RELATIVE ERROR BOUNDS FOR NONNEGATIVE MATRIX FACTORIZATION UNDER A GEOMETRIC ASSUMPTION

Zhaoqiang Liu, Vincent Y. F. Tan

Department of Mathematics, National University of Singapore

ABSTRACT

We propose a geometric assumption on nonnegative data matrices such that under this assumption, we are able to provide upper bounds (both deterministic and probabilistic) on the relative error of nonnegative matrix factorization (NMF). The algorithm we propose first uses the geometric assumption to obtain an exact clustering of the columns of the data matrix; subsequently, it employs several rank-one NMFs to obtain the final decomposition. Furthermore, when combined with the classical alternating nonnegative least-squares algorithm, we show on synthetic examples that our proposed algorithm outperforms the standard algorithm based on multiplicative updates.

Index Terms— Nonnegative matrix factorization, Relative error bound, Clusterability, Separability

1. INTRODUCTION

Nonnegative matrix factorization (NMF) problem can be formulated as follows: given a nonnegative data matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, and a positive integer K, we seek nonnegative factor matrices $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$, such that some distance of $\mathbf{V}$ and $\mathbf{WH}$ is minimized. Due to its non-subtractive, part-based property which enhances interpretability, NMF has been widely used in the fields of machine learning [1] and signal processing [2]. In addition, there are many fundamental algorithms to approximately solve the NMF problem, including multiplicative update algorithms [3], alternating (nonnegative) least-squares-type algorithms [4–6] and rank-one residual iteration [7]. However, it is proved in [8] that NMF problem is NP-hard and all the basic algorithms simply ensure that the sequence of objective function is non-increasing and that the algorithm converges to the set of stationary points [7, 9, 10]. To the best of our knowledge, none of them is suitable for analyzing the bound on the approximation error of NMF.

In an effort to find computationally tractable algorithms for NMF and to provide theoretical guarantees on the errors of these algorithms, researchers have revisited the so-called *separability assumption* proposed by Donoho and Stodden in 2003 [11]. An exact nonnegative factorization $\mathbf{V} = \mathbf{WH}$ is *separable* if for any $k \in \{1, 2, \cdots, K\}$, there is an $n(k) \in \{1, 2, \cdots, N\}$, such that $\mathbf{W}_{n(k),j} = 0$ for all $j \neq k$ and $\mathbf{W}_{n(k),k} > 0$. That is, an exact nonnegative factorization is separable if all the features can be represented as nonnegative linear combinations of K features. It is proved in [12] that under the separability condition, there is an algorithm that runs in time polynomial in F, N and K and outputs a separable nonnegative factorization $\mathbf{V} = \mathbf{W}^*\mathbf{H}^*$ with the number of columns of $\mathbf{W}^*$ being at most K. Furthermore, a perturbation analysis of their algorithm is presented. The authors showed that if $\mathbf{V}$ has a separable nonnegative factorization $\mathbf{V} = \mathbf{WH}$, and each row of $\mathbf{V}$ is perturbed by adding a vector of small l_1 norm to obtain a new data matrix $\mathbf{V}'$, then with additional assumptions on the noise and $\mathbf{H}$, their algorithm leads to an approximate nonnegative matrix factorization $\mathbf{W}'\mathbf{H}'$ of $\mathbf{V}'$ with a provable error bound for the l_1 norm of each row of $\mathbf{V}' - \mathbf{W}'\mathbf{H}'$. To develop more efficient and scalable algorithms and to extend the basic formulation to more general noise models, a collection of elegant papers dealing with NMF under various separability conditions has emerged [13–17].

1.1. Main Contributions

We introduce a geometric assumption on the data matrix $\mathbf{V}$ that allows us to correctly group columns of $\mathbf{V}$ into disjoint subsets. This then naturally suggests an algorithm that first clusters the columns and subsequently uses a rank-one approximate NMF algorithm [18] to obtain the final decomposition. We analyze the error performance and provide an upper bound on the relative error. We also show that this algorithm performs well in practice. In fact, when combined with the alternating (nonnegative) least-squares (ALS) algorithm implemented in `nnmf` function of Matlab, it outperforms the standard algorithm based on multiplicative updates in terms of the relative error. We note that our geometric assumption can be considered as a special case of the separability assumption [11]. However, there are certain differences: first, because our assumption is based on a notion of clusterability [19], our proof technique is different from techniques in the existing literature that leverage the separability condition; second, unlike most of the papers considering separability [13–17], we mainly exploit the 2-norm of vectors instead of 1-norm of vectors; third, the data matrix does not need to be assumed to be normalized; and finally, we assume all the samples can be approximately represented by certain special samples (e.g., centroids) instead of using a small set of salient features to represent all the features. For the final point, although mathematically, these two approximations seem to be equivalent, our data reduction technique enables us to provide a tighter probabilistic relative error bound for the NMF approximation.

2. PROBLEM FORMULATION

2.1. Notations

We use capital boldface letters to denote matrices and we use lowercase boldface letters to denote vectors. We use Matlab-style notation for indexing, for example, $\mathbf{V}(i, j)$ denotes the entry of $\mathbf{V}$ in the i-th row and j-th column, $\mathbf{V}(i, :)$ denotes the i-th row of $\mathbf{V}$, $\mathbf{V}(:, j)$ denotes the j-th column of $\mathbf{V}$ and $\mathbf{V}(:, \mathscr{K})$ denotes the columns of $\mathbf{V}$ indexed by $\mathscr{K}$. $\|\mathbf{V}\|_F$ represents the Frobenius norm of $\mathbf{V}$ and $[N]$ represents $\{1, 2, \cdots, N\}$ for any positive integer N. Inequality $\mathbf{v} \geq 0$ or $\mathbf{V} \geq 0$ denotes element-wise nonnegativity. Let $\mathbf{V}_1 \in \mathbb{R}^{F\times N_1}$ and $\mathbf{V}_2 \in \mathbb{R}^{F\times N_2}$, we denote by $[\mathbf{V}_1, \mathbf{V}_2]$ the horizontal

The work of Z. Liu (a0109983@u.nus.edu, zqliu12@gmail.com) is supported by an National University of Singapore (NUS) Research Scholarship. The work of V. Y. F Tan (vtan@nus.edu.sg), who is also affiliated to the Department of Electrical and Computer Engineering, NUS, is supported by an NUS Young Investigator Award R-263-000-B37-133.

concatenation of the two matrices. Let $\mathbf{V}_1 \in \mathbb{R}^{F_1 \times N}$ and $\mathbf{V}_2 \in \mathbb{R}^{F_2 \times N}$. We denote by $[\mathbf{V}_1; \mathbf{V}_2]$ the vertical concatenation of the two matrices. We use $\mathbb{R}_{++}$ to represent the set of positive numbers.

2.2. Our Geometric Assumption on V

We assume all the columns of our data matrix $\mathbf{V}$ are sampled from K circular cones which satisfy a geometric assumption presented in (1) to follow. We define *circular cones* as follows.

Definition 1 *Given $\mathbf{u} \in \mathbb{R}_+^F$ with unit l_2 norm and an angle $\alpha \in (0, \pi/2)$, the nonnegative circular cone with respect to $\mathbf{u}$ and α is defined as*

$$\mathcal{C}(\mathbf{u}, \alpha) := \left\{ \mathbf{x} \in \mathbb{R}_+^F \setminus \{0\} : \frac{\mathbf{x}^T \mathbf{u}}{\|\mathbf{x}\|_2} \geq \cos \alpha \right\}.$$

In other words, $\mathcal{C}(\mathbf{u}, \alpha)$ contains all $\mathbf{x} \in \mathbb{R}_+^F \setminus \{0\}$ for which the angle between $\mathbf{u}$ and $\mathbf{x}$ is not larger than α. We say that α and $\mathbf{u}$ are respectively the *size angle* and *basis vector* of the circular cone.

We suppose that there are K such circular cones $C_1, \cdots, C_K$ with corresponding unit vectors and angles, i.e., $C_k := \mathcal{C}(\mathbf{u}_k, \alpha_k)$ for $k \in [K]$. Let $\alpha_{ij} := \arccos\left(\mathbf{u}_i^T \mathbf{u}_j\right)$, we make the geometric assumption that the columns of our data matrix $\mathbf{V}$ are sampled from K circular cones which satisfy

$$\min_{i,j \in [K], i \neq j} \alpha_{ij} > \max_{i,j \in [K], i \neq j} \{\max\{\alpha_i + 3\alpha_j, 3\alpha_i + \alpha_j\}\}. \quad (1)$$

In the proof of Lemma 1, we show that we can correctly group the data points generated from these K circular cones into K sets corresponding to the K circular cones under our assumption on $\mathbf{V}$.

3. OUR ALGORITHM AND MAIN THEOREMS

In this section we first provide several useful lemmas and demonstrate our algorithm for approximate NMF under our geometric assumption. Subsequently, we present our main theorems that present upper bounds on the relative errors of the NMF approximation.

Lemma 1 *Under the geometric assumption given in Section 2.2 for generating $\mathbf{V} \in \mathbb{R}_+^{F \times N}$, if Algorithm 1 is applied to $\mathbf{V}$, then the columns of $\mathbf{V}$ are partitioned into K subsets, such that the data points in the same subset are generated from the same circular cone.*

Proof We normalize $\mathbf{V}$ to obtain $\mathbf{V}'$, such that all the columns of $\mathbf{V}'$ have unit l_2 norm. From the definition, we know if a data point is in a circular cone, then the normalized data point is also in the circular cone. Then for any two columns $\mathbf{x}$, $\mathbf{y}$ of $\mathbf{V}'$ that are in the same circular cone $C_k, k \in [K]$, we have the largest possible angle between them is $2\alpha_k$, and thus the largest possible distance $\|\mathbf{x} - \mathbf{y}\|_2$ between these two data points is $\sqrt{2(1 - \cos(2\alpha_k))}$. On the other hand, for any two columns $\mathbf{x}$, $\mathbf{y}$ of $\mathbf{V}'$ that are in two circular cones $C_i, C_j, i \neq j$, we have the smallest possible angle between them is $\alpha_{ij} - \alpha_i - \alpha_j$, thus the smallest possible distance between them is $\sqrt{2(1 - \cos(\alpha_{ij} - \alpha_i - \alpha_j))}$. Then under the geometric assumption (1), the distance between any two unit data points in distinct circular cones is larger than the distance between any two unit data points in the same circular cone. Thus Algorithm 1 returns the correct clusters. ■

Now we present the following two useful lemmas. Lemma 2 provides an upper bound for the sum of the squares of perturbations of singular values. Lemma 3 shows that we can directly obtain best order-one nonnegative matrix factorization from best rank-one singular value decomposition.

Algorithm 1 Greedy clustering method with geometric assumption in (1)

Input: Data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$, $K \in \mathbb{N}$
Output: A set of non-empty, pairwise disjoint index sets $\mathscr{I}_1, \mathscr{I}_2, \cdots, \mathscr{I}_K \subseteq [N]$ such that their union is $[N]$
1) Normalize $\mathbf{V}$ to obtain $\mathbf{V}'$, such that all the columns of $\mathbf{V}'$ have unit l_2 norm.
2) Arbitrarily pick a point $\mathbf{z}_1 \in \mathbf{V}'$ (i.e., $\mathbf{z}_1$ is a column in $\mathbf{V}'$) as the first centroid.
3) **for** k = 1 to $K - 1$ **do**

$$\mathbf{z}_{k+1} := \arg\min_{\mathbf{z} \in \mathbf{V}'} \{\max\{\mathbf{z}_i^T \mathbf{z}, i \in [k]\}\} \quad (2)$$

and set $\mathbf{z}_{k+1}$ be the $(k+1)$-st centroid.
4) $\mathscr{I}_k := \{n \in [N] : k = \arg\max_{j \in [K]} \mathbf{z}_j^T \mathbf{V}'(:, n)\}$ for all $k \in [K]$.

Lemma 2 *(Wielandt-Hoffman Theorem for singular values) [20] If $\mathbf{A}$ and $\mathbf{A} + \mathbf{E}$ are in $\mathbb{R}^{F \times N}$, and denote $P = \min(F, N)$, then*

$$\sum_{p=1}^{P} (\sigma_p(\mathbf{A} + \mathbf{E}) - \sigma_p(\mathbf{A}))^2 \leq \|\mathbf{E}\|_F^2. \quad (3)$$

where $\sigma_p(\mathbf{A})$ is the p-th largest singular value of $\mathbf{A}$.

Lemma 3 *(Rank-one approximate NMF [18]) Let $\sigma \mathbf{u}\mathbf{v}^T$ be the rank-one singular value decomposition of a matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$. Then $\mathbf{u}' = \sigma|\mathbf{u}|$, $\mathbf{v}' = |\mathbf{v}|$ solves*

$$\min_{\mathbf{x} \in \mathbb{R}_+^F, \mathbf{y} \in \mathbb{R}_+^N} \|\mathbf{V} - \mathbf{x}\mathbf{y}^T\|_F.$$

We now state and prove a relative error bound of the proposed approximate NMF algorithm detailed in Algorithm 2 under our geometric assumption. We can see that if the size angles of all circular cones are small compared to the angle between the basis vectors of any two circular cones, then exact clustering is possible, and thus the relative error of the best approximate NMF of an arbitrary nonnegative matrix generated from these circular cones can be appropriately controlled by these size angles.

Theorem 4 *Under the geometric assumption given in Section 2.2 for generating $\mathbf{V} \in \mathbb{R}_+^{F \times N}$, Algorithm 2 outputs $\mathbf{W}^* \in \mathbb{R}_+^{F \times K}$, $\mathbf{H}^* \in \mathbb{R}_+^{K \times N}$, such that*

$$\frac{\|\mathbf{V} - \mathbf{W}^* \mathbf{H}^*\|_F}{\|\mathbf{V}\|_F} \leq \max_{k \in [K]} \{\sin \alpha_k\}. \quad (4)$$

Proof From Lemma 1, under the geometric assumption in Section 2.2, we can obtain a set of non-empty, pairwise disjoint index sets $\mathscr{I}_1, \mathscr{I}_2, \cdots, \mathscr{I}_K \subseteq [N]$ such that their union is $[N]$ and two data points $\mathbf{V}(:, j_1)$ and $\mathbf{V}(:, j_2)$ are in the same circular cones if and only if j_1 and j_2 are in the same index set. Denote $\mathbf{V}_k = \mathbf{V}(:, \mathscr{I}_k)$ and without loss of generality, suppose $\mathbf{V}_k \in C_k, k \in [K]$. For any $k \in [K]$, any column $\mathbf{z}$ of $\mathbf{V}_k$, suppose the angle between $\mathbf{z}$ and $\mathbf{u}_k$ is β, we have $\beta \leq \alpha_k$ and $\mathbf{z} = \|\mathbf{z}\|_2(\cos\beta)\mathbf{u}_k + \mathbf{y}$, with $\|\mathbf{y}\|_2 = \|\mathbf{z}\|_2 \sin\beta \leq \|\mathbf{z}\|_2 \sin\alpha_k$. Thus $\mathbf{V}_k$ can be written as the sum of a rank-one matrix $\mathbf{A}_k$ and a perturbation matrix $\mathbf{E}_k$. By Lemma 3, we can find the best rank-one approximate NMF of $\mathbf{V}_k$ from the singular value decomposition of $\mathbf{V}_k$. Suppose $\mathbf{w}_k^* \in \mathbb{R}_+^F$ and $\mathbf{h}_k \in \mathbb{R}_+^{|\mathscr{I}_k|}$ solves the best rank-one approximate NMF, and

Algorithm 2 Approximate NMF under the geometric assumption

Input: Data matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, $K \in \mathbb{N}$
Output: Factor matrices $\mathbf{W}^* \in \mathbb{R}_+^{F\times K}$, $\mathbf{H}^* \in \mathbb{R}_+^{K\times N}$
1) Use Algorithm 1 to find a set of non-empty, pairwise disjoint index sets $\mathscr{I}_1, \mathscr{I}_2, \cdots, \mathscr{I}_K \subseteq [N]$.
2) **for** k = 1 to K **do**

$$\begin{aligned}&\mathbf{V}_k := \mathbf{V}(:, \mathscr{I}_k),\\ &[\mathbf{U}_k, \mathbf{\Sigma}_k, \mathbf{X}_k] := \mathrm{svd}(\mathbf{V}_k),\\ &\mathbf{w}_k^* := \mathbf{\Sigma}_k(1,1)\,|\mathbf{U}_k(:,1)|, \quad \mathbf{h}_k := |\mathbf{X}_k(:,1)|,\\ &\mathbf{h}_k^* := \mathrm{zeros}(N,1), \mathbf{h}_k^*(\mathscr{I}_k) = \mathbf{h}_k.\end{aligned}$$

3) $\mathbf{W}^* := [\mathbf{w}_1^*, \cdots, \mathbf{w}_K^*]$, $\mathbf{H}^* := \left[(\mathbf{h}_1^*)^T; \cdots; (\mathbf{h}_K^*)^T\right]$.

denote $\mathbf{S}_k := \mathbf{w}_k^*\mathbf{h}_k^T$ as the best rank-one approximation matrix of $\mathbf{V}_k$. Let $P_k = \min(F, |\mathscr{I}_k|)$, then by Lemma 2, we have

$$\|\mathbf{V}_k - \mathbf{S}_k\|_F^2 = \sum_{p=2}^{P_k}\sigma_p^2(\mathbf{V}_k) = \sum_{p=2}^{P_k}\sigma_p^2(\mathbf{A}_k + \mathbf{E}_k) \le \|\mathbf{E}_k\|_F^2. \tag{5}$$

From the previous result, we know that

$$\frac{\|\mathbf{E}_k\|_F^2}{\|\mathbf{V}_k\|_F^2} = \frac{\sum_{\mathbf{z}\in\mathbf{V}_k}\|\mathbf{z}\|_2^2\sin^2\beta_{\mathbf{z}}}{\sum_{\mathbf{z}\in\mathbf{V}_k}\|\mathbf{z}\|_2^2} \le \sin^2\alpha_k, \tag{6}$$

where $\beta_{\mathbf{z}}$ denotes the angle between $\mathbf{z}$ and $\mathbf{u}_k$ and $\mathbf{z} \in \mathbf{V}_k$ runs over all columns of the matrix $\mathbf{V}_k$.

Define $\mathbf{h}_k^* \in \mathbb{R}_+^N$ as $\mathbf{h}_k^*(n) = \mathbf{h}_k(n)$, if $n \in \mathscr{I}_k$ and $\mathbf{h}_k^*(n) = 0$ if $n \notin \mathscr{I}_k$. Let $\mathbf{W}^* := [\mathbf{w}_1^*, \cdots, \mathbf{w}_K^*]$ and $\mathbf{H}^* := \left[(\mathbf{h}_1^*)^T; \cdots; (\mathbf{h}_K^*)^T\right]$, then we have

$$\frac{\|\mathbf{V}-\mathbf{W}^*\mathbf{H}^*\|_F^2}{\|\mathbf{V}\|_F^2} = \frac{\sum_{k=1}^K\|\mathbf{V}_k-\mathbf{w}_k^*\mathbf{h}_k^T\|_F^2}{\|\mathbf{V}\|_F^2} \le \frac{\sum_{k=1}^K\|\mathbf{V}_k\|_F^2\sin^2\alpha_k}{\sum_{k=1}^K\|\mathbf{V}_k\|_F^2}.$$

Thus we have (4) as desired. ■

We now provide a tighter relative error bound assuming a probabilistic model. For simplicity, we assume a straightforward and easy-to-implement statistical model for the sampling procedure.

Theorem 5 *Suppose the K circular cones $C_k := \mathcal{C}(\mathbf{u}_k, \alpha_k) \in \mathbb{R}_+^F$ for $k \in [K]$ satisfy the geometric assumption given by* (1)*. Let $\boldsymbol{\lambda} := (\lambda_1; \lambda_2; \cdots; \lambda_K) \in \mathbb{R}_{++}^K$. We generate a data matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$ from the following generative process for each column $\mathbf{v}$ of $\mathbf{V}$:*
1. sample $k \in [K]$ with equal probability $1/K$;
2. sample l from the exponential distribution with parameter λ_k, i.e., the distribution $\mathrm{Exp}(\lambda_k)$;
3. uniformly sample a unit vector $\mathbf{z} \in C_k$ with respect to the angle between $\mathbf{z}$ and $\mathbf{u}_k$; if $\mathbf{z} \notin \mathbb{R}_+^F$, set all the negative entries of $\mathbf{z}$ to zero, and rescale $\mathbf{z}$ to be a unit vector;
4. let $\mathbf{v} = \sqrt{l}\mathbf{z}$;
Let $f(\alpha) := 0.5 - (\sin 2\alpha)/(4\alpha)$, then for small $\epsilon > 0$, with probability at least $1 - 8\exp\left(-\xi N\epsilon^2\right)$, one has

$$\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_F}{\|\mathbf{V}\|_F} \le \sqrt{\frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k}} + \epsilon, \tag{7}$$

where the constant $\xi > 0$ depends only on λ_k and $f(\alpha_k)$ for all $k \in [K]$.

The relative error bound produced by Theorem 5 is better than that of Theorem 4, i.e., the former is more conservative. This can be seen from (10) to follow or by the inequality $\alpha \le \tan\alpha$ for $\alpha \in [0, \pi/2)$. We also observe this in the experiments.

Definition 2 *A* sub-exponential random variable X *is one that satisfies one of the following equivalent properties*
1. Tails: $\mathbb{P}(|X| > t) \le \exp(1 - t/K_1)$ for all $t \ge 0$;
2. Moments: $(\mathbb{E}|X|^p)^{1/p} \le K_2 p$ for all $p \ge 1$;
3. $\mathbb{E}[\exp(X/K_3)] \le e$;
where $K_i, i = 1, 2, 3$ are positive constants. The sub-exponential norm *of X, denoted $\|X\|_{\Psi_1}$, is defined to be*

$$\|X\|_{\Psi_1} := \sup_{p\ge 1} p^{-1}\left(\mathbb{E}|X|^p\right)^{1/p}.$$

Lemma 6 *(Bernstein-type inequality) [21] Let $X_1, \cdots, X_N$ be independent sub-exponential random variables with zero expectations, and $M = \max_i \|X_i\|_{\Psi_1}$. Then for every $\epsilon \ge 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N X_i\right| \ge \epsilon N\right) \le 2\exp\left[-c\cdot\min\left(\frac{\epsilon^2}{M^2}, \frac{\epsilon}{M}\right)N\right], \tag{8}$$

where $c > 0$ is an absolute constant.

By combining Lemma 6 with the upper bound result shown in the proof of Theorem 4, we can obtain the proof of Theorem 5.

Proof of Theorem 5 From (5) and (6) in the proof of Theorem 4, to obtain an upper bound for the square of the relative error, we can consider the following random variable

$$D_N := \frac{\sum_{n=1}^N L_n^2\sin^2 B_n}{\sum_{n=1}^N L_n^2}, \tag{9}$$

where L_n is the random variable corresponding to the length of the n-th point, and B_n is the random variable corresponding to the angle between the n-th point and $\mathbf{u}_k$ for some $k \in [K]$ such that the point is in C_k. We first consider estimating the above random variable with the assumption that all the data points are generated from a single circular cone $C := \mathcal{C}(\mathbf{u}, \alpha)$ (i.e., assume $K = 1$), and the square of lengths are generated according to the exponential distribution $\mathrm{Exp}(\lambda)$. Because we assume the angle $\beta_n, n \in [N]$ is sampled from a uniform distribution on $[0, \alpha]$, the expectation of $\sin^2 B_n$ is

$$\mathbb{E}\left[\sin^2 B_n\right] = \int_0^\alpha \frac{1}{\alpha}\sin^2\beta\,\mathrm{d}\beta = 0.5 - \frac{\sin 2\alpha}{4\alpha} = f(\alpha). \tag{10}$$

Here we only need to consider vectors $\mathbf{z} \in \mathbb{R}_+^F$ whose angles with $\mathbf{u}$ are not larger than α. Otherwise, we have $\mathbb{E}\left[\sin^2 B_n\right] \le f(\alpha)$. Our probabilistic upper bound also holds in this case.

Since the length and the angle are independent, we have

$$\mathbb{E}[D_N] = \mathbb{E}\left[\mathbb{E}[D_N|L_1, \cdots, L_N]\right] = f(\alpha), \tag{11}$$

and we also have

$$\mathbb{E}\left[L_n^2\sin^2 B_n\right] = \mathbb{E}\left[L_n^2\right]\mathbb{E}\left[\sin^2 B_n\right] = \frac{f(\alpha)}{\lambda}. \tag{12}$$

Denote $X_n = L_n^2$ for all $n \in [N]$. Let

$$H_N := \frac{\sum_{n=1}^N X_n}{N}, \quad \text{and} \quad G_N := \frac{\sum_{n=1}^N X_n\sin^2 B_n}{N}. \tag{13}$$

We have for all $n \in [N]$,

$$\mathbb{E}[X_n^p] = \lambda^{-p}\Gamma(p+1) \le \lambda^{-p}p^p, \qquad \forall\, p \ge 1.$$

Thus $\|X_n\|_{\Psi_1} \leq \lambda^{-1}$, and X_n is sub-exponential. By the triangle inequality, we have $\|X_n - \mathbb{E}X_n\|_{\Psi_1} \leq \|X_n\|_{\Psi_1} + \|\mathbb{E}X_n\|_{\Psi_1} \leq 2\|X_n\|_{\Psi_1}$. Hence, by Lemma 6, for all $\epsilon > 0$, we have (8) where M can be taken as $M = 2/\lambda$. Because

$$\left(\mathbb{E}\left[\left(X_n \sin B_n^2\right)^p\right]\right)^{1/p} \leq \lambda^{-1} p \sin^2 \alpha \leq \lambda^{-1} p,$$

we have similar large deviation result for G_N.

On the other hand, for all $\epsilon > 0$

$$\mathbb{P}\left(|D_N - f(\alpha)| \geq \epsilon\right) = \mathbb{P}\left(\left|\frac{G_N}{H_N} - f(\alpha)\right| \geq \epsilon\right)$$
$$\leq \mathbb{P}\left(|\lambda G_N - f(\alpha)| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(\left|\frac{G_N}{H_N} - \lambda G_N\right| \geq \frac{\epsilon}{2}\right)$$

For the second term, by fixing small $\delta_1, \delta_2 > 0$, we have

$$\mathbb{P}\left(\left|\frac{G_N}{H_N} - \lambda G_N\right| \geq \frac{\epsilon}{2}\right) = \mathbb{P}\left(\frac{|1 - \lambda H_N| G_N}{H_N} \geq \epsilon\right)$$
$$\leq \mathbb{P}\left(\frac{|1 - \lambda H_N| G_N}{H_N} \geq \epsilon, H_N \geq \frac{1}{\lambda} - \delta_1, G_N \leq \frac{f(\alpha)}{\lambda} + \delta_2\right)$$
$$+ \mathbb{P}\left(H_N < \frac{1}{\lambda} - \delta_1\right) + \mathbb{P}\left(G_N > \frac{f(\alpha)}{\lambda} + \delta_2\right)$$

Combining the bounds for H_N and G_N in (8) with the above inequalities, if we set $\delta_1 = \delta_2 = \epsilon$ and take ϵ sufficiently small,

$$\mathbb{P}\left(|D_N - f(\alpha)| \geq \epsilon\right) \leq 8 \exp\left(-\xi N \epsilon^2\right), \tag{14}$$

where ξ is a positive constant depending on λ and $f(\alpha)$.

Now we turn to the general case in which $K \in \mathbb{N}$. We have

$$\mathbb{E}[X_n] = \frac{\sum_{k=1}^K 1/\lambda_k}{K}, \quad \mathbb{E}\left[X_n \sin^2 B_n\right] = \frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{K},$$

and for all $p \geq 1$,

$$\left(\mathbb{E}[X_n^p]\right)^{1/p} = \left(\frac{\sum_{k=1}^K \lambda_k^{-p} \Gamma(p+1)}{K}\right)^{1/p} \leq p \left(\min_k \lambda_k\right)^{-1}.$$

Similar to (14), we have

$$\mathbb{P}\left(\left|D_N - \frac{\sum_{k=1}^K f(\alpha_k/\lambda_k)}{\sum_{k=1}^K 1/\lambda_k}\right| \geq \epsilon\right) \leq 8 \exp\left(-\xi N \epsilon^2\right),$$

and thus, if we let $\Delta := \sqrt{\frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k}}$, we have

$$\mathbb{P}\left(\left|\sqrt{D_N} - \Delta\right| \leq \epsilon\right) \geq \mathbb{P}\left(\left|D_N - \Delta^2\right| \leq \Delta\epsilon\right)$$
$$\geq 1 - 8 \exp\left(-\xi N \Delta^2 \epsilon^2\right).$$

This completes the proof of (7). ■

4. EXPERIMENTS

To verify the correctness of our bounds and to observe the computational efficacy of the proposed algorithm, we perform numerical simulations on synthetic datasets. To generate the columns of $\mathbf{V}$, given an integer $k \in [K]$ and an angle $\beta \in [0, \alpha_k]$, we need to find a unit nonnegative vector $\mathbf{z}$ such that the angle between $\mathbf{z}$ and $\mathbf{u}_k$ is β. Let $b = \cos\beta$, we take $\mathbf{x} = b\mathbf{u}_k$ ($\mathbf{x}$ satisfies $\mathbf{u}_k^T \mathbf{x} = b$) and let $i := \arg\min_{f \in [F]} \mathbf{u}_k(f)$. Let $\mathbf{y}$ be a vector satisfying $\mathbf{u}_k^T \mathbf{y} = 0$, $\mathbf{y}(i) = 1$ and $\hat{\mathbf{y}} \in \text{span}\{\hat{\mathbf{u}}_k\}$, where $\hat{\mathbf{y}}$ denotes $\mathbf{y}$ without the i-th entry and similar for $\hat{\mathbf{u}}_k$. Then let $\mathbf{t} = \sqrt{1 - b^2}\mathbf{y}/\|\mathbf{y}\|_2$ and set $\mathbf{z} = \mathbf{x} + \mathbf{t}$. The vector $\mathbf{z}$ is one of the vectors we desire. This

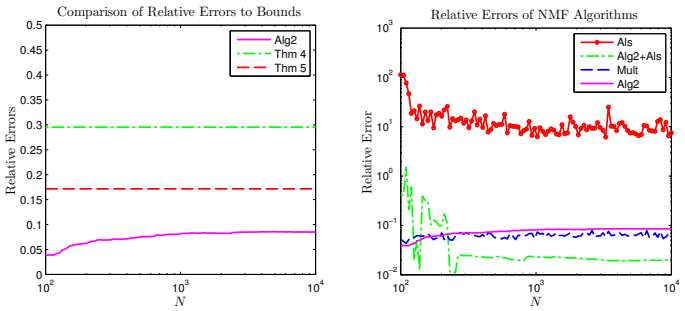

Fig. 1. Error bounds and performances of various algorithms

Table 1. Running time(s) of various algorithms

N	Alg2	Als	Mult	Alg2+Als
10^2	0.4470	0.1950	0.6150	0.5160
10^3	1.5340	0.2410	1.9300	2.5970
10^4	13.8670	4.1930	25.3410	28.7450

generation procedure does not affect the upper bounds in Theorem 4 and Theorem 5 because we only use the length of samples and angle between samples with the basis vectors to derive the upper bounds.

For simplicity of presentation, we set the size angles α to be the same for all the circular cones. The angle between any two basis vectors is set to be $(4\alpha + \Delta\alpha)$ where $\Delta\alpha := 0.01$. The parameter for the exponential distribution $\boldsymbol{\lambda} := 1./(1:K)$. We increase N from 10^2 to 10^4 logarithmically. We fix the parameters $F = 10^3$, $K = 50$ and $\alpha = 0.3$, and obtain the results shown in Fig. 1. All the results are averaged over 10 runs. In the left plot of Fig. 1, we compare the relative errors of our algorithm with the derived relative error bounds. In the right plot, we compare the relative errors of our algorithm (denoted as "Alg2") with the relative errors of two classical algorithms, namely, the multiplicative update algorithm (denoted as "Mult") and the alternating (nonnegative) least-squares algorithm (denoted as "Als"); these two algorithms can be implemented using the `nnmf` function of Matlab. Furthermore, we consider "Alg2+Als", i.e., running the "Als" algorithm with the initial factor matrices set to be the results obtained by "Alg2". In addition, the running times for the algorithms are shown in Table 4. All experiments we run on an Intel Core i7 CPU at 2.50GHz and 16GB of memory, and the Matlab version is 8.3.0.532 (R2014a).

From Fig. 1, we observe that the numerical relative errors are truly smaller than our theoretical relative error bounds. The "Als" algorithm is not stable and may lead to large relative errors for certain sample sizes. The performance of "Alg2" is comparable to "Mult". However, when combined with "Als", "Alg2" achieves noticeable smaller relative errors for large sample sizes N. In addition, the running time of "Alg2+Als" is comparable to "Mult" so we conclude that for large sample sizes, "Alg2+Als" is a useful alternative to the ubiquitous "Mult" algorithm for NMF.

5. FUTURE WORK

We hope to extend Theorem 5 by providing an even tighter error bound for the relative error using the theory of singular values of random matrices [21–23]. We also would like to provide relative error bounds when K is unknown (cf. [24]). Finally, we will apply our algorithm to real datasets in future.

6. REFERENCES

[1] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*, John Wiley & Sons, 2009.

[2] Ioan Buciu, "Non-negative matrix factorization, a new tool for feature extraction: theory and applications," *Intl J. Computers, Comm. and Control*, vol. 3, pp. 67–74, 2008.

[3] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562.

[4] M. Chu, F. Diele, R. Plemmons, and S. Ragni, "Optimality, computation, and interpretation of nonnegative matrix factorizations," *SIAM J. Matrix Anal.*, 2004.

[5] Hyunsoo Kim and Haesun Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J. Matrix Anal. A.*, vol. 30, no. 2, pp. 713–730, July 2008.

[6] Jingu Kim and Haesun Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM J. Sci. Comput.*, vol. 33, no. 6, pp. 3261–3281, 2011.

[7] Ngoc-Diep Ho, Paul Van Dooren, and Vincent D. Blondel, *Descent methods for Nonnegative Matrix Factorization*, 2011.

[8] Stephen A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization*, vol. 20, pp. 1364–1377, 2009.

[9] Chih-Jen Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, Oct. 2007.

[10] Chih-Jen Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1589–1596, Nov 2007.

[11] David Donoho and Victoria Stodden, "When does non-negative matrix factorization give correct decomposition into parts?," in *NIPS*. 2004, pp. 1141–1148, MIT Press.

[12] S. Arora, R. Ge, R. Kannan, and A. Moitra, "Computing a nonnegative matrix factorization–provably," in *ACM STOC*, May 2012, pp. 145–162.

[13] Gillis Nicolas and Stephen A. Vavasis, "Fast and robust recursive algorithmsfor separable nonnegative matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 698–714, 2014.

[14] V. Bittorf, B. Recht, and C. Ré, "Factoring nonnegative matrices with linear programs," in *NIPS*, 2014, pp. 1214–1222.

[15] A. Kumar, V. Sindhwani, and P. Kambadur, "Fast conical hull algorithms for near-separable non-negative matrix factorization," in *ICML*, Jun 2013, pp. 231–239.

[16] A. Benson, B. Rajwa J. Lee, and D. Gleich, "Scalable methods for nonnegative matrix factorizations of near-separable tall-and-skinny matrice," in *NIPS*, 2014, pp. 945–953.

[17] Gillis Nicolas and Robert Luce, "Robust near-separable nonnegative matrix factorization using linear optimization," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1249–1280, 2014.

[18] Edward F. Gonzalez, *Efficient alternating gradient-type algorithms for the approximate non-negative matrix factorization problem*, ProQuest, 2009.

[19] Margareta Ackerman and Shai Ben-David, "Clusterability: A theoretical study," in *AISTATS*, 2009, vol. 5, pp. 1–8.

[20] Gene H. Golub and Charles F. Van Loan, *Matrix computations*, JHU Press, 1989.

[21] Roman Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint*, 2010.

[22] Mark Rudelson and Roman Vershynin, "Non-asymptotic theory of random matrices: extreme singular values," *arXiv preprint*, 2010.

[23] Terence Tao and Van Vu, "Random matrices: universality of local spectral statistics of non-hermitian matrices," *The Annals of Probability*, vol. 43, no. 2, pp. 782–874, 2015.

[24] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1592–1605, 2013.