsymbol{\beta}'^{\top})_j\| = \|\mathbf{R}'_i - \sum_{j\neq i} c_j\mathbf{R}'_j\| + \|(\mathbf{R}'\mathbf{J})_i - \sum_{j\neq i} c_j(\mathbf{R}'\mathbf{J})_j\| \geq \gamma$. So it is at least γ simplicial. □

A Computational Complexity of the Index Passing scheme Alg.-Index

Proposition 1 in Sec. 3 of the main paper (Time complexity of the Alg.-Index) *The running time of Alg.-Index is $\mathcal{O}(HN+W)$ for each node and $\mathcal{O}(WL+K^2)$ for the fusion center. Meanwhile, the total communication cost of Alg.-Index is $\mathcal{O}(\log(W))$ per node with an additional cost of $\mathcal{O}(LK^2)$ if no document is archived on the fusion center.*

The steps to analysis is very similar to the next section and are omitted.

Remark: We use the total number of floats to be transmitted to measure the communication cost throughout the paper and the supplementary. This would ignore the network structure of how the nodes are connected and implicitly assumes all the message can be transmitted between each node and the fusion center within constant hoops. In fact, we can assume that each node is connected to the fusion center within $\mathcal{O}(\log(L))$ hoops which most network structure satisfies. And such log factor can be viewed as a constant in the Big O notation, compared to other terms.

B Computational Complexity of Projection value Passing scheme Alg.-Value

Proposition 2 in Sec. 3 of the main paper (Time complexity of the Alg.-Value) *The running time of Alg.-Value is $\mathcal{O}(NHP)$ for each distributed node and $\mathcal{O}(WP+WL+K^2)$ for the fusion center. Meanwhile, the total communication cost per node is $\mathcal{O}(WP)$.*

We decompose this algorithm into steps for analysis. We assume $N < W$ and $W \gg K$ in most cases. The number of novel words is at most $\mathcal{O}(K)$, i.e., each topic has constant number of novel words. When one make use of sparsity in matrix computation via say Hashing tricks, there is an additional $\log(W)$ factor in computation complexity, which we would take as constant compared to the other polynomial factors in the complexity analysis.

The matrix $\mathbf{C}$ here represents $\hat{\mathbf{E}}$ in the main text, i.e., the estimated normalized second order word co-occurrence matrix.

Step 1: Normalization. We need to normalized the word-by-document matrices as $\widetilde{\mathbf{X}}_{w,d} = \frac{\mathbf{X}_{w,d}}{\sum_d \mathbf{X}_{w,d}}, \widetilde{\mathbf{X}}'_{w,d} = \frac{\mathbf{X}'_{w,d}}{\sum_d \mathbf{X}_{w,d}}$. It can be decomposed since $N_w := \sum_d \mathbf{X}_{w,d} = \sum_{l=1}^{L} \sum_{d\in \text{bin } l} \mathbf{X}_{w,d} = \sum_{l=1}^{L} N_{w,l}$ and so is for $\widetilde{\mathbf{X}}'$.

- Center Action : For $w = 1, \dots, W$, aggregate $N_{w,l}$ from each node. Calculate N_w. Broadcast the total words counts N_w's to each node.
- Node Action : For $w = 1, \dots, W$, aggregate partial total words count $N_{w,l}$. Then get N_w from the Fusion Center, normalize rows of $\mathbf{X}_{(l)}$ and $\mathbf{X}'_{(l)}$.
- Comm. Cost per node : $\mathcal{O}(W)$.
- Comp. Cost, Fusion Center : $\mathcal{O}(WL)$ as required to calculate the summations.
- Comp. Cost, Single Node : $\mathcal{O}(HN)$. This is achievable since only there are at most $N \ll W$ non-zero element in each column of $\mathbf{X}$. We can calculate the sum by exploiting such sparsity. So in sum, only HN non-zero elements are stored in each node hence to get the summation which is a $W \times 1$ vector, at most HN summation is needed.

Step 2: Random Projections. We compute projection values $\mathbf{v}_r = \widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}^\top \mathbf{d}_r$ for projections $r = 1, \dots, P$. It can be decomposed as $\mathbf{v}_r = \sum_{l=1}^{L} \widetilde{\mathbf{X}}'_{(l)}\widetilde{\mathbf{X}}_{(l)}^\top \mathbf{d}_r = \sum_l \mathbf{v}_{r,l}$.

- Center Action: Aggregate partial projection values $\mathbf{v}_{r,l}$ from each node for every projections. Obtain $\mathbf{v}_r$, and then find the maximums and minimums for each projection.
- Node Action: Calculate $\widetilde{\mathbf{X}}'_{(l)}\widetilde{\mathbf{X}}_{(l)}^\top \mathbf{d}_r$ for each projections $r = 1, \dots, P$, and then transmit the information to the center node.
- Comm. Cost per node : $\mathcal{O}(WP)$.
- Comp. Cost, Fusion Center : $\mathcal{O}(WPL)$ adding, $\mathcal{O}(WP)$ finding max/min.
- Comp. Cost, Single Node : $\mathcal{O}(HNP)$. This computational complexity can be achieved by exploiting the sparsity, i.e., there are at most $N \ll W$ non-zero elements in each column of $\mathbf{X}_{(l)}$. Consider calculating $\mathbf{X}_{(l)}^\top \mathbf{y}$ where $\mathbf{y}$ is any $W \times 1$ vector. For each of H rows of $\mathbf{X}_{(l)}^\top$, at most N components are non-zero. So the inner product can be calculated within $\mathcal{O}(HN)$ time complexity. It is similar to calculate $\mathbf{X}'\mathbf{z}$ where $\mathbf{z}$ is a $H \times 1$ vector in the same amount of time. Therefore, the key component, i.e., the projection values $\widetilde{\mathbf{X}}'_{(l)}\widetilde{\mathbf{X}}_{(l)}^\top \mathbf{d}_r$ can be calculated efficiently in $\mathcal{O}(HN)$ for one random projection, hence $\mathcal{O}(HNP)$ for P number of projections.

Step 3: Find the nearest neighbors of the selected max/min We need to find all the d neighbors of max/min selected in step 2. Say they are word $(1), \dots, (\hat{P})$ and should be to the order of $\mathcal{O}(K)$. Then $\mathbf{C}_{(i),w}, i = 1, \dots, \hat{P}, w = 1, \dots, W$ has to be computed. Note that $\mathbf{C}_{(i),w} = \widetilde{\mathbf{X}}'_{(i)}\widetilde{\mathbf{X}}_w^\top$ hence the inner product can be decomposed as summation of partial inner product from each node.

- Center Action: Aggregate $\mathbf{C}_{(i),w}^{(l)}$ from each
- Node Action: Calculate $\mathbf{C}_{(i),w}^{(l)} = \widetilde{\mathbf{X}}'_{(l),(i)}\widetilde{\mathbf{X}}_{(l),w}^\top$ for $i = 1, \dots, \hat{P}$ and $w = 1, \dots, W$.
- Comm. Cost per node : $\mathcal{O}(WK)$.
- Comp. Cost, Fusion Center : $\mathcal{O}(WKL)$ adding, $\mathcal{O}(WK)$ finding neighbors.
- Comp. Cost, Single Node : $\mathcal{O}(HNK)$. To achieve this complexity, the same trick as in Step. 2 can be used to exploit the sparsity.

Step 4: Clustering. This requires work from only the Fusion Center, following the steps in Alg.-Index. All the statistics required in this step, namely, $\mathbf{C}_{i,j}$ where i, j are for the top $\mathbf{O}(K)$ words with high $\hat{q}_i$. These has been calculated and transmitted to the fusion center at the previous step.

- Center Action: Perform clustering as in Alg. Alg.-Value .
- Node Action: No.
- Comm. Cost per node : 0.
- Comp. Cost, Fusion Center : $\mathcal{O}(K^2)$.
- Comp. Cost, Single Node : 0

Step 5: Estimation. Estimate topics using L2 regression as in Section B.1. Note that the inner products has all been calculated and aggregated in Step 4 hence no further communications and additional calculation are required.

- Center Action: Solving W number of constraint regressions.
- Node Action: No.
- Comm. Cost per node : 0
- Comp. Cost, Fusion Center : $\mathcal{O}(WK^3)$.
- Comp. Cost, Single Node : 0.
- **Remark:** At some additional communication cost of $\mathcal{O}(K^2 + WK/L)$, the W regressions can be taken in a paralleled way. The computational cost for each single node would be reduced to $\mathcal{O}(\frac{W}{L}K^3)$. In the largest NYT dataset we considered in the main paper, by setting $L = 300$ and using CVX Matlab implementationGrant and Boyd [2013], it only takes 9.3s. Note that typically, we have $L > K$ hence the additional communication cost is really small.

So in sum, including the topic estimation step (parallelized), the total communication cost per node is $\mathcal{O}(WP + K^2 + WK/L)$. The computation cost for a single node is $\mathcal{O}(HNP)$. The computation cost for the fusion center is $\mathcal{O}(WPL + K^2)$. This conclude the computational efficiency properties of our proposed Alg.-Value .

B.1 Topic Estimation

Algorithm 1 EstimateTopics

```
Input: $\mathcal{I} = \{i_1, \ldots, i_K\}$, $\mathbf{X}$, $\mathbf{X}'$, precision $\epsilon$
Output: $\widehat{\boldsymbol{\beta}}$, which is the estimation of $\boldsymbol{\beta}$ matrix
  $\mathbf{Y} = (\widetilde{\mathbf{X}}_{j_1}^\top, \ldots, \widetilde{\mathbf{X}}_{j_K}^\top)^\top$,$\mathbf{Y}' = (\widetilde{\mathbf{X}}_{j_1}'^\top, \ldots, \widetilde{\mathbf{X}}_{j_K}'^\top)^\top$
  for all $1 \le i \le W$ do
    $\widehat{\boldsymbol{\beta}}_i \leftarrow \underset{b_j \ge 0, \sum_{j=1}^K b_j = 1}{\arg\min} M(\widetilde{\mathbf{X}}_i - \mathbf{bY})(\widetilde{\mathbf{X}}_i' - \mathbf{bY}')^\top$
    (with stopping precision $\epsilon$)
    $\widehat{\boldsymbol{\beta}}_i \leftarrow (\frac{1}{M}\mathbf{X}_i\mathbf{1})\widehat{\boldsymbol{\beta}}_i$
  end for
  column normalize $\widehat{\boldsymbol{\beta}}$
```

The topic estimation step inherits similar ideas from the previous work [Ding et al., 2013, Arora et al., 2013, Kumar et al., 2013]. It is summarized in Alg. 1. The objective functions $\mathbf{bYY}^\top\mathbf{b}^\top - 2\widetilde{\mathbf{X}}_i\mathbf{Y}^\top\mathbf{b}^\top$ shares the same part $\mathbf{YY}^\top$ so it doesn't have to be recalculated for all the regressions. Moreover, the $\hat{E}_{i,j}$ needed form $\mathbf{YY}^\top$ and $\widetilde{\mathbf{X}}_i\mathbf{Y}$ have already been computed in say Alg.-Value, in step 3 and 4. Therefore, these regressions can be computed in parallel on each node, with an additional $\mathcal{O}(K^2 + WK/L)$ float number to be communicated per node.

C Analysis of Alg.-Value

Say $\mathbf{C} = \widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}^\top$. Recall P number of directions $\mathbf{d}^r = \widetilde{\mathbf{X}}^\top\mathbf{u}^r$, $r = 1, \ldots, P$ are generated i.i.d. Two isotropic distributions are considered: $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$ or $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$. The consistency of Algorithm Alg.-Value depends on the convergences of following quantities :

$$\begin{aligned} q_i &= \Pr\{\forall j \in \bar{\mathcal{N}}(i)^*,\ \mathbf{E}_i\mathbf{u} \ge \mathbf{E}_j\mathbf{u}\} && (1) \\ p_i(\mathbf{E}) &= \Pr_{\mathbf{u}}\{\forall j \in \bar{\mathcal{N}}(i),\ \mathbf{E}_i\mathbf{u} \ge \mathbf{E}_j\mathbf{u}|\mathbf{C}\} && (2) \\ p_i(\mathbf{C}) &= \Pr_{\mathbf{u}}\{\forall j \in \bar{\mathcal{N}}(i),\ \mathbf{C}_i\mathbf{u} \ge \mathbf{C}_j\mathbf{u}|\mathbf{C}\} && (3) \\ \hat{p}_i &= \frac{1}{P}\sum_{r=1}^{P} \mathbb{I}\{\forall j \in \bar{\mathcal{N}}(i),\ \mathbf{C}_i\mathbf{u}^r \ge \mathbf{C}_j\mathbf{u}^r\} && (4) \end{aligned}$$

The proof consists of showing a sequence of convergence :

- Convergence of $\hat{p}_i$ to $p_i(\mathbf{C})$ by Hoeffeding lemma as P increase.
- Convergence of $p_i(\mathbf{C})$ to $p_i(\mathbf{E})$ since $\mathbf{C}$ converges to $\mathbf{E}$ as M increase.

To begin with, $\hat{p}_i$ is the winning frequency of word i as defined in Algorithm Alg.-Value . Based on the topic geometry Lemma 1, we have the following claim,

Lemma 2. *Suppose $\mathbf{R}'$ is γ-simplicial and topic matrix $\boldsymbol{\beta}$ is separable. Let $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$ or $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$. Then*

1. *For all novel words, exist $q_\wedge$, such that $\min_{i \in \mathcal{C}_k} q_i := q_\wedge > 0$.*
2. *For all non-novel words $i \in \mathcal{C}_0$, $q_i = 0$.*

$p_i(\mathbf{E})$ and $p_i(\mathbf{C})$ are function of $\mathbf{C}$. They has the following properties,

Lemma 3. *For $\bar{\mathcal{N}}(i)$ constructed in Algorithm Alg.-Value ,*

1. *For a novel word $i \in \mathcal{C}_k$, if $\bar{\mathcal{N}}(i) \subseteq \bar{\mathcal{N}}(i)^*$, then $p_i(\mathbf{E}) \ge q_i \ge q_\wedge$.*
2. *For a non-novel word $i \in \mathcal{C}_0$, if $\bar{\mathcal{N}}(i) \supseteq \bar{\mathcal{N}}(i)^*$, then $p_i(\mathbf{E}) \le q_i = 0$*

This is a direct result of the definitions of $p_i(\mathbf{E})$ and q_i. $\hat{p}_i$ is a random variable as function of $\mathbf{C}$ and $\mathbf{u}$'s. Moreover, $\mathbb{E}_{\mathbf{u}}(\hat{p}_i|\mathbf{C}) = p_i(\mathbf{C})$. Hence we have the following lemma on convergence of $\hat{p}_i$ to $p_i(\mathbf{C})$,

Lemma 4. *$\forall t > 0$,*

$$\Pr\{|\hat{p}_i - p_i(\mathbf{C})| \ge t\} \le 2\exp(-2Pt^2) \qquad (5)$$

Proof. Note that $\hat{p}_i^{(r)}$ are i.i.d 0-1 random variables conditioned on $\mathbf{C}$. By Heoffding's lemma, $\forall i$, $\forall$ $\mathbf{C}$, $\Pr_{\mathbf{u}}(|\hat{p}_i - p_i(\mathbf{C})| \ge t|\mathbf{C}) \le 2\exp(-2Pt^2)$. By marginalizing over $\mathbf{C}$, we conclude the lemma. □

For convenient, $\forall\epsilon > 0$, we define a good set of $\mathbf{C}$ as,

$$\begin{aligned} \mathcal{G}(\epsilon) = \{\ \mathbf{C} : &\forall i \in \mathcal{C}_k, 1 \le k \le K,\ \bar{\mathcal{N}}(i) \subseteq \bar{\mathcal{N}}(i)^*; \\ &\forall i \in \mathcal{C}_0,\ \bar{\mathcal{N}}(i) \supseteq \bar{\mathcal{N}}(i)^*; \\ &\forall 1 \le i, j \le W,\ |\mathbf{C}_{i,j} - \mathbf{E}_{i,j}| \le \epsilon\ \}. \end{aligned} \qquad (6)$$

This set has asymptotically overwhelming mass as $M, N \to \infty$. To be precise,

Lemma 5. *Consider $\mathcal{G}(\epsilon)$ defined in equation (6)*

$$\begin{aligned} \Pr(\mathcal{G}(\epsilon)^c) \le &c_1 W^2 \exp(-c_2 MN\epsilon^2\phi^2\eta^4) \\ &+ c_3 W^2 \exp(-MNc_4 d^2\phi^2\eta^4) \end{aligned}$$

where $\eta = \min_{1 \le i \le W} \boldsymbol{\beta}_i\mathbf{a}$, $\phi = \min_{i,j} \frac{a_i a_j}{R_{i,j}}$ and $d = \gamma^2 \min_{j \notin \mathcal{C}_k, k \ge 1}(1 - \beta_{j,k}')^2$. c_1 to c_4 are constants.

Proof. Following the result in Ding et al. [2013], we have $\Pr(|\mathbf{C}_{i,j} - \mathbf{E}_{i,j}| \geq \epsilon) \leq 10\exp(-MN\epsilon^2\phi^2\eta^4/32)$. Use the short notation $e_{i,j} = C_{i,i} - 2C_{i,j} + C_{i,i}$ and similarly $\Pr(|e_{i,j} - (\boldsymbol{\beta}'_i - \boldsymbol{\beta}'_j)\mathbf{R}'(\boldsymbol{\beta}'_i - \boldsymbol{\beta}'_j)^\top| \geq d) \leq c_1\exp(-c_2MNd^2\phi^2\eta^4)$.

Now suppose that $i \in \mathcal{C}_k, j \notin \mathcal{C}_k$. By proposition 1, $\|(\boldsymbol{\beta}'_i - \boldsymbol{\beta}'_j)\mathbf{R}'\|_2 \geq \gamma(1 - \beta'_{j,k})$. Therefore we have $(\boldsymbol{\beta}'_i - \boldsymbol{\beta}'_j)\mathbf{R}'(\boldsymbol{\beta}'_i - \boldsymbol{\beta}'_j)^\top \geq \gamma^2(1-\beta'_{j,k})^2/\lambda_\vee$, where $\lambda_\vee$ is the maximum eigenvalue of the positive semi-definite matrix $\mathbf{R}'$.

Set $d = \gamma^2\min_{j\notin\mathcal{C}_k,k\geq 1}(1 - \beta'_{j,k})^2/\lambda^2_\wedge$. Then by union bound, we obtain for any novel words i, $\Pr(\bar{\mathcal{N}}(i) \nsubseteq \bar{\mathcal{N}}(i)^*) \leq Wc_1\exp(-MNc_2d^2\phi^2\eta^4)$. Similarly, for any non-novel words i, $\Pr(\bar{\mathcal{N}}(i) \nsupseteq \bar{\mathcal{N}}(i)^*) < Wc_3\exp(-MNc_4d^2\phi^2\eta^4)$. Hence we obtain the conclusion. □

The following lemmas shows the convergence of $p_i(\mathbf{C})$ to $p_i(\mathbf{E})$ under the considered distributions. We start with an useful proposition,

Proposition 2. *Let $\mathbf{v}^n, \mathbf{v} \in \mathbb{R}^m$ be two random vectors, $\mathbf{x}, \boldsymbol{\epsilon} \in \mathbb{R}^m$ are two vectors and $\boldsymbol{\epsilon} > \mathbf{0}$, then*

$$\begin{aligned} &|\Pr\{\mathbf{v}^n \leq \mathbf{x}\} - \Pr\{\mathbf{v} \leq \mathbf{x}\}| \\ &\leq \Pr(\exists i : |v^n_i - v_i| \geq \epsilon_i) + \Pr(\mathbf{x} - \epsilon \leq \mathbf{v} \leq \mathbf{x} + \epsilon) \end{aligned} \tag{7}$$

The inequalities is element-wise.

Lemma 6. *Suppose $\mathbf{R}'$ is γ-simplicial and topic matrix $\boldsymbol{\beta}$ is separable. Let $\mathbf{u} \sim$ Uniform$(\mathcal{B}^W)$. Then, $\forall\ \epsilon_0 > 0$, $\exists\ \epsilon > 0$ such that $\forall\ \mathbf{C} \in \mathcal{G}(\epsilon)$, $|p_i(\mathbf{C}) - p_i(\mathbf{E})| \leq \epsilon_0$.*

Proof. Recall the definition of $p_i(\mathbf{C})$ and $p_i(\mathbf{E})$,

$$\begin{aligned} p_i(\mathbf{E}) &= \Pr_{\mathbf{u}}\{\forall j \in \bar{\mathcal{N}}(i),\ \mathbf{E}_i\mathbf{u} \geq \mathbf{E}_j\mathbf{u}|\mathbf{C}\} \\ p_i(\mathbf{C}) &= \Pr_{\mathbf{u}}\{\forall j \in \bar{\mathcal{N}}(i),\ \mathbf{C}_i\mathbf{u} \geq \mathbf{C}_j\mathbf{u}|\mathbf{C}\} \end{aligned}$$

Given any $\mathbf{C}$ and $\forall\epsilon > 0$, by proposition 2, we have,

$$\begin{aligned} |p_i(\mathbf{C}) - p_i(\mathbf{E})| \leq &\Pr_{\mathbf{u}}(\exists j \in \bar{\mathcal{N}}(i) : |\mathbf{e}_{i,j}\mathbf{u}| \geq 2\epsilon) \\ &+ \Pr_{\mathbf{u}}(\forall j \in \bar{\mathcal{N}}(i) : |\mathbf{z}_{ij}\mathbf{u}| \leq 2\epsilon) \end{aligned} \tag{8}$$

where $\mathbf{e}_{i,j} = \mathbf{E}_i - \mathbf{C}_i + \mathbf{C}_j - \mathbf{E}_j$ and $\mathbf{z}_{ij} = \mathbf{E}_i - \mathbf{E}_j$. Note that $\mathbf{u} \sim$ Uniform$(\mathcal{B}^W)$ and $|\mathbf{e}_{i,j}\mathbf{u}| \leq \|\mathbf{e}_{i,j}\|\|\mathbf{u}\|$. By Cauchy-Schwartz inequality, the first term in equation (8) is 0 if $\mathbf{C} \in \mathcal{G}(\epsilon)$.

For the second term in equation (8), consider for any novel words $i \in \mathcal{C}_k$, and $\mathbf{C} \in \mathcal{G}(\epsilon)$, we have $\bar{\mathcal{N}}(i) \subseteq \bar{\mathcal{N}}(i)^*$, so it can be union bounded by $\sum_{j\in\bar{\mathcal{N}}(i)^*}\Pr_{\mathbf{u}}(|\mathbf{z}_{ij}\mathbf{u}| \leq 2\epsilon)$. For a non-novel word $i \in \mathcal{C}_0$ and $\mathbf{C} \in \mathcal{G}(\epsilon)$, $\bar{\mathcal{N}}(i) \supseteq \bar{\mathcal{N}}(i)^*$ contains all the novel words. So for $\forall j \in \mathcal{C}_0 \bigcap \bar{\mathcal{N}}(i)$, $\mathbf{z}_{ij} = \sum_{k=1}^K e_k\mathbf{z}_{i,l(k)}$ where $l(k) \in \mathcal{C}_k$ is some novel word of k-th topic and e_k are convex combination weights. Therefore, $|\mathbf{z}_{i,l(k)}\mathbf{u}| \leq 2\epsilon, k = 1, \ldots, K \Rightarrow |\mathbf{z}_{ij}\mathbf{u}| \leq 2\epsilon$. Hence it can be union bounded by the same term. In sum, we have, given any $\mathbf{C} \in \mathcal{G}(\epsilon)$,

$$\begin{aligned} |p_i(\mathbf{C}) - p_i(\mathbf{E})| &\leq \sum_{j\in\bar{\mathcal{N}}(i)^*}\Pr_{\mathbf{u}}(|\mathbf{z}_{ij}\mathbf{u}| \leq 2\epsilon|\mathbf{C}) \\ &\leq \frac{4WG(W)\epsilon}{\min_{j\in\bar{\mathcal{N}}(i)^*}\|\mathbf{z}_{ij}\|_2} \end{aligned}$$

Last inequality is true since that a strip of width $|2\epsilon|/\|\mathbf{z}_{ij}\|$ in a unit ball has a fraction of $G(W)|2\epsilon|/\|\mathbf{z}_{ij}\|$ of the total volume, where $G(W) \triangleq \frac{\Gamma\left(\frac{W}{2}+1\right)}{\sqrt{\pi}\Gamma\left(\frac{W+1}{2}\right)}$.

So by defining $\rho \triangleq \min_{j\in\bar{\mathcal{N}}(i)^*}\|\ (\boldsymbol{\beta}'_j - \boldsymbol{\beta}'_i)\mathbf{R}'\boldsymbol{\beta}'^\top\|$, for $i = 1, \ldots, W,\ \ \forall\ \mathbf{C} \in \mathcal{G}(\rho\epsilon_0/(4WG(W))),\ \ |p_i(\mathbf{C}) - p_i(\mathbf{E})| \leq \epsilon_0$.

We could further find explicit expression for ρ. Using the similar argument in Lemma 5 we have $\rho = \gamma\min_{j\notin\mathcal{C}_k,k\geq 1}(1 - \beta'_{j,k})$. □

Remark : When W is large, the function $G(W)$ behave like $G(W) = \mathcal{O}(\sqrt{W})$.

Lemma 7. *Suppose $\mathbf{R}'$ is γ-simplicial and topic matrix $\boldsymbol{\beta}$ is separable. Let $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$. Then, $\forall\ \epsilon_0 > 0$, $\exists\ \epsilon > 0$ such that $\forall\ \mathbf{C} \in \mathcal{G}(\epsilon)$, $|p_i(\mathbf{C}) - p_i(\mathbf{E})| \leq \epsilon_0$.*

Proof. We could follow the same argument and notation as in the proof of lemma 6 up to equation (8) and obtain $\forall\mathbf{C} \in \mathcal{G}(\epsilon)$, $\forall\delta > 0$,

$$\begin{aligned} &|p_i(\mathbf{C}) - p_i(\mathbf{E})| \\ &\leq \Pr_{\mathbf{u}}(\exists j \in \bar{\mathcal{N}}(i), |\mathbf{e}_{i,j}\mathbf{u}| \geq \delta) + \Pr_{\mathbf{u}}(\forall j \in \bar{\mathcal{N}}(i), |\mathbf{z}_{ij}\mathbf{u}| \leq \delta) \\ &\leq \sum_j \Pr_{\mathbf{u}}(|\mathbf{e}_{i,j}\mathbf{u}| \geq \delta) + \sum_{j\in\bar{\mathcal{N}}(i)^*}\Pr_{\mathbf{u}}(|\mathbf{z}_{ij}\mathbf{u}| \leq \delta) \end{aligned}$$

where the second union bound follow the same argument as in Lemma 6 for both novel and non-novel words. Note that $\mathbf{z}_{ij}\mathbf{u} \sim \mathcal{N}(0, \|\mathbf{z}_{ij}\|_2^2)$ and $\mathbf{a}_{ij}\mathbf{u} \sim \mathcal{N}(0, \|\mathbf{a}_{ij}\|_2^2)$ conditioned on $\mathbf{C}$, we obtain,

$$\begin{aligned} \Pr_{\mathbf{u}}(|\mathbf{z}_{ij}\mathbf{u}| \leq \delta|\mathbf{C}) &= \int_{-\delta}^{\delta}\frac{1}{\sqrt{2\pi}\|\mathbf{z}_{ij}\|}e^{-t^2/2\|\mathbf{z}_{ij}\|^2}dt \\ &\leq \frac{\sqrt{2/\pi}}{\|\mathbf{z}_{ij}\|}\delta \end{aligned}$$

$$\Pr_{\mathbf{u}}(|\mathbf{a}_{i,j}\mathbf{u}| \geq \delta) = 2Q(\delta/\|\mathbf{a}_{i,j}\|) \leq \exp(-\delta^2/8\epsilon^2)$$

where the second bound is by the property of the Q function. In sum we obtain that

$$|p_i(\mathbf{C}) - p_i(\mathbf{E})| \leq W(\frac{\sqrt{2/\pi}}{\rho}\delta + \exp(-\delta^2/8\epsilon^2))$$

for any $\delta > 0$ and ρ as defined in Lemma 6. One possible choice is to set $\delta = \frac{\epsilon_0 \rho}{2W\sqrt{2/\pi}}$. Then for $\epsilon \le \frac{\sqrt{\pi}\epsilon_0\rho}{4W\sqrt{\log(2W/\epsilon_0)}}$, then $|p_i(\mathbf{C}) - p_i(\mathbf{E})| \le \epsilon_0$. □

Now we state and prove the main theorem which summarize the consistency and sample complexity of the random projection algorithm.

Theorem 1. *Suppose* $\mathbf{R}'$ *is* γ*-simplicial and topic matrix* $\boldsymbol{\beta}$ *is separable. Let* $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$ *or* $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$. *Then Algorithm Alg.-Value would output novel words of all distinct topics consistently as number of document* $M = H \times L \to \infty$ *and number of projections* $P \to \infty$. *Further more,* $\forall \delta > 0$, *for*

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

when $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$ *or*

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

when $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$*; and*

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

Algorithm Alg.-Value fails with probability at most δ.

Proof. For the success of detection and clustering, we require for all novel word i, $\hat{p}_i$ would be ranked topmost. Therefore the error event we are interested in is $\{\exists\ i \in \mathcal{C}_k, \exists\ j \in \mathcal{C}_0 \text{ s.t. } \hat{p}_i - \hat{p}_j < 0\}$. Consider the set $\mathcal{G}(\epsilon)$ defined in (6). Then $\forall i \in \mathcal{C}_k,\ \forall j \in \mathcal{C}_0,\ \epsilon > 0$,

$$\Pr(\hat{p}_i - \hat{p}_j < 0) \le \Pr(\hat{p}_i - \hat{p}_j < 0 \mid \mathbf{C} \in \mathcal{G}(\epsilon)) + \Pr(\mathcal{G}(\epsilon)^c)$$

For the first term, we have following decomposition.

$$\begin{aligned} \hat{p}_i - \hat{p}_j \quad = \quad & (\hat{p}_i - p_i(\mathbf{C})) + (p_i(\mathbf{C}) - p_i(\mathbf{E})) \\ & + (p_j(\mathbf{E}) - p_j(\mathbf{C})) + (p_j(\mathbf{C}) - \hat{p}_j) \\ & + (p_i(\mathbf{E}) - p_j(\mathbf{E})) \end{aligned}$$

By Lemma 3, $\forall\ \mathbf{C} \in \mathcal{G}(\epsilon)$, $(p_i(\mathbf{E}) - p_j(\mathbf{E})) \ge q_i - 0 = q_i \ge q_\wedge$. Therefore

$$\begin{aligned} & \Pr\{\hat{p}_i - \hat{p}_j < 0 | \mathbf{C} \in \mathcal{G}(\epsilon)\} \\ \le\ & \Pr\{(p_i(\mathbf{C}) - \hat{p}_i) + (p_i(\mathbf{E}) - p_i(\mathbf{C})) \\ & \quad + (p_j(\mathbf{C}) - p_j(\mathbf{E})) + (\hat{p}_j - p_j(\mathbf{C})) > q_\wedge |\ \mathcal{G}(\epsilon)\} \\ \le\ & \Pr\{p_i(\mathbf{C}) - \hat{p}_i \ge \frac{q_\wedge}{4} |\ \mathcal{G}(\epsilon)\} \\ & + \Pr\{\hat{p}_j - p_j(\mathbf{C}) \ge \frac{q_\wedge}{4} |\ \mathcal{G}(\epsilon)\} \\ & + \Pr\{p_j(\mathbf{C}) - p_j(\mathbf{E}) \ge \frac{q_\wedge}{4}\ |\mathcal{G}(\epsilon)\} \\ & + \Pr\{p_i(\mathbf{E}) - p_i(\mathbf{C}) \ge \frac{q_\wedge}{4} |\ \mathcal{G}(\epsilon)\} \end{aligned}$$

By lemma 4, the first two term is upper-bounded by $\exp(-2P(\frac{q_\wedge}{4})^2)$. When $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$, by lemma 6, if we set $\epsilon \le \frac{\rho q_\wedge}{8WG(W)}$, then $|p_i(\mathbf{E}) - p_i(\mathbf{C})| \le \frac{q_\wedge}{4}$. Therefore last two terms are exactly zero and in sum

$$\Pr\{\hat{p}_i - \hat{p}_j < 0\} \le 2\exp(-2P(\frac{q_\wedge}{4})^2) + \Pr\{\mathcal{G}(\frac{\rho q_\wedge}{8WG(W)})^c\}$$

By lemma 5 and union bound, we obtain

$$\begin{aligned} & \Pr(\exists i \in \mathcal{C}_k, \exists j \in \mathcal{C}_0,\ \hat{p}_i - \hat{p}_j < 0) \\ \le & e_1 W^4 \exp(-e_2 MNd^2\phi^2\eta^4) \\ & + e_3 W^4 \exp(-e_4 \frac{MN\rho^2 q_\wedge^2 \phi^2 \eta^4}{W^2 G(W)^2}) \\ & + e_5 W^2 \exp(-e_6 P (q_\wedge)^2) \end{aligned}$$

where $d = \gamma^2 \min_{j \notin \mathcal{C}_k} (1 - \beta'_{j,k})^2 / \lambda_\wedge^2$, $\rho = \gamma \min_{j \notin \mathcal{C}_k} (1 - \beta'_{j,k})$. By setting

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

and $P \ge c_3 \frac{\log(3W/\delta)}{q_\wedge^2}$, the error probability would be less than δ. When $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$, by applying lemma 7 instead of 6, we have

$$\begin{aligned} & \Pr(\exists i \in \mathcal{C}_k, \exists j \in \mathcal{C}_0,\ \hat{p}_i - \hat{p}_j < 0) \\ \le & e_1 W^4 \exp(-e_2 MNd^2\phi^2\eta^4) \\ & + e_3 W^4 \exp(-e_4 \frac{MN\rho^2 q_\wedge^2 \phi^2 \eta^4}{W^2 \log(2W/q_\wedge)}) \\ & + e_5 W^2 \exp(-e_6 P (q_\wedge)^2) \end{aligned}$$

And we obtain the conclusion. □

D Analysis of Alg.-Index

The proof for binning algorithm Alg.-Index has similar structure. But we have full independence across bins other than conditional independence across projections in algorithm Alg.-Value .

Recall that $\mathbf{E} = \boldsymbol{\beta}'\mathbf{R}'\boldsymbol{\beta}'^\top$. $\mathbf{C}^{(l)} = \widetilde{\mathbf{X}}^{(l)}\widetilde{\mathbf{X}}^{(l)\top}$. We denote $\mathbf{C}$ and $\bar{\mathcal{N}}(i)$ as the dummy variable but not to be confused with the one defined in previous proof. $\bar{\mathcal{N}}(i)^* = \mathcal{C}_k{}^c$ for a novel word $i \in \mathcal{C}_k$ and $\bar{\mathcal{N}}(i)^* = \mathcal{C}_0{}^c$ for non-novel word $i \in \mathcal{C}_0$. For each bin $l = 1, \ldots, L$, **one** random direction $\mathbf{d}^l = \widetilde{\mathbf{X}}^{(l)\top}\mathbf{u}^l$ is generated. The consistency of algorithm Alg.-Index depends on the convergence of following quantities,

$$\hat{p}_i = \frac{1}{L}\sum_{l=1}^{L} \mathbb{I}\{\forall j \in \bar{\mathcal{N}}(i)^{(l)}, \mathbf{C}_i^{(l)}\mathbf{u}^l \ge \mathbf{C}_j^{(l)}\mathbf{u}^l\} \quad (9)$$

$$p_i = \Pr\{\forall j \in \bar{\mathcal{N}}(i), \mathbf{C}_i\mathbf{u} - \mathbf{C}_j\mathbf{u} \ge 0\} \quad (10)$$

$$q_i = \Pr\{\forall j \in \bar{\mathcal{N}}(i)^*, \mathbf{E}_i\mathbf{u} - \mathbf{E}_j\mathbf{u} \ge 0\} \quad (11)$$

We will show in sequence that

- $\hat{p}_i$ converges to p_i as L increase.

- $\mathbf{C}$ and $\bar{\mathcal{N}}(i)$ converges to $\mathbf{E}$ and $\bar{\mathcal{N}}(i)^*$ as H increase hence,

- p_i converges to q_i.

To start with, the topic geometry indicate some minimum solid angle for novel words as extreme points. More precise,

Lemma 8. *Suppose* $\mathbf{R}'$ *is* γ*-simplicial and topic matrix* $\boldsymbol{\beta}$ *is separable. Let* $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$ *or* $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$*. Then*

1. *For all novel words,* $\exists$ $q_\wedge$ *such that* $\min_{i \in \mathcal{C}_k, 1 \le k \le K} q_i := q_\wedge > 0$*.*

2. *For all non-novel words* $i \in \mathcal{C}_0$*,* $q_i = 0$*.*

Lemma 9. $\forall t > 0$*,* $\Pr(|\hat{p}_i - p_i| \ge t) \le 2\exp(-2Lt^2)$*.*

Proof. For any word i, $\hat{p}_i^{(l)} = \mathbb{I}\{\forall j \in \bar{\mathcal{N}}(i)^{(l)},\ \mathbf{C}_i^{(l)}\mathbf{u}^l \ge \mathbf{C}_j^{(l)}\mathbf{u}^l\}$ are i.i.d 0-1 random variables whose expectation is p_i. Hence by Heoffding's lemma, we obtain the conclusion. □

Similarly as $\mathcal{G}(b)$ defined in equation (6), for a word i and $\forall b > 0$, we define

$$\begin{aligned}\mathcal{G}_i(b) = \{\ \mathbf{C} : \bar{\mathcal{N}}(i) &= \bar{\mathcal{N}}(i)^* \text{ if } i \in \mathcal{C}_k;\\ \bar{\mathcal{N}}(i) &\supseteq \bar{\mathcal{N}}(i)^* \text{ if } i \in \mathcal{C}_0;\\ \forall j,\ |\mathbf{C}_{i,j} &- \mathbf{E}_{i,j}| \le b\ \}.\end{aligned}$$

By Lemma 5 its complement is vanishing as,

$$\begin{aligned}\Pr(\mathcal{G}_i(b)^c) \le & c_1 W \exp(-c_2 H N b^2 \phi^2 \eta^4)\\ & + c_3 W \exp(-c_4 H N d^2 \phi^2 \eta^4)\end{aligned} \tag{12}$$

with the same set of parameters d, ϕ, η.

Lemma 10. *Suppose* $\mathbf{R}'$ *is* γ*-simplicial and topic matrix* $\boldsymbol{\beta}$ *is separable. Let* $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$*. Then as* $H \to \infty$*,* $|p_i - q_i| \to 0$*.*

Proof. We can decompose $p_i - q_i$ as,

$$\begin{aligned}&|p_i - q_i|\\ \le & |\Pr\{\forall j \in \bar{\mathcal{N}}(i), \mathbf{C}_i\mathbf{u} \ge \mathbf{C}_j\mathbf{u} \ \&\ \mathbf{C} \in \mathcal{G}(b)\} - q_i| + \Pr(\mathcal{G}(b)^c)\end{aligned} \tag{13}$$

For a novel word $i \in \mathcal{C}_k$, $\mathbf{C} \in \mathcal{G}_i(b)$, $\bar{\mathcal{N}}(i) = \bar{\mathcal{N}}(i)^*$. So by applying proposition 2, the first term in equation (13) can be upper bounded by

$$\begin{aligned}&|\Pr\{\forall j \in \bar{\mathcal{N}}(i), \mathbf{C}_i\mathbf{u} \ge \mathbf{C}_j\mathbf{u} \ \&\ \mathbf{C} \in \mathcal{G}(b)\} - q_i|\\ \le & \Pr(\exists j \in \bar{\mathcal{N}}(i)^* : |\mathbf{e}_{i,j}\mathbf{u}| \ge a \ \&\ \mathbf{C} \in \mathcal{G}(b))\\ & + \Pr(\forall j \in \bar{\mathcal{N}}(i)^* : |\mathbf{z}_{ij}\mathbf{u}| \le a)\\ \le & \sum_{j \in \bar{\mathcal{N}}(i)^*} \Pr(|\mathbf{e}_{i,j}\mathbf{u}| \ge a \ \&\ \mathbf{C} \in \mathcal{G}(b)) + \Pr(|\mathbf{z}_{ij}\mathbf{u}| \le a)\\ \overset{(i)}{\le} & W\left\{\exp(-\frac{a^2}{2b^2}) + \frac{2a}{\sqrt{2\pi}\rho}\right\}\end{aligned}$$

for $\forall a > 0$, where $\mathbf{e}_{i,j} = \mathbf{E}_i - \mathbf{C}_i + \mathbf{C}_j - \mathbf{E}_j$ and $\mathbf{z}_{ij} = \mathbf{E}_i - \mathbf{E}_j$. Inequality (i) is true since conditioned on $\mathbf{C}$, $\mathbf{e}_{i,j}\mathbf{u}$ and $\mathbf{z}_{i,j}\mathbf{u}$ are Gaussian.

For a novel word $i \in \mathcal{C}_k$, $\mathbf{C} \in \mathcal{G}_i(b)$, $\bar{\mathcal{N}}(i) \supseteq \bar{\mathcal{N}}(i)^* = \mathcal{C}_0$. Note that $\max_{j \in \bar{\mathcal{N}}(i)^*} \mathbf{z}_{ij}\mathbf{u} = \max_{j \in \bar{\mathcal{N}}(i)} \mathbf{z}_{ij}\mathbf{u}$ so in analog to the above bounding for novel words, we have,

$$\begin{aligned}&|\Pr\{\forall j \in \bar{\mathcal{N}}(i), \mathbf{C}_i\mathbf{u} \ge \mathbf{C}_j\mathbf{u} \ \&\ \mathbf{C} \in \mathcal{G}(b)\} - q_i|\\ \le & \Pr(\exists j \in \bar{\mathcal{N}}(i) : |\mathbf{e}_{i,j}\mathbf{u}| \ge a \ \&\ \mathbf{C} \in \mathcal{G}(b))\\ & + \Pr(\forall j \in \bar{\mathcal{N}}(i) : |\mathbf{z}_{ij}\mathbf{u}| \le a)\\ \overset{(i)}{\le} & \sum_{j \in \bar{\mathcal{N}}(i)} \Pr(|\mathbf{e}_{i,j}\mathbf{u}| \ge a \ \&\ \mathbf{C} \in \mathcal{G}(b)) + \sum_{j \in \bar{\mathcal{N}}(i)^*} \Pr(|\mathbf{z}_{ij}\mathbf{u}| \le a)\\ \le & W\left\{\exp(-\frac{a^2}{2b^2}) + \frac{2a}{\sqrt{2\pi}\rho}\right\}\end{aligned}$$

The union bounds for the second tern in (i) is true since for any $j \in \bar{\mathcal{N}}(i) \bigcap \mathcal{C}_0^c$, $|\mathbf{z}_{i,k}\mathbf{u}| \le a$ for all novel words implies $|\mathbf{z}_{ij}\mathbf{u}| \le a$.

In sum, combining equation (13) and (12), we have $\forall a > 0, b > 0$

$$\begin{aligned}|p_i - q_i| \le W\Bigg\{ & \exp(-\frac{a^2}{2b^2}) + c_1 \exp(-c_2 H N d^2 \phi^2 \eta^4)\\ & + c_3 \exp(-c_4 b^2 H N \phi^2 \eta^4) + \frac{2}{\rho\sqrt{2\pi}} a\Bigg\}\end{aligned} \tag{14}$$

So for any $\epsilon_0 > 0$, we can chose $a > 0, b > 0$ and find $H_0(t, a, b) > 0$, such that $\forall H > H_0$, the right hand side of (14) is less than ϵ_0. This conclude the proof. □

Lemma 11. *Suppose* $\mathbf{R}'$ *is* γ*-simplicial and topic matrix* $\boldsymbol{\beta}$ *is separable. Let* $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$*. Then as* $H \to \infty$*,* $|p_i - q_i| \to 0$*.*

Proof. The proof is the same as in Lemma 11. Note that when $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$

$$\Pr(|\mathbf{e}_{i,j}\mathbf{u}| \ge a) \le c_3 \exp(-c_4 a^2 H N \phi^2 \eta^4)$$

$$\Pr(|\mathbf{z}_{i,j}\mathbf{u}| \leq a) \leq \frac{2G(W)a}{\rho}, \quad j \in \bar{\mathcal{N}}(i)^*$$

where $G(W) \triangleq \frac{\Gamma\left(\frac{W}{2}+1\right)}{\sqrt{\pi}\Gamma\left(\frac{W+1}{2}\right)}$ by lemma 6 . In sum, we have $\forall a \geq 0$

$$|p_i - q_i| \leq W\Bigg\{c_1 \exp(-c_2 HNd^2\phi^2\eta^4) + c_3 \exp(-c_4 a^2 HN\phi^2\eta^4) + \frac{2G(W)a}{\rho}\Bigg\} \tag{15}$$

We could come up with different strategy of setting free variables a and find the minimum document number H_0 required. □

Now we state and prove the main theorem which summarize the consistency and sample complexity of the binning algorithm.

Theorem 2. *Suppose $\mathbf{R}'$ is γ-simplicial and topic matrix $\boldsymbol{\beta}$ is separable. Let $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$ or $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$. Then Algorithm Alg.-Index would output novel words of all distinct topics consistently as both $H, L \to \infty$. (Hence total number of document $M = H \times L \to \infty$). Furthermore, $\forall \delta > 0$, for*

$$H \geq \max\Bigg\{c_1 \frac{\log(8W/q_\wedge)}{Nd^2\phi^2\eta^4},\ c_2 \frac{W^2 G(W)^2 \log(4W/q_\wedge)}{N\rho^2 q_\wedge^2 \phi^2 \eta^4}\Bigg\}$$

when $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$ or

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

when $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$ and for

$$L \geq c_3 \frac{\log(4W)/\delta}{q_\wedge^2}$$

Algorithm Alg.-Index fails with probability at most δ.

Proof. Recall that the error event we are interested in is $\{\exists i \in \mathcal{C}_k,\ \exists j \in \mathcal{C}_0,\ \hat{p}_i - \hat{p}_j < 0\}$. We have decomposition

$$\hat{p}_i - \hat{p}_j = (\hat{p}_i - q_i) + (q_j - \hat{p}_j) + (q_i - q_j)$$

and $q_i - q_j \geq q_\wedge$ for $i \in \mathcal{C}_k, j \in \mathcal{C}_0$. Therefore,

$$\begin{aligned} &\Pr\{\exists i \in \mathcal{C}_k,\ \exists j \in \mathcal{C}_0,\ \hat{p}_i - \hat{p}_j < 0\} \\ \leq &\sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_0} \Pr\{\hat{p}_i - \hat{p}_j < 0\} \\ \leq &\sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_0} \Pr\{\hat{p}_i - q_i > q_\wedge/2\} + \Pr\{\hat{p}_j - q_j > q_\wedge/2\} \end{aligned}$$

For each term, by lemma 9

$$\begin{aligned} &\Pr(|\hat{p}_i - q_i| \geq q_\wedge/2) \\ \leq &\Pr(|\hat{p}_i - p_i| \geq q_\wedge/2 - |p_i - q_i|) \\ \leq &2\exp(-2L(q_\wedge/2 - t)^2) \text{ if } |p_i - q_i| \leq t \leq q_\wedge/2 \end{aligned}$$

By lemma 10, when $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$, $\forall a > 0, b > 0$

$$|p_i - q_i| \leq W\Bigg\{\exp(-\frac{a^2}{2b^2}) + e_1 \exp(-e_2 HNd^2\phi^2\eta^4) + e_3 \exp(-e_4 b^2 HN\phi^2\eta^4) + \frac{2}{\rho\sqrt{2\pi}}a\Bigg\}$$

We could set $t = q_\wedge/4$, $a = \frac{\sqrt{2\pi}\rho t}{8W}$, $b = \frac{\rho\sqrt{\pi}t}{8W\sqrt{\log(8W/t)}}$ so for

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

and

$$L \geq c_3 \frac{\log(4W)/\delta}{q_\wedge^2}$$

by union bound the error probability $\Pr\{\exists i \in \mathcal{C}_k,\ \exists j \in \mathcal{C}_0,\ \hat{p}_i - \hat{p}_j < 0\} \leq \delta$. Similarly, by lemma 11 for $\mathbf{u} \sim \text{Uniform}(\mathcal{B}^W)$, $\forall a > 0$

$$|p_i - q_i| \leq W\Bigg\{c_1 \exp(-c_2 HNd^2\phi^2\eta^4) + c_3 \exp(-c_4 a^2 HN\phi^2\eta^4) + \frac{2G(W)a}{\rho}\Bigg\}$$

We could set $t = q_\wedge/4$, $a = \frac{t\rho}{2WG(W)}$, so for

$$H \geq \max\Bigg\{c_1 \frac{\log(8W/q_\wedge)}{Nd^2\phi^2\eta^4},\ c_2 \frac{W^2 G(W)^2 \log(4W/q_\wedge)}{N\rho^2 q_\wedge^2 \phi^2 \eta^4}\Bigg\}$$

and $L \geq c_3 \frac{\log(4W)/\delta}{q_\wedge^2}$, error probability is less than δ. □

Remark: We should point out that in general, L and H can be some function of $M = L \times H$. On one hand, H should be large as indicated by the above bounds for H are much higher. On the other hand, L should go to infinity as well so H cannot be too large.

For example, by setting $L = H = \sqrt{M}$, the sample complexity is almost squared compared to the one for M in random projection Theorem 1. On the other hand, if we set $L = \log(M)$, $H = M/\log(M)$ is almost the same order as M. But for the success of the algorithm, one has to get at least $L = \log(M) \geq K$ bins, which implies that $M = \exp(K)$ which is impractical.

References

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.

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.

Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.0 beta. `http://cvxr.com/cvx`, September 2013.

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.