

---

# Intersecting Faces: Non-negative Matrix Factorization With New Guarantees

---

**Rong Ge**

Microsoft Research New England

RONGGE@MICROSOFT.COM

**James Zou**

Microsoft Research New England

JAZO@MICROSOFT.COM

## Abstract

Non-negative matrix factorization (NMF) is a natural model of admixture and is widely used in science and engineering. A plethora of algorithms have been developed to tackle NMF, but due to the non-convex nature of the problem, there is little guarantee on how well these methods work. Recently a surge of research have focused on a very restricted class of NMFs, called separable NMF, where provably correct algorithms have been developed. In this paper, we propose the notion of subset-separable NMF, which substantially generalizes the property of separability. We show that subset-separability is a natural necessary condition for the factorization to be unique or to have minimum volume. We developed the Face-Intersect algorithm which provably and efficiently solves subset-separable NMF under natural conditions, and we prove that our algorithm is robust to small noise. We explored the performance of Face-Intersect on simulations and discuss settings where it empirically outperformed the state-of-art methods. Our work is a step towards finding provably correct algorithms that solve large classes of NMF problems.

## 1. Introduction

In many settings in science and engineering the observed data are admixtures of multiple latent sources. We would typically want to infer the latent sources as well as the admixture distribution given the observations. Non-negative matrix factorization (NMF) is a natural mathematical framework to model many admixture problems.

In NMF we are given an observation matrix  $M \in \mathbb{R}^{n \times m}$ , where each row of  $M$  corresponds to a data-point in  $\mathbb{R}^m$ .

---

*Proceedings of the 32<sup>nd</sup> International Conference on Machine Learning*, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

We assume that there are  $r$  latent sources, modeled by the unobserved matrix  $W \in \mathbb{R}^{r \times m}$ , where each row of  $W$  characterizes one source. Each observed data-point is a linear combination of the  $r$  sources and the combination weights are encoded in a matrix  $A \in \mathbb{R}^{n \times r}$ . Moreover, in many natural settings, the sources are non-negative and the combinations are additive. The computational problem is then to factor a given matrix  $M$  as  $M = AW$ , where all the entries of  $M$ ,  $A$  and  $W$  are non-negative. We call  $r$  the inner-dimension of the factorization, and the smallest possible  $r$  is usually called the nonnegative rank of  $M$ . NMF was first purposed by (Lee & Seung, 1999), and has been widely applied in computer vision (Lee & Seung, 2000), document clustering (Xu et al., 2003), hyperspectral unmixing (Nascimento & Dias, 2004; Gomez et al., 2007), computational biology (Devarajan, 2009), etc. We give two concrete examples

**Example 1.** In topic modeling,  $M$  is the  $n$ -by- $m$  word-by-document matrix, where  $n$  is the vocabulary size and  $m$  is the number of documents. Each column of  $M$  corresponds to one document and the entry  $M(i, j)$  is the frequency with which word  $i$  appears in document  $j$ . The topics are the columns of  $A$ , and  $A(i, k)$  is the probability that topic  $k$  uses word  $i$ .  $W$  is the topic-by-document matrix and captures how much each topic contributes to each document. Since all the entries of  $M$ ,  $A$  and  $W$  are frequencies, they are all non-negative. Given  $M$  from a corpus of documents, we would like to factor  $M = AW$  and recover the relevant topics in these documents. (Note that in this example  $A$  is the matrix of “sources” and  $W$  is the matrix of mixing weights, so it is the transpose of what we just introduced. We use this notation to be consistent with previous works (Arora et al., 2012).)

**Example 2.** In many bio-medical applications, we collect samples and for each sample perform multiple measurements (e.g. expression of  $10^4$  genes or DNA methylation at  $10^6$  positions in the genome; all the values are non-negative).  $M$  is the sample-by-measurement matrix, where  $M(i, j)$  is the value of the  $j$ th measurement in sam-

ple  $i$ . Each sample, whether taken from humans or animals, is typically a composition of several cell-types that we do not directly observe. Each row of  $W$  corresponds to one cell-type, and  $W(k, j)$  is the value of cell-type  $k$  in measurement  $j$ . The entry  $A(i, k)$  is the fraction of sample  $i$  that consists of cell-type  $k$ . Experiments give us the matrix  $M$ , and we would like to factor  $M = AW$  to identify the relevant cell-types and their compositions in our samples.

Despite the simplicity of its formulation, NMF is a challenging problem. First, the NMF problem may not be identifiable, and hence we can not hope to recover the *true*  $A$  and  $W$ . Moreover, even ignoring the identifiability Vavasis (2009) showed that finding any factorization  $M = AW$  with inner-dimension  $r$  is an *NP*-hard problem. Arora et al. (2012) showed under reasonable assumptions we cannot hope to find a factorization in time  $(mn)^{o(r)}$ , and the best algorithm known is Moitra (2013) that runs in time  $O(2^r mn)^{O(r^2)}$ .

Many heuristic algorithms have been developed for NMF but they do not have guarantees for when they would actually converge to the true factorization (Lee & Seung, 2000; Lin, 2007). More recently, there has been a surge of interest in constructing practical NMF algorithms with strong theoretical guarantees. Most of this activity (e.g. Arora et al. (2012); Bittorf et al. (2012); Kumar et al. (2012); Gillis (2012); Gillis & Vavasis (2014), see more in Gillis (2014)) are based on the notion of *separability* (Donoho & Stodden, 2003) which is a very strict condition that requires that all the rows of  $W$  appear as rows in  $M$ . While this might hold in some document corpus, it is unlikely to be true in other engineering and bio-medical applications.

**Our Results** In this paper, we develop the notion of *subset separability*, which is a significantly weaker and more general condition than separability. In topic models, for example, separability states that there is a word that is unique to each topic. Subset separability means that there is a combination of words that is unique to each topic. We show that subset separability arise naturally as a necessary condition when the NMF is identifiable or when we are seeking the minimal volume factorization. We characterize settings when subset-separable NMF can be solved in polynomial-time, and this include the separable setting as a special case. We construct the Face-Intersect algorithm which provably and robustly solves the NMF even in the presence of adversarial noise. We use simulations to explore conditions where our algorithm achieves more accurate inference than current state-of-art algorithms.

**Organization** We first describe the geometric interpretation of NMF (Sec. 2), which leads us to the notion of subset-separable NMF (Sec. 3). We then develop our Face-Intersect algorithm and analyze its robustness (Sec. 4). Our main result, Theorem 4.2, states that for subset-separable NMF, if the facets are *properly filled* in a way that de-

pends on the magnitude of the adversarial noise, then Face-Intersect is guaranteed to find a factorization that is close to the true factorization in polynomial time. We discuss the algorithm in more detail in Sections 5 and 6, and analyze a generative model that give rise to properly filled facets in Section 7. Finally we present experiments to explore settings where Face-Intersect outperforms state-of-art NMF algorithms (Sec. 8). Due to space constraints, all the proofs are presented in the appendix. Throughout the paper, we give intuitions behind proofs of the main results.

## 2. Geometric intuition

For a matrix  $M \in \mathbb{R}^{n \times m}$ , we use  $M^i \in \mathbb{R}^m$  to denote the  $i$ -th row of  $M$ , but it is viewed as a column vector. Given a factorization  $M = AW$ , without loss of generality we can assume the rows of  $M, A, W$  all sum up to 1 (this can always be done, see (Arora et al., 2012)). In this way we can view the rows of  $W$  as vertices of an unknown simplex, and the rows of  $M$  are all in the convex hull of these vertices. The NMF is then equivalent to the following geometric problem:

**NMF, Geometric Interpretation** There is an unknown  $W$ -simplex whose vertices are the rows of  $W \in \mathbb{R}^m$ ,  $W^1, \dots, W^r$ . We observe  $n$  points  $M^1, M^2, \dots, M^n \in \mathbb{R}^m$  (corresponding to rows of  $M$ ) that lie in the  $W$ -simplex. The goal is to identify the vertices of the  $W$ -simplex.

When clear from context, we also call the  $W$  matrix as the simplex, and the goal is to find the vertices of this simplex. There is one setting where it is easy to identify all the vertices.

**Definition 2.1** (separability). *A NMF is separable if all the vertices  $W^j$ 's appear in the points  $M^i$ 's that we observe.*

Separability was introduced in Donoho & Stodden (2003). When the NMF is separable, the problem simplifies as we only need to identify which of the points  $M^j$ 's are vertices of the simplex. This can be done in time polynomial in  $n, m$  and  $r$  (Arora et al., 2012). Separability is a highly restrictive condition and it takes advantage of only the 0-dimensional structure (vertices) of the simplex. In this work, we use higher dimensional structures of the simplex to solve the NMF. We use the following standard definition of facets:

**Definition 2.2** (facet). *A facet  $S \subset [r]$  of the  $W$ -simplex is the convex hull of vertices  $\{W^j : j \in S\}$ . We call  $S$  a filled facet if there is at least one point  $M^i$  in the interior of  $S$  (or if  $|S| = 1$  and there is one point  $M^i$  that is equal to that vertex; such  $M^i$  is called an anchor).*

**Conventions** When it's clear from context, we interchangeably represent a facet  $S$  both by the indices of its

vertices and by the convex hull of these vertices. A facet also corresponds to a unique linear subspace  $Q_S$  with dimension  $|S|$  that is the span of  $\{W^j : j \in S\}$ . In the rest of the paper, it's convenient to use linear algebra to quantify various geometric ideas. We will represent a  $d$ -dimensional subspace of  $\mathbb{R}^m$  using a matrix  $U \in \mathbb{R}^{m \times d}$ , the columns of matrix  $U$  is an arbitrary orthonormal basis for the subspace (hence the representation is not unique). We use  $P_U = UU^T$  to denote the projection matrix to subspace  $U$ , and  $U^\perp \in \mathbb{R}^{m \times (m-d)}$  to denote an arbitrary representation of the orthogonal subspace. For two subspaces  $U$  and  $V$  of the same dimension, we define their distance to the the sin of the principle angle between the two subspaces (this is the largest angle between vectors  $u, v$  for  $u \in U$  and  $v \in V$ ). This distance can be computed as the spectral norm  $\|P_{U^\perp} V\|$  (and has many equivalent formulations).

### 3. Subset Separability

NMF is not identifiable up to scalings and permutations of the rows of  $W$ . Ignoring such transformations, there can still be multiple non-negative factorizations of the same matrix  $M$ . This arise when there are different sets of  $r$  vertices in the non-negative orthant that contain all the points  $M^i$  in its convex hull. For example, suppose  $M = AW$  and the  $A$  matrix has all positive entries. All the points  $M^i$  are in the interior of the  $W$ -simplex. Then it is possible to perturb the vertices of  $W$  while still maintaining all of the  $M^i$ 's in its convex hull. This give rise to a different factorization  $M = \hat{A}\hat{W}$ . When the factorization is not unique, we may want find a solution where the  $W$ -simplex has minimal volume, in the sense that it is impossible to move a single vertex and shrink the volume while maintaining the validity of the solution.

It's clear that in order for  $W$  to be the minimal volume solution to the NMF, there must be some points  $M^i$  that lie on the boundary of the  $W$ -simplex. We show that a necessary condition for  $W$  to be volume minimizing is for the filled facets (facets of  $W$  with points in its interior) to be subset-separable. Intuitively, this means that each vertex of  $W$  is the unique intersection point of a subset of filled facets.

**Definition 3.1** (subset-separable). *A NMF  $M = AW$  is subset-separable if there is a set of filled facets  $S_1, \dots, S_k \subset [r]$  such that  $\forall j \in [r]$ , there is a subset of  $S_{j_1}, S_{j_2}, \dots, S_{j_{k_j}}$  whose intersection is exactly  $j$ .*

**Proposition 3.1.** *Suppose  $W$  is a minimal volume rank  $r$  solution of the NMF  $M = AW$ . Then  $W$  is subset-separable.*

It is easy to see that the factorization  $M = AW$  is subset-separable is equivalent to the property that for every  $j_1 \neq j_2 \in [r]$ , there is a row  $i$  of  $A$  such that

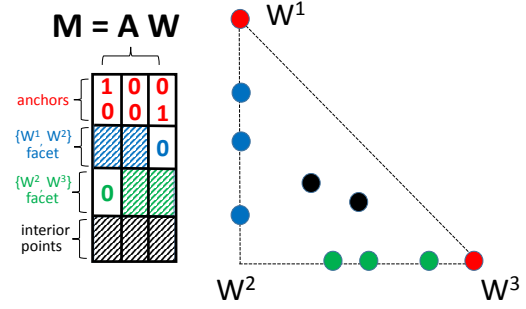


Figure 1. Illustration of the NMF geometry.

$A_{i,j_1} = 0$  and  $A_{i,j_2} \neq 0$ . The previously proposed separability condition corresponds to the special case where the filled facets  $S_1, \dots, S_k$  correspond to the singleton sets  $\{W^1\}, \dots, \{W^r\}$ .

**Example.** We illustrate the subset-separable condition in Figure 1. In this figure, the circles correspond to data points  $M^i$ 's and they are colored according to the facet that they belong. The filled facets are  $S_1 = \{1\}$ ,  $S_2 = \{3\}$ ,  $S_3 = \{1, 2\}$  and  $S_4 = \{2, 3\}$ . The facet  $\{W^1, W^3\}$  is not filled since there are no points in its interior. The singleton facets  $S_1$  and  $S_2$  are also called anchors. This NMF is subset-separable since  $W^2$  is the unique intersection of  $S_3$  and  $S_4$ , but it is not separable. The figure also illustrates the corresponding  $A$  matrix, where the rows are grouped by facets and the shaded entries denote the support of each row.

The geometry of the simplex suggests an intuitive meta-algorithm for solving subset-separable NMFs, which is the basis of our Face-Intersect algorithm.

1. Identify the filled facets,  $S_1, \dots, S_k, r \leq k \leq n$ .
2. Take intersections of the facets to recover all the rows of  $W$  (vertices of the simplex).
3. Use  $M$  and  $W$  to solve for  $A$ .

### 4. Robust algorithm for subset-separable NMF

In order to carry out the meta-algorithm, the key computational challenge is to efficiently and correctly identify the filled facets of the  $W$  simplex. Finding filled facets is related to well-studied problems in subspace clustering (Vidal, 2010) and subspace recovery (Hardt & Moitra, 2013). In subspace clustering we are given points in  $k$  different subspaces and the goal is to cluster the points according to which subspace it belong to. This problem is in general NP-hard (Elhamifar & Vidal, 2009) and can only be solved under strong assumptions. Subspace recovery tries to find a unique subspace a fraction  $p$  of the points. Hardt

& Moitra (2013) showed this problem is hard unless  $p$  is large compared to the ratio of the dimensions. Techniques and algorithms from subspace clustering and recovery typically make strong assumptions about the independence of subspaces or the generative model of the points, and cannot be directly applied to our problem. Moreover, our filled facets have the useful property that they are on the boundary of the convex hull of the data points, which is not considered in general subspace clustering/discovery methods. We identified a general class of filled facets, called *properly filled facets* that are computationally efficient to find.

**Definition 4.1** (properly filled facets). *Given a NMF  $M = AW$ , a set of facets  $S_1, \dots, S_k \in [r]$  of  $W$  is properly filled if it satisfies the following properties:*

1. *For any facet  $|S_i| > 1$ , the rows of  $A$  with support equal to  $S_i$  (i.e. points that lie on this facet) has a  $|S_i| - 1$ -dimensional convex hull. Moreover, there is at least one row of  $A$  that is in the interior of the convex hull.*
2. *(General positions property.) For any subspace of dimension  $1 < t < r$ , if it contains more than  $t$  rows in  $M$ , then the subspace contains at least one  $S_i$  which is not a singleton facet.*

Condition 1 ensures that each  $S_i$  has sufficiently many points to be non-degenerate. Condition 2 says that points that are not in the lower dimensional facets  $S_1, \dots, S_k$  are in general positions, so that no random subspace look like a properly filled facet. A set of properly filled facets  $S_1, \dots, S_k$  may contain singleton sets corresponding some of the rows  $W^j$  if these rows also appear as rows in  $M$ . We first state the main results and then state the Face-Intersect algorithm.

**Theorem 4.1.** *Suppose  $M = AW$  is subset separable by  $S_1, \dots, S_k$  and these facets are properly filled, then given  $M$  the Face-Intersect algorithm computes  $A$  and  $W$  in time polynomial in  $n$ ,  $m$  and  $r$  (and in particular the factorization is unique).*

In many applications, we have to deal with noisy NMF  $\hat{M} = AW + \text{noise}$  where (potentially correlated) noise is added to rows of the data matrix  $M$ . Suppose every row is perturbed by a small noise  $\epsilon$  (in  $\ell_2$  norm), we would like the algorithm to be robust to such additive noise. We need a generalization of properly filled facets.

**Definition 4.2** ( $(N, H, \gamma)$  properly filled facets). *Given a NMF  $M = AW$ , a set of facets  $S_1, \dots, S_k \in [r]$  of  $W$  is  $(N, H, \gamma)$  properly filled if it satisfies the following properties:*

1. *In any set  $|S_i| > 1$ , there is a row  $i^*$  in  $A$  whose support is equal to  $S_i$ , and is in the convex hull*

*of other rows of  $A$ . There exists a convex combination  $M^{i^*} = \sum_{i \in [n] \setminus i^*} w_i M^i$ , such that the matrix  $\sum_{i \in [n] \setminus i^*} w_i (M^i)(M^i)^T$  has rank  $|S_i|$ , and the smallest nonzero singular value is at least  $\gamma$ . We call this special point  $M^{i^*}$  the center for this facet.*

2. *For any set  $|S_i| > 1$ , there are at least  $N$  rows in  $A$  whose support is exactly equal to  $S_i$ .*
3. *For any subspace  $Q$  of dimension  $1 < t < r$ , if there are at least  $N$  rows of  $M$  in an  $\epsilon$ -neighborhood of  $Q$ , then there exists a non-singleton set  $S_i$  with corresponding subspace  $Q_i$  such that  $\|P_{Q^\perp} Q_i\| \leq H\epsilon$ .*

Intuitively, if we represent the center point as a convex combination of other points, the only points that have a nonzero contribution must be on the same facet as the center. Condition 1 then ensures there is a “nice” convex combination that allows us to robustly recover the subspace corresponding to the facet even in presence of noise. Condition 2 shows every properly filled facets contain many points, which is why they are different from other subspaces and are the facets of the true solution. Condition 3 is a generalization of the general position property, which essentially says “every subspace that contains many points must be close to a properly filled facet”. In Section 7 we show that under a natural generative model, the NMF has  $(N, H, \gamma)$ -properly filled facets with high probability.

Properly filled facets is a property of how the points  $M^i$  are distributed on the facets of  $W$ . The geometry of the  $W$ -simplex itself also affects the accuracy of our Face-Intersect algorithm.

**Definition 4.3.** *A matrix  $W \in \mathbb{R}^{r \times m}$  ( $r \leq m$ ) is  $\alpha$ -robust if its rows have norm bounded by 1, and its  $r$ -th singular value is at least  $\alpha$ .*

Under these assumptions we prove that Face-Intersect robustly learns the unknown simplex  $W$ .

**Theorem 4.2.** *Suppose  $M = AW$  is subset separable by  $S_1, \dots, S_k$  and these facets are  $(N, H, \gamma)$  properly filled, and the matrix  $W$  is  $\alpha$ -robust. Then given  $\hat{M}$  whose rows are within  $\ell_2$  distance  $\epsilon$  to  $M$ , with  $\epsilon < o(\alpha^4 \gamma / Hr^3)$ , Algorithm Face-Intersect finds  $\hat{W}$  such that there exists a permutation  $\pi$  and for all  $i$   $\|\hat{W}_i - W_{\pi(i)}\| \leq O(Hr^2 \epsilon / \alpha^2 \gamma)$ . The running time is polynomial in  $n$ ,  $m$  and  $r$ .*

A vertex  $j \in [r]$  is an intersection vertex if there exists a subset of properly filled facets  $\{S_{j_k} : |S_{j_k}| \geq 2\}$  such that  $j = \cap_k S_{j_k}$ . Since the first module of Face-Intersect, Algorithm 3, only finds non-singleton facets, the intersection vertices are all the vertices that we could find using these facets. The last module of Face-Intersect finds all the remaining vertices of the simplex.

---

**Algorithm 1** Face-Intersect

Run Algorithm 3 to find subspaces that correspond to properly filled facets  $S_1, S_2, \dots, S_k$  where  $|S_i| \geq 2$ .  
 Run Algorithm 5 to find the *intersection vertices*  $P$ .  
 Run Algorithm 5 (similar to Algorithm 4 in (Arora et al., 2013)) to find the singleton points (anchors).  
 Given  $\hat{M}, \hat{W}$ , compute  $\hat{A}$ .

---

**Our approach** The main idea of our algorithm is to first find the subspaces corresponding properly filled facets, then take the intersections of these facets to find the intersection vertices. Finally we adapt the algorithm from (Arora et al., 2013) to find the remaining vertices that correspond to singleton sets.

- **Finding facets** For each row of  $M$ , we try to represent it as the convex combination of other rows of  $M$ . We use an iterative algorithm to make sure the span of points used in this convex combination is exactly the subspace corresponding to the facet.
- **Removing false positives** The previous step will generate subspaces that correspond to properly filled facets, but it might also generate false positives (subspaces that do not correspond to any properly filled facets). Condition 3 in Definition 4.2 allows us to filter out these false positives as these subspaces will not contain enough nearby points.
- **Finding intersection vertices** We design an algorithm that systematically tries to take the intersections of subspaces in order to find the intersection vertices. This relies on the subset-separable property and robustness properties of the simplex. This step computes at most  $O(nr)$  subspace intersection operations.
- **Finding remaining vertices** The remaining vertices correspond to the singleton sets. This is similar to the separable case and we use an algorithm from Arora et al. (2013).

## 5. Finding properly filled facets

In this section we show how to find properly filled facets  $S_i$  with  $|S_i| \geq 2$ . The singleton facets (anchors) are not considered in this section, since they will be found through a separate algorithm. We first show how to find a properly filled facet if we know its center (Condition 1 in Definition 4.2). Then to find all the properly filled facets we enumerate points to be the center and remove false positives.

**Finding one properly filled facet** Given the center point, if there is no noise then when we represent this point as convex combinations of other points, all the points with positive weight will be on the same facet. Intuitively the span of these points should be equal to the subspace cor-

responding to the facet. However there are two key challenges here: first we need to show that when there is noise, points with large weights in the convex combination are close to the true facet; second, it is possible that points with large weights only span a lower dimensional subspace of the facet. Condition 1 in Definition 4.2 guarantees that there exists a *nice* convex combination that spans the entire subspace (and robustly so because the smallest singular value is large compared to noise). In Algorithm 2, we iteratively improve our convex combination and eventually converge to this nice combination.

---

**Algorithm 2** Finding a properly filled facet

**input** points  $\hat{v}^1, \hat{v}^2, \dots, \hat{v}^n$ , and center point  $\hat{v}^0$  (Condition 1 in Definition 4.2).

**output** the proper facet containing  $\hat{v}^0$ .

- 1: Maintain a subspace  $\hat{Q}$  (initially empty)
- 2: Iteratively solve the following optimization program:

$$\begin{aligned} \max \quad & \text{tr}(P_{\hat{Q}^\perp} \sum_{i=1}^n w_i \hat{v}^i (\hat{v}^i)^T P_{\hat{Q}^\perp}) \\ & \forall i \in [n] \quad w_i \geq 0 \\ & \sum_{i=1}^n w_i = 1 \\ & \|\hat{v}^0 - \sum_{i=1}^n w_i \hat{v}^i\| \leq 2\epsilon \\ \text{diag}(\hat{Q}^T \left( \sum_{i=1}^n w_i \hat{v}^i (\hat{v}^i)^T \right) \hat{Q}) & \geq \gamma/2. \end{aligned}$$

- 3: Let  $\hat{Q}$  be the top singular space of  $(\sum_{i=1}^n w_i \hat{v}^i (\hat{v}^i)^T)$  for singular values larger than  $\gamma/2d$ .
  - 4: Repeat until the dimension of  $\hat{Q}$  does not increase.
- 

**Theorem 5.1.** *Suppose  $\|\hat{v}^i - v^i\| \leq \epsilon$ ,  $v^0$  is the center point of a properly filled facet  $S \subset [r]$  with  $|S| = d$ , and the unknown simplex  $W$  is  $\alpha$ -robust, when  $d\sqrt{r}\epsilon/\alpha\gamma \ll 1$  Algorithm 2 stops within  $d$  iterations, and the subspace  $\hat{Q}$  is within distance  $O(\sqrt{r}\epsilon/\alpha\gamma)$  to the true subspace  $Q_S$ .*

The intuition of Algorithm 2 is to maintain a convex combination for the center point. We show for any convex combination, the top singular space associated with the combination,  $\hat{Q}$ , is always close to a subspace of the true space  $Q_S$ . The algorithm then tries to explore other directions by maximizing the projection that is outside the current subspace  $\hat{Q}$  (the objective function of the convex optimization), while maintaining that the current subspace have large singular values (the last constraint). In the proof we show since there is a nice solution, the algorithm will always be able to make progress until the final solution is a *nice* convex combination.

**Finding all subsets** Algorithm 2 can find one properly filled facet, if we have its center point (Condition 1 in Definition 4.2). In order to find all the properly filled facets, we enumerate through rows of  $M$  and prune false positives using Condition 3 in Definition 4.2

---

**Algorithm 3** Finding all proper facets
 

---

**input**  $\hat{M}$  whose factorization is subset-separable with  $(N, H, \gamma)$ -properly filled facets.

- 1: **for**  $i = 1$  to  $n$  **do**
- 2: Let  $\hat{v}^0 = \hat{M}^i$  and  $\hat{v}^1, \dots, \hat{v}^{n-1}$  be the rest of vertices.
- 3: Run Algorithm 2 to get a subspace  $Q$ .
- 4: If  $\dim(Q) < r$ , and there are at least  $N$  points that are within distance  $O(\sqrt{r}\epsilon/\alpha\gamma)$  add it to the collection of subspaces.
- 5: **end for**
- 6: Let  $Q$  be a subspace in the collection, remove  $Q$  if there is a subspace  $Q'$  with  $\dim(Q') < \dim(Q)$  and  $\|P_{Q^\perp} Q'\| \leq O(H\sqrt{r}\epsilon/\alpha\gamma)$
- 7: Merge all subspaces that are within distance  $O(H\sqrt{r}\epsilon/\alpha\gamma)$  to each other.

---

**Theorem 5.2.** *If  $H\sqrt{r}\epsilon/\alpha\gamma = o(\alpha)$ , then the output of Algorithm 3 contains only subspaces that are  $\epsilon_S = O(H\sqrt{r}\epsilon/\alpha\gamma)$ -close to the properly filled facets, and for every properly filled facet there is a subspace in the output that is  $\epsilon_S$  close.*

## 6. Finding intersections

Given an subset-separable NMF with  $(N, H, \gamma)$ -properly filled facets, let  $Q_i$  denote the subspace associated with a set  $S_i$  of vertices:  $Q_i = \text{span}(W^{S_i})$ . For all properly filled facets with at least two vertices, Algorithm 3 returns noisy versions of the subspaces  $\hat{Q}_i$  that are  $\epsilon_S$  close to the true subspaces. Without loss of generality, assume the first  $h$  facets are non-singletons. Our goal is to find all the intersection vertices  $\{W^i : i \in P\}$ . Recall that intersection vertices are the unique intersections of subsets of  $S_1, \dots, S_h$ . We can view this as a set intersection problem:

**Set Intersections** We are given sets  $S_1, S_2, \dots, S_h \subset [r]$ . There is an unknown set  $P \subset [r]$  such that  $\forall i \in P$  there exists  $\{S_{i_k}\}$  and  $i = \cap_k S_{i_k}$ . Our goal is to find the set  $P$ .

This problem is simple if we know the subsets of  $W^j$  in each facet. However, since what we really have access to are subspaces, it is impossible to identify the vertices unless we have a subspace of dimension 1. On the other hand, we can perform intersection and linear-span for the subspaces, which correspond to intersection and union for the sets. We also know the size of a set by looking at the dimension of the subspace. The main challenge here is that we cannot afford to enumerate all the possible combinations of the sets,

---

**Algorithm 4** Finding Intersection
 

---

**input**  $k$  sets  $S_1, \dots, S_h$ .

**output** A set  $P$  that has all the intersection vertices.

Initialize  $P = \emptyset, R = \emptyset$ .

**for**  $i = 1$  to  $r$  **do**

Let  $S = [r]$

**for**  $j = 1$  to  $h$  **do**

**if**  $|S \cap S_j| < |S|$  and  $S \cap S_j \not\subseteq R$  **then**

$S = S \cap S_j$

**end if**

**end for**

$R = R \cup S$

Add  $S$  to  $P$  if  $|S| = 1$ .

**end for**

---

and also there are vertices that are not intersection vertices and they may or may not appear in the sets we have. The idea of the algorithm is to keep vertices that we have already found in  $R$ , and try to avoid finding the same vertices by making sure  $S$  is never a subset of  $R$ . We show after every inner-loop one of the two cases can happen: in the first case we find an element in  $P$ ; in the second case  $S$  is a set that satisfies  $(S \setminus R) \cap P = \emptyset$ , so by adding  $S$  to  $R$  we remove some of the vertices that are not in  $P$ . Since the size of  $R$  increases by at least 1 in every iteration until  $R = [r]$ , the algorithm always ends in  $r$  iterations and finds all the vertices in  $P$ . In practice, we implement all the set operations in 4 using the analogous subspace operations (see Algorithm 6 in Appendix). We prove the following :

**Theorem 6.1.** *When  $W$  is  $\alpha$ -robust and  $\epsilon_S < o(\alpha^3/r^{2.5})$ , Algorithm 6 finds all the intersection vertices of  $W$ , with error at most  $\epsilon_v = 4r^{1.5}\epsilon_S/\alpha$ .*

---

**Algorithm 5** Finding remaining vertices
 

---

**input** matrix  $\hat{M}$ , intersection vertices  $\hat{W}^1, \dots, \hat{W}^{|P|}$ .

**output** remaining vertices  $\hat{W}^{|P|+1}, \dots, \hat{W}^r$ .

**for**  $i = |P| + 1$  TO  $r$  **do**

Let  $Q = \text{span}\{\hat{W}^1, \dots, \hat{W}^{i-1}\}$ .

Pick the point  $\hat{M}^j$  with largest  $\|P_{Q^\perp} \hat{M}^j\|$ , let  $\hat{W}^i = \hat{M}^j$ .

**end for**

---

**Finding the remaining vertices** The remaining vertices correspond to singleton sets in subset-separable assumption. They appear in rows of  $M$ . The situation is very similar to the separable NMF and we use an algorithm from (Arora et al., 2013) to find the remaining vertices. For completeness we describe the algorithm here. By Lemma 4.5 in (Arora et al., 2013) we directly get the following theorem:

**Theorem 6.2.** *If vertices already found have accuracy  $\epsilon_v$  such that  $\epsilon_v \leq \alpha/20r$ , Algorithm 5 outputs the remaining vertices with accuracy  $O(\epsilon/\alpha^2) < \epsilon_v$ .*

**Running time.** Face-Intersect (Algorithm 1) has 3 parts: find facets (Algorithm 3), find intersections (Algorithm 4) and find remaining anchors (Algorithm 5). We discuss the runtime of each part. We first do dimension reduction to map the  $n$  points to an  $r$ -dimensional subspace to improve the running time of later steps. The dimension reduction takes  $O(nmr)$  time, where  $n, m$  are the number of rows and columns of  $M$ , respectively, and  $r$  is the rank of the factorization. Algorithm 3's runtime is  $O(nd \cdot \text{OPT})$ , where  $d$  is the max dimension of properly filled facets (typically  $d < r \ll m$ ).  $\text{OPT}$  is the time to solve the convex optimization problem in Algorithm 2.  $\text{OPT}$  is essentially equivalent to solving an LP with  $n$  nonnegative variables and  $r + d$  constraints. Algorithm 4's runtime is  $O(kr^4)$  where  $k$  is the number of properly filled facets; typically  $k \ll n$ . Algorithm 5's runtime is  $O(nr^3)$ . The overall runtime of Face-Intersect is  $O(mnr + nd \cdot \text{OPT} + kr^4 + nr^3)$ . Calling the  $\text{OPT}$  routine is the most expensive part of the algorithm. Empirically, we find that the algorithm converges after  $\sim k \ll nd$  calls to  $\text{OPT}$ .

## 7. Generative model of NMF naturally creates properly filled facets

To better understand the generality of our approach, we analyzed a simple generative model of subset-separable NMFs and showed that properly filled facets naturally arise with high probability.

**Generative Model** Given a simplex  $W$  that is  $\alpha$ -robust and a subset of facets  $S_1, S_2, \dots, S_k$  that is subset separable. Let  $p_i$  be the probability associated with facet  $i$ , and let  $p_{\min} = \min_{i \leq k} p_i$  and  $d = \max_{i \in [k]} |S_i|$ . For convenience, denote  $S_0 = [r]$  and  $p_0 = 1 - \sum_{i=1}^k p_i$ . To generate a sample, first sample facet  $S_i$  with probability  $p_i$ , and then uniformly randomly sample a point within the convex hull of the points  $\{W^j : j \in S_i\}$ . Here we think of  $d$  as a small constant or  $O((\log n) / \log \log n)$  (in general  $d$  can be much smaller than  $r$ ). For example, separability assumption implies  $d = 1$ , and it is already nontrivial when  $d = 2$ .

**Theorem 7.1.** *Given  $n = \Omega(\max\{(4d)^d \log(d/\eta), kr^2 \log(d/p_{\min}\eta)\} / p_{\min})$  samples from the model, with high probability the facets  $S_1, \dots, S_k$  are  $(p_{\min}n/2, 200r^{1.5}/p_{\min}\alpha, \alpha^2/16d)$  properly filled.*

The proof relies on the following two lemmas. The first lemma shows that once we have enough points in a simplex, then there is a center point with high probability.

**Lemma 7.2.** *Given  $n = \Omega((4d)^d \log d / \eta)$  uniform points  $v^1, v^2, \dots, v^n$  in a standard  $d$ -dimensional simplex (with vertices  $e_1, e_2, \dots, e_d$ ), with probability  $1 - \eta$  there exists a point  $v_i$  such that  $v_i = \sum_{j \neq i} w_j v^j$  ( $w_j \geq 0, \sum_{j \neq i} w_j =$*

*1), and  $\sigma_{\min}(\sum_{j \neq i} w_j (v^j)(v^j)^T) \geq 1/16d$ .*

The next lemma shows unless a subspace contains a properly filled facet, it cannot contain too many points in its neighborhood.

**Lemma 7.3.** *Given  $n = \Omega(d^2 \log(d/p_{\min}\eta) / p_{\min})$  uniform points  $v^1, v^2, \dots, v^n$  in a standard  $d$ -dimensional simplex (with vertices  $e_1, e_2, \dots, e_d$ ), with probability  $1 - \eta$  for all matrices  $A$  whose largest column norm is equal to 1, there are at most  $p_{\min}n/4$  points with  $\|Av^i\| \leq p_{\min}/200d$ .*

## 8. Experiments

While our algorithm has strong theoretical guarantees, we additionally performed proof-of-concept experiments to show that when the noise is relatively small, our algorithm can outperform the state-of-art NMF algorithms. We simulated data according to the generative NMF model described in Section 7. We first randomly select  $r$  nonnegative vectors in  $\mathbb{R}^m$  as rows of the  $W$  matrix. We grouped the vertices  $W^j$  into  $r$  groups,  $S_1, \dots, S_r$  of three elements each, such that each vertex is the unique intersection of two groups. Each  $S_i$  then corresponds to a 2-dim facet. To generate the  $A$  matrix, for each  $S_i$ , we randomly sampled  $n_1$  rows of  $A$  with support  $S_i$ , where each entry is an i.i.d. from  $\text{Unif}(0, 1)$ . An additional  $n_2$  rows of  $A$  were sampled with full support. These correspond to points in the interior of the simplex. We tested a range of settings with  $m$  between 5 to 100,  $r$  between 3 to 10, and  $n_1$  and  $n_2$  between 100 and 500. We generated the true data as  $M = AW$  and added i.i.d. Gaussian noise to each entry of  $M$  to generate the observed data  $\tilde{M}$ .

There are many algorithms for solving NMF, most of them are either iterative algorithms that have no guarantees, or algorithms that work only under separability condition. We choose two typical algorithms: the Anchor-Words algorithm (Arora et al., 2013) for separable NMF, and Projected Gradient (Lin, 2007) for iterative algorithms. For each simulated NMF, we evaluated the output factors  $\hat{A}, \hat{W}$  of these algorithms on three criteria: accuracy of the reconstructed anchors to the true anchors,  $\|W - \hat{W}\|_2$ ; accuracy of the reconstructed data matrix to the *observed* data,  $\|\tilde{M} - \hat{A}\hat{W}\|_2$ ; accuracy of the reconstructed data to the *true* data,  $\|M - \hat{A}\hat{W}\|_2$ . In Figure 2, we show the results for the three methods under the setting  $n_1 = 100, n_2 = 100, m = 10, r = 5$ . We grouped the results by the noise level of the experiment, which is defined to be the ratio of the average magnitude of the noise vectors to the average magnitude of the data points in  $\mathbb{R}^m$ . Face-Intersect is substantially more accurate in reconstructing the  $W$  matrix compared to Anchor-Words and Projected Gradient. In terms of reconstructing the  $\tilde{M}$  and  $M$  matrices, Face-Intersect slightly outperforms Anchor-Words ( $p < 0.05$  t-test), and

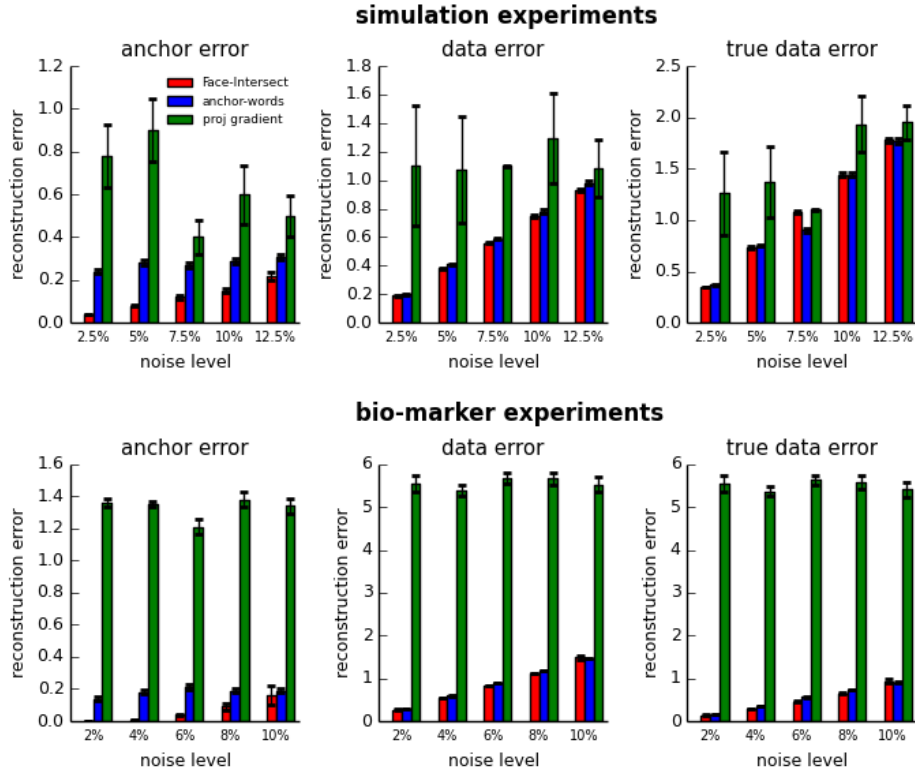


Figure 2. Reconstruction accuracy of the three NMF algorithm as a function of data noise. Standard error shown in the error bars.

they both were substantially more accurate than Projected Gradient. As noise level increased, the accuracy of Face-Intersect and Anchor-Words degrades and at noise around 12.5%, the accuracy of the three methods converged. In many applications, we are more interested in accurate reconstruction of the latent  $W$  than of  $M$ . For example, in bio-medical applications, each row of  $M$  is a sample and each column is the measurement of that sample at a particular bio-marker. Each sample is typically a mixture of  $r$  cell-types, and each cell-type corresponds to a row of  $W$ . The  $A$  matrix gives the mixture weights of the cell-types into the samples. Given measurement on a set of samples,  $M$ , an important problem is to infer the values of the latent cell-types at each bio-marker,  $W$  (Zou et al., 2014). To create a more realistic simulation of this setting, we used DNA methylation values measures at 100 markers in 5 cell-types (Monocytes, B-cells, T-cells, NK-cells and Granulocytes) as the true  $W$  matrix (Zou et al., 2014). From these 5 anchors we generated 600 samples—which is a typical size of such datasets—using the same procedure as above. Both Face-Intersect and Anchor-Words substantially outperformed Projected Gradient across all three reconstruction criteria. In terms of reconstructing the biomarker matrix  $W$ , Face-Intersect was significantly more accurate than Anchor-Words. For reconstructing the data matrices  $M$  and  $\tilde{M}$ , Face-Intersect was statistically more accurate than

Anchor-Words when the noise is less than 8% ( $p < 0.05$ ), though the magnitude of the difference is small.

**Discussion** We have presented the notion of subset separability, which substantially generalizes separable NMFs and is a necessary condition for the factorization to be unique or to have minimal volume. This naturally led us to develop the Face-Intersect algorithm, and we showed that when the NMF is subset separable and have properly filled facets, this algorithm provably recovers the true factorization. Moreover, it is robust to small adversarial noise. We show that the requirements for Face-Intersect to work are satisfied by simple generative models of NMFs. The original theoretical analysis of separable NMF led to a burst of research activity. Several highly efficient NMF algorithms were inspired by the theoretical ideas. We are hopeful that the idea of subset-separability will similarly lead to practical and theoretically sound algorithms for a much larger class of NMFs. Our Face-Intersect algorithm and its analysis is a first proof-of-concept that this is a promising direction. In exploratory experiments, we showed that under some settings where the relative noise is low, the Face-Intersect algorithm can outperform state-of-art NMF solvers. An important agenda of research will be to develop more robust and scalable algorithms motivated by our subset-separability analysis.



## References

- Arora, S., Ge, R., Kannan, R., and Moitra, A. Computing a nonnegative matrix factorization – provably. In *STOC*, pp. 145–162, 2012.
- Arora, Sanjeev, Ge, Rong, Halpern, Yoni, Mimno, David M., Moitra, Ankur, Sontag, David, Wu, Yichen, and Zhu, Michael. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 28 (2), pp. 280–288. JMLR: W&CP, 2013.
- Bittorf, V., Recht, B., Re, C., and Tropp, J. Factoring non-negative matrices with linear programs. In *NIPS*, 2012.
- Devarajan, K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol*, 2009.
- Donoho, D. and Stodden, V. When does non-negative matrix factorization give the correct decomposition into parts? In *NIPS*, 2003.
- Dwyer, Rex A. On the convex hull of random points in a polytope. *Journal of Applied Probability*, 25 (4):pp. 688–699, 1988. ISSN 00219002. URL <http://www.jstor.org/stable/3214289>.
- Efron, Bradley. The convex hull of a random set of points. *Biometrika*, 52(3-4):331–343, 1965.
- Elhamifar, Ehsan and Vidal, René. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2790–2797. IEEE, 2009.
- Gillis, N. Robustness analysis of hottopixx, a linear programming model for factoring nonnegative matrices. 2012. <http://arxiv.org/abs/1211.6687>.
- Gillis, N. The why and how of nonnegative matrix factorization, 2014. <http://arxiv.org/abs/1401.5226>.
- Gillis, N. and Vavasis, S.A. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(4):698–714, April 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.226.
- Gomez, C., Borgne, H. Le, Allemand, P., Delacourt, C., and Ledru, P. N-findr method versus independent component analysis for lithological identification in hyperspectral imagery. *Int. J. Remote Sens.*, 28(23), January 2007.
- Hardt, Moritz and Moitra, Ankur. Algorithms and hardness for robust subspace recovery. In *COLT*, pp. 354–375, 2013.
- Kumar, A., Sindhvani, V., and Kambadur, P. Fast conical hull algorithms for near-separable non-negative matrix factorization. 2012. <http://arxiv.org/abs/1210.1190v1>.
- Lee, D. and Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature*, pp. 788–791, 1999.
- Lee, D. and Seung, H. Algorithms for non-negative matrix factorization. In *NIPS*, pp. 556–562, 2000.
- Lin, Chih-Jen. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.
- Moitra, Ankur. An almost optimal algorithm for computing nonnegative rank. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pp. 1454–1464, 2013.
- Nascimento, J.M. P. and Dias, J. M. B. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE TRANS. GEOSCI. REM. SENS*, 43:898–910, 2004.
- Stewart, G.W. and Sun, J. *Matrix perturbation theory*, volume 175. Academic press New York, 1990.
- Vavasis, S. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, pp. 1364–1377, 2009.
- Vidal, René. A tutorial on subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2010.
- Xu, W., Liu, X., and Gong, Y. Document clustering based on non-negative matrix factorization. In *SIGIR*, pp. 267–273, 2003.
- Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. Genome-wide association studies without the need for cell-type composition. *Nature Methods*, pp. 309–311, 2014.

## A. Subset Separability and minimal volume

In this section we prove Proposition 3.1 subset separability condition is necessary for a minimal volume solution.

*Proof.* Suppose  $M = AW$  is a rank- $r$  nonnegative matrix factorization with minimal volume. If this decomposition does not satisfy the subset-separable condition, then there exists  $i \neq j \in R$  such that for every row  $A^t$ , the two entries  $A_{t,i}, A_{t,j}$  are either all zero or all nonzero. That is, the columns  $A_i$  and  $A_j$  have the same support. Consider a new factorization  $A'W'$ , where the columns of  $A'$  are the same as columns of  $A$  except for columns  $i, j$ , and rows of  $W'$  are the same as rows of  $W$  except for row  $i$ .

Let  $A'_i = \frac{1}{1-\epsilon}A_i$ ,  $A'_j = A_j - \frac{\epsilon}{1-\epsilon}A_i$ , and  $(W')^i = (1-\epsilon)W^i + \epsilon W^j$ , it is easy to verify that  $A'W' = AW = M$ , and  $W'$  is still nonnegative for  $\epsilon \in [0, 1]$ .

Since the support of  $A_i$  and  $A_j$  are the same, there exists a positive  $\epsilon$  such that  $A'_j$  is still a nonnegative vector. In that case  $A'W'$  is a valid nonnegative matrix factorization where only one row of  $W'$  is different from  $W$ . By construction it is clear that the volume of  $W'$  is equal to  $(1-\epsilon)$  times the volume of  $W$ , so this contradicts with the assumption that  $M = AW$  is a factorization with minimal volume.  $\square$

## B. Detailed analysis for finding properly filled facets

In this section we analyze Algorithms 2 and 3.

### B.1. Finding one properly filled facet

We first prove Theorem 5.1. For this algorithm, it is more natural to use the following robustness condition, which is a corollary of  $\alpha$ -robustness.

**Lemma B.1.** *Suppose the vertices of the unknown simplex are rows of  $W \in \mathbb{R}^r \times n$ , and  $W$  is  $\alpha$ -robust. For any face  $S$  of  $W$  with corresponding subspace  $Q$ , there exists a unit vector  $h \perp Q$ , let  $v$  be any vector in the simplex and  $v^\perp$  be its component that is orthogonal to  $Q$ , then  $\frac{|h \cdot v^\perp|}{\|v^\perp\|} \geq \frac{\alpha}{\sqrt{r}}$ .*

*Proof.* Suppose  $Q$  has dimension  $d$  (we know  $d < r$ ). Let  $B$  be the projection of  $W$  to the orthogonal subspace of  $Q$ , and remove the 0-columns in  $B$ . The matrix  $B$  is a  $n \times (r-d)$  matrix whose smallest singular value is at least  $\alpha$  (the smallest singular value in a projection is at least the smallest singular value of the matrix). We construct  $h$  as  $h = \frac{B^\dagger \mathbf{1}}{\|B^\dagger \mathbf{1}\|}$ . By the property of  $B$  we know  $h \cdot B_i = \frac{1}{\|B^\dagger \mathbf{1}\|} = \alpha/\sqrt{r}$ .

For any vector  $v$  in the simplex, its orthogonal component  $v^\perp$  is equal to  $P_{Q^\perp}(\sum_{i=1}^r w_i W^i)$ , which is a nonnegative combination of columns in  $B$ . Therefore  $\frac{|h \cdot v^\perp|}{\|v^\perp\|} = \frac{\sum_i w_i h_i \cdot B_i}{\|\sum_i w_i B_i\|} \geq \max_i \frac{h_i \cdot B_i}{\|B_i\|} = \alpha/\sqrt{r}$  (here we used the fact that  $h \cdot B_i$  are all positive).  $\square$

As we explained, there are two challenges in proving Theorem 5.1: 1). the observations are noisy. We would like to show even with the noisy  $\hat{v}$ 's, the subspace  $\hat{Q}$  is always close to a subspace of the true space  $Q$ ; 2). the convex combination may not find the entire space  $Q$ , for which we show the dimension of  $\hat{Q}$  will increase until it is equal to the dimension of  $Q$ . Throughout this section we will use  $d$  to denote the dimension of true space  $Q$ .

We first show that in every step of the algorithm all the vectors in  $\hat{Q}$  are close to the subspace  $Q$ . We start by proving a general perturbation lemma for singular subspaces:

**Lemma B.2.** *Let  $\hat{F} = F + E$  where both  $\hat{F}$  and  $F$  are positive semidefinite,  $F$  is a rank  $d$  matrix with column span  $Q$ , and  $U$  is the top  $t$  ( $t \leq d$ ) singular space of  $\hat{F}$  with the  $t$ -th singular value  $\sigma_t(\hat{F}) > \|E\|$ , then  $\|P_{Q^\perp} U\| \leq \|E\|/\sigma_t(\hat{F})$ .*

*Proof.* Let  $UDU^T$  be the truncated top  $t$  SVD of  $\hat{F}$ , and  $\hat{U}\hat{D}\hat{U}^T$  be the full SVD. We know

$$\|P_{Q^\perp} \hat{U} \hat{D}\| = \|P_{Q^\perp} \hat{F}\| = \|P_{Q^\perp} E\| \leq \|E\|,$$

where the first equality is because  $\hat{U}$  is an orthonormal matrix, and the second equality is because the column span of  $F$  is inside  $Q$ .

On the other hand,  $P_{Q^\perp}UD$  is a submatrix of  $P_{Q^\perp}\hat{U}\hat{D}$ , so we know  $\|P_{Q^\perp}UD\| \leq \|P_{Q^\perp}\hat{U}\hat{D}\| \leq \|E\|$ . Since all the entries in  $D$  are at least  $\sigma_t(\hat{F})$ , this implies  $\|P_{Q^\perp}U\| \leq \|E\|/\sigma_t(\hat{F})$ .  $\square$

In the later proofs we usually think of  $\hat{F}$  as  $(\sum_{i=1}^n w_i \hat{v}^i (\hat{v}^i)^T)$ , and  $F$  as  $(\sum_{i=1}^n w_i P_Q v^i (v^i)^T P_Q)$ . The next lemma shows that for any feasible solution of the optimization program (even just considering the first three constraints), the matrix  $\hat{F}$  is close to  $F$ :

**Lemma B.3.** *For any feasible solution that satisfies the first three constraint, let  $\hat{F} = (\sum_{i=1}^n w_i \hat{v}^i (\hat{v}^i)^T)$  and  $F = (\sum_{i=1}^n w_i P_Q v^i (v^i)^T P_Q)$ , we have  $\|E\| = \|\hat{F} - F\| \leq O(\sqrt{r}\epsilon/\alpha)$ . In fact, even the nuclear norm<sup>1</sup>  $\|\hat{F} - F\|_* \leq O(\sqrt{r}\epsilon/\alpha)$ .*

*Proof.* Let  $\tilde{F} = (\sum_{i=1}^n w_i v^i (v^i)^T)$ , we show  $\tilde{F}$  is close to both  $F$  and  $\hat{F}$ . Let  $\delta_i = \hat{v}^i - v^i$ . By assumption we know  $\|\delta_i\| \leq \epsilon$ . Also, by assumption  $\|v^i\| \leq 1$  (normalization) so  $\|\hat{F} - \tilde{F}\| \leq \sum_{i=1}^n w_i \|\delta_i (v^i)^T + v^i \delta_i^T + \delta_i \delta_i^T\| \leq (2\epsilon + \epsilon^2) \sum_{i=1}^n w_i = O(\epsilon)$ .

On the other hand, by the third constraint we know  $\|\hat{v}^0 - \sum_{i=1}^n w_i \hat{v}^i\| \leq 2\epsilon$ , which implies  $\|v^0 - \sum_{i=1}^n w_i v^i\| \leq 4\epsilon$  (because  $\|v^i - \hat{v}^i\| \leq \epsilon$  and  $w_i$ 's form a probability distribution). Using the robustness condition, let  $v^{i\perp} = v^i - P_Q v_i$ , then

$$4\epsilon > \left\| \sum_i w_i v^i - v^0 \right\| \geq \sum_i w_i h_I \cdot v^{i\perp} \geq \Delta \sum_{i=1}^n w_i \|v^{i\perp}\|$$

Therefore we know  $\|\tilde{F} - F\| \leq \sum_{i=1}^n w_i \|v^{i\perp} (v^i)^T P_Q + P_Q v_i (v^{i\perp})^T + (v^{i\perp})(v^{i\perp})^T\| \leq O(\sqrt{r}\epsilon/\alpha)$  (note that  $\|v^{i\perp}\| \leq 1$  by normalization).

The nuclear norm bound follows from exactly the same proof.  $\square$

The previous two lemmas guarantee that at any time of the algorithm, the subspace  $\hat{Q}$  is always close to a subspace of  $Q$ . In the next lemma we show that the algorithm makes progress

**Lemma B.4.** *If  $\dim(\hat{Q}) = t < d$ , then in the next iteration the dimension of  $\hat{Q}$  increases by at least 1.*

*Proof.* Since  $v^0$  is a center of the facet, we know there exists a ‘‘nice’’ solution  $w^*$  such that  $v^0 = \sum_{i=1}^n w_i^* v^i$  and  $\sum_{i=1}^n w_i^* (v^i)(v^i)^T$  has  $d$ -th singular value  $\gamma$ .

We first show that this guaranteed good solution  $w^*$  is always a feasible solution. Clearly it satisfies the first three constraints (by triangle inequality). For the last constraint, let  $F^*$  be the  $F$ -matrix constructed by  $w^*$  and  $\hat{F}^*$  be the corresponding  $\hat{F}$  matrix. By assumption we know  $\hat{F}^*$  has  $\sigma_d(\hat{F}^*) \geq \gamma$ , so in particular for any direction  $u$  in subspace  $Q$ ,  $u^T F^* u \geq \gamma$ . Since by previous two lemmas we have  $\|P_{Q^\perp} \hat{Q}\| \leq O(d\sqrt{r}\epsilon/\alpha\gamma)$ , in particular every column of  $\hat{Q}$  is within  $O(d\sqrt{r}\epsilon/\alpha\gamma)$  with its projection in  $Q$ , we know  $\text{diag}(\hat{Q}^T F^* \hat{Q}) \geq \frac{3}{4}\gamma$  (when  $\sqrt{r}\epsilon/\alpha\gamma$  is smaller than some universal constant). Now by Lemma B.3  $F^*$  and  $\hat{F}^*$  are close (in spectral norm) we have  $\text{diag}(\hat{Q}^T \hat{F}^* \hat{Q}) \geq \gamma/2$ .

Since the solution  $w^*$  is feasible, the optimal solution must have objective value no less than the objective value of  $w^*$ . By the nuclear norm bound, for any subspace  $\hat{Q}$  we know

$$\text{tr}(P_{\hat{Q}^\perp} F P_{\hat{Q}^\perp}) - \text{tr}(P_{\hat{Q}^\perp} \hat{F} P_{\hat{Q}^\perp}) = \text{tr}(P_{\hat{Q}^\perp} (F - \hat{F}) P_{\hat{Q}^\perp}) \leq \|P_{\hat{Q}^\perp} (F - \hat{F}) P_{\hat{Q}^\perp}\|_* \leq \|F - \hat{F}\|_* \leq O(\sqrt{r}\epsilon/\alpha),$$

where we used the fact that the trace of a matrix is always bounded by its nuclear norm, and nuclear norm of a projection is always smaller than nuclear norm of the original matrix<sup>2</sup>.

On the other hand  $\text{tr}(P_{\hat{Q}^\perp} F^* P_{\hat{Q}^\perp}) \geq \text{tr}(F^*) - \sum_{i=1}^t \sigma_i(F^*) \geq \sum_{i=t+1}^d \sigma_i(F^*) \geq \gamma(d-t)$ . So the optimal objective value must be at least  $\gamma(d-t) - O(\sqrt{r}\epsilon/\alpha)$ .

Let  $w$  be the optimal solution and  $F, \hat{F}$  be the corresponding matrices, by the same argument we know

$$\text{tr}(P_{\hat{Q}^\perp} F P_{\hat{Q}^\perp}) \geq \text{tr}(P_{\hat{Q}^\perp} \hat{F} P_{\hat{Q}^\perp}) - O(\sqrt{r}\epsilon/\alpha) \geq \gamma(d-t) - O(\sqrt{r}\epsilon/\alpha).$$

<sup>1</sup>Nuclear norm  $\|M\|_*$  is equal to the sum of singular values of  $M$ , it is also the dual norm of spectral norm in the sense that  $\|M\|_* = \max_{\|A\| \leq 1} \langle A, M \rangle$ .

<sup>2</sup>This follows from the fact that  $\|A\|_* = \max_{\|B\| \leq 1} \langle A, B \rangle$  and spectral norms do not increase after projection.

However,  $F$  is a matrix of rank at most  $d$ , therefore  $\|P_{\hat{Q}}^\perp F P_{\hat{Q}}^\perp\| \geq \gamma/d - O(\sqrt{r}\epsilon/\alpha d)$ . For the  $\hat{F}$  matrix we also have  $\|P_{\hat{Q}}^\perp \hat{F} P_{\hat{Q}}^\perp\| \geq \gamma/2d$  because  $\|\hat{F} - F\|$  is small.

Now for the matrix  $\hat{F}$ , there are  $t+1$  orthogonal directions ( $t$  from  $\hat{Q}$ , and at least one orthogonal to  $\hat{Q}$ ) with singular value at least  $\gamma/2d$ , hence  $\sigma_{t+1}(\hat{F}) \geq \gamma/2d$ . As a result in the next step the dimension of  $\hat{Q}$  increases by at least 1.  $\square$

Now we are ready to prove Theorem 5.1.

*Proof.* (of Theorem 5.1). By Lemma B.4 and Lemma B.3, we know when the algorithm ends we must have  $\dim(\hat{Q}) \geq \dim(Q)$ .

Now by the last constraint, we know  $\sigma_r(\sum_{i=1}^n w_i \hat{v}^i (\hat{v}^i)^T) \geq \gamma/2$ . Combined with Lemma B.3 and Lemma B.2 this implies the final subspace is within distance  $O(\sqrt{r}\epsilon/\alpha\gamma)$ .  $\square$

## B.2. Finding all properly filled facets

Theorem 5.2 follows immediately from Theorem 5.1, for completeness we provide the proof here.

*Proof.* (of Theorem 5.2) By Theorem 5.1, and by Condition 2 in Definition 4.2, when we run Algorithm 2 on a correct center point, the resulting subspace will always be added to the collection. Therefore at the end of the loop for each facet  $S_i$  with at least two vertices, and its corresponding subspace  $Q_i$ , there must be a  $\hat{Q}_i$  in the collection that is  $O(\sqrt{r}\epsilon/\alpha\gamma)$ -close.

On the other hand, by Condition 3 in Definition 4.2 we know every subspace  $\hat{Q}$  that is in the collection must satisfy  $\|P_{\hat{Q}^\perp} Q_i\| \leq O(H\sqrt{r}\epsilon/\alpha\gamma)$  for some true subspace  $Q_i$ . If  $\hat{Q}$  has dimension larger than  $Q_i$ , then  $\|P_{\hat{Q}^\perp} \hat{Q}_i\| \leq \|P_{\hat{Q}^\perp} Q_i\| + \|P_{Q_i^\perp} \hat{Q}_i\| \leq O(H\sqrt{r}\epsilon/\alpha\gamma)^3$ . Therefore all the false positives with higher dimension are removed. The remaining subspaces must be  $O(H\sqrt{r}\epsilon/\alpha\gamma)$ -close to one of the true subspaces.

By the  $\alpha$ -robustness condition, two subspaces corresponding to different facets must have distance at least  $\alpha$ , so when  $H\sqrt{r}\epsilon/\alpha\gamma \leq o(\alpha)$  the subspaces  $\hat{Q}$  close to a true subset cannot be removed. Also, in the last step it is easy to identify the subspaces  $\hat{Q}$  that are close to one true space  $Q_i$ , any one of those will be  $\epsilon_S$ -close to the true subsets.  $\square$

## C. Detailed Analysis for finding intersections

In this section we first prove Theorem 6.1, then we discuss how to apply Algorithm 5 from (Arora et al., 2013) to find the remaining vertices.

The main idea of the implementation is that the subspace  $Z$  will always be close to the span of  $\{W^i : i \in S\}$  where  $S = \bigcap_{j \in U} S_j$ . The subspace  $\Gamma$  will correspond to span of  $\{W^i : i \in R\}$  where  $R$  is the set of points that we have already found. If  $\|P_{\Gamma^\perp} Z\|$  is large then it means  $S$  is not a subset of  $R$ .

For this step we also need a particular corollary of the  $\alpha$ -robustness condition.

**Lemma C.1.** *Suppose the vertices of the unknown simplex are rows of  $W \in \mathbb{R}^{r \times m}$  and  $W$  is  $\alpha$ -robust. Let  $Q_1, Q_2, \dots, Q_t$  be a set of faces that has intersection  $S \subset [r]$ , and  $Q_i^\perp$  be an arbitrary basis for the orthogonal subspace of  $Q_i$ . The matrix  $\Sigma = [Q_1^\perp, Q_2^\perp, \dots, Q_t^\perp]^T$  has a null-space equal to  $\text{span}\{W^i : i \in S\}$  and  $\sigma_{n-|S|}(\Sigma) \geq \alpha/\sqrt{r}$ .*

*Proof.* Clearly all the vectors  $\{W^i : i \in S\}$  are in the null-space of  $\Sigma$  as  $W^i \in Q_j$  for all  $j \in [t]$ . For vectors that are orthogonal to the span of columns of  $W$ , they have projection 1 in all of  $P_{Q_j^\perp}$ 's, and they do not influence the projections within the row span of  $W$ . We only need to prove that within the row span of  $W$ , for all the directions orthogonal to  $\{W^i : i \in S\}$  the matrix still has large singular values.

Let  $S_j$  be the set of vertices that  $Q_j$  contains, we define  $S'_j$  as follows:  $S'_1 = [r] \setminus S_1$ , for all  $j > 1$   $S'_j = [r] \setminus ((\cup_{j' < j} S'_{j'}) \cup S_j)$ . Since  $S$  is the intersection of the vertices, we know  $\cup S'_j = [r] \setminus S$ . Also by construction we know the  $S'_j$ 's are disjoint. For each  $S'_j$ , let  $Q'_j$  be the span of rows of  $W$  with indices in  $[r] \setminus S'_j$ . Since  $[r] \setminus S'_j$  is a superset of  $S_j$ , we know  $Q_j$  is a subspace of  $Q'_j$  and hence  $P_{(Q'_j)^\perp} \preceq P_{Q_j^\perp}$ . For each  $j$  construct  $B^j$  to be the matrix that is an

<sup>3</sup>This uses the variational characterization of  $P_{U^\perp} V = \max_{u \in U, v \in V} \sin \theta(u, v)$  where  $\theta(u, v)$  is the angle between  $u, v$ .

**Algorithm 6** Finding Intersection

**input**  $k$  subspaces  $\hat{Q}_1, \dots, \hat{Q}_h$ .

**output** intersection vertices  $\hat{W}^i$  that corresponds to  $\{W^i : i \in P\}$ 

 Let  $\epsilon_v = 4r^{1.5}\epsilon_S/\alpha$ ,  $\epsilon_Y = 2r\epsilon_v/\alpha$ .

 Maintain list of vertices  $\{\hat{W}^i\}$ , matrix  $Y$ , and subspace  $\Gamma$  that correspond to the left singular space of  $Y$  with singular values larger than  $\alpha/2$ .

**for**  $i = 1$  TO  $r$  **do**

 Maintain set  $U \subset [k]$ ,  $\Sigma = [\hat{Q}_i^\perp : i \in U]$  and  $Z$  be the space of left singular vectors of  $\Sigma$  with singular values at most  $r\epsilon_S$ .

 Initialize  $S = \emptyset$ 
**for**  $j = 1$  TO  $h$  **do**

 Let  $\Sigma' = \Sigma + P_{\hat{Q}_j^\perp}$ , Let  $Z'$  be the space of eigenvectors of  $\Sigma$  with eigenvalues at most  $r\epsilon_S$ .

**if**  $\dim(Z') < \dim(Z)$  and  $\|P_{\Gamma^\perp} Z'\| > \alpha/2$  **then**

 let  $U = U \cup \{j\}$ , replace  $\Sigma$  and  $Z$  by  $\Sigma'$ ,  $Z'$ .

**end if**
**end for**

 Append  $Z$  to  $Y$  ( $Y = [Y, Z]$ ), update  $\Gamma$ .

 If  $\dim(Z) = 1$  then add the direction to list of  $\hat{W}$ 
**end for**

(arbitrary) orthogonal basis of the orthogonal subspace of  $Q'_j$  in span of  $W$ . Let  $B = [B^1, B^2, \dots, B^t] \in \mathbb{R}^{n \times (r-1)}$ . We know  $BB^T \preceq \sum_j P_{(Q'_j)^\perp} \preceq \Sigma \Sigma^T$ . Therefore we only need to show the matrix  $B$  has large smallest singular value.

Now consider the product  $WB$ . By construction of  $B$ , this is a block diagonal matrix (with blocks correspond to  $S'_j$ 's). Since  $W$  is  $\alpha$ -robust we know each block has smallest singular value  $\alpha$ . Therefore  $\sigma_{\min}(WB) \geq \alpha$ , and  $\sigma_{\min}(B) \geq \alpha/\|W\| \geq \alpha/\sqrt{r}$ . By the relationship between  $\Sigma$  and  $B$  we know  $\sigma_{n-1}(\Sigma) \geq \sigma_{\min}(B) \geq \alpha/\sqrt{r}$ .  $\square$

This lemma allows us to take the intersections of subspaces robustly.

We prove the theorem by induction. The induction hypothesis is

**Claim C.1.** *At the end of every outer-loop,  $\hat{W}^i$ 's are  $\epsilon_v = 4r^{1.5}\epsilon_S/\alpha$  close to some vertices in  $W$ ,  $\Gamma$  is  $\epsilon_Y = 2r\epsilon_v/\alpha$ -close to a subspace spanned by  $W^R$  where  $R \subset [r]$  is a subset of vertices. The set  $R$  never contains any vertex in  $P$  that is not already close to one of the elements in the list  $\hat{W}^i$ .*

Clearly this hypothesis is true before the first iteration (everything was empty). Next we analyze the inner-loop of the algorithm. During the inner-loop the algorithm maintains the following properties:

**Lemma C.2.** *The set  $U$  always has size at most  $r - 1$ , the subspace  $Z$  is always  $\epsilon_v = 4r^{1.5}\epsilon_S/\alpha$ -close to the subspace spanned by  $\{A_i : i \in \cap_{j \in U} S_j\}$ .*

*Proof.* After the first element is added to  $U$ , the dimension of  $Z$  is equal to the dimension of some  $\hat{Q}_j$ , which is at most  $r - 1$ . Every time we add an element to  $U$  the dimension of  $Z$  decreases by 1, and when  $\dim(Z)$  becomes 1 the algorithm stops. So there must be at most  $r - 1$  elements in  $U$ . By Lemma C.1 we know if the matrix consist of the true  $Q_j^\perp$ , then it has nullspace equal to the span of  $\{W^i : i \in \cap_{j \in U} S_j\}$ , and all the other directions have eigenvalue at least  $\alpha/\sqrt{r}$ . The difference between  $\Sigma$  and the true matrix is at most  $2r\epsilon_S$ , so when  $2r\epsilon_S < \alpha/4\sqrt{r}$  by matrix perturbation bounds (Wedin's Theorem(Stewart & Sun, 1990)) we know  $Z$  is always  $4r^{1.5}\epsilon_S/\alpha$ -close.  $\square$

Another property is that in the intersection  $\cap_{j \in U} S_j$  there is always an element that is not already found.

**Lemma C.3.**  *$\|P_{\Gamma^\perp} Z\| > \alpha/2$  if and only if  $\cap_{j \in U} S_j$  contains at least one element outside of  $R$ . Further, this ensures  $S = \cap_{j \in U} S_j$  always contains at least one element outside of  $R$  during the inner-loop.*

*Proof.* This is because by induction hypothesis  $\Gamma$  is  $\epsilon_Y$ -close to the row span of  $W^R$ . On the other hand by Lemma C.2 we know  $Z$  is  $\epsilon_v$  close to the row span of  $W^{\cap_{j \in U} S_j}$ . If  $\cap_{j \in U} S_j \subset R$  then the row span of  $W^{\cap_{j \in U} S_j}$  is a subspace of the row span of  $W^R$ , and  $\|P_{\Gamma^\perp} Z\| \leq \epsilon_Y + \epsilon_v \ll \alpha/2$ .

On the other hand, if  $\cap_{j \in U} S_j$  has an element that is outside  $R$ , then since  $W$  is  $\alpha$ -robust, there is a direction in  $W^{\cap_{j \in U} S_j}$  that has distance at least  $\alpha$  to the row span of  $W^R$ . By triangle inequality the distance between  $P_{\Gamma^\perp} Z \geq \alpha - \epsilon_Y - \epsilon_v > \alpha/2$ .

The last statement of the lemma then follows directly because this is true initially ( $S = [r]$  initially) and the conditions in the if-statement ensures this property is preserved.  $\square$

Using these two properties we know whenever the inner-loop adds a point to the list  $\hat{W}^i$  then it must be  $\epsilon_v$  close to one of the unfound  $W^i$ 's (which is the first part of the induction hypothesis). Next we prove if at the end of the inner-loop  $\dim(Z)$  is more than 1, then  $\cap_{j \in U} S_j$  does not contain any vertices in  $P \setminus R$ .

**Lemma C.4.** *If  $\dim(Z)$  is more than 1 after the inner-loop, then  $\cap_{j \in U} S_j$  does not contain any vertices in  $P \setminus R$ .*

*Proof.* Assume towards contradiction that  $\dim(Z) > 1$  and there is an element  $i \in P \setminus R$  and  $i \in \cap_{j \in U} S_j$ . Let  $S = \cap_{j \in U} S_j$  after the inner-loop. By assumption and by Lemma C.3 we know  $S$  has at least two elements, one of them must be  $i$ , and call another  $i'$ . By the property of  $P$  we know there exists a set  $S_j$  where  $i \in S_j$  and  $i' \notin S_j$ . Clearly  $j > p$  (where  $p$  is the initial element) as it contains an element outside of  $R$  and  $S_p$  must contain  $i'$  (otherwise  $i'$  will not be in  $S$ ).

When the inner-loop goes to  $j$ , by Lemma C.2 the dimension of  $Z'$  will be smaller than  $Z$ . Also, by the robustness we know the set at that point contains an element (namely  $i$ ) that is not in  $R$ , so by Lemma C.3 we know  $\|P_{\Gamma^\perp} Z'\| > \alpha/2$ . As a result  $j$  must be added to  $U$  and this contradicts with the fact that in the end  $i'$  is still in  $S$ .  $\square$

Let  $S = \cap_{j \in U} S_j$  after the inner-loop, finally we show in the next iteration  $\Gamma$  will be  $\epsilon_Y$ -close to the span of vertices in  $R \cup S$ .

**Lemma C.5.** *Let  $S = \cap_{j \in U} S_j$  after the inner-loop, then in the next iteration,  $\Gamma$  is  $\epsilon_Y$  close to row span of  $W^{R \cup S}$ .*

*Proof.* Based on the hypothesis all the matrices appended to the matrix  $Y$  are  $\epsilon_v$ -close to the span of subset of rows of  $W$ , and the union of all the previous subsets equal to  $R$ . Let  $B$  be a matrix that corresponds to the matrix  $Y$  with the true spans, then  $\|B - Y\| \leq r\epsilon_v$ , and on the other hand the span of  $B$  is equal to row span of  $W^{R \cup S}$ , with smallest nonzero singular value at least  $\alpha$  (because  $W$  is  $\alpha$ -robust). Therefore by Wedin's theorem we know since  $r\epsilon_v \ll \alpha$   $\Gamma$  must be  $\epsilon_Y = 2r\epsilon_v/\alpha$ -close to the row span of  $W^{R \cup S}$ .  $\square$

The last two lemmas proved the second half of induction hypothesis. Finally it is easy to see that the algorithm will not stop as long as  $P \setminus R$  is not empty, and it must stop after  $r$  iterations because the size of  $R$  increases by at least 1 in every iteration. This concludes the proof of Theorem 6.1

**Finding the remaining vertices** The proof of Theorem 6.2 follows directly from Lemma 4.5 in (Arora et al., 2013), for completeness we explain the proof here. ( $\epsilon_v \leq \alpha/20r$ ,  $O(\epsilon/\alpha^2)$ )

*Proof.* (of Theorem 6.2) First observe that  $\alpha$ -robust implies  $\alpha$ -robust in (Arora et al., 2013), because for any vertex  $W^i$ , let  $v$  be the direction of  $W^i$  projected to the orthogonal subspace of  $W^{-i}$  (all the other rows). By  $\alpha$ -robust condition of this paper we know  $\|Wv\| \geq \alpha$ , which in particular implies  $\|P_{(W^{-i})^\perp} W^i\| \geq \alpha$ .

By Lemma 4.5 in (Arora et al., 2013), as long as the previously found vertices are at least  $\alpha/20r$ -close, and all the points are  $\epsilon$ -close, the new vertex found by the algorithm must be  $O(\epsilon/\alpha^2)$  close. Since  $\epsilon/\alpha^2 \ll \epsilon_v$  we can find all the remaining vertices.

Note that we are not running the clean-up phase of Algorithm 4 FastAnchorWords, this is because the vertices we find in this phase is already more accurate than the intersection vertices and the clean-up phase cannot improve the quality of the intersection vertices (as they don't appear in  $M^i$ ).  $\square$

## D. Generative model for subset-separable NMF

In this section we prove under natural generative model an NMF problem can have  $(N, H, \gamma)$ -properly filled facets with high probability.

In order to prove Theorem 7.1, We use the following two lemmas. The first lemma shows with enough uniform points in a simplex, with high probability one of them will be a center for the simplex.

**Lemma D.1** (Restating Lemma 7.2). *Given  $n = \Omega((4d)^d \log d/\eta)$  uniform points  $v^1, v^2, \dots, v^n$  in a standard  $d$ -dimensional simplex (with vertices  $e_1, e_2, \dots, e_d$ ), with probability  $1 - \eta$  there exists a point  $v_i$  such that  $v_i = \sum_{j \neq i} w_j v^j$  ( $w_j \geq 0, \sum_{j \neq i} w_j = 1$ ), and  $\sigma_{\min}(\sum_{j \neq i} w_j (v^j)(v^j)^T) \geq 1/16d$ .*

*Proof.* Consider  $d + 1$  subsets of the  $d$ -dimensional simplex: let  $S_0$  be the set of points that satisfy  $v_i \geq 1/2d$  for all  $i \in [d]$ ; let  $S_j$  ( $j \in [d]$ ) be the set of points that satisfy  $v_j \geq 1 - 1/4d$ . The volume of these sets are at least  $(4d)^{-d}$ . By simple Chernoff bound we know when there are  $n = \Omega((4d)^d \log d/\eta)$  samples, with probability at least  $1 - \eta$  there is a point in each of these sets.

Next we shall prove the point in  $S_0$  is in the convex hull of the points in  $S_j$ , and the convex hull satisfies the smallest singular value requirement. First we relabel the points, let  $v^0$  be any point in  $S_0$  and  $v^j$  be any point in  $S^j$  ( $j \neq 0$ ). Let  $V \in \mathbb{R}^{d \times d}$  be the matrix whose columns are  $v^j$ 's ( $j \in [d]$ ). We can apply Gershgorin's Disk Theorem to the matrix  $V^T V$  (this is a matrix with diagonal entries at least  $1 - 1/2d$  and off-diagonal entries at most  $1/2d$ ), and conclude that  $\sigma_{\min}(V^T V) \geq 1/4$ .

Since in particular  $V$  is full rank, let  $w = V^{-1}v^0$ . Let  $w_i$  be the smallest entry. If  $w_i < 0$  then since  $\sum_{j=1}^d w_j = 1$  (all the columns of  $V$  and  $v^0$  sum up to 1),  $\sum_{j=1}^d |w_j| \leq 1 - 2dw_i$ . The  $i$ -th coordinate  $(Vu)_i = \sum_{j=1}^d w_j v_i^j \leq (1 - 1/4d)w_i + (1 - 2dw_i)/4d \leq 1/4d$ , which cannot be equal to  $v_i^0$ , therefore  $w_i \geq 0$ . In this case since  $1/2d \leq v_i^0 = (Vw)_i = \sum_{j=1}^d w_j v_i^j \leq w_i + \frac{1}{4d}$ , we know  $w_i \geq 1/4d$ . Therefore  $\sigma_{\min}(\sum_{j \neq i} w_j (v^j)(v^j)^T) \geq \frac{1}{4d} \sigma_{\min}(VV^T) \geq \frac{1}{16d}$ .  $\square$

Next lemma shows only subspaces that contains a properly filled facet can have many points.

**Lemma D.2** (Restating Lemma 7.3). *Given  $n = \Omega(rd \log(d/p_{\min}\eta)/p_{\min})$  uniform points  $v^1, v^2, \dots, v^n$  in a standard  $d$ -dimensional simplex (with vertices  $e_1, e_2, \dots, e_d$ ), with probability  $1 - \eta$  for all matrices  $A \in \mathbb{R}^{r \times d}$  whose largest column norm is equal to 1, there are at most  $p_{\min}n/4$  points with  $\|Av^i\| \leq p_{\min}/200d$ .*

*Proof.* We first prove this for a particular matrix  $A$ , then we will construct an  $\epsilon$ -net and do union bound over all possible matrices  $A$ .

Let  $u = A_i$  where  $A_i$  is the column with norm 1. For random  $v$  that is uniform in the standard  $d$  dimensional simplex, we will show  $\Pr[|u^T Av| \leq \dots] \leq p_{\min}/8$ . By property of uniform distribution on a simplex, we know  $v_i$  is independent of  $v_{-i}/(1 - v_i)$  (where  $v_{-i}$  is the vector  $v$  with  $i$ -th coordinate removed), and  $v_i$  is distributed as a Beta distribution  $Beta(1, d - 1)$ . Let  $q = u^T Av_{-i}/(1 - v_i)$ , then we know  $u^T Av = v_i + (1 - v_i)q$  and  $q \in [-1, 1]$ . The density function of  $v_i$  is bounded by  $d - 1$ , therefore for any value  $q$ , the probability that  $|u^T Av| \leq p_{\min}/100d$  is at most  $p_{\min}/8$ . When the number of samples is at least  $n = \Omega(\log(1/\eta')/p_{\min})$ , with probability  $1 - \eta'$  there are at most  $p_{\min}n/4$  points that satisfy  $|u^T Av| \leq p_{\min}/100d$ .

Now we construct an  $\epsilon$ -net so that for any matrix  $A$  with largest column norm 1, there is a matrix  $A'$  in the  $\epsilon$ -net that is column-wise  $\epsilon$ -close to  $A$ . Set  $\epsilon = p_{\min}/200d^2$ , by standard construction the number of matrices in the  $\epsilon$ -net is  $O(d^2/p_{\min})^{rd}$ . Let  $\eta' = \eta/O(d^2/p_{\min})^{rd}$  (and hence  $n = \Omega(rd \log(d/p_{\min}\eta)/p_{\min})$ ), by union bound we know with probability  $1 - \eta$ , there are at most  $p_{\min}n/4$  points with  $\|Av^i\| \leq p_{\min}/100d$  for all matrices  $A$  in the  $\epsilon$ -net. For a matrix  $A$  that is not in the  $\epsilon$ -net, let  $A'$  be the matrix in the net that is column-wise  $\epsilon$ -close, clearly  $\|Av^i\| - \|A'v^i\| \leq p_{\min}/200d$ . If there are more than  $p_{\min}n/4$  points with  $\|Av^i\| \leq p_{\min}/200d$  then all these points will have  $\|A'v^i\| \leq p_{\min}/100d$  and that is impossible.  $\square$

With these two lemmas we can now prove the theorem:

*Proof.* (of Theorem 7.1) In order to satisfy Condition 1, we apply Lemma 7.2. For any proper facet the points are equal to the rows of  $W^{S_i}$  multiplied by uniform random points, since  $\sigma_{\min}(W^{S_i}) \geq \sigma_{\min}(W) \geq \alpha$ , we know if the facet has more than  $\Omega((4d)^d \log d/\eta)$  points the convex combination has smallest singular value  $\alpha^2/16d$ . This is ensured when the number of samples is at least  $n = \Omega((4d)^d \log(kd/\eta)/p_{\min})$  by Chernoff bound.

Condition 2 is satisfied whenever  $n = \Omega(\log(k/\eta)/p_{\min})$  by simple Chernoff bound.

Condition 3 follows from Lemma 7.3. Suppose  $Q$  is a subspace that for any proper facet  $Q_i$  we have  $\|P_{Q^\perp} Q_i\| > H\epsilon$ . Since  $W$  is  $\alpha$ -robust this means  $\|P_{Q^\perp} W^{S_i}\| \geq H\alpha\epsilon$ , therefore there is always a column that has norm  $H\alpha\epsilon/\sqrt{r}$ . By

Lemma 7.3 we know no matter which subspace the point is chosen from, with probability at most  $p_{min}/8$  it will be  $\epsilon/2$ -close to the subspace. Now we can apply union bound to the product of all the  $\epsilon$ -nets constructed for different proper facets, so the size of the net is  $\exp(kdr \log d)$ . Therefore we know when  $n = \Omega(kr^2 \log(r/p_{min}\eta)/p_{min})$  (here for simplicity we used  $d = r$  because in particular the interior points are in a space of dimension  $r$ ) with high probability there will be at most  $p_{min}n/4$  points for this subspace  $Q$ .  $\square$