Diversifying Restricted Boltzmann Machine for Document Modeling

Pengtao Xie
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213
pengtaox@cs.cmu.edu

Yuntian Deng
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213
yuntiand@cs.cmu.edu

Eric P. Xing
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213
epxing@cs.cmu.edu

ABSTRACT

Restricted Boltzmann Machine (RBM) has shown great effectiveness in document modeling. It utilizes hidden units to discover the latent topics and can learn compact semantic representations for documents which greatly facilitate document retrieval, clustering and classification. The popularity (or frequency) of topics in text corpora usually follow a power-law distribution where a few dominant topics occur very frequently while most topics (in the long-tail region) have low probabilities. Due to this imbalance, RBM tends to learn multiple redundant hidden units to best represent dominant topics and ignore those in the long-tail region, which renders the learned representations to be redundant and non-informative. To solve this problem, we propose Diversified RBM (DRBM) which diversifies the hidden units, to make them cover not only the dominant topics, but also those in the long-tail region. We define a diversity metric and use it as a regularizer to encourage the hidden units to be diverse. Since the diversity metric is hard to optimize directly, we instead optimize its lower bound and prove that maximizing the lower bound with projected gradient ascent can increase this diversity metric. Experiments on document retrieval and clustering demonstrate that with diversification, the document modeling power of DRBM can be greatly improved.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications, Data Mining

General Terms

Algorithms, Experiments

Keywords

Diversified Restricted Boltzmann Machine, Diversity, Power-law Distribution, Document Modeling, Topic Modeling

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783264.

1. INTRODUCTION

Learning lower-dimensional semantic representations [14, 3, 13, 25, 18, 26, 8, 9, 11, 15, 19, 7] for documents is essential for many tasks such as document retrieval, clustering and classification, to name a few. It is assumed that there exist a set of hidden topics underlying the observed words and each document has a representation vector in the latent topic space. In general, the latent topic space is able to capture more compact and semantically meaningful characteristics of the documents than the observed word space. Performing text mining and natural language understanding tasks on the learned semantic representations can yield better performance than on the words based representations.

Restricted Boltzmann Machine (RBM) [24, 13, 26] has been successfully used for document representation learning. RBM is a two-layer undirected graphical model where one layer of hidden units are used to capture latent topics and one layer of visible units are used to represent observed words. The correlation between hidden units and visible units are modeled with undirected weighted edges. Each hidden unit is characterized by a weight vector (to be learned) where each component in the vector corresponds to a word in the vocabulary. Compared with directed topic models such as Probabilistic Latent Semantic Analysis (PLSA) [14] and Latent Dirichlet Allocation (LDA) [3], RBM is very efficient in inferring the latent representations of documents, which is crucial for applications having real-time requirements such as retrieval. And RBM has been demonstrated to perform more accurately than LDA in document retrieval and classification [13].

In most (if not all) document collections, the popularity of topics usually follows a power-law distribution. A few dominant topics appear with very high frequency while most topics are of low probabilities. For example, in a news corpus, topics like *politics*, *economics*, *sports* occur a lot while topics like *garden*, *furniture*, *poem* are rarely mentioned. Due to this skewness of topic popularity, RBM tends to learn multiple redundant hidden units to cover dominant topics as best as possible and pay little attention to topics in the long-tail region. Failing to capture long-tail topics incurs significant information loss. Though the probability of each long-tail topic is low, the total probability mass of the long-tail region is large because the number of topics in this region is large. These long-tail topics with such a large total probability mass are of equal importance with, if not more importance than, the few dominant topics. Besides, the latent representations learned by RBM are redundant in that many dimensions are actually of the same or similar meaning

and the resultant high dimensionality incurs computational inefficiency.

In this paper, we aim to address this problem by proposing to enhance the diversity of the hidden units in RBM to make them cover as many topics as possible, not only the dominant ones, but also those in the long-tail region. To combat the phenomenon that many redundant hidden units are learned to characterize the dominant topics as best as possible with the price of ignoring long-tail topics, we impose a diversity regularizer over these hidden units to reduce their redundancy and improve their coverage of long-tail topics. We define a diversity metric to formally measure how diverse a set of hidden units are: given the weight vectors $\mathbf{A}$ where each column in $\mathbf{A}$ corresponds to one hidden unit, the diversity metric is defined as $\Omega(\mathbf{A}) = \Psi(\mathbf{A}) - \Pi(\mathbf{A})$, where $\Psi(\mathbf{A})$ is the mean of the angles between all pairs of weights vectors and $\Pi(\mathbf{A})$ is the variance of these angles. A larger $\Omega(\mathbf{A})$ indicates that the weight vectors in $\mathbf{A}$ are more diverse. This diversity metric is defined with the rationale that a set of weight vectors possess a larger diversity if the pairwise angles between them have a larger mean $\Psi(\mathbf{A})$ and a smaller variance $\Pi(\mathbf{A})$. A larger mean implies that these vectors share larger angles on the whole, hence are more different from each other. A smaller variance indicates that these vectors uniformly spread out to different directions and each vector is evenly different from all other vectors. We employ this diversity metric to regularize the learning of RBM to encourage the hidden units to be diverse. Considering that the diversity metric is hard to optimize, we seek a lower bound of it which is much more amenable for optimization. We prove that maximizing the lower bound with projected gradient ascent can increase the diversity metric. Experiments on three datasets demonstrate that with diversification, RBM can learn much more powerful latent document representations that boost the performance of document retrieval and clustering greatly.

The major contributions of this paper are:

- We propose the problem of diversifying Restricted Boltzmann Machine to make the learned hidden units not only to cover dominant topics, but also to capture long-tail topics effectively.
- We define a diversity metric to measure how diverse the hidden units are and use it to regularize RBM to encourage diversity.
- To make optimization easier, we derive a lower bound of the diversity metric and show that maximizing the lower bound with projected gradient ascent can increase the diversity metric.
- On two tasks — document retrieval and clustering — and three datasets, we empirically demonstrate that through diversification, RBM can learn much more effective document representations.

The rest of the paper is organized as follows. In Section 2, we review related works. In Section 3, we propose the Diversified RBM (DRBM). Section 4 gives experimental results and Section 5 concludes the paper.

2. RELATED WORKS

Document modeling aims to discover the latent semantics underlying document corpora and learn representations for documents in the latent semantic space. Directed topic models such as Probabilistic Latent Semantic Analysis (PLSA) [14] and Latent Dirichlet Allocation (LDA) [3] represent each topic with a multinomial distribution over the words in the vocabulary and assume each document has a multinomial distribution over topics. The observed words are generated from the latent topics in a hierarchical way. In directed topic models, inferring documents' distributions over latent topics usually involve iterative optimization or sampling, which is time-consuming and limits the applicability of these models in real-time applications such as document retrieval.

Restricted Boltzmann Machine (RBM) [13, 26] has shown great success in document modeling. [13] proposed a Replicated Softmax RBM to model word counts. [26] introduced Deep Boltzmann Machine to extract distributed semantic representations for documents. Another paradigm of works are based on neural network (NN) [2, 25, 18, 20, 8, 9, 11, 15, 19, 7], which aim to learn word/document representations that are able to capture the latent semantics. Compared with directed topic models [14, 3], RBM and NN based methods enable fast inference of the latent representations, which is key to real-time applications. Our work is built on top of the existing RBM methods, with a special focus on how to capture low-frequency topics/semantics. It is worth noting that the techniques developed in this paper can be also applied to NN based approaches.

There have been several works [21, 22, 27, 1] attempting to capture the power-law behavior of topic popularity, using priors that encourage the distributions over topics to be long-tailed, such as Pitman-Yor Process prior [21, 22], asymmetric Dirichlet prior [27] and Indian Buffet Process compound Dirichlet Process prior [1]. These methods have no regularization over topics' distributions over words. While these priors encourage new topics to be generated, the newly generated topics may be alike to the existing ones used to model dominant topics. In the end, the learned topics are still redundant and unable to capture the long-tail topics effectively. Notably, [17] studied the diversification of Latent Dirichlet Allocation with the goal to learn diverse topics, by imposing a Determinantal Point Process (DPP) [16] prior over the topics-word multinomial vectors. The DPP prior allows one to specify a preference for diversity using a positive definite kernel function. In this paper, we study the diversification of RBM and propose an alternative diversity measure that has properties complementary to the DPP prior, such as (1) invariant to scaling of the weight vectors; (2) encouraging the weight vectors to evenly spread out.

3. DIVERSIFY RESTRICTED BOLTZMANN MACHINE

In this section, we propose to diversify the hidden units in RBM to reduce their redundancy and improve the coverage of long-tail topics.

3.1 Restricted Boltzmann Machine

Figure 1 shows the basic Restricted Boltzmann Machine (RBM). RBM [24] is an undirected graphical model consisting of a set of hidden units $\mathbf{h} = \{h_k\}_{k=1}^{K}$ and a set of visible units $\mathbf{v} = \{v_j\}_{j=1}^{J}$. Both h_k and v_j are assumed to be binary. Each hidden unit is connected with all visible units with undirected edges and there is no connection between two hidden units or two visible units. The energy function

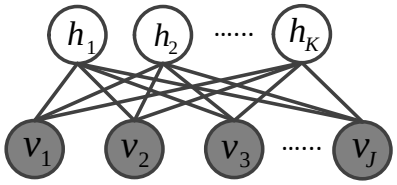

Figure 1: Restricted Boltzmann Machine

defined over $\mathbf{h}$ and $\mathbf{v}$ is

$$E(\mathbf{h},\mathbf{v}) = -\sum_{j=1}^{J}\alpha_j v_j - \sum_{k=1}^{K}\beta_k h_k - \sum_{j=1}^{J}\sum_{k=1}^{K} A_{jk} v_j h_k \quad (1)$$

where $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^J$ and $\boldsymbol{\beta} = \{\beta_j\}_{j=1}^J$ are the biases associated with the visible and hidden units respectively. $\mathbf{A}$ are the weights on the edges connecting two set of units. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{A}$ are model parameters.

To better model word counts, [13] proposed Replicated Softmax RBM. Let $\mathbf{V}$ be a $D \times J$ observed binary matrix of a document containing D tokens. J is the vocabulary size. Row i of $\mathbf{V}$ is the 1-of-J coding vector of the ith token in this document. $V_{ij} = 1$ if the ith token is the jth word in the vocabulary. Under this representation, the energy function $E(\mathbf{h},\mathbf{V})$ is defined as

$$-\sum_{i=1}^{D}\sum_{j=1}^{J}\alpha_j V_{ij} - D\sum_{k=1}^{K}\beta_k h_k - \sum_{i=1}^{D}\sum_{j=1}^{J}\sum_{k=1}^{K} A_{jk} V_{ij} h_k \quad (2)$$

Given the observed tokens $\mathbf{V}$, inferring the latent representation $\mathbf{h}$ can be done very efficiently

$$p(h_k = 1|\mathbf{V}) = \sigma(D\beta_k + \sum_{i=1}^{D}\sum_{j=1}^{J} A_{jk} V_{ij}) \quad (3)$$

where $\sigma(x) = 1/(1+\exp(-x))$ is the logistic function. The model parameters can be learned by maximizing the data likelihood $\mathcal{L}(\{\mathbf{V}_n\}_{n=1}^N; \mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ using the contrastive divergence [12] method, which is essentially a gradient ascent approach.

3.2 Diversify Restricted Boltzmann Machine

The first step towards diversifying Restricted Boltzmann Machine is to measure the diversity of the hidden units. In RBM, each hidden unit possesses a weight vector where the weights are associated with the edges connecting this hidden unit and the visible units. These weight vectors are the parameters to be learned in RBM. Given the weight vectors $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_K]$ of K hidden units, where each column in matrix $\mathbf{A}$ corresponds to one hidden unit, their diversity can be informally described as how different each vector is from others. While there are many ways to measure the difference between vectors $\mathbf{a}_i$ and $\mathbf{a}_j$, we prefer to use the angle between them since the angle is invariant to translation and scaling (with positive factors) of the two vectors. In addition, we do not care about the orientation of vectors, thus preferring the angle to be acute or right. If the angle θ is obtuse, we replace it with $\pi - \theta$. To sum up, we define the angle $\theta(\mathbf{a}_i, \mathbf{a}_j)$ between vector $\mathbf{a}_i$ and $\mathbf{a}_j$ to be $\theta(\mathbf{a}_i, \mathbf{a}_j) = \arccos(\frac{|\mathbf{a}_i \cdot \mathbf{a}_j|}{\|\mathbf{a}_i\|\|\mathbf{a}_j\|})$. Given a set of weight vectors $\mathbf{A}$, we define the diversity metric as $\Omega(\mathbf{A}) = \Psi(\mathbf{A}) - \Pi(\mathbf{A})$, where $\Psi(\mathbf{A}) = \frac{1}{K^2}\sum_{i=1}^K\sum_{j=1}^K \theta(\mathbf{a}_i, \mathbf{a}_j)$ is the mean of the angles between all pairs of weights vectors and $\Pi(\mathbf{A}) = \frac{1}{K^2}\sum_{i=1}^K\sum_{j=1}^K (\theta(\mathbf{a}_i, \mathbf{a}_j) - \Psi(\mathbf{A}))^2$ is the variance of these angles. A larger $\Omega(\mathbf{A})$ indicates that the weight vectors in $\mathbf{A}$ are more diverse. Intuitively, the set of weight vectors possess a larger diversity if the pairwise angles between them have a larger mean and a smaller variance. A larger mean implies that these vectors share larger angles on the whole, hence are more different from each other. A smaller variance indicates that these vectors uniformly spread out to different directions and each vector is evenly different from all other vectors. Encouraging the variance to be small can prevent the phenomenon that the vectors fall into several groups where vectors in the same group have small angles and vectors between groups have large angles. Such a phenomenon renders the vectors to be redundant and less diverse, and hence should be prohibited. Throughout the paper, we assume the weight vectors are linearly independent. Later we present a justification of the validity of this assumption.

To diversify the hidden units in RBM, we use the diversity metric described above to regularize the learning of the weight vectors and define a Diversified RBM (DRBM) problem as

$$(\mathbf{P1}) \quad \max_{\mathbf{A}} \quad \mathcal{L}(\{\mathbf{V}_n\}_{n=1}^N; \mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda\Omega(\mathbf{A}) \quad (4)$$

where $\lambda > 0$ is a tradeoff parameter between the data likelihood and the diversity regularizer. The term $\lambda\Omega(\mathbf{A})$ in the new objective function encourages the weight vectors in $\mathbf{A}$ to be diverse. λ plays an important role in balancing the fitness of $\mathbf{A}$ to the data likelihood and its diversity. Under a small λ, $\mathbf{A}$ is learned to best maximize the data likelihood and its diversity is ignored. As discussed earlier, such a $\mathbf{A}$ has high redundancy and may not be able to cover long-tail topics effectively. Under a large λ, $\mathbf{A}$ is learned with high diversity, but may not be well fitted to the data likelihood and hence lose the capability to properly model documents. To sum up, a proper λ needs to be chosen to achieve the optimal balance.

3.3 Optimization

In this section, we study how to solve the problem (**P1**) defined in Eq.(4). For the ease of optimization, we first reformulate this problem. Let $\mathbf{A} = \text{diag}(\mathbf{g})\widetilde{\mathbf{A}}$, where $\mathbf{g}$ is a vector and g_i denotes the L_2 norm of the ith column of $\mathbf{A}$, then the L_2 norm of each column vector $\tilde{\mathbf{a}}$ in $\widetilde{\mathbf{A}}$ is 1. Based on the definition of the diversity metric, we have $\Omega(\mathbf{A}) = \Omega(\widetilde{\mathbf{A}})$. Accordingly, (**P1**) can be reformulated as

$$(\mathbf{P2}) \quad \begin{array}{ll} \max_{\widetilde{\mathbf{A}},\mathbf{g}} & \mathcal{L}(\{\mathbf{V}_n\}_{n=1}^N; \text{diag}(\mathbf{g})\widetilde{\mathbf{A}}, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda\Omega(\widetilde{\mathbf{A}}) \\ s.t. & \forall i = 1, \cdots, K, \|\tilde{\mathbf{a}}_i\| = 1 \end{array} \quad (5)$$

(**P2**) can be solved by alternating between $\mathbf{g}$ and $\widetilde{\mathbf{A}}$: optimizing $\mathbf{g}$ with $\widetilde{\mathbf{A}}$ fixed and optimizing $\widetilde{\mathbf{A}}$ with $\mathbf{g}$ fixed. With $\widetilde{\mathbf{A}}$ fixed, the problem defined over $\mathbf{g}$ is

$$\max_{\mathbf{g}} \quad \mathcal{L}(\{\mathbf{V}_n\}_{n=1}^N; \text{diag}(\mathbf{g})\widetilde{\mathbf{A}}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (6)$$

which can be efficiently solved with contrastive divergence (CD) based gradient ascent method. Fixing $\mathbf{g}$, the problem defined over $\widetilde{\mathbf{A}}$ is non-smooth and non-convex, which is hard to solve. Now we focus on how to tackle it. As discussed earlier, the data likelihood of RBM is usually maximized with gradient ascent (GA) method. To be consistent, it is desirable to optimize the diversity regularizer with projected[1]

[1]Projection is needed due to the constraints in (**P2**).

gradient ascent (PGA) as well. Since $\Omega(\widetilde{\mathbf{A}})$ is non-smooth and non-convex, (sub)gradient methods are not applicable. To solve this problem, we derive a smooth lower bound $\Gamma(\widetilde{\mathbf{A}})$ of $\Omega(\widetilde{\mathbf{A}})$ and require $\Gamma(\widetilde{\mathbf{A}})$ to have the following two traits: 1) the gradient of $\Gamma(\widetilde{\mathbf{A}})$ is easy to compute; 2) optimizing $\Gamma(\widetilde{\mathbf{A}})$ with PGA can increase $\Omega(\widetilde{\mathbf{A}})$. The lower bound is given in Lemma 1[2].

LEMMA 1. *Let* $\det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})$ *denote the determinant of the Gram matrix of* $\widetilde{\mathbf{A}}$*, then* $0 < \det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}) \leq 1$*. Let* $\Gamma(\widetilde{\mathbf{A}}) = \arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})}) - (\frac{\pi}{2} - \arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})}))^2$*, then* $\Gamma(\widetilde{\mathbf{A}})$ *is a lower bound of* $\Omega(\widetilde{\mathbf{A}})$*.* $\Gamma(\widetilde{\mathbf{A}})$ *and* $\Omega(\widetilde{\mathbf{A}})$ *have the same global optimal.*

Next we prove that maximizing $\Gamma(\widetilde{\mathbf{A}})$ using PGA can increase the diversity metric $\Omega(\widetilde{\mathbf{A}})$. The statement is formally described in Theorem 1.

THEOREM 1. *Let* $\mathbf{G}^{(t)}$ *be the gradient of* $\Gamma(\widetilde{\mathbf{A}})$ *w.r.t* $\widetilde{\mathbf{A}}^{(t)}$ *at iteration* t*.* $\exists\tau > 0$*, such that* $\forall\eta \in (0, \tau)$*,* $\Omega(\widetilde{\mathbf{A}}^{(t+1)}) \geq \Omega(\widetilde{\mathbf{A}}^{(t)})$*, where* $\widetilde{\mathbf{A}}^{(t+1)} = \mathcal{P}(\widetilde{\mathbf{A}}^{(t)} + \eta\mathbf{G}^{(t)})$ *and* $\mathcal{P}(\cdot)$ *denotes the projection to the unit sphere.*

Note that $\Omega(\widetilde{\mathbf{A}})$ consists of two parts $\Psi(\widetilde{\mathbf{A}})$ and $\Pi(\widetilde{\mathbf{A}})$. To prove Theorem 1, we prove that maximizing $\Gamma(\widetilde{\mathbf{A}})$ using PGA can increase the mean $\Psi(\widetilde{\mathbf{A}})$ and reduce the variance $\Pi(\widetilde{\mathbf{A}})$ simultaneously, which are stated in Theorem 2 and 3.

THEOREM 2. *Let* $\mathbf{G}^{(t)}$ *be the gradient of* $\Gamma(\widetilde{\mathbf{A}})$ *w.r.t* $\widetilde{\mathbf{A}}^{(t)}$ *at iteration* t*.* $\exists\tau_1 > 0$*, such that* $\forall\eta \in (0, \tau_1)$*,* $\Psi(\widetilde{\mathbf{A}}^{(t+1)}) \geq \Psi(\widetilde{\mathbf{A}}^{(t)})$*, where* $\widetilde{\mathbf{A}}^{(t+1)} = \mathcal{P}(\widetilde{\mathbf{A}}^{(t)} + \eta\mathbf{G}^{(t)})$*.*

THEOREM 3. *Let* $\mathbf{G}^{(t)}$ *be the gradient of* $\Gamma(\widetilde{\mathbf{A}}^\top)$ *w.r.t* $\widetilde{\mathbf{A}}^{(t)}$ *at iteration* t*.* $\exists\tau_2 > 0$*, such that* $\forall\eta \in (0, \tau_2)$*,* $\Pi(\widetilde{\mathbf{A}}^{(t+1)}) \leq \Pi(\widetilde{\mathbf{A}}^{(t)})$*, where* $\widetilde{\mathbf{A}}^{(t+1)} = \mathcal{P}(\widetilde{\mathbf{A}}^{(t)} + \eta\mathbf{G}^{(t)})$*.*

These two theorems immediately imply Theorem 1. To prove Theorem 2, the following lemma is needed.

LEMMA 2. *Let the weight vector* $\tilde{\mathbf{a}}_i$ *of hidden unit* i *be decomposed into* $\tilde{\mathbf{a}}_i = \mathbf{x}_i + l_i\mathbf{e}_i$*, where* $\mathbf{x}_i = \sum_{j=1,j\neq i}^{K} \alpha_j\tilde{\mathbf{a}}_j$ *lies in the subspace* L *spanned by* $\{\tilde{\mathbf{a}}_1, \cdots, \tilde{\mathbf{a}}_K\}\backslash\{\tilde{\mathbf{a}}_i\}$*,* $\mathbf{e}_i$ *is in the orthogonal complement of* L*,* $\|\mathbf{e}_i\| = 1$*,* $\mathbf{e}_i \cdot \tilde{\mathbf{a}}_i > 0$*,* l_i *is a scalar. Then the gradient of* $\Gamma(\widetilde{\mathbf{A}})$ *w.r.t* $\mathbf{a}_i$ *is* $k_i\mathbf{e}_i$*, where* k_i *is a positive scalar.*

Given Lemma 2, we can justify why the weight vectors can be always linearly independent during the projected gradient ascent procedure. First, we assume they are initialized to be linearly independent. From Lemma 2, we know that the gradient direction of each weight vector $\mathbf{a}_i$ is orthogonal to all other weight vectors. So moving along such a direction prevents each vector to fall into the span of all other vectors. And projecting the vectors onto the unit sphere does not change the directions of these vectors. Thereby, these weight vectors would be always linearly independent.

The proofs of Lemma 2 and Theorem 2 are presented in Appendix B and C respectively. To prove Theorem 3, we need the following lemma:

[2]The proof of Lemma 1 is given in Appendix A.

	#categories	#samples	vocab. size
TDT	30	9394	5000
20-News	20	18846	5000
Reuters	9	7195	5000

Table 1: Statistics of Datasets

K	25	50	100	200	500
RBM	11.2	11.4	11.9	12.1	47.4
DRBM	**78.4**	**84.2**	**78.6**	**79.9**	**77.6**

Table 2: Precision@100 on TDT dataset

LEMMA 3. *Given a nondecreasing sequence* $b = (b_i)_{i=1}^n$ *and a strictly decreasing function* $g(x)$ *which satisfies* $0 \leq g(b_i) \leq min\{b_{i+1} - b_i : i = 1, 2, \cdots, n-1, b_{i+1} \neq b_i\}$*, we define a sequence* $c = (c_i)_{i=1}^n$ *where* $c_i = b_i + g(b_i)$*. If* $b_1 < b_n$*, then* $var(c) < var(b)$*, where* $var(\cdot)$ *denotes the variance of a sequence. Furthermore, let* $n' = max\{j : b_j \neq b_n\}$*, we define a sequence* $b' = (b'_i)_{i=1}^n$ *where* $b'_i = b_i + g(b_n) + (g(b_{n'}) - g(b_n))\mathbb{I}(i \leq n')$ *and* $\mathbb{I}(\cdot)$ *is the indicator function, then* $var(c) \leq var(b') < var(b)$*.*

The proofs of Lemma 3 and Theorem 3 are given in Appendix D and E respectively.

4. EXPERIMENTS

In this section, we present experimental results on two tasks and on three datasets, which demonstrate that with diversification, DRBM can learn much more effective representations than RBM.

4.1 Experimental Setup

Three datasets were used in the experiments. The first one [5] is a subset of the Nist Topic Detection and Tracking (TDT) corpus which contains 9394 documents from the largest 30 categories. 70% documents were used for training and 30% were used for testing. The second dataset is the 20 Newsgroups (20-News), which contains 18846 documents from 20 categories. 60% documents were used for training and 40% were used for testing. The third dataset [4] is the Reuters-21578 (Reuters) dataset. Categories with less than 100 documents were removed, which left us 9 categories and 7195 documents. 70% documents were used for training and 30% were used for testing. Each dataset used a vocabulary of 5000 words with the largest document frequency. Table 1 summarizes the statistics of three datasets.

We used gradient methods to train RBM [13] and DRBM. The mini-batch size was set to 100. The learning rate was set to 1e-4. The number of gradient ascent iterations was set to 1000 and the number of Gibbs sampling iterations in contrastive divergence was fixed to 1. The tradeoff parameter λ in DRBM was tuned with 5-fold cross validation. We compared with the following baselines methods: (1) bag-of-words (BOW); (2) Latent Dirichlet Allocation (LDA) [3]; (3) LDA regularized with Determinantal Point Process prior (DPP-LDA) [17]; (4) Pitman-Yor Process Topic Model (PYTM) [22]; (5) Latent IBP Compound Dirichlet Allocation (LIDA) [1]; (6) Neural Autoregressive Topic Model (DocNADE) [18]; (7) Paragraph Vector (PV) [19]; (8) Restricted Boltzmann Machine [13]. The parameters in baseline methods were tuned using 5-fold cross validation.

Category ID	1	2	3	4	5	6	7	8	9
Number of Documents	3713	2055	321	298	245	197	142	114	110
Precision@100 (%) of RBM	68.5	44.4	9.1	10.1	6.3	4.4	3.8	3.0	2.6
Precision@100 (%) of DRBM	89.7	80.2	31.3	39.5	26.5	22.7	9.4	14.0	12.9
Relative Improvement	31%	81%	245%	289%	324%	421%	148%	366%	397%

Table 5: Precision@100 on each category in Reuters dataset

K	25	50	100	200	500
RBM	6	6.1	5.7	9.2	**22.3**
DRBM	**14.5**	**24.9**	**15.4**	**20.3**	21.1

Table 3: Precision@100 on 20-News dataset

K	25	50	100	200	500
RBM	37.7	38.1	50.1	64.0	**70.1**
DRBM	**67.8**	**73.3**	**75.9**	**70.3**	66.2

Table 4: Precision@100 on Reuters dataset

4.2 Document Retrieval

In this section, we evaluate the effectiveness of the learned representations on retrieval. Precision@100 is used to evaluate the retrieval performance. For each test document, we retrieve 100 documents from the training set that have the smallest Euclidean distance with the query document. The distance is computed on the learned representations. Precision@100 is defined as $n/100$, where n is the number of retrieved documents that share the same class label with the query document.

Table 2, 3 and 4 show the precision@100 under different number K of hidden units on TDT, 20-News and Reuters dataset respectively. As can be seen from these tables, DRBM with diversity regularization largely outperforms RBM which has no diversity regularization under various choices of K, which demonstrates that diversifying the hidden units can greatly improve the effectiveness of document representation learning. The improvement is especially significant when K is small. For example, on TDT dataset, under $K = 25$, DRBM improves the precision@100 from 11.2% to 78.4%. Under a small K, RBM allocates most (if not all) hidden units to cover dominant topics, thus long-tail topics have little chance to be modeled effectively. DRBM solves this problem by increasing diversity of these hidden units to enforce them to cover not only the dominant topics, but also the long-tail topics. Thereby, the learned representations are more effective in capturing the long-tail semantics and the retrieval performance is greatly improved. As K increases, the performance of RBM increases. This is because under a larger K, some hidden units can be spared to model topics in the long-tail region. In this case, enforcing diversity still improves performance, though the significance of improvement diminishes as K increases.

To further examine whether DRBM can effectively capture the long-tail semantics, we show the precision@100 on each of the 9 categories in Reuters dataset in Table 5. The 2nd row shows the number of documents in each category. The distribution of document frequency is in a power-law fashion, where dominant categories (such as 1 and 2) have a lot of documents while most categories (called long-tail categories) have a small amount of documents. The 3rd and 4th row show the precision@100 achieved by RBM and DRBM on each category. The 5th row shows the relative improvement of DRBM over RBM. The relative improvement is defined as $\frac{P_{drbm}-P_{rbm}}{P_{rbm}}$, where P_{drbm} and P_{rbm} denote the precision@100 achieved by DRBM and RBM respectively. While DRBM improves RBM over all the categories, the improvements on long-tail categories are much more significant than dominant categories. For example, the relative improvements on category 8 and 9 are 366% and 397% while the improvements are 31% and 81% on category 1 and 2. This indicates that DRBM can effectively capture the long-tail topics, thereby improve the representations learned for long-tail categories significantly.

	TDT	20-News	Reuters
BOW	40.9	7.4	69.3
LDA [3]	79.4	19.6	68.5
DPP-LDA [17]	81.9	18.2	69.9
PYTM [22]	78.7	20.1	70.6
LIDA [1]	77.9	21.8	71.4
DocNADE [18]	80.3	16.8	72.6
PV [19]	81.7	19.1	**76.9**
RBM [13]	47.4	22.3	70.1
DRBM	**84.2**	**24.9**	**75.9**

Table 6: Precision@100 on three datasets

K	25	50	100	200	500
RBM	19.7	19.1	14.4	13.0	23.3
DRBM	**52.4**	**46.2**	**46.5**	**41.4**	**39.5**

Table 7: Clustering accuracy (%) on TDT test set

One great merit of DRBM is that it can achieve notable performance under a small K, which is of key importance for fast retrieval. On the TDT dataset, DRBM can achieve a precision@100 of 78.4% with $K = 25$, which cannot be achieved by RBM even when K is raised to 500. This indicates that with DRBM, one can perform retrieval on low-dimensional representations, which is usually much easier than on high-dimensional representations. For example, in KD tree [10] based nearest neighbor search, while building a KD tree on feature vectors with hundreds of dimensions is extremely hard, feature vectors whose dimension is less than one hundred are much easier to handle.

Table 6 presents the comparison with the state of the art document representation learning methods. As can be seen from this table, our method achieves the best performances on the TDT and 20-News datasets and achieves the second best performance on the Reuters dataset. The bag-of-word (BOW) representation cannot capture the underlying semantics of documents, thus its performance is inferior. LDA, RBM and neural network based approaches including DocNADE [18] and PV [19] can represent documents into the latent topic space where document retrieval can be performed more accurately. However, they lack the mechanisms

K	25	50	100	200	500
RBM	6.4	6.8	**21.5**	12.7	22.7
DRBM	**18.2**	**29.4**	19.8	**25.9**	**25.6**

Table 8: Clustering accuracy (%) on 20-News test set

K	25	50	100	200	500
RBM	45.0	41.7	38.4	46.8	47.6
DRBM	**51.4**	**58.6**	**60.9**	**53.4**	**48.5**

Table 9: Clustering accuracy (%) on Reuters test set

to cover long-tail topics, hence the resultant representations are less effective. PYTM [22] and LIDA [1] use power-law priors to encourage new topics to be generated, however, the newly generated topics may still be used to model the dominant semantics rather than those in the long-tail region. DPP-LDA [17] uses a correlation kernel Determinantal Point Process (DPP) prior to diversify the topics in LDA. However, it does not improve LDA too much.

4.3 Clustering

Another task we study is to do k-means clustering on the learned representations. In our experiments, the input cluster number of k-means was set to the ground truth number of categories in each dataset. In each run, k-means was repeated 10 times with different random initializations and the solution with lowest loss value was returned. Following [6], we used accuracy to measure the clustering performance. Please refer to [6] for the definition of accuracy. Table 7, 8 and 9 show the clustering accuracy on TDT, 20-News and Reuters test set respectively under different number K of hidden units.

As can be seen from these tables, with diversification, DRBM achieves significantly better clustering accuracy than the standard RBM. On TDT test data, the best accuracy of RBM is 23.3%. DRBM dramatically improves the accuracy to 52.4%. On Reuters test set, the best accuracy achieved by DRBM is 60.9%, which is largely better than the 47.6% accuracy achieved by RBM. The great performance gain achieved by DRBM attributes to the improved effectiveness of the learned representations. RBM lacks the ability to learn hidden units to cover long-tail topics, which largely inhibits its ability to learn rich and expressive representations. DRBM uses the diversity metric to regularize the hidden units to enhance their diversity. The learned hidden units under DRBM can not only represent dominant topics effectively, but also cover long-tail topics properly. Accordingly, the resultant representations can cover diverse semantics and dramatically improve the performance of k-means clustering.

DRBM can achieve a high accuracy with a fairly small number of hidden units, which greatly facilitates computational efficiency. For example, on TDT dataset, with 25 hidden units, DRBM can achieve an accuracy of 52.4%, which cannot be achieved by RBM with even 500 hidden units. The computational complexity of k-means is linear to the feature dimension. Thus, on this dataset, with the latent representations learned by DRBM, k-means can achieve a significant speed up. Similar observations can be seen from the other two datasets.

	TDT	20-News	Reuters
BOW	51.3	21.3	49.7
LDA [3]	45.2	21.9	51.2
DPP-LDA [17]	46.3	10.9	49.3
PYTM [22]	46.9	21.5	51.7
LIDA [1]	47.3	17.4	53.1
DocNADE [18]	45.7	18.7	48.7
PV [19]	48.2	24.3	52.8
RBM [13]	23.3	22.7	47.6
DRBM	**52.4**	**29.4**	**60.9**

Table 10: Clustering accuracy (%) on three datasets

K	25	50	100	200	500
RBM	2602	2603	2606	2609	2350
DRBM	**1699**	**1391**	**1658**	**1085**	**859**

Table 11: Perplexity on TDT test set

Table 10 presents the comparison of DRBM with the baseline methods on clustering accuracy. As can be seen from this table, our method consistently outperforms the baselines across all three datasets. The analysis of why DRBM is better than the baseline methods follows that presented in Section 4.2.

4.4 Perplexity on Testing Data

Following [13], we computed perplexity on the held out test set to assess the document modeling power of RBM and DRBM. Table 11, 12 and 13 compare the perplexity of RBM and DRBM computed on TDT, 20-News and Reuters dataset respectively. As can be seen from these tables, DRBM can achieve significant lower (better) perplexity than RBM, which corroborates that by diversifying the hidden units, the document modeling power of RBM can be dramatically improved.

4.5 Sensitivity to Parameters

We study the sensitivity of DRBM to the tradeoff parameter λ. Figure 2 shows how precision@100 in retrieval varies as λ increases on the TDT, 20-News and Reuters dataset respectively. The number of hidden units was fixed to 100. As can be seen from the figures, starting from 0, increasing λ improves precision@100. That is because a larger λ induces more diversity of the hidden units, enabling them to better cover long-tail topics. However, further increasing λ causes the precision to drop. This is because, if λ is too large, too much emphasis is paid to the diversify regularizer and the data likelihood of RBM is ignored.

4.6 Topic Visualization

Other than evaluating the learned hidden units of DRBM quantitatively, we also visualize and evaluate them in a qualitative way. We can interpret each hidden unit as a topic and its weight vector as a pseudo distribution over the vocabulary. To visualize each hidden unit, we pick up the top 10 representative words which correspond to the ten largest values in the weight vector. Table 14 shows 5 topics learned by RBM and 5 topics learned by DRBM. As can be seen from the table, topics learned by RBM have many near-duplicates and are very redundant. In contrast, the topics learned by DRBM are much more diverse, with a broad cov-

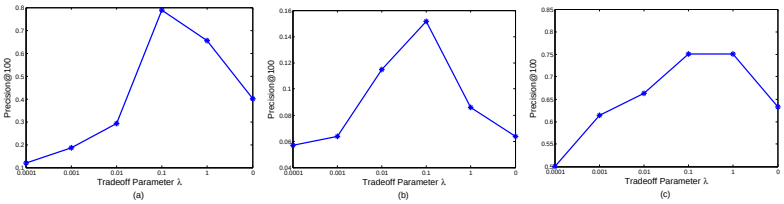

Figure 2: Sensitivity of DRBM to tradeoff parameter λ on (a) TDT dataset (b) 20-News dataset (c) Reuters dataset

Table 14: Exemplar topics learned By RBM and DRBM

RBM					DRBM				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
president	iraq	iraq	olympic	spkr	president	olympic	iraq	lawyers	students
clinton	united	un	games	voice	iraq	games	united	kaczynski	japanese
iraq	un	iraqi	nagano	tobacco	clinton	olympics	un	ms	japan
united	weapons	lewinsky	olympics	olympic	united	nagano	weapons	defense	school
spkr	iraqi	saddam	game	games	million	team	iraqi	trial	ms
house	nuclear	clinton	team	people	lewinsky	gold	baghdad	judge	united
people	india	baghdad	gold	olympics	thailand	game	council	people	yen
lewinsky	minister	inspectors	japan	nagano	spkr	hockey	inspectors	prosecutors	gm
government	saddam	weapons	medal	game	government	medal	nations	kaczynskis	tokyo
white	military	white	hockey	gold	jones	winter	military	government	south

K	25	50	100	200	500
RBM	764.9	765.1	765	741	633
DRBM	**713**	**623**	**659**	**558**	**497**

Table 12: Perplexity on 20-News test set

K	25	50	100	200	500
RBM	1147	1129	1130	881	849
DRBM	**1028**	**859**	**746**	**734**	**848**

Table 13: Perplexity on Reuters test set

erage of various topics including American politics, sports, Iraq war, law and Japanese education. This demonstrates the ability of DRBM to learn diverse topics.

5. CONCLUSIONS

In this paper, we study the problem of diversifying Restricted Boltzmann Machine. Due to the skewed distribution of topic popularity, existing RBM tends to learn redundant hidden units to represent dominant topics with best efforts and fails to learn hidden units to effectively cover long-tail topics. To solve this problem, we propose to diversify the hidden units to make them to cover not only dominant topics, but also long-tail topics. We define a diversity metric and use it to regularize the learning of the hidden units. Considering the diversity metric is not amenable for optimization, we instead optimize its lower bound. We prove that maximizing the lower bound with projected gradient ascent can increase the diversity metric. Experiments on document retrieval and clustering show that through diversification, the representations learned by DRBM can be greatly improved.

6. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for the helpful suggestions. This work is supported by the following grants to Eric P. Xing: ASFOR FA95501010247; NSF IIS1111142, IIS447676.

7. REFERENCES

[1] C. Archambeau, B. Lakshminarayanan, and G. Bouchard. Latent ibp compound dirichlet allocation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015.

[2] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*. Springer, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 2003.

[4] D. Cai and X. He. Manifold adaptive experimental design for text categorization. *Knowledge and Data Engineering, IEEE Transactions on*, 2012.

[5] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *Knowledge and Data Engineering, IEEE Transactions on*, 2005.

[6] D. Cai, X. He, and J. Han. Locally consistent concept factorization for document clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 2011.

[7] Z. Cao, S. Li, Y. Liu, W. Li, and H. Ji. A novel neural topic model and its supervised extension. *AAAI Conference on Artificial Intelligence*, 2015.

[8] Z. Chen and B. Liu. Mining topics in documents: standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.

[9] M. Denil, A. Demiraj, N. Kalchbrenner, P. Blunsom, and N. de Freitas. Modelling, visualising and summarising documents with a single convolutional neural network. *arXiv preprint arXiv:1406.3830*, 2014.

[10] J. H. Friedman, J. L. Bentley, and R. A. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 1977.

[11] R. Guha. Towards a model theory for distributed representations. *arXiv preprint arXiv:1410.5859*, 2014.

[12] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 2002.

[13] G. E. Hinton and R. R. Salakhutdinov. Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, 2009.

[14] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 1999.

[15] C. Huang, X. Qiu, and X. Huang. Text classification with document embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, 2014.

[16] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

[17] J. T. Kwok and R. P. Adams. Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems*. 2012.

[18] H. Larochelle and S. Lauly. A neural autoregressive topic model. In *Advances in Neural Information Processing Systems*, 2012.

[19] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *International Conference on Machine Learning*, 2014.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

[21] J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 1997.

[22] I. Sato and H. Nakagawa. Topic models with power-law using pitman-yor process. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.

[23] I. R. Shafarevich, A. Remizov, D. P. Kramer, and L. Nekludova. *Linear algebra and geometry*. Springer Science & Business Media, 2012.

[24] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. 1986.

[25] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning*, 2011.

[26] N. Srivastava, R. Salakhutdinov, and G. Hinton. Modeling documents with a deep boltzmann machine. In *Uncertainty in Artificial Intelligence*, 2013.

[27] Y. Wang, X. Zhao, Z. Sun, H. Yan, L. Wang, Z. Jin, L. Wang, Y. Gao, C. Law, and J. Zeng. Peacock: Learning long-tail topic features for industrial applications. *arXiv preprint arXiv:1405.4402*, 2014.

APPENDIX

A. PROOF OF LEMMA 1

To prove Lemma 1, the following lemma is needed.

LEMMA 4. *Let the weight vector $\tilde{\mathbf{a}}_i$ of hidden unit i be decomposed into $\tilde{\mathbf{a}}_i = \mathbf{x}_i + l_i\mathbf{e}_i$, where $\mathbf{x}_i = \sum_{j=1,j\neq i}^{K} \alpha_j \tilde{\mathbf{a}}_j$ lies in the subspace L spanned by $\{\tilde{\mathbf{a}}_1, \cdots, \tilde{\mathbf{a}}_K\} \backslash \{\tilde{\mathbf{a}}_i\}$, $\mathbf{e}_i$ is in the orthogonal complement of L, $\|\mathbf{e}_i\| = 1$, $\mathbf{e}_i \cdot \tilde{\mathbf{a}}_i > 0$, l_i is a scalar. Then $det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}}) = det(\widetilde{\mathbf{A}}_{-i}^\intercal \widetilde{\mathbf{A}}_{-i})(l_i \mathbf{e}_i \cdot \tilde{\mathbf{a}}_i)$, where $\widetilde{\mathbf{A}}_{-i} = [\tilde{\mathbf{a}}_1, \cdots, \tilde{\mathbf{a}}_{i-1}, \tilde{\mathbf{a}}_{i+1}, \cdots, \tilde{\mathbf{a}}_K]$ with $\tilde{\mathbf{a}}_i$ excluded..*

PROOF. Part of the proof follows [23].

$$\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}}) = \begin{vmatrix} \tilde{\mathbf{a}}_1 \cdot \tilde{\mathbf{a}}_1 & \cdots & \tilde{\mathbf{a}}_1 \cdot \tilde{\mathbf{a}}_i & \cdots & \tilde{\mathbf{a}}_1 \cdot \tilde{\mathbf{a}}_K \\ \tilde{\mathbf{a}}_2 \cdot \tilde{\mathbf{a}}_1 & \cdots & \tilde{\mathbf{a}}_2 \cdot \tilde{\mathbf{a}}_i & \cdots & \tilde{\mathbf{a}}_2 \cdot \tilde{\mathbf{a}}_K \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{a}}_K \cdot \tilde{\mathbf{a}}_1 & \cdots & \tilde{\mathbf{a}}_K \cdot \tilde{\mathbf{a}}_i & \cdots & \tilde{\mathbf{a}}_K \cdot \tilde{\mathbf{a}}_K \end{vmatrix} \tag{7}$$

Let $\mathbf{c}_i$ denote the ith column of the Gram matrix $\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}}$. Subtracting $\sum_{j=1,j\neq i}^{k} \alpha_j \mathbf{c}_j$ from $\mathbf{c}_i$ [23], where α_j is the linear coefficient in $\mathbf{x}_i$, we get

$$\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}}) = \begin{vmatrix} \tilde{\mathbf{a}}_1 \cdot \tilde{\mathbf{a}}_1 & \cdots & 0 & \cdots & \tilde{\mathbf{a}}_1 \cdot \tilde{\mathbf{a}}_K \\ \tilde{\mathbf{a}}_2 \cdot \tilde{\mathbf{a}}_1 & \cdots & 0 & \cdots & \tilde{\mathbf{a}}_2 \cdot \tilde{\mathbf{a}}_K \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{a}}_i \cdot \tilde{\mathbf{a}}_i & \cdots & l_i \mathbf{e}_i \cdot \tilde{\mathbf{a}}_i & \cdots & \tilde{\mathbf{a}}_i \cdot \tilde{\mathbf{a}}_K \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{a}}_K \cdot \tilde{\mathbf{a}}_1 & \cdots & 0 & \cdots & \tilde{\mathbf{a}}_K \cdot \tilde{\mathbf{a}}_K \end{vmatrix} \tag{8}$$

Expanding the determinant according to the ith column, we get $\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}}) = \det(\widetilde{\mathbf{A}}_{-i}^\intercal \widetilde{\mathbf{A}}_{-i})(l_i \mathbf{e}_i \cdot \tilde{\mathbf{a}}_i)$. □

Now we proceed to prove Lemma 1.

PROOF. The diversity metric $\Omega(\widetilde{\mathbf{A}})$ comprises of two terms: $\Omega(\widetilde{\mathbf{A}}) = \Psi(\widetilde{\mathbf{A}}) - \Pi(\widetilde{\mathbf{A}})$, in which $\Psi(\widetilde{\mathbf{A}})$ and $\Pi(\widetilde{\mathbf{A}})$ measure the mean and variance of the pairwise angles respectively. We bound the two terms separately. We first bound the mean $\Psi(\widetilde{\mathbf{A}})$. Since the weight vectors are assumed to be linearly independent, we have $\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}}) > 0$. As $l_i \mathbf{e}_i \cdot \tilde{\mathbf{a}}_i \leq \|l_i \mathbf{e}_i\| \|\tilde{\mathbf{a}}_i\| \leq 1$ and $\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}}) = \det(\widetilde{\mathbf{A}}_{-i}^\intercal \widetilde{\mathbf{A}}_{-i})(l_i \mathbf{e}_i \cdot \tilde{\mathbf{a}}_i)$ (according to Lemma 4), we have $\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}}) \leq \det(\widetilde{\mathbf{A}}_{-i}^\intercal \widetilde{\mathbf{A}}_{-i})$. As $\forall j, \det(\tilde{\mathbf{a}}_j^\intercal \tilde{\mathbf{a}}_j) = 1$, we can eliminate the columns of $\widetilde{\mathbf{A}}_{-i}$ and apply the inequality repeatedly to draw the conclusion that $\det(\widetilde{\mathbf{A}}_{-i}^\intercal \widetilde{\mathbf{A}}_{-i}) \leq 1$ (and $\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}}) \leq 1$). So $l_i \mathbf{e}_i \cdot \tilde{\mathbf{a}}_i = \|l_i \mathbf{e}_i\|^2 \geq \det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})$. For any $j \neq i$, the pairwise angle between $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$ is:

$$\begin{aligned} &\theta(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j) = \arccos(|\tilde{\mathbf{a}}_i \cdot \tilde{\mathbf{a}}_j|) = \arccos(|\mathbf{x}_i \cdot \tilde{\mathbf{a}}_j|) \\ &\leq \arccos(\|\mathbf{x}_i\| \|\tilde{\mathbf{a}}_j\|) = \arccos(\|\mathbf{x}_i\|) = \arccos(\sqrt{1 - \|l_i \mathbf{e}_i\|^2}) \\ &\geq \arccos(\sqrt{1 - \det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})}) = \arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})}) \end{aligned} \tag{9}$$

Thus $\Psi(\widetilde{\mathbf{A}}) \geq \arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})})$.

Now we bound the variance $\Pi(\widetilde{\mathbf{A}})$. For any $i \neq j$, we have proved that $\theta(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j) \geq \arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})})$. From the definition of $\theta(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j)$, we also have $\theta(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j) \leq \frac{\pi}{2}$. As $\Psi(\widetilde{\mathbf{A}})$ is the mean value of all pairwise $\theta(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j)$, we have $\arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})}) \leq \Psi(\widetilde{\mathbf{A}}) \leq \frac{\pi}{2}$. So $|\theta(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j) - \Psi(\widetilde{\mathbf{A}})| \leq \frac{\pi}{2} - \arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})})$. So $\Pi(\widetilde{\mathbf{A}}) \leq (\frac{\pi}{2} - \arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})}))^2$.

Combining the lower bound of $\Psi(\widetilde{\mathbf{A}})$ and upper bound of $\Pi(\widetilde{\mathbf{A}})$, we have $\Omega(\widetilde{\mathbf{A}}) \geq \Gamma(\widetilde{\mathbf{A}}) = \arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})}) - (\frac{\pi}{2} - \arcsin(\sqrt{\det(\widetilde{\mathbf{A}}^\intercal \widetilde{\mathbf{A}})}))^2$. Both $\Omega(\widetilde{\mathbf{A}})$ and $\Gamma(\widetilde{\mathbf{A}})$ obtain the optimal value of $\pi/2$ when the weight vectors in $\widetilde{\mathbf{A}}$ are orthogonal to each other. The proof completes. □

B. PROOF OF LEMMA 2

PROOF. According to chain rule, the gradient of $\Gamma(\widetilde{\mathbf{A}})$ w.r.t $\tilde{\mathbf{a}}_\mathbf{i}$ can be written as $g'(\det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}))\frac{\partial \det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})}{\partial \tilde{\mathbf{a}}_\mathbf{i}}$, where $g(x) = \arcsin(\sqrt{x}) - (\frac{\pi}{2} - \arcsin(\sqrt{x}))^2$. It is easy to check that $g(x)$ is an increasing function and $g'(x) > 0$. Now we discuss the $\frac{\partial \det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})}{\partial \tilde{\mathbf{a}}_\mathbf{i}}$ term. According to Lemma 4, we have $\det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}) = \det(\widetilde{\mathbf{A}}^\top_{-i}\widetilde{\mathbf{A}}_{-i})l_i\mathbf{e}_i \cdot \tilde{\mathbf{a}}_\mathbf{i}$. From this equation, we have $\frac{\partial \det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})}{\partial \tilde{\mathbf{a}}_\mathbf{i}} = \det(\widetilde{\mathbf{A}}^\top_{-i}\widetilde{\mathbf{A}}_{-i})l_i\mathbf{e}_i$. As assumed earlier, the weight vectors in $\widetilde{\mathbf{A}}$ are linearly independent and hence $\det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}) > 0$ and $\det(\widetilde{\mathbf{A}}^\top_{-i}\widetilde{\mathbf{A}}_{-i}) > 0$. From $\det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}) = \det(\widetilde{\mathbf{A}}^\top_{-i}\widetilde{\mathbf{A}}_{-i})l_i\mathbf{e}_i \cdot \tilde{\mathbf{a}}_i$ and $\mathbf{e}_i \cdot \tilde{\mathbf{a}}_i > 0$, we know $l_i > 0$. Overall, the gradient of $\Gamma(\widetilde{\mathbf{A}})$ w.r.t $\tilde{\mathbf{a}}_\mathbf{i}$ can be written as $k_i\mathbf{e}_i$, where $k_i = g'(\det(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}))\det(\widetilde{\mathbf{A}}^\top_{-i}\widetilde{\mathbf{A}}_{-i})l_i > 0$. The proof completes. □

C. PROOF OF THEOREM 2

To prove Theorem 2, we first introduce some notations. Let $V = \{(i,j)|1 \le i,j \le K, i \ne j, \tilde{\mathbf{a}}_i^{(t)} \cdot \tilde{\mathbf{a}}_j^{(t)} = 0\}$, $N = \{(i,j)|1 \le i,j \le K, i \ne j, \tilde{\mathbf{a}}_i^{(t)} \cdot \tilde{\mathbf{a}}_j^{(t)} \ne 0\}$, where $\tilde{\mathbf{a}}_i^{(t)}$ is the ith column of $\mathbf{A}^{(t)}$. Let $\Delta V = \sum_{(i,j)\in V}(\theta(\tilde{\mathbf{a}}_i^{(t+1)}, \tilde{\mathbf{a}}_j^{(t+1)}) - \theta(\tilde{\mathbf{a}}_i^{(t)}, \tilde{\mathbf{a}}_j^{(t)}))$, $\Delta N = \sum_{(i,j)\in N}(\theta(\tilde{\mathbf{a}}_i^{(t+1)}, \tilde{\mathbf{a}}_j^{(t+1)}) - \theta(\tilde{\mathbf{a}}_i^{(t)}, \tilde{\mathbf{a}}_j^{(t)}))$, then $\Omega(\widetilde{\mathbf{A}}^{(t+1)}) - \Omega(\widetilde{\mathbf{A}}^{(t)}) = \Delta V + \Delta N$. Let $x_{ij}^{(t)} = |\tilde{\mathbf{a}}_i^{(t)} \cdot \tilde{\mathbf{a}}_j^{(t)}|$, $x_{ij}^{(t+1)} = |\tilde{\mathbf{a}}_i^{(t+1)} \cdot \tilde{\mathbf{a}}_j^{(t+1)}|$. According to Lemma 2 presented in the main paper, $\tilde{\mathbf{a}}_i^{(t+1)} = \frac{\tilde{\mathbf{a}}_i^{(t)} + \eta k_i \mathbf{e}_i}{\|\tilde{\mathbf{a}}_i^{(t)} + \eta k_i \mathbf{e}_i\|}$, $\tilde{\mathbf{a}}_j^{(t+1)} = \frac{\tilde{\mathbf{a}}_j^{(t)} + \eta k_j \mathbf{e}_j}{\|\tilde{\mathbf{a}}_j^{(t)} + \eta k_j \mathbf{e}_j\|}$ and $\mathbf{e}_i \cdot \tilde{\mathbf{a}}_j^{(t)} = 0$, $\mathbf{e}_j \cdot \tilde{\mathbf{a}}_i^{(t)} = 0$. According to the definition $\tilde{\mathbf{a}}_i = \mathbf{x}_i + l_i\mathbf{e}_i$ in Lemma 4, we have $\|\tilde{\mathbf{a}}_i^{(t)} + \eta k_i \mathbf{e}_i\| = \sqrt{1 + 2l_ik_i\eta + k_i^2\eta^2}$, and $x_{ij}^{(t+1)} = \frac{|\tilde{\mathbf{a}}_i^{(t)} \cdot \tilde{\mathbf{a}}_j^{(t)} + \eta^2 k_i k_j \mathbf{e}_i^{(t)} \cdot \mathbf{e}_j^{(t)}|}{\sqrt{1+2l_ik_i\eta+k_i^2\eta^2}\sqrt{1+2l_jk_j\eta+k_j^2\eta^2}}$. The following lemmas are needed for proving Theorem 2.

LEMMA 5. *$\forall (i,j) \in V$, we have $\theta(\tilde{\mathbf{a}}_i^{(t+1)}, \tilde{\mathbf{a}}_j^{(t+1)}) - \theta(\tilde{\mathbf{a}}_i^{(t)}, \tilde{\mathbf{a}}_j^{(t)}) = o(\eta)$, where $\lim_{\eta\to 0}\frac{o(\eta)}{\eta} = 0$.*

PROOF. For $(i,j) \in V$, $\tilde{\mathbf{a}}_i^{(t)} \cdot \tilde{\mathbf{a}}_j^{(t)} = 0$, thereby $x_{ij}^{(t)} = 0$ and $\theta(\tilde{\mathbf{a}}_i^{(t+1)}, \tilde{\mathbf{a}}_j^{(t+1)}) - \theta(\tilde{\mathbf{a}}_i^{(t)}, \tilde{\mathbf{a}}_j^{(t)}) = \arccos(x_{ij}^{(t+1)}) - \arccos(x_{ij}^{(t)}) = \arccos(x_{ij}^{(t+1)}) - \frac{\pi}{2}$. Now we prove $\lim_{\eta\to 0}\frac{\arccos(x_{ij}^{(t+1)}) - \frac{\pi}{2}}{\eta} = 0$. Plugging in $\tilde{\mathbf{a}}_i^{(t)} \cdot \tilde{\mathbf{a}}_j^{(t)} = 0$ into $x_{ij}^{(t+1)}$, we have $x_{ij}^{(t+1)} = \frac{|\eta^2 k_i k_j \mathbf{e}_i^{(t)} \cdot \mathbf{e}_j^{(t)}|}{\sqrt{1+2l_ik_i\eta+k_i^2\eta^2}\sqrt{1+2l_jk_j\eta+k_j^2\eta^2}}$. Thereby $\lim_{\eta\to 0}\frac{x_{ij}^{(t+1)}}{\eta} = 0$ (equivalently, $x_{ij}^{(t+1)} = o(\eta)$) and $\lim_{\eta\to 0} x_{ij}^{(t+1)} = 0$. According to the Taylor expansion of $\arccos(x)$ at $x = 0$, $\arccos(x) = \frac{\pi}{2} - x + o(x)$, so $\lim_{x\to 0}\frac{\arccos(x) - \frac{\pi}{2}}{x} = -1$. Since $\lim_{\eta\to 0} x_{ij}^{(t+1)} = 0$, $\lim_{\eta\to 0}\frac{\arccos(x_{ij}^{(t+1)}) - \frac{\pi}{2}}{x_{ij}^{(t+1)}} = \lim_{x_{ij}^{(t+1)}\to 0}\frac{\arccos(x_{ij}^{(t+1)}) - \frac{\pi}{2}}{x_{ij}^{(t+1)}} = -1$. Since $\lim_{\eta\to 0}\frac{x_{ij}^{(t+1)}}{\eta} = 0$, we have $\lim_{\eta\to 0}\frac{\arccos(x_{ij}^{(t+1)}) - \frac{\pi}{2}}{\eta}$ $= \lim_{\eta\to 0}\frac{\arccos(x_{ij}^{(t+1)}) - \frac{\pi}{2}}{x_{ij}^{(t+1)}}\frac{x_{ij}^{(t+1)}}{\eta} = 0$. The proof completes. □

LEMMA 6. *$\forall (i,j) \in N$, $\exists c_{ij} > 0$, such that $\theta(\tilde{\mathbf{a}}_i^{(t+1)}, \tilde{\mathbf{a}}_j^{(t+1)}) - \theta(\tilde{\mathbf{a}}_i^{(t)}, \tilde{\mathbf{a}}_j^{(t)}) = c_{ij}\eta + o(\eta)$, where $\lim_{\eta\to 0}\frac{o(\eta)}{\eta} = 0$.*

PROOF. Using the Taylor expansion of $\arccos(x)$ at $x = x_{ij}^{(t)}$, we have

$$\begin{aligned}&\arccos(x_{ij}^{(t+1)}) - \arccos(x_{ij}^{(t)})\\ &= -\frac{1}{\sqrt{1-{x_{ij}^{(t)}}^2}}(x_{ij}^{(t+1)} - x_{ij}^{(t)}) + o(x_{ij}^{(t+1)} - x_{ij}^{(t)})\end{aligned} \tag{10}$$

According to the definition of $x_{ij}^{(t+1)}$, we have

$$x_{ij}^{(t+1)} = x_{ij}^{(t)}\frac{|1+\eta^2\frac{k_ik_j\mathbf{e}_i^{(t)}\cdot\mathbf{e}_j^{(t)}}{\tilde{\mathbf{a}}_i^{(t)}\cdot\tilde{\mathbf{a}}_j^{(t)}}|}{\sqrt{1+2l_ik_i\eta+k_i^2\eta^2}\sqrt{1+2l_jk_j\eta+k_j^2\eta^2}} \tag{11}$$

Using the Taylor expansion of $\frac{1}{\sqrt{1+x}}$ at $x = 0$, we can obtain that $\frac{1}{\sqrt{1+2l_ik_i\eta+k_i^2\eta^2}} = 1 - \frac{1}{2}(2l_ik_i\eta + k_i^2\eta^2) + o(2l_ik_i\eta + k_i^2\eta^2)$. As $\eta^2 = o(\eta)$ and $o(2l_ik_i\eta + k_i^2\eta^2) = o(\eta)$, we can obtain that $\frac{1}{\sqrt{1+2l_ik_i\eta+k_i^2\eta^2}} = 1 - l_ik_i\eta + o(\eta)$. Similarly, $\frac{1}{\sqrt{1+2l_jk_j\eta+k_j^2\eta^2}} = 1 - l_jk_j\eta + o(\eta)$. When $\eta \to 0$, $|1+\eta^2\frac{k_ik_j\mathbf{e}_i^{(t)}\cdot\mathbf{e}_j^{(t)}}{\tilde{\mathbf{a}}_i^{(t)}\cdot\tilde{\mathbf{a}}_j^{(t)}}| = 1$. Thereby when η is small enough, $|1+\eta^2\frac{k_ik_j\mathbf{e}_i^{(t)}\cdot\mathbf{e}_j^{(t)}}{\tilde{\mathbf{a}}_i^{(t)}\cdot\tilde{\mathbf{a}}_j^{(t)}}| = 1 + \eta^2\frac{k_ik_j\mathbf{e}_i^{(t)}\cdot\mathbf{e}_j^{(t)}}{\tilde{\mathbf{a}}_i^{(t)}\cdot\tilde{\mathbf{a}}_j^{(t)}}$, so $|1+\eta^2\frac{k_ik_j\mathbf{e}_i^{(t)}\cdot\mathbf{e}_j^{(t)}}{\mathbf{a}_i^{(t)}\cdot\mathbf{a}_j^{(t)}}| = 1 + o(\eta)$. Substituting the above equations to $x_{ij}^{(t+1)}$, we can obtain that $x_{ij}^{(t+1)} - x_{ij}^{(t)} = x_{ij}^{(t)}((1+o(\eta))(1 - l_ik_i\eta + o(\eta))(1 - l_jk_j\eta + o(\eta)) - 1) = -x_{ij}^{(t)}(l_ik_i + l_jk_j)\eta + o(\eta)$. So $\lim_{\eta\to 0}\frac{x_{ij}^{(t+1)} - x_{ij}^{(t)}}{\eta} = -x_{ij}^{(t)}(l_ik_i + l_jk_j)$. As $\lim_{\eta\to 0}\frac{o(x_{ij}^{(t+1)} - x_{ij}^{(t)})}{x_{ij}^{(t+1)} - x_{ij}^{(t)}} = \lim_{x_{ij}^{(t+1)}\to x_{ij}^{(t)}}\frac{o(x_{ij}^{(t+1)} - x_{ij}^{(t)})}{x_{ij}^{(t+1)} - x_{ij}^{(t)}} = 0$, we can draw the conclusion that $\lim_{\eta\to 0}\frac{o(x_{ij}^{(t+1)} - x_{ij}^{(t)})}{\eta} = \lim_{\eta\to 0}\frac{o(x_{ij}^{(t+1)} - x_{ij}^{(t)})}{x_{ij}^{(t+1)} - x_{ij}^{(t)}}\frac{x_{ij}^{(t+1)} - x_{ij}^{(t)}}{\eta} = 0$, hence $o(x_{ij}^{(t+1)} - x_{ij}^{(t)}) = o(\eta)$. So $\arccos(x_{ij}^{(t+1)}) - \arccos(x_{ij}^{(t)}) = -\frac{1}{\sqrt{1-{x_{ij}^{(t)}}^2}}(x_{ij}^{(t+1)} - x_{ij}^{(t)}) + o(x_{ij}^{(t+1)} - x_{ij}^{(t)}) = -\frac{1}{\sqrt{1-{x_{ij}^{(t)}}^2}}(-x_{ij}^{(t)}(l_ik_i + l_jk_j)\eta + o(\eta)) + o(\eta) = \frac{x_{ij}^{(t)}(l_ik_i + l_jk_j)}{\sqrt{1-{x_{ij}^{(t)}}^2}}\eta + o(\eta)$. Let $c_{ij} = \frac{x_{ij}^{(t)}(l_ik_i + l_jk_j)}{\sqrt{1-{x_{ij}^{(t)}}^2}}$, clearly $c_{ij} > 0$. The proof completes. □

Given these two lemmas, we can prove Theorem 2 now.

PROOF.

$$\begin{aligned}&\Psi(\widetilde{\mathbf{A}}^{(t+1)}) - \Omega(\widetilde{\mathbf{A}}^{(t)}) = \Delta V + \Delta N\\ &= \textstyle\sum_{(i,j)\in V} o(\eta) + \sum_{(i,j)\in N}(c_{ij}\eta + o(\eta))\\ &= o(\eta) + \textstyle\sum_{(i,j)\in N} c_{ij}\eta\end{aligned} \tag{12}$$

$\lim_{\eta\to 0}\frac{\Psi(\widetilde{\mathbf{A}}^{(t+1)}) - \Psi(\widetilde{\mathbf{A}}^{(t)})}{\eta} = \lim_{\eta\to 0}\frac{o(\eta) + \sum_{(i,j)\in N} c_{ij}\eta}{\eta} = \sum_{(i,j)\in N} c_{ij} > 0$. So $\exists \tau > 0$ such that $\forall \eta \in (0,\tau)$ we have $\frac{\Psi(\widetilde{\mathbf{A}}^{(t+1)}) - \Psi(\widetilde{\mathbf{A}}^{(t)})}{\eta} \ge \frac{1}{2}\sum_{(i,j)\in N} c_{ij} > 0$. The proof completes. □

D. PROOF SKETCH OF LEMMA 3

Due to space limit, we present the proof sketch of Lemma 3 here. Please refer to the external supplementary material[3] for the detailed proof.

[3]The proofs are available at `http://www.cs.cmu.edu/~pengtaox/papers/kdd15_supp.pdf`

PROOF. First, we construct a sequence of sequences with decreasing variance, in which the variance of the first sequence is var(b) and the variance of the last sequence is var(c). We sort the unique values in b in ascending order and denote the resultant sequence as $d = (d_j)_{j=1}^m$. Let $l(j) = \max\{i : b_i = d_j\}$, $u(i) = \{j : d_j = b_i\}$, we construct a sequence of sequences $h^{(j)} = (h_i^{(j)})_{i=1}^n$ where $j = 1, 2, \cdots, m+1$, in the following way:

- $h_i^{(1)} = b_i$, $i = 1, \cdots, n$;
- $h_i^{(j+1)} = h_i^{(j)}$, $j = 1, \cdots, m$ and $l(m-j+1) < i \leq n$;
- $h_i^{(2)} = h_i^{(1)} + g(d_m)$, $1 \leq i \leq l(m)$;
- $h_i^{(j+1)} = h_i^{(j)} + g(d_{m-j+1}) - g(d_{m-j+2})$, $j = 2, \cdots, m$ and $1 \leq i \leq l(m-j+1)$.

From the definition of $h^{(j)}$, we know var($h^{(1)}$) = var(b). As $b_1 < b_n$, we have $m \geq 2$. We can prove that var($h^{(m+1)}$) = var(c) and $\forall j = 1, 2, \cdots, m$, var($h^{(j+1)}$) < var($h^{(j)}$), which further imply var(c) < var($h^{(1)}$) = var(b). Furthermore, let $n' = \max\{j : b_j \neq b_n\}$, then $\forall i$,

$$\begin{array}{l} h_i^{(2)} = h_i^{(1)} + g(d_m) = b_i + g(b_n) \\ h_i^{(3)} = h_i^{(2)} + (g(d_{m-1}) - g(d_m))\mathbb{I}(i \leq l(m-1)) \\ = b_i + g(b_n) + (g(b_{n'}) - g(b_n))\mathbb{I}(i \leq n') = b'_i \end{array} \quad (13)$$

so var(c) $\leq$ var(b') < var(b). The proof completes. □

E. PROOF OF THEOREM 3

The intuition is when the stepsize η is sufficiently small, we can make sure the changes of smaller angles (between consecutive iterations) are larger than the changes of larger angles, then Lemma 3 can be used to prove that the variance decreases. The proof of Theorem 3 utilizes Lemma 5 and 6.

PROOF. Let $\theta_{ij}^{(t)}$ denote $\theta(\tilde{\mathbf{a}}_i^{(t)}, \tilde{\mathbf{a}}_j^{(t)})$. We sort $\theta_{ij}^{(t)}$ in nondecreasing order and denote the resultant sequence as $\theta^{(t)} = (\theta_k^{(t)})_{k=1}^n$, then var($(\theta_{ij}^{(t)})$) = var($\theta^{(t)}$). We use the same order to index $\theta_{ij}^{(t+1)}$ and denote the resultant sequence as $\theta^{(t+1)} = (\theta_k^{(t+1)})_{k=1}^n$, then var($(\theta_{ij}^{(t+1)})$) = var($\theta^{(t+1)}$). Let $g(\theta_{ij}^{(t)}) = \frac{2\cos(\theta_{ij}^{(t)})}{\sqrt{1-\cos(\theta_{ij}^{(t)})^2}}\eta$ if $\theta_{ij}^{(t)} < \frac{\pi}{2}$ and 0 if $\theta_{ij}^{(t)} = \frac{\pi}{2}$, then $g(\theta_{ij}^{(t)})$ is a strictly decreasing function. Let $\tilde{\theta}_k^{(t)} = \theta_k^{(t)} + c_k\eta = \theta_k^{(t)} + g(\theta_k^{(t)})$. It is easy to see when η is sufficiently small, $0 \leq g(\theta_k^{(t)}) \leq \min\{\theta_{k+1}^{(t)} - \theta_k^{(t)} : k = 1, 2, \cdots, n-1, \theta_{k+1}^{(t)} \neq \theta_k^{(t)}\}$. We continue the proof from two complementary cases: (1) $\theta_1^{(t)} < \theta_n^{(t)}$; (2) $\theta_1^{(t)} = \theta_n^{(t)}$. If $\theta_1^{(t)} < \theta_n^{(t)}$, then according to Lemma 3, we have var($\tilde{\theta}^{(t)}$) < var($\theta^{(t)}$), where $\tilde{\theta}^{(t)} = (\tilde{\theta}_k^{(t)})_{k=1}^n$. Furthermore, let $n' = \max\{j : \theta_j^{(t)} \neq \theta_n^{(t)}\}$, $\theta_k'^{(t)} = \theta_k^{(t)} + g(\theta_n^{(t)}) + (g(\theta_{n'}^{(t)}) - g(\theta_n^{(t)}))\mathbb{I}(k \leq n')$, then var($\tilde{\theta}^{(t)}$) $\leq$ var($\theta'^{(t)}$) < var($\theta^{(t)}$), where $\theta'^{(t)} = (\theta_k'^{(t)})_{k=1}^n$. var($\theta'^{(t)}$) can be written as:

$$\begin{array}{rcl} & & \\ & = & \frac{1}{n}\sum_{i=1}^n(\theta_i'^{(t)} - \frac{1}{n}\sum_{j=1}^n\theta_j'^{(t)})^2 \\ & = & \frac{1}{n}\sum_{i=1}^n(\theta_i^{(t)} + (g(\theta_{n'}^{(t)}) - g(\theta_n^{(t)}))\mathbb{I}(i \leq n') \\ & & -\frac{1}{n}\sum_{j=1}^n\theta_j^{(t)} - \frac{n'}{n}(g(\theta_{n'}^{(t)}) - g(\theta_n^{(t)})))^2 \\ & = & \mathrm{var}(\theta^{(t)}) + 2(\frac{1}{n}\sum_{i=1}^n(\theta_i^{(t)} - \frac{1}{n}\sum_{j=1}^n\theta_j^{(t)})(g(\theta_{n'}^{(t)}) \\ & & -g(\theta_n^{(t)}))(\mathbb{I}(i \leq n') - \frac{n'}{n}) + \frac{1}{n}\sum_{j=1}^n(g(\theta_{n'}^{(t)}) \\ & & -g(\theta_n^{(t)}))^2(\mathbb{I}(i \leq n') - \frac{n'}{n})^2 \end{array} \quad (14)$$

Let $\lambda = 2(\frac{1}{n}\sum_{i=1}^n(\theta_i^{(t)} - \frac{1}{n}\sum_{j=1}^n\theta_j^{(t)})(\mathbb{I}(i \leq n') - \frac{n'}{n})$, it can be further written as

$$\begin{array}{rcl} & & \\ & = & \frac{2}{n}(\sum_{i=1}^{n'}(\theta_i^{(t)} - \frac{1}{n}\sum_{j=1}^n\theta_j^{(t)})(1 - \frac{n'}{n}) \\ & & +\frac{2}{n}(\sum_{i=n'+1}^n(\theta_i^{(t)} - \frac{1}{n}\sum_{j=1}^n\theta_j^{(t)})(-\frac{n'}{n}) \\ & = & \frac{2}{n}((\sum_{i=1}^{n'}\theta_i^{(t)} - \frac{n'}{n}\sum_{j=1}^n\theta_j^{(t)})(1 - \frac{n'}{n}) \\ & & +\frac{2}{n}((\sum_{i=n'+1}^n\theta_i^{(t)} - \frac{n-n'}{n}\sum_{j=1}^n\theta_j^{(t)})(-\frac{n'}{n}) \\ & = & \frac{2n'(n-n')}{n^2}(\frac{1}{n'}\sum_{i=1}^{n'}\theta_i^{(t)} - \frac{1}{n-n'}\sum_{i=n'+1}^n\theta_i^{(t)}) \end{array} \quad (15)$$

As $\theta_k^{(t)}$ is nondecreasing and $\theta_n^{(t)} \neq \theta_{n'}^{(t)}$, we have $\lambda < 0$. Let $\mu = \frac{2\cos(\theta_{n'}^{(t)})}{\sqrt{1-\cos(\theta_{n'}^{(t)})^2}} - \frac{2\cos(\theta_n^{(t)})}{\sqrt{1-\cos(\theta_n^{(t)})^2}}$ when $\theta_n^{(t)} < \frac{\pi}{2}$ and $\mu = \frac{2\cos(\theta_{n'}^{(t)})}{\sqrt{1-\cos(\theta_{n'}^{(t)})^2}}$ when $\theta_n^{(t)} = \frac{\pi}{2}$, then $g(\theta_{n'}^{(t)}) - g(\theta_n^{(t)}) = \mu\eta$ and $\mu > 0$. Substituting λ and μ into var($\theta'^{(t)}$), we can obtain:

$$\begin{array}{rcl} \mathrm{var}(\theta'^{(t)}) & = & \mathrm{var}(\theta^{(t)}) + \lambda\mu\eta + \frac{1}{n}\sum_{j=1}^n(\mathbb{I}(i \leq n') - \frac{n'}{n})^2\mu^2\eta^2 \\ & = & \mathrm{var}(\theta^{(t)}) + \lambda\mu\eta + o(\eta) \end{array}$$

Note that $\lambda < 0$ and $\mu > 0$, so $\exists\delta_1$, such that $\eta < \delta_1 \Rightarrow$ var($\theta'^{(t)}$) < var($\theta^{(t)}$) + $\frac{\lambda\mu}{2}\eta$. As var($\tilde{\theta}^{(t)}$) < var($\theta'^{(t)}$), we can draw the conclusion that var($\tilde{\theta}^{(t)}$) < var($\theta^{(t)}$) + $\frac{\lambda\mu}{2}\eta$. On the other hand,

$$\begin{array}{rcl} \mathrm{var}(\theta^{(t+1)}) & = & \frac{1}{n}\sum_{i=1}^n(\theta_i^{(t+1)} - \frac{1}{n}\sum_{j=1}^n\theta_j^{(t+1)})^2 \\ & = & \frac{1}{n}\sum_{i=1}^n(\tilde{\theta}_i^{(t)} + o(\eta) - \frac{1}{n}\sum_{j=1}^n\tilde{\theta}_j^{(t)} + o(\eta))^2 \\ & = & \frac{1}{n}\sum_{i=1}^n(\tilde{\theta}_i^{(t)} - \frac{1}{n}\sum_{j=1}^n\tilde{\theta}_j^{(t)})^2 + o(\eta) \\ & = & \mathrm{var}(\tilde{\theta}^{(t)}) + o(\eta) \end{array}$$

So $\exists\delta_2 > 0$ such that $\eta < \delta_2 \Rightarrow$ var($\theta^{(t+1)}$) < var($\tilde{\theta}^{(t)}$) − $\frac{\lambda\mu}{4}\eta$. Let $\delta = \min\{\delta_1, \delta_2\}$, then

$$\begin{array}{rcl} \eta < \delta & \Rightarrow & \mathrm{var}(\theta^{(t+1)}) < \mathrm{var}(\theta^{(t)}) + \frac{\lambda\mu}{4}\eta < \mathrm{var}(\theta^{(t)}) \\ & \Rightarrow & \mathrm{var}((\theta^{(t+1)})) < \mathrm{var}(\theta^{(t)}) \end{array}$$

For the second case $\theta_1^{(t)} = \theta_n^{(t)}$, i.e., $\forall(i_1, j_1), (i_2, j_2) \in N \cup V$, $\theta_{i_1j_1}^{(t)} = \theta_{i_2j_2}^{(t)}$, we prove that var($\theta^{(t+1)}$) = var($\theta^{(t)}$). In this case, $\forall(i_1, j_1), (i_2, j_2) \in N \cup V$, $((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)})_{i_1j_1} = \tilde{\mathbf{a}}_{i_1}^{(t)} \cdot \tilde{\mathbf{a}}_{j_1}^{(t)} = \tilde{\mathbf{a}}_{i_2}^{(t)} \cdot \tilde{\mathbf{a}}_{j_2}^{(t)} = ((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)})_{i_2j_2}$. Denote $p_1 = \tilde{\mathbf{a}}_i^{(t)} \cdot \tilde{\mathbf{a}}_j^{(t)}$ for $i \neq j$ and $p_2 = \tilde{\mathbf{a}}_i^{(t)} \cdot \tilde{\mathbf{a}}_j^{(t)}$ for $i = j$. As $\widetilde{\mathbf{A}}^{(t+1)} = \widetilde{\mathbf{A}}^{(t)} + c\widetilde{\mathbf{A}}^{(t)}((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)})^{-1}$, where $c = 2\eta g'(\det((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)}))$ $\det((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)})$ and $g(x) = \arcsin(\sqrt{x}) - (\frac{\pi}{2} - \arcsin(\sqrt{x}))^2$, we have $(\widetilde{\mathbf{A}}^{(t+1)})^\intercal\widetilde{\mathbf{A}}^{(t+1)} = (\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)} + 2c\mathbf{I} + c^2((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)})^{-1}$. It is clear that $\forall(i_1, j_1), (i_2, j_2) \in N \cup V$, $((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)} + 2c\mathbf{I})_{i_1j_1} = ((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)} + 2c\mathbf{I})_{i_2j_2}$. For $c^2((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)})^{-1}$, write it as $c^2((p_2 - p_1)\mathbf{I}_K + p_1\mathbf{1}_K\mathbf{1}_K^\intercal)^{-1}$, where $\mathbf{I}_K$ is the identity matrix and $\mathbf{1}_K$ is a vector of 1s whose length is K. Applying Sherman-Morrison formula, we can obtain that $((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)})^{-1} = ((p_2 - p_1)^{-1}\mathbf{I}_K - \frac{(p_2-p_1)^{-1}\mathbf{1}_K\mathbf{1}_K^T}{1+K(p_2-p_1)})$ which implies that $\forall(i_1, j_1), (i_2, j_2) \in N \cup V$, $((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)})_{i_1j_1}^{-1} = ((\widetilde{\mathbf{A}}^{(t)})^\intercal\widetilde{\mathbf{A}}^{(t)})_{i_2j_2}^{-1}$, so $((\widetilde{\mathbf{A}}^{(t+1)})^\intercal\widetilde{\mathbf{A}}^{(t+1)})_{i_1j_1}$ $= ((\widetilde{\mathbf{A}}^{(t+1)})^\intercal\widetilde{\mathbf{A}}^{(t+1)})_{i_2j_2}$, so var($\theta^{(t+1)}$) = 0 = var($\theta^{(t)}$).

Putting these two cases together, we conclude that $\exists\tau_2 > 0$, such that $\forall\eta \in (0, \tau_2)$, $\Pi(\widetilde{\mathbf{A}}^{(t+1)}) \leq \Pi(\widetilde{\mathbf{A}}^{(t)})$. □