IMPROVED ALGORITHM AND BOUNDS FOR SUCCESSIVE PROJECTION

Jiashun Jin & Gabriel Moryoussef
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{jiashun, gmoryous}@andrew.cmu.edu

Zheng Tracy Ke & Jiajun Tang & Jingming Wang
Department of Statistics
Harvard University
Cambridge, MA 02138, USA
{zke,jiajuntang,jingmingwang}@fas.harvard.edu

ABSTRACT

Given a K-vertex simplex in a d-dimensional space, suppose we measure n points on the simplex with noise (hence, some of the observed points fall outside the simplex). Vertex hunting is the problem of estimating the K vertices of the simplex. A popular vertex hunting algorithm is successive projection algorithm (SPA). However, SPA is observed to perform unsatisfactorily under strong noise or outliers. We propose pseudo-point SPA (pp-SPA). It uses a projection step and a denoise step to generate pseudo-points and feed them into SPA for vertex hunting. We derive error bounds for pp-SPA, leveraging on extreme value theory of (possibly) high-dimensional random vectors. The results suggest that pp-SPA has faster rates and better numerical performances than SPA. Our analysis includes an improved non-asymptotic bound for the original SPA, which is of independent interest.

CONTENTS

1 INTRODUCTION

Fix $d \geq 1$ and suppose we observe n vectors $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$, where

$$X_i = r_i + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2 I_d). \tag{1}$$

The Gaussian assumption is for technical simplicity and can be relaxed. For an integer $1 \leq K \leq d+1$, we assume that there is a simplex with K vertices $\mathcal{S}_0$ on the hyperplane $\mathcal{H}_0$ such that each r_i falls within the simplex (note that a simplex with K vertices always falls on a $(K-1)$-dimensional hyperplane of $\mathbb{R}^d$). In other words, let $v_1, v_2, \ldots, v_K \in \mathbb{R}^d$ be the vertices of the simplex and let $V = [v_1, v_2, \ldots, v_K]$. We assume that for each $1 \leq i \leq n$, there is a K-dimensional weight vector π_i (a weight vector is vector where all entries are non-negative with a unit sum) such that

$$r_i = \sum_{k=1}^{K} \pi_i(k) v_k = V\pi_i. \tag{2}$$

Here, π_i's are unknown but are of major interest, and to estimate π_i, the key is vertex hunting (i.e., estimating the K vertices of the simplex $\mathcal{S}_0$). In fact, once the vertices are estimated, we can estimate $\pi_1, \pi_2, \ldots, \pi_n$ by the relationship of $X_i \approx r_i = V\pi_i$. Motivated by these, the primary interest of this paper is vertex hunting (VH). The problem may arise in many application areas. *(1) Hyperspectral unmixing*: Hyperspectral unmixing (Bioucas-Dias et al., 2012) is the problem of separating the pixel spectra from a hyperspectral image into a collection of constituent spectra. X_i contains the spectral measurements of pixel i at d different channels, $v_1, \ldots, v_K$ are the constituent spectra (called *endmembers*), and π_i contains the fractional *abundances* of endmembers at pixel i. It is of great interest to identify the endmembers and estimate the abundances. *(2) Archetypal analysis*. Archytypal analysis (Cutler & Breiman, 1994) is a useful tool for representation learning. Take its

application in genetics for example (Satija et al., 2015). Each X_i is the gene expression of cell i, and each v_k is an archetypal expression pattern. Identifying these archetypal expression patterns is useful for inferring a transcriptome-wide map of spatial patterning. *(3) Network membership estimation*. Let $A \in \mathbb{R}^{n,n}$ be the adjacency matrix of an undirected network with n nodes and K communities. Let $(\hat{\lambda}_k, \hat{\xi}_k)$ be the k-th eigenpair of A, and write $\widehat{\Xi} = [\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K]$. Under certain network models (e.g., Huang et al. (2023); Airoldi et al. (2008); Zhang et al. (2020); Ke & Jin (2023); Rubin-Delanchy et al. (2022)), there is a K-vertex simplex in $\mathbb{R}^K$ such that for each $1 \leq i \leq n$, the i-th row of $\widehat{\Xi}$ falls (up to noise corruption) inside the simplex, and vertex hunting is an important step in community analysis. *(4) Topic modeling*. Let $D \in \mathbb{R}^{n,p}$ be the frequency of word counts of n text documents, where p is the dictionary size. If D follows the Hoffman's model with K topics, then there is also simplex in the spectral domain (Ke & Wang, 2022)), so vertex hunting is useful.

Existing vertex hunting approaches can be roughly divided into two lines: constrained optimizations and stepwise algorithms. In the first line, one proposes an objective function and estimates the vertices by solving an optimization problem. The minimum volume transform (MVT) (Craig, 1994), archetypal analysis (AA) (Cutler & Breiman, 1994; Javadi & Montanari, 2020), and N-FINDER (Winter, 1999) are approaches of this line. In the second line, one uses a stepwise algorithm which iteratively identifies one vertex of the simplex at a time. This includes the popular successive projection algorithm (SPA) (Araújo et al., 2001). SPA is a stepwise greedy algorithm. It does not require an objective function (how to select the objective function may be a bit subjective), is computationally efficient, and has a theoretical guarantee. This makes SPA especially interesting.

Our contributions. Our primary interest is to improve SPA. Despite many good properties aforementioned, SPA is a greedy algorithm, which is vulnerable to noise and outliers, and may be significantly inaccurate. Below, we list two reasons why SPA may underperform. First, typically in the literature (e.g., Araújo et al. (2001)), one apply the SPA directly to the d-dimensional data points $X_1, X_2, \ldots, X_n$, regardless of what (K, d) are. However, since the true vertices $v_1, \ldots, v_K$ lie on a $(K-1)$-dimensional hyperplane, if we directly apply SPA to $X_1, X_2, \ldots, X_n$, the resultant hyperplane formed by the estimated simplex vertices is likely to deviate from the true hyperplane, due to noise corruption. This will cause inefficiency of SPA. Second, since the SPA is a greedy algorithm, it tends to be biased outward bound. When we apply SPA, it is frequently found that most of the estimated vertices fall outside of true simplex (and some of them are faraway from the true simplex).

For illustration, Figure 1 presents an example, where $X_1, X_2, \ldots, X_n$ are generated from Model (1) with $(n, K, d, \sigma) = (1000, 3, 2, 1)$, and r_i are uniform samples over T (T is the triangle with vertices $(1, 1)$, $(2, 4)$, and $(5, 2)$). In this example, the true vertices (large black points) form a triangle (dashed black lines) on a 2-dimensional hyperplane. The green and cyan-colored triangles are estimated by SPA and pp-SPA (our main algorithm to be introduced; since d is equal to $K-1$, the hyperplane projection is skipped), respectively. In this example, the estimated simplex by SPA is significantly biased outward bound, suggesting a large room for improvement. Such outward bound bias of SPA is related to the design of the algorithm and is frequently observed (Gillis, 2019).

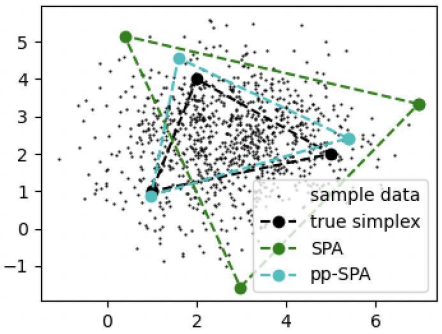

Figure 1: A numerical example (d=2, K=3).

To fix the issues, we propose pseudo-point SPA (pp-SPA) as a new approach to vertex hunting. It contains two novel ideas as follows. First, since the simplex $\mathcal{S}_0$ is on the hyperplane $\mathcal{H}_0$, we first use all data $X_1, \ldots, X_n$ to estimate the hyperplane, and then project all these points to the hyperplane. Second, since SPA is vulnerable to noise and outliers, a reasonable idea is to add a denoise step before we apply SPA. We propose a *pseudo-point (pp) approach* for denoising, where for each data point, we replace it by a pseudo point, computed as the average of *all of its neighbors within a radius of* Δ. Utilizing information in the nearest neighborhood is a known idea in classification (Hastie et al., 2009), and the well-known k-nearest neighborhood (KNN) algorithm is such an approach. However, KNN or similar ideas were never used as a denoise step for vertex hunting. Compared with KNN, the idea of pseudo-point approach is motivated by the underlying geometry and is for a different purpose. For these reasons, the idea is new at least to some extent.

We have two theoretical contributions. First, Gillis & Vavasis (2013) derived a non-asymptotic error bound for SPA, but the bound is not always tight. Using a very different proof, we derive a sharper non-asymptotic bound for SPA. The improvement is substantial in the following case. Recall that $V = [v_1, v_2, \ldots, v_K]$ and let $s_k(V)$ be the k-th largest singular value of V. The bound in Gillis & Vavasis (2013) is proportional to $1/s_K^2(V)$, while our our bound is proportional to $1/s_{K-1}^2(V)$. Since all vertices lie on a $(K-1)$-dimensional hyperplane, $s_{K-1}(V)$ is bounded away from 0, as long as the volume of true simplex is lower bounded. However, $s_K(V)$ may be 0 or nearly 0; in this case, the bound in Gillis & Vavasis (2013) is too conservative, but our bound is still valid. Second, we use our new non-asymptotic bound to derive the rate for pp-SPA, and show that the rate is much faster than the rate of SPA, especially when $d \gg K$. Even when $d = O(K)$, the bound we get for pp-SPA is still sharper than the bound of the original SPA. The main reason is that, for those points far away outside the true simplex, the corresponding pseudo-points we generate are much closer to the true simplex. This greatly reduces the *outward bound biases* of SPA (see Figure 1).

Related literature. It was observed that SPA is susceptible to outliers, motivating several variants of SPA (Gillis & Vavasis, 2015; Mizutani & Tanaka, 2018; Gillis, 2019). For example, Bhattacharyya & Kannan (2020); Bakshi et al. (2021); Nadisic et al. (2023) modified SPA by incorporating smoothing at each iteration. In contrast, our approach involves generating all pseudo points through neighborhood averaging before executing all successive projection steps. Additionally, we exploit the fact that the simplex resides in a low-dimensional hyperplane and apply a hyperplane projection step prior to the denoising and successive projection steps. Our theoretical results surpass those existing works for several reasons: (a) we propose a new variant of SPA; (b) our analyses build upon a better version of the non-asymptotic bound than the commonly-used one in Gillis & Vavasis (2013); and (c) we incorporate delicate random matrix and extreme value theory in our analysis.

2 A NEW VERTEX HUNTING ALGORITHM

The successive projection algorithm (SPA) (Araújo et al., 2001) is a popular vertex hunting method. This is an iterative algorithm that estimates one vertex at a time. At each iteration, it first projects all points to the orthogonal complement of those previously found vertices and then takes the point with the largest Euclidean norm as the next estimated vertex. See Algorithm 1 for a detailed description.

Algorithm 1 The (orthodox) Successive Projection Algorithm (SPA)

Input: $X_1, X_2, \ldots, X_n$, and K.
Initialize $u = \mathbf{0}_p$ and $y_i = X_i$, for $1 \leq i \leq n$. For $k = 1, 2, \ldots, K$,

- Update y_i to $(I_d - uu')y_i$. Obtain $i_k = \arg\max_{1 \leq i \leq n} \|y_i\|$. Update $u = \|y_{i_k}\|^{-1} y_{i_k}$.

Output: $\hat{v}_k = X_{i_k}$, for $1 \leq k \leq K$.

We propose pp-SPA as an improved version of the (orthodox) SPA, containing two main ideas: a *hyperplane projection* step and a *pseudo-point denoise* step. We now discuss the two steps separately.

Consider the *hyperplane projection* step first. In our model (2), the noiseless points $r_1, \ldots, r_n$ live in a $(K-1)$-dimensional hyperplane. However, with noise corruption, the observed data $X_1, \ldots, X_n$ are not exactly contained in a hyperplane. Our proposal is to first use data to find a 'best-fit' hyperplane and then project all data points to this hyperplane. Fix $d \geq K \geq 2$. Given a point $x_0 \in \mathbb{R}^d$ and a projection matrix $H \in \mathbb{R}^{d \times d}$ with rank $K-1$, the $(K-1)$-dimensional hyperplane associated with (x_0, H) is $\mathcal{H} = \{x \in \mathbb{R}^d : (I_d - H)(x - x_0) = 0\}$. For any $x \in \mathbb{R}^d$, the Euclidean distance between x and the hyperplane is equal to $\|(I_d - H)(x - x_0)\|$. Given $X_1, X_2, \ldots, X_n$, we aim to find a hyperplane to minimize the sum of square distances:

$$\min_{(x_0, H)} \{S(x_0, H)\}, \quad \text{where} \quad S(x_0, H) = \sum_{i=1}^{n} \|(I_d - H)(X_i - x_0)\|^2. \tag{3}$$

Let $Z = [Z_1, \ldots, Z_n]$, where $Z_i = X_i - \bar{X}$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. For each k, let $u_k \in \mathbb{R}^d$ be the kth left singular vector of Z. Write $U = [u_1, \ldots, u_{K-1}]$. The next lemma is proved in the appendix.

Lemma 1. *$S(x_0, H)$ is minimized by $x_0 = \bar{X}$ and $H = UU'$.*

For each $1 \leq i \leq n$, we first project each X_i to $\tilde{X}_i$ and then transform $\tilde{X}_i$ to Y_i, where

$$\tilde{X}_i := \bar{X} + H(X_i - \bar{X}), \qquad Y_i := U'\tilde{X}_i; \qquad \text{note that } H = UU' \text{ and } Y_i \in \mathbb{R}^{K-1}. \tag{4}$$

These steps reduce noise. To see this, we note that the true simplex lives in a hyperplane with a projection matrix $H_0 = U_0U_0'$. It can be shown that $U \approx U_0$ (up to a rotation) and $Y_i \approx r_i^* + U_0'\epsilon_i$, with $r_i^* = U_0'\bar{X} + U_0'r_i$. These points r_i^* still live in a simplex (in dimension $(K-1)$). Comparing this with the original model $X_i = r_i + \epsilon_i$, we see that $U_0'\epsilon_i$ are iid samples from $N(0, \sigma^2 I_{K-1})$, and ϵ_i are iid samples from $N(0, \sigma^2 I_d)$. Since $K - 1 \ll d$ in may applications, the projection may significantly reduce the dimension of the noise variable. Later in Section 4, we see that this implies a significant improvement in the convergence rate.

Next, consider the *neighborhood denoise* step. Fix an $\Delta > 0$ and an integer $N \geq 1$. Define the Δ-neighborhood of Y_i by $B_\Delta(Y_i) = \{x \in \mathbb{R}^{K-1} : \|x - Y_i\| \leq \Delta\}$. When there fewer than N points in $B_\Delta(Y_i)$ (including Y_i itself), remove Y_i for the vertex hunting step next. Otherwise, replace Y_i by the average of all points in $B_\Delta(Y_i)$ (denoted by Y_i^*). The main effect of the denoise effect is on the points that are *far outside the simplex*. For these points, we either delete them for the vertex hunting step (see below), or replace it by a point closer to the simplex. This way, we pull all these points "towards" the simplex, and thus reduce the estimation error in the subsequent vertex hunting step.

Finally, we apply the (orthodox) successive projection algorithm (SPA) to $Y_1^*, Y_2^*, \cdots, Y_n^*$ and let $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$ be the estimated vertices. Let $\hat{V} = [\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K]$. See Algorithm 2.

Algorithm 2 Pseudo-Point Successive Projection Algorithm (pp-SPA)

Input: $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$, the number of vertices K, and tuning parameters (N, Δ).

Step 1 *(Projection)*. Obtain $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $Z = X - \bar{X}\mathbf{1}_n'$. Let $U = [u_1, \ldots, u_{K-1}]$ contain the first $(K-1)$ singular vectors of Z. For $1 \leq i \leq n$, let $Y_i = U'X_i \in \mathbb{R}^{K-1}$.

Step 2 *(Denoise)*. Let $B_\Delta(Y_i) = \{x \in \mathbb{R}^{K-1} : \|x - Y_i\| \leq \Delta\}$ denote the Δ-neighborhood of Y_i.

- If there are fewer than N points (including Y_i itself) in $B_\Delta(Y_i)$, delete this point.
- Otherwise, replace Y_i by Y_i^*, which is the average of all points in $B_\Delta(Y_i)$.

Step 3 *(VH)*. Let $\mathcal{J} \subset \{1, \ldots, n\}$ be the set of retained points in Step 2. Apply Algorithm 1 to $\{Y_i^*\}_{i \in \mathcal{J}}$ to get $\hat{v}_1^*, \hat{v}_2^*, \ldots, \hat{v}_K^* \in \mathbb{R}^{K-1}$. Let $\hat{v}_k = (I_d - H)\bar{X} + U\hat{v}_k^* \in \mathbb{R}^d$, $1 \leq k \leq K$.

Output: The estimated vertices $\hat{v}_1, \ldots, \hat{v}_K$.

Remark 1: The complexity of the orthodox SPA is $O(ndK)$. Regarding the complexity of pp-SPA, it applies SPA on $(K-1)$-dimensional pseudo-points, so the complexity is $O(nK^2)$. To obtain these pseudo points, we need a projection step and a denoise step. The projection step extracts the first $(K-1)$ singular vectors of a matrix $Z(n \times d)$. Performing the whole SVD decomposition would result in $O(\min(n^2 d, nd^2))$ time complexity. However, faster approach exists such as the truncated SVD which would decrease this complexity to $O(ndK)$. In the denoise step, we need to find the Δ-neighborhoods for all n points $Y_1, Y_2, \ldots, Y_n$. This can be made computationally efficient using the KD-Tree. The construction of KD-Tree takes $O(n \log n)$, and the search of neighbors typically takes $O\big(n^{(2-\frac{1}{K-1})} + nm\big)$, where m is the maximum number of points in a neighborhood.

Remark 2: Algorithm 2 has tuning parameters (N, Δ), where Δ is the radius of the neighborhood, and N is used to prune out points far away from the simplex. For N, we typically take $N = \log(n)$ in theory and $N = 3$ in practice. Concerning Δ, we use a heuristic choice $\Delta = \max_i \|Y_i - \bar{Y}\|/5$, where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. It works satisfactorily in simulations.

Remark 3 *(P-SPA and D-SPA)*: We can view pp-SPA as a generic algorithm, where we may either replace the projection step by a different dimension reduction step, or replace the denoise step by a different denoise idea, or both. In particular, it is interesting to consider two special cases: (i) ***P-SPA***, which skips the denoise step and only uses the projection and VH steps; (ii) ***D-SPA***, which skips the projection step and only uses the denoise and VH steps. We analyze these algorithms, together with pp-SPA (see Table 1 and Section C of the appendix). In this way, we can better understand the respective improvements of the projection step and the denoise step.

3 AN IMPROVED BOUND FOR SPA

Recall that $V = [v_1, v_2, \ldots, v_K]$, whose columns are the K vertices of the true simplex $\mathcal{S}_0$. Let

$$\gamma(V) = \max_{1\le k\le K}\{\|v_k\|\}, \qquad g(V) = 1 + 80\frac{\gamma^2(V)}{s_K^2(V)}, \qquad \beta(X) = \max_{1\le i\le n}\{\|\epsilon_i\|\}. \tag{5}$$

Lemma 2 (Gillis & Vavasis (2013), orthodox SPA)**.** *Consider d-dimensional vectors $X_1, \ldots, X_n$, where $X_i = r_i + \epsilon_i$, $1 \le i \le n$ and r_i satisfy model (2). For each $1 \le k \le K$ there is an i such that $\pi_i = e_k$. Suppose $\max_{1\le i\le n}\|\epsilon_i\| \le \frac{s_K(V)}{1+80\gamma^2(V)/s_K^2(V)}\min\{\frac{1}{2\sqrt{K-1}}, \frac{1}{4}\}$. Apply the orthodox SPA to $X_1, \ldots, X_n$ and let $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$ be the output. Up to a permutation of these K vectors,*

$$\max_{1\le k\le K}\{\|\hat{v}_k - v_k\|\} \le \Big[1 + 80\frac{\gamma^2(V)}{s_K^2(V)}\Big]\max_{1\le i\le n}\|\epsilon_i\| := g(V)\cdot\beta.$$

Lemma 2 is among the best known results for SPA, but this bound is still not satisfying. One issue is that $s_K(V)$ depends on the location (i.e., center) of $\mathcal{S}_0$, but how well we can do vertex hunting should not depend on its location. We expect that vertex hunting is difficult only if $\mathcal{S}_0$ has a small volume (so the simplex is nearly flat). To see how these insights connect to singular values of V, let $\bar{v} = K^{-1}\sum_{k=1}^K v_k$ be the center of $\mathcal{S}_0$, define $\tilde{V} = [v_1 - \bar{v}, \ldots, v_K - \bar{v}]$, and let $s_k(\tilde{V})$ be the k-th singular value of $\tilde{V}$. The next lemma is proved in the appendix:

Lemma 3. $\text{Volume}(\mathcal{S}_0) = \frac{\sqrt{K}}{(K-1)!}\prod_{k=1}^{K-1} s_k(\tilde{V})$, *$s_{K-1}(V) \ge s_{K-1}(\tilde{V})$, and $s_K(V) \le \sqrt{K}\|\bar{v}\|$.*

Lemma 3 yields several observations. First, as we shift the location of $\mathcal{S}_0$ so that its center gets close to the origin, $\|\bar{v}\| \approx 0$, and $s_K(V) \approx 0$. In this case, the bound in Lemma 2 becomes almost useless. Second, the volume of $\mathcal{S}_0$ is determined by the first $(K-1)$ singular values of $\tilde{V}$, irrelevant to the Kth singular value. Finally, if the volume of $\mathcal{S}_0$ is lower bounded, then we immediately get a lower bound for $s_{K-1}(V)$. These observations motivate us to modify $g(V)$ in (5) to a new quantity that depends on $s_{K-1}(V)$ instead of $s_K(V)$; see (6) below.

Another issue of the bound in Lemma 2 is that $\beta(X)$ depends on the maximum of $\|\epsilon_i\|$, which is too conservative. Consider a toy example in Figure 2, where $\mathcal{S}_0$ is the dashed triangle, the red stars represent r_i's and the black points are X_i's. We observe that X_2 and X_5 are deeply in the interior of $\mathcal{S}_0$, and they should not affect the performance of SPA. We hope to modify $\beta(X)$ to a new quantity that does not depend on $\|\epsilon_2\|$ and $\|\epsilon_5\|$. One idea is to modify $\beta(X)$ to $\beta^*(X,V) = \max_i \text{Dist}(X_i, \mathcal{S}_0)$, where $\text{Dist}(\cdot, \mathcal{S}_0)$ is the Euclidean distance from a point to the simplex. For any point inside the simplex, this Euclidean distance is exactly zero. Hence, for this toy example, $\beta^*(X,V) \le \max_{i\notin\{1,2,5\}}\|\epsilon_i\|$. However, we cannot simply replace $\beta(X)$ by $\beta^*(X,V)$, because $\|\epsilon_1\|$ also affects the performance of SPA and should not be left out. Note that r_1 is the only point located at the top vertex. When X_1 is far away from r_1, no matter whether X_1 is inside or outside $\mathcal{S}_0$, SPA still makes a large error in estimating this vertex. This inspires us to define $\beta^\dagger(X,V) = \max_k \min_{\{i:r_i=v_k\}}\|\epsilon_i\|$. When $\beta^\dagger(X,V)$ is small, it means for each v_k, there exists at least one X_i that is close enough to v_k. To this end, let $\beta_{\text{new}}(X,V) = \max\{\beta^*(X,V), \beta^\dagger(X,V)\}$. Under this definition, $\beta_{\text{new}}(X) \le \max_{i\notin\{2,5\}}\|\epsilon_i\|$, which is exactly as hoped.

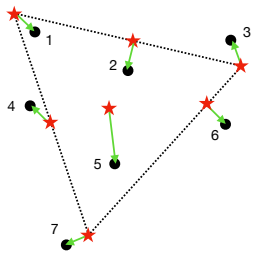

Figure 2: A toy example to show the difference between $\beta(X)$ and $\beta_{\text{new}}(X,V)$, where $\beta(X) = \max_i\|\epsilon_i\|$, and $\beta_{\text{new}}(X,V) \le \max_{i\notin\{2,5\}}\|\epsilon_i\|$.

Inspired by the above discussions, we introduce (for a point $x \in \mathbb{R}^d$, $\text{Dist}(x, \mathcal{S}_0)$ is the Euclidean distance from x to $\mathcal{S}_0$; this distance is zero if $x \in \mathcal{S}_0$)

$$\begin{aligned} g_{\text{new}}(V) &= 1 + \frac{30\gamma(V)}{s_{K-1}(V)}\max\Big\{1, \frac{\gamma(V)}{s_{K-1}(V)}\Big\}, \\ \beta_{\text{new}}(X) &= \max\Big\{\max_{1\le i\le n}\text{Dist}(X_i, \mathcal{S}_0), \max_{1\le k\le K}\min_{\{i:r_i=v_k\}}\|X_i - v_k\|\Big\}. \end{aligned} \tag{6}$$

Theorem 1. *Consider d-dimensional vectors $X_1, \ldots, X_n$, where $X_i = r_i + \epsilon_i$, $1 \le i \le n$ and r_i satisfy model (2). For each $1 \le k \le K$ there is an i such that $\pi_i = e_k$. Suppose for a properly small universal constant $c^* > 0$, $\max\{1, \frac{\gamma(V)}{\sigma_{K-1}(V)}\}\beta_{\rm new}(X, V) \le c^* \frac{s^2_{K-1}(V)}{\gamma(V)}$. Apply the orthodox SPA to $X_1, \ldots, X_n$ and let $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$ be the output. Up to a permutation of these K vectors,*

$$\max_{1\le k\le K}\{\|\hat{v}_k - v_k\|\} \le g_{\rm new}(V)\beta_{\rm new}(X, V).$$

Note that $g_{\rm new}(V) \le g(V)$ and $\beta_{\rm new}(X, V) \le \beta(X)$. The non-asymptotic bound in Theorem 1 is always better than the bound in Lemma 2. We use an example to illustrate that the improvement can be substantial. Let $K = d = 3$, $v_1 = (20, 20, 10)$, $v_2 = (20, 30, 10)$, and $v_3 = (30, 22, 10)$. We put r_1, r_2, r_3 at each of the three vertices, r_4, r_5, r_6 at the mid-point of each edge, and r_7 at the center of the simplex (which is $\bar{v}$). We sample $\epsilon_1^*, \epsilon_2^*, \ldots, \epsilon_7^*$ i.i.d., from the unit sphere in $\mathbb{R}^3$. Let $\epsilon_i = 0.01\epsilon_i^*$, for $1 \le i \le 6$, and $\epsilon_7 = 0.05\epsilon_i^*$. By straightforward calculations, $g(V) = 4.3025\times 10^4$, $g_{\rm new}(V) = 6.577\times 10^2$, $\beta(X) = 0.05$, $\beta_{new}(X, V) = 0.03$. Therefore, the bound in Lemma 2 gives $\max_k \|\hat{v}_k - v_k\| \le 2151.3$, while the improved bound in Theorem 1 gives $\max_k \|\hat{v}_k - v_k\| \le 18.7$. A more complicated version of this example can be found in Section D of the supplementary material.

The main reason we can achieve such a significant improvement is that our proof idea is completely different from the one in Gillis & Vavasis (2013). The proof in Gillis & Vavasis (2013) is driven by *matrix norm inequalities* and does not use any geometry. This is why they need to rely on quantities such as $s_K(V)$ and $\max_i \|\epsilon_i\|$ to control the norms of various matrices in their analysis. It is very difficult to modify their proof to obtain Theorem 1, as the quantities in (6) are insufficient to provide strong matrix norm inequalities. In contrast, our proof is guided by *geometric insights*. We construct a *simplicial neighborhood* near each true vertex and show that the estimate $\hat{v}_k$ in each step of SPA must fall into one of these simplicial neighborhoods.

4 THE BOUND FOR PP-SPA AND ITS IMPROVEMENT OVER SPA

We focus on the orthodox SPA in Section 3. In this section, we show that we can further improve the bound significantly if we use pp-SPA for vertex hunting. Recall that we have also introduced P-SPA and D-SPA in Section 2 as simplified versions of pp-SPA. We establish error bounds for P-SPA, D-SPA, and pp-SPA, under the Gaussian noise assumption in (1). A high-level summary is in Table 1. Recall that P-SPA, D-SPA, and pp-SPA all create pseudo-points and then feed them into SPA. Different ways of creating pseudo-points only affect the term $\beta_{\rm new}(X, V)$ in the bound in Theorem 1. Assuming that $g_{\rm new}(V) \ge C$, the order of $\beta_{\rm new}(X, V)$ fully captures the error bound. Table 1 lists the sharp orders of $\beta_{\rm new}(X, V)$ (including the constant).

Table 1: The sharp orders of $\beta_{\rm new}(X, V)$ (settings: $K \ge 3$, d satisfies (7), $s_{K-1}(V) > C$, and m satisfies the condition in Theorem 3). P-SPA and D-SPA use *the projection only* and *the denoise only*, respectively. The constant $c_0 \in (0, 1)$ comes from m, and the constant $a_1 > 2$ is as in Lemma 5.

	$d \ll \log(n)$	$d = a_0\log(n)$	$\log(n) \ll d \ll n^{1-\frac{2(1-c_0)}{K-1}}$	$d \gg n^{1-\frac{2(1-c_0)}{K-1}}$
SPA	$\sqrt{2\log(n)}$	$\sqrt{a_1\log(n)}$	$\sqrt{d}$	$\sqrt{d}$
P-SPA	$\sqrt{2\log(n)}$	$\sqrt{2\log(n)}$	$\sqrt{2\log(n)}$	$\sqrt{2\log(n)}$
D-SPA	$\sqrt{2c_0\log(n)}$	NA	NA	NA
pp-SPA	$\sqrt{2c_0\log(n)}$	$\sqrt{2c_0\log(n)}$	$\sqrt{2c_0\log(n)}$	$\sqrt{2\log(n)}$

The results suggest that pp-SPA always has a strictly better error bound than SPA. When $d \gg \log(n)$, the improvement is a factor of $o(1)$; the larger d, the more improvement. When $d = O(\log(n))$, the improvement is a constant factor that is strictly smaller than 1. In addition, by comparing P-SPA and D-SPA with SPA, we have some interesting observations:

- *The projection effect*. From the first two rows of Table 1, the error bound of P-SPA is never worse than that of SPA. In many cases, P-SPA leads to a significant improvement. When $d \gg \log(n)$, the rate is faster by a factor of $\sqrt{\log(n)/d}$ (which is a huge improvement for high-dimensional data). When $d \asymp \log(n)$, there is still a constant factor of improvement.

- *The denoise effect*. We compare the error bounds for P-SPA and pp-SPA, where the difference is caused by the denoise step. In three out of the four cases of d in Table 1, pp-SPA strictly improves P-SPA by a constant factor $c_0 < 1$.

  We note that pp-SPA applies denoise to the projected data in $\mathbb{R}^{K-1}$. We may also apply denoise to the original data in $\mathbb{R}^d$, which gives D-SPA. By Table 1, when $d \ll \sqrt{\log(n)}$, D-SPA improves SPA by a constant factor. However, for $d \gg \log(n)$, we always recommend applying denoise to the projected data. In such cases, the leading term in the extreme value of chi-square (see Lemma 5) is d, so the denoise is not effective if applied to original data.

Table 1 and the above discussions are for general settings. In a slightly more restrictive setting (see Theorem 2 below), both projection and denoise can improve the error bounds by a factor of $o(1)$.

We now present the rigorous statements. Owing to space constraint, we only state the error bounds of pp-SPA in the main text. The error bounds of P-SPA and D-SPA can be found in the appendix.

4.1 Some useful preliminary results

Recall that $V = [v_1, \ldots, v_K]$ and $r_i = V\pi_i$, $1 \le i \le n$. Let $\bar{v}$, $\bar{r}$, and $\bar{\pi}$ be the empirical means of v_k's, r_i's, and π_i's, respectively. Introduce $\tilde{V} = [v_1 - \bar{v}, \ldots, v_K - \bar{v}]$, $R = n^{-1/2}[r_1 - \bar{r}, \ldots, r_n - \bar{r}]$, and $G = (1/n)\sum_{i=1}^n (\pi_i - \bar{\pi})(\pi_i - \bar{\pi})'$. Lemma 4 relates singular values of R to those of G and V and is proved in the appendix ($A \preceq B$: $B - A$ is positive semi-definite. Also, $\lambda_k(G)$ is the k-th largest (absolute value) eigenvalue of G, $s_k(V)$ is the k-th largest singular value of V; same below).

Lemma 4. *The following statements are true: (a) $RR' = VGV'$, (b) $\lambda_{K-1}(G) \cdot \tilde{V}\tilde{V}' \preceq VGV' \preceq \lambda_1(G) \cdot \tilde{V}\tilde{V}'$, and (c) $\lambda_{K-1}(G) \cdot s^2_{K-1}(\tilde{V}) \preceq \sigma^2_{K-1}(R) \preceq \lambda_1(G) \cdot s^2_{K-1}(\tilde{V})$.*

To analyze SPA and pp-SPA, we need precise results on the extreme values of chi-square variables. Lemma 5 is proved in the appendix.

Lemma 5. *Let M_n be the maximum of n iid samples from $\chi^2_d(0)$. As $n \to \infty$, (a) if $d \ll \log(n)$, then $M_n/(2\log(n)) \to 1$, (b) if $d \gg \log(n)$, then $M_n/d \to 1$, and (c) if $d = a_0 \log(n)$ for a constant $a_0 > 0$, then $M_n/(a_1 \log(n)) \to 1$ where $a_1 > 2$ is unique solution of the equation $a_1 - a_0 \log(a_1) = 2 + a_0 - a_0 \log(a_0)$ (convergence in three cases are convergence in probability).*

4.2 Regularity conditions and main theorems

We assume

$$K = o(\log(n)/\log\log(n)), \qquad d = o(\sqrt{n}). \tag{7}$$

These are mild conditions. In fact, in practice, the dimension of the true simplex is usually relatively low, so the first condition is mild. Also, when the (low-dimensional) true simplex is embedded in a high dimensional space, it is not preferable to directly apply vertex hunting. Instead, one would use tools such as PCA to significantly reduce the dimension first and then perform vertex hunting. For this reason, the second condition is also mild. Moreover, recall that $G = n^{-1}\sum_{i=1}^n (\pi_i - \bar{\pi})(\pi_i - \bar{\pi})'$ is the empirical covariance matrix of the (weight vector) π_i and $\gamma(V) = \max_{1 \le k \le K}\{\|v_k\|\}$. We assume for some constant $C > 0$,

$$\lambda_{K-1}(G) \ge C^{-1}, \qquad \lambda_1(G) \le C, \qquad \gamma(V) \le C. \tag{8}$$

The first two items are a mild balance condition on π_i and the last one is a natural condition on V. Finally, in order for the (orthodox) SPA to perform well, we need

$$\sigma\sqrt{\log(n)}/s_{K-1}(\tilde{V}) \to 0. \tag{9}$$

In many applications, vertex hunting is used as a module in the main algorithm, and the data points fed into VH are from previous steps of some algorithm and satisfy $\sigma = o(1)$ (for example, see Jin et al. (2023); Ke & Wang (2022)). Hence, this condition is reasonable.

We present the main theorems (which are used to obtain Table 1). In what follows, Theorem 3 is for a general setting, and Theorem 2 concerns a slightly more restrictive setting. For each setting, we will specify explicitly the theoretically optimal choices of thresholds (t_n, ϵ_n) in pp-SPA.

For $1 \le k \le K$, let $J_k = \{i : r_i = v_k\}$ be the set of r_i located at vertex v_k, and let $n_k = |J_k|$, for $1 \le k \le K$. Let $\Gamma(\cdot)$ denote the standard Gamma function. Define

$$m = \min\{n_1, n_2, \ldots, n_K\}, \qquad c_2 = 0.5(2e^2)^{-\frac{1}{K-1}}\sqrt{2/(K-1)}\big[\Gamma(\frac{K+1}{2})\big]^{\frac{1}{K-1}}. \tag{10}$$

Note that as $K \to \infty$, $c_2 \to 0.5/\sqrt{e}$. We also introduce

$$\alpha_n = \frac{\sqrt{d}}{\sqrt{n}s^2_{K-1}(\tilde{V})}\big(1 + \sigma\sqrt{\max\{d, 2\log(n)\}}\big), \qquad b_n = \frac{2\sigma}{\sqrt{n}}\sqrt{\max\{d, 2\log(n)\}}. \tag{11}$$

The following theorem is proved in the appendix.

Theorem 2. *Suppose $X_1, X_2, \ldots, X_n$ are generated from model (1)-(2) where $m \ge c_1 n$ for a constant $c_1 > 0$ and conditions (7)-(9) hold. Fix δ_n such that $(K-1)/\log(n) \ll \delta_n \ll 1$, and let $t_n = \sqrt{K-1}\big(\frac{\log(n)}{n^{1-\delta_n}}\big)^{\frac{1}{K-1}}$. We apply pp-SPA to $X_1, X_2, \ldots, X_n$ with (N, Δ) to be determined below. Let $\hat{V} = [\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K]$, where $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$ are the estimated vertices.*

- *In the first case, $\alpha_n \ll t_n$. We take $N = \log(n)$ and $\Delta = c_3 t_n \sigma$ in pp-SPA, for a constant $c_3 \le c_2$. Up to a permutation of $\hat{v}_1, \ldots, \hat{v}_K$, $\max_{1\le k\le K}\{\|\hat{v}_k - v_k\|\} \le \sigma g_{\rm new}(V)[\sqrt{\delta_n} \cdot \sqrt{2\log(n)} + C\alpha_n] + b_n$.*
- *In the second case, $t_n \ll \alpha_n \ll 1$. We take $N = \log(n)$ and $\Delta = \sigma\alpha_n$ in pp-SPA. Up to a permutation of $\hat{v}_1, \ldots, \hat{v}_K$, $\max_{1\le k\le K}\{\|\hat{v}_k - v_k\|\} \le \sigma g_{\rm new}(V) \cdot (1 + o_{\mathbb{P}}(1))\sqrt{2\log(n)}$.*

To interpret Theorem 2, we consider a special case where $K = O(1)$, $s_{K-1}(\tilde{V})$ is lower bounded by a constant, and we set $\delta_n = \log\log(n)/\log(n)$. By our assumption (7), $d = o(\sqrt{n})$. It follows that $\alpha_n \asymp \max\big\{d, \sqrt{d\log(n)}\big\}/\sqrt{n}$, $b_n \asymp \sigma\sqrt{\max\{d, \log(n)\}/n}$, and $t_n \asymp [\log(n)]^{\frac{1}{K-1}}/n^{\frac{1-o(1)}{K-1}}$. We observe that α_n always dominates b_n/σ. Whether α_n dominates t_n is determined by d/n. When d/n is properly small so that $\alpha_n \ll t_n$, using the first case in Theorem 2, we get $\max_k\{\|\hat{v}_k - v_k\|\} \le C\big(\sqrt{\log(\log(n))} + \max\big\{d, \sqrt{d\log(n)}\big\}/\sqrt{n}\big) = O(\sqrt{\log\log(n)})$. When d/n is properly large so that $\alpha_n \gg t_n$, using the second case in Theorem 2, we get $\max_k\{\|\hat{v}_k - v_k\|\} = O\big(\sqrt{\log(n)}\big)$. We then combine these two cases and further plug in the constants in Theorem 2. It yields

$$\max_{1\le k\le K}\{\|\hat{v}_k^{\rm ppspa} - v_k\|\} \le \sigma g_{\rm new}(V) \cdot \begin{cases} \sqrt{\log\log(n)} & \text{if } d/n \text{ is properly small;} \\ \sqrt{[2+o(1)]\log(n)} & \text{if } d/n \text{ is properly large.} \end{cases} \tag{12}$$

It is worth comparing the error bound in Theorem 2 with that of the orthodox SPA (where we directly apply SPA on the original data points $X_1, X_2, \ldots, X_n$). Recall that $\beta(X)$ is as defined in (6). Note that $\beta(X) \le \max_{1\le i\le n}\|\epsilon_i\|$, where $\|\epsilon_i\|^2$ are i.i.d. variables from $\chi^2_d(0)$. Combining Lemma 5 and Theorem 1, we immediately obtain that for the (orthodox) SPA estimates $\hat{v}_1^{spa}, \hat{v}_2^{spa}, \ldots, \hat{v}_K^{spa}$, up to a permutation of these vectors (the constant a_1 is as in Lemma 5 and satisfies $a_1 > 2$):

$$\max_{1\le k\le K}\{\|\hat{v}_k^{\rm spa} - v_k\|\} \le \sigma g_{\rm new}(V) \cdot \begin{cases} \sqrt{\max\{d,\ 2\log(n)\}} & \text{if } d \ll \log(n) \text{ or } d \gg \log(n); \\ \sqrt{a_1\log(n)} & \text{if } d = a_0\log(n). \end{cases} \tag{13}$$

This bound is tight (e.g., when all r_i fall into vertices). We compare (13) with Theorem 2. If $d \gg \log(n)$, the improvement is a factor of $\sqrt{\log(n)/d}$, which is huge when d is large. If $d = O(\log(n))$, the improvement can still be a factor of $o(1)$ sometimes (e.g., in the first case of Theorem 2).

Theorem 2 assumes that there are a constant fraction of r_i falling at each vertex. This can be greatly relaxed. The following theorem is proved in the appendix.

Theorem 3. *Fix $0 < c_0 < 1$ and a sufficiently small constant $0 < \delta < c_0$. Suppose $X_1, X_2, \ldots, X_n$ are generated from model (1)-(2) where $m \ge n^{1-c_0+\delta}$ and conditions (7)-(9) hold. Let $t_n^* = \sqrt{K-1}\big(\frac{\log(n)}{n^{1-c_0}}\big)^{\frac{1}{K-1}}$. We apply pp-SPA to $X_1, X_2, \ldots, X_n$ with (N, Δ) to be determined below. Let $\hat{V} = [\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K]$, where $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$ are the estimated vertices.*

- *In the first case, $\alpha_n \ll t_n^*$. We take $N = \log(n)$ and $\Delta = c_3 t_n \sigma$ in pp-SPA, for a constant $c_3 \le e^{c_0/(K-1)}c_2$. Up to a permutation of $\hat{v}_1, \ldots, \hat{v}_K$, $\max_{1\le k\le K}\{\|\hat{v}_k - v_k\|\} \le \sigma g_{\rm new}(V)[\sqrt{c_0} \cdot \sqrt{2\log(n)} + C\alpha_n] + b_n$.*

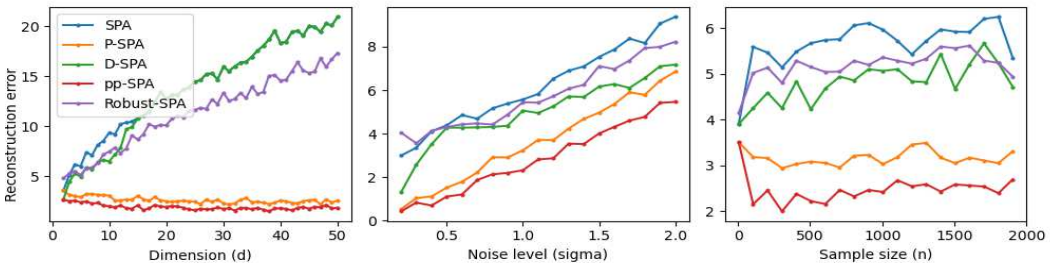

Figure 3: Performances of SPA, P-SPA, D-SPA, and pp-SPA in Experiment 1-3.

- *In the second case,* $\alpha_n \gg t_n^*$. *Suppose* $\alpha_n = o(1)$. *We take* $N = \log(n)$ *and* $\Delta = \alpha_n$ *in pp-SPA. Up to a permutation of* $\hat{v}_1, \ldots, \hat{v}_K$, $\max_{1\le k\le K}\{\|\hat{v}_k - v_k\|\} \le \sigma g_{\text{new}}(V) \cdot (1 + o_{\mathbb{P}}(1))\sqrt{2\log(n)}$.

Comparing Theorem 3 with Theorem 2, the difference is in the first case, where the $o(1)$ factor of δ_n is replaced by a constant factor of $c_0 < 1$. Similarly as in (12), we obtain

$$\max_{1\le k\le K}\{\|\hat{v}_k^{\text{ppspa}} - v_k\|\} \le \sigma g_{\text{new}}(V) \cdot \begin{cases} \sqrt{2c_0\log(n)} & \text{if } d/n \text{ is properly small;} \\ \sqrt{[2+o(1)]\log(n)} & \text{if } d/n \text{ is properly large.} \end{cases} \tag{14}$$

In this relaxed setting, we also compare Theorem 3 with (13): (a) When $d \gg \log(n)$, the improvement is a factor of $\sqrt{\log(n)/d}$. (b) When $d = O(\log(n))$, the improvement is at the constant order. It is interesting to further compare these "constants". Note that $g_{\text{new}}(V)$ is the same for all methods. It suffices to compare the constants in the bound for $\beta_{\text{new}}(V)$. In Case (b), the error bound of pp-SPA is smaller than that of SPA by a factor of $c_0 \in (0, 1)$. For the practical purpose, even the improvement of a constant factor can have a huge impact, especially when the data contain strong noise and potential outliers. Our simulations in Section 5 further confirm this point.

5 Numerical Study

We compare SPA, pp-SPA, and two simplified versions P-SPA and D-SPA (for illustration). We also compared these approaches with robust-SPA (Gillis, 2019) from `bit.ly/robustSPA` (with default tuning parameters). For pp-SPA and D-SPA, we need to specify tuning parameters (N, Δ). We use the heuristic choice in Remark 2. Fix $K = 3$ and three points $\{y_1, y_2, y_3\}$ in $\mathbb{R}^2$. Given (n, d, σ), we first draw $(n-30)$ points uniformly from the 2-dimensional simplex whose vertices are y_1, y_2, y_3, and then put 10 points on each vertex of this simplex. Denote these points by $w_1, w_2, \ldots, w_n \in \mathbb{R}^2$. Next, we fix a matrix $A \in \mathbb{R}^{d\times 2}$, whose top 2×2 block is equal to I_d and the remaining entries are zero. Let $r_i = Aw_i$, for all i. Finally, we generate $X_1, X_2, \ldots, X_n$ from model (1). We consider three experiments. In Experiment 1, we fix $(n, \sigma) = (1000, 1)$ and let d range in $\{1, 2, \ldots, 49, 50\}$. In Experiment 2, we fix $(n, d) = (1000, 4)$ and let σ range in $\{0.2, 0.3, \ldots, 2\}$. In Experiment 3, we fix $(d, \sigma) = (4, 1)$ and let n range in $\{500, 600, \ldots, 1500\}$. We evaluate the vertex hunting error $\max_k\{\|\hat{v}_k - v_k\|\}$ (subject to a permutation of $\hat{v}_1, \ldots, \hat{v}_K$). For each set of parameters, we report the average error over 20 repetitions. The results are in Figure 3. They are consistent with our theoretical insights: The performances of P-SPA and D-SPA are both better than that of SPA, and the performance of pp-SPA is better than those of P-SPA and D-SPA. It suggests that both the projection and denoise steps are effective in reducing noise, and it is beneficial to combine them. When $d \le 10$, pp-SPA, P-SPA and D-SPA all outperform robust-SPA; when $d > 10$, both pp-SPA and P-SPA outperform robust-SPA, and D-SPA (the simplified version without hyperplain projection) underperforms robust-SPA. The code to reproduce these experiments is available at `https://github.com/Gabriel78110/VertexHunting`.

6 DISCUSSION

Vertex hunting is a fundamental problem found in many applications. The Successive Projection algorithm (SPA) is a popular approach, but may behave unsatisfactorily in many settings. We propose pp-SPA as a new approach to vertex hunting. Compared to SPA, the new algorithm provides much improved theoretical bounds and encouraging improvements in a wide variety of numerical study. We also provide a sharper non-asymptotic bound for the orthodox SPA. For technical simplicity, our model assumes Gaussian noise, but our results are readily extendable to subGaussian noise. Also, our non-asymptotic bounds do not require any distributional assumption, and are directly applicable to different settings. For future work, we note that an improved bound on vertex hunting frequently implies improved bounds for methods that *contains vertex hunting as an important step*, such as Mixed-SCORE for network analysis (Jin et al., 2023; Bhattacharya et al., 2023), Topic-SCORE for text analysis (Ke & Wang, 2022), and state compression of Markov processes (Zhang & Wang, 2019), where vertex hunting plays a key role. Our algorithm and bounds may also be useful for related problems such as estimation of convex density support (Brunel, 2016).

A PROOF OF PRELIMINARY LEMMAS

A.1 PROOF OF LEMMA 1

This is a quite standard result, which can be found at tutorial materials (e.g., `https://people.math.wisc.edu/~roch/mmids/roch-mmids-llssvd-6svd.pdf`). We include a proof here only for convenience of readers.

We start by introducing some notation. Let $Z_i = X_i - \bar{X}$ and let $Z = [Z_1, \ldots, Z_n] \in \mathbb{R}^{d,n}$. Suppose the singular value decomposition of Z is given by $Z = U_Z D_Z V_Z'$. Since H is a rank-$(K-1)$ projection matrix, we have $H = QQ'$, where $Q \in \mathbb{R}^{d,K-1}$ is such that $Q'Q = I_{K-1}$. Hence, we rewrite the optimization in (3) as follows:

$$\text{minimize } \sum_{i=1}^{n}(X_i - x_0)'(I_d - QQ')(X_i - x_0), \quad \text{subject to} \quad Q'Q = I_{K-1}.$$

For $\lambda \in \mathbb{R}$, consider the Lagrangian objective function

$$\widetilde{S}(x_0, Q, \lambda) = \sum_{i=1}^{n}(X_i - x_0)'(I_d - QQ')(X_i - x_0) + \lambda(Q'Q - I_{K-1}). \tag{A.1}$$

Setting its gradients w.r.t. x_0 and Q to be 0 yields

$$\nabla_{x_0}\widetilde{S}(x_0, Q, \lambda) = -2(I_d - QQ')\sum_{i=1}^{n}(X_i - x_0) = 0, \tag{A.2}$$

$$\nabla_{Q}\widetilde{S}(x_0, Q, \lambda) = -2Q'\sum_{i=1}^{n}(X_i - x_0)(X_i - x_0)' + 2\lambda Q' = 0. \tag{A.3}$$

Firstly, we deduce from (A.2) that $\hat{x}_0 = \bar{X}$, which in view of (A.3) implies that $Q'(ZZ' - \lambda I_d) = 0$. The above equations also implies that the $(K-1)$ columns of $\widehat{Q}$ should be the distinct columns of U_Z. Now, the objective function in (A.1) is given by

$$\begin{aligned}\widetilde{S}(x_0, Q, \lambda) &= \sum_{i=1}^{n} Z_i'(I_d - QQ')Z_i = \mathrm{tr}[(I_d - QQ')ZZ'] = \mathrm{tr}[(I_d - QQ')U_Z D_Z^2 U_Z'] \\ &= \mathrm{tr}(D_Z)^2 - \mathrm{tr}[Q'U_Z D_Z^2 U_Z' Q] = \mathrm{tr}(D_Z^2) - \|D_Z U_Z' Q\|_{\mathrm{F}}^2. \end{aligned} \tag{A.4}$$

Note that for each column of $U_Z'Q \in \mathbb{R}^{d,K-1}$, it has exactly one entry being 1 and its other entries are all 0. Therefore, taking $\widehat{Q} = U$ maximizes $\|D_Z U_Z' Q\|_{\mathrm{F}}^2$ and hence minimizes the objective function $\widetilde{S}$ in (A.1), that is, $\widehat{H} = UU'$. The proof is complete.

A.2 PROOF OF LEMMA 3

For the simplex formed by $V \in \mathbb{R}^{d\times K}$, we can always find an orthogonal matrix $O \in \mathbb{R}^{d\times d}$ and a scalar a such that

$$OV = \begin{pmatrix} x_1 & x_2 & \dots & x_K \\ a & a & \dots & a \\ 0 & 0 & \dots & 0 \end{pmatrix}, \quad \text{where} \quad x_k \in \mathbb{R}^{K-1} \text{ for } k = 1, \dots, K.$$

Denote $\bar{x} = K^{-1}\sum_{k=1}^{K} x_k$. Further we can represent

$$O\tilde{V} = \begin{pmatrix} x_1 - \bar{x} & x_2 - \bar{x} & \dots & x_K - \bar{x} \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

We write $\tilde{X} := (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_K - \bar{x})$. Since rotation and location do not change the volume,

$$\text{Volume}(\mathcal{S}_0) = \text{Volume}(\mathcal{S}(\tilde{X})).$$

where $\mathcal{S}(\tilde{X})$ represents the simplex formed by $\tilde{X}$. By Stein (1966), we have

$$\text{Volume}(\mathcal{S}_0) = \frac{\det(\tilde{A})}{(K-1)!}, \quad \text{with} \quad \tilde{A} = \begin{bmatrix} 1 & (x_1 - \bar{x})' \\ 1 & (x_2 - \bar{x})' \\ \vdots & \vdots \\ 1 & (x_K - \bar{x})' \end{bmatrix}$$

We also define

$$A = \begin{bmatrix} 1 & (v_1 - \bar{v})' \\ 1 & (v_2 - \bar{v})' \\ \vdots & \vdots \\ 1 & (v_K - \bar{v})' \end{bmatrix} = [\mathbf{1}_K, \tilde{V}'],$$

Since $(\tilde{A}, 0) = A\begin{pmatrix} 1 & 0 \\ 0 & O \end{pmatrix}$, it follows that $\tilde{A}\tilde{A}' = AA'$ and $\text{Volume}(\mathcal{S}_0) = \frac{\sqrt{\det(AA')}}{(K-1)!} = \frac{\sqrt{\det(A'A)}}{(K-1)!}$. Note that $A'A = \begin{pmatrix} K & 0 \\ 0 & \tilde{V}\tilde{V}' \end{pmatrix}$ by the fact that $\tilde{V}\mathbf{1}_K = 0$. Then $\det(A'A) = K\det(\tilde{V}\tilde{V}')$. Further notice that $\text{rank}(\tilde{V}\tilde{V}') = K - 1$. We thus conclude that

$$\text{Volume}(\mathcal{S}_0) = \frac{\sqrt{K}}{(K-1)!}\prod_{k=1}^{K-1} s_k(\tilde{V}).$$

This proves the first claim.

For the second and last claims, we first notice that $V = \tilde{V} - \bar{v}\mathbf{1}_K'$. Then $VV' = \tilde{V}\tilde{V}' + K\bar{v}\bar{v}'$ again by $\tilde{V}\mathbf{1}_K = 0$. Because both $\tilde{V}\tilde{V}'$ and $K\bar{v}\bar{v}'$ are positive semi-definite, by Weyl's inequality (see, for example Horn & Johnson (1985)), it follows that $s_{K-1}(V) \geq s_{K-1}(\tilde{V})$ and $s_K(V) = \sqrt{\lambda_{\min}(VV')} \leq \sqrt{K\|\bar{v}\|^2} = \sqrt{K}\|\bar{v}\|$.

A.3 PROOF OF LEMMA 4

We first prove claim (a). Let $\Pi = [\pi_1 - \bar{\pi}, \dots, \pi_n - \bar{\pi}] \in \mathbb{R}^{K,n}$. Recalling the definitions of G and V, we have $G = n^{-1}\Pi\Pi'$ and $R = n^{-1/2}V\Pi$, so that $RR' = n^{-1}V\Pi\Pi'V' = VGV'$.

Next, we prove claim (b). Recall that $\tilde{V} = V - \bar{v}1_K'$, so that $\tilde{V}\tilde{V}' = (V - \bar{v}1_K')(V - \bar{v}1_K')' = VV' - K\bar{v}\bar{v}'$. Note that Since $\pi_i'1_K = \bar{\pi}'1_K = 1$, we have $\Pi'1_K = 0$, which implies that $G1_K = n^{-1}\Pi(\Pi'1_K) = 0$. We deduce from this observation that $\lambda_K(G) = 0$ and its associated eigenvector is $K^{-1/2}\mathbf{1}_K$. Therefore, $G - \lambda_{K-1}(G)I_K + K^{-1}\lambda_{K-1}(G)\mathbf{1}_K\mathbf{1}_K'$ is a positive semi-definite matrix, so that

$$\begin{aligned} VGV' - \lambda_{K-1}(G)\tilde{V}\tilde{V}' &= VGV' - \lambda_{K-1}(G)VV' + \lambda_{K-1}(G)K\bar{v}\bar{v}' \\ &= V[G - \lambda_{K-1}(G)I_K + K^{-1}\lambda_{K-1}(G)\mathbf{1}_K\mathbf{1}_K']V' \geq 0. \end{aligned}$$

In addition, observing that $\Pi' 1_K = 0$ due to the fact that $\|\pi_i\|_1 = \|\bar{\pi}\|_1 = 1$, we obtain that

$$\tilde{V}G\tilde{V}' = (V - \bar{v}1_K')G(V - \bar{v}1_K')' = n^{-1}(V - \bar{v}1_K')\Pi\Pi'(V - \bar{v}1_K')' = VGV'.$$

Therefore,

$$\lambda_1(G)\tilde{V}\tilde{V}' - VGV' = \lambda_1(G)\tilde{V}\tilde{V}' - \tilde{V}G\tilde{V}' = \tilde{V}[\lambda_1(G)I_K - G]\tilde{V}' \geq 0,$$

which completes the proof of claim (b).

Finally, for claim (c), we obtain from (a) that $\sigma_{K-1}^2(R) = \lambda_{K-1}(RR') = \lambda_{K-1}(VGV')$, which by Weyl's inequality (see, for example, Horn & Johnson (1985)) and in view of claim (b) implies that $\lambda_{K-1}(G)\lambda_{K-1}(\tilde{V}\tilde{V}') \leq \sigma_{K-1}^2(R) \leq \lambda_1(G)\lambda_{K-1}(\tilde{V}\tilde{V}')$. The proof is therefore complete.

A.4 Proof of Lemma 5

Recall that $z_1 \sim \chi_d^2(0)$. Let b_n be the value such that

$$\mathbb{P}(z_1 \geq b_n) = 1/n.$$

By basic extreme value theory, it is known that

$$\frac{\max_{1\leq i\leq n}\{z_i\}}{b_n} \to 1, \qquad \text{in probability.}$$

We now solve for b_n. It is seen that $b_n \geq d$. Recall that the density of $\chi_d^2(0)$ is

$$\frac{1}{2^{d/2}\Gamma(d/2)} x^{d/2-1}e^{-x/2}, \qquad x > 0.$$

Note that for any $x_0 \geq d$,

$$\int_{x_0}^{\infty} x^{d/2-1}e^{-x/2}dx = 2x_0^{d/2-1}e^{-x_0/2} + \int_{x_0}^{\infty}(d-2)x^{d/2-2}e^{-x/2}dx \tag{A.5}$$

where the RHS is no greater than

$$\leq 2x_0^{d/2-1}e^{-x_0/2} + \frac{(d-2)}{x_0}\int_{x_0}^{\infty} x^{d/2-1}e^{-x/2}dx.$$

It follows that for all $x_0 \geq d$,

$$2x_0^{d/2-1}e^{-x_0/2} \leq \int_{x_0}^{\infty} x^{d/2-1}e^{-x/2}dx \leq x_0 \cdot x_0^{d/2-1}e^{-x_0/2}, \tag{A.6}$$

where we have used

$$\frac{x_0}{x_0 - d + 2} \leq x_0/2.$$

It now follows that there is a term $a(x)$ such that when $x \geq d$,

$$1 \leq a(x) \leq x/2$$

and

$$\mathbb{P}(z_1 \geq x) = a(x)\frac{1}{2^{d/2}\gamma(d/2)}2x^{d/2-1}e^{-x/2}.$$

Combining these, b_n is the solution of

$$a(x)\frac{1}{2^{d/2}\gamma(d/2)}2x^{d/2-1}e^{-x/2} = \frac{1}{n}. \tag{A.7}$$

We now solve the equation in (A.7). Consider the case d is even. The case where d is odd is similar, so we omit it. When d is even, using

$$\Gamma(d/2) = (d/2-1)! = (2/d)(d/2)! = (2/d)\theta(\frac{d}{2e})^{d/2},$$

where θ is the factor in the Stirling's formula which is $\leq C\sqrt{\log(d)}$. Plugging this into the left hand side of (A.7) and re-arrange, we have

$$\log(d/x) + (d/2)\log(\frac{ex}{d}) - x/2 = -\log(n) + o(\log(n)). \tag{A.8}$$

We now consider three cases below separately.

- Case 1. $d \ll \log(n)$.
- Case 2. $d = a_0 \log(n)$ for a constant $a_0 > 0$.
- Case 3. $d \gg \log(n)$.

Consider Case 1. In this case, it is seen that when

$$x = O(\log(n)),$$

the LHS of (A.8) is

$$-x/2 + o(\log(n)).$$

Therefore, the solution of (A.8) is seen to be

$$b_n = (1 + o(1)) \cdot 2\log(n).$$

Consider Case 2. In this case, $d = a_0 \log(n)$. Let $x = b_1 \log(n)$. Plugging these into (A.8) and rearranging,

$$a_1 - a_0 \log(a_1) = 2 + a_0 - a_0 \log(a_0) + o(1). \tag{A.9}$$

Now, consider the equation

$$a_1 - a_0 \log(a_1) = 2 + a_0 - a_0 \log(a_0).$$

It is seen that the equation has a unique solution (denoted by b_0) that is bigger than 2. Therefore, in this case,

$$b_n = (1 + o(1))b_0,$$

Consider Case 3. In this case, $d \gg \log(n)$. Consider again the equation

$$\log(d/x) + (d/2)\log(\frac{ex}{d}) - x/2 = -\log(n) + o(\log(n)).$$

Letting $y = x/d$ and rearranging, it follows that

$$y - \log(y) - 1 = o(1), \tag{A.10}$$

where for sufficiently large n, $o(1) > 0$ and $o(1) \to 0$. Note that the function $g(y) = y - \log(y) - 1$ is a convex function with a minimum of 0 reached at $y = 1$, it follows

$$y = 1 + o(1).$$

Recalling $y = x/d$, this shows

$$b_n = (1 + o(1))d.$$

This completes the proof of Lemma 5.

B Analysis of the SPA algorithm

Fix $d \geq K - 1$. For any $V = [v_1, v_2, \ldots, v_K] \in \mathbb{R}^{d\times K}$, let $\sigma_k(V)$ denote the kth singular value of V, and define

$$\gamma(V) = \min_{v_0\in\mathbb{R}^d} \max_{1\leq k\leq K} \|v_k - v_0\|, \qquad d_{\max}(V) = \max_{x\in\mathcal{S}} \|x\|.$$

To capture the error bound for SPA, we introduce a useful quantity in the main paper:

$$\beta(X, V) := \max\Big\{ \max_{1\leq i\leq n} \mathrm{Dist}(X_i, \mathcal{S}), \quad \max_{1\leq k\leq K} \min_{i:r_i=v_k} \|X_i - v_k\| \Big\}. \tag{B.11}$$

We note that when $\max_i \mathrm{Dist}(X_i, \mathcal{S})$ is small, no point is too far away from the simplex; and when $\max_k \min_{i:r_i=v_k} \|X_i - v_k\|$ is small, there is at least one point near each vertex.

Let's denote $\gamma = \gamma(V)$, $d_{\max} = d_{\max}(V)$, $\beta = \beta(X, V)$, and $\sigma_* = \sigma_{K-1}(V)$ for brevity. We shall prove the following theorem, which is a slightly stronger version of Theorem 1 in the main paper.

Theorem B.1. *Suppose for each $1 \leq k \leq K$, there exists $1 \leq i \leq n$ such that $\pi_i = e_k$. Suppose $\beta(X, V)$ satisfies that $450 d_{\max} \max\{1, \frac{d_{\max}}{\sigma_*}\}\beta \leq \sigma_*^2$. Let $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_r$ be the output of SPA. Up to a permutation of these r vectors,*

$$\max_{1\leq k\leq r} \|\hat{v}_k - v_k\| \leq \Big(1 + \frac{30\gamma}{\sigma_*} \max\{1, \frac{d_{\max}}{\sigma_*}\}\Big)\beta(X, V).$$

B.1 SOME PRELIMINARY LEMMAS IN LINEAR ALGEBRA

To establish Theorem B.1, it is necessary to develop a few lemmas in linear algebra. First, we notice that the vertex matrix V defines a mapping from the standard probability simplex $\mathcal{S}^*$ to the target simplex $\mathcal{S}$. The following lemma gives some properties of the mapping:

Lemma B.1. *Let $\mathcal{S}^* \subset \mathbb{R}^K$ be the standard probability simplex consisting of all weight vectors. Let $F : \mathcal{S}^* \to \mathcal{S}$ be the mapping with $F(\pi) = V\pi$. For any π and $\tilde{\pi}$ in $\mathcal{S}^*$,*

$$\sigma_{K-1}(V) \cdot \|\pi - \tilde{\pi}\| \leq \|F(\pi) - F(\tilde{\pi})\| \leq \gamma(V) \cdot \|\pi - \tilde{\pi}\|_1. \tag{B.12}$$

Fix $1 \leq s \leq K-2$. If π and $\tilde{\pi}$ share at least s common entries, then

$$\|F(\pi) - F(\tilde{\pi})\| \geq \sigma_{K-1-s}(V)\|\pi - \tilde{\pi}\|. \tag{B.13}$$

The first claim of Lemma B.1 is about the case where $\mathcal{S}$ is non-degenerate. In this case,

$$\sigma_{K-1}(V) > 0.$$

Hence, we can upper/lower bound the distance between any two points in $\mathcal{S}$ by the distance between their barycentric coordinates. The second claim considers the case where $\mathcal{S}$ can be degenerate (i.e., $\sigma_{K-1}(V) = 0$ is possible) but

$$\sigma_{K-1-s}(V) > 0.$$

We can still use (B.12) to upper bound the distance between two points in $\mathcal{S}$ but the lower bound there is ineffective. Fortunately, if the two points share s common entries in their barycentric coordinates (which implies that the two points are on the same face or edge), then we can still lower bound the distance between them.

Second, we study the Euclidean norm of a convex combination of m points. Let $w_1, \ldots, w_m$ be the convex combination weights. By the triangle inequality,

$$\Big\|\sum_{i=1}^m w_i x_i\Big\| \leq \sum_{i=1}^m w_i \|x_i\| \leq \max_{1\leq k\leq K} \|v_k\|.$$

This explains why $\max_{x\in\mathcal{S}} \|x\|$ is always attained at a vertex. Write

$$\delta := \sum_{i=1}^m w_i \|x_i\| - \Big\|\sum_{i=1}^m w_i x_i\Big\|.$$

Knowing $\delta \geq 0$ is not enough for showing Theorem B.1. We need to have an explicit lower bound for δ, as given in the following lemma.

Lemma B.2. *Fix $m \geq 2$ and $x_1, \ldots, x_m \in \mathbb{R}^d$. Let $a = \min_{i\neq j} \|x_i - x_j\|$ and $b = \max_{i\neq j} |\|x_i\| - \|x_j\||$. For any $w_1, \ldots, w_m \geq 0$ such that $\sum_{i=1}^m w_i = 1$,*

$$\Big\|\sum_{i=1}^m w_i x_i\Big\| \leq L - \frac{a^2 - b^2}{4L} \sum_{i=1}^m w_i(1 - w_i), \quad \textit{with } L := \sum_{i=1}^m w_i \|x_i\|. \tag{B.14}$$

By Lemma B.2, the lower bound for δ has the expression $\frac{a^2-b^2}{4L}\sum_{i=1}^m w_i(1-w_i)$. This lower bound is large if $a = \min_{i\neq j} \|x_i - x_j\|$ is properly large, and $b = \max_{i\neq j} |\|x_i\| - \|x_j\||$ is properly small, and $\sum_i w_i(1 - w_i)$ is properly large.

- A large a means that these m points are sufficiently 'different' from each other.
- A small b means that the norms of these m points are sufficiently close.
- A large $\sum_i w_i(1 - w_i)$ prevents each of w_i from being too close to 1, implying that the convex combination is sufficiently 'mixed'.

Later in Section B.2, we will see that Lemma B.2 plays a critical role in the proof of Theorem B.1.

Third, we explore the projection of $\mathcal{S}$ into a lower-dimensional space. Let $H \in \mathbb{R}^{d\times d}$ be an arbitrary projection matrix with rank s. We use $(I_d - H)$ to project $\mathcal{S}$ into the orthogonal complement of H, where the projected vertices are the columns of

$$V^\perp = (I_d - H)V.$$

Since the projected simplex is not guranteed to be non-degenerate, it is possible that $\sigma_{K-1}(V^\perp) = 0$. However, we have a lower bound for $\sigma_{K-1-s}(V^\perp)$, as given in the following lemma:

Lemma B.3. *Fix $1 \le s \le K-2$. For any projection matrix $H \in \mathbb{R}^{d\times d}$ with rank s,*

$$\sigma_{K-1-s}((I_d - H)V) \ge \sigma_{K-1}(V). \tag{B.15}$$

Finally, we present a lemma about

$$d_{\max} = \max_{x\in\mathcal{S}} \|x\| = \max_{1\le k\le K} \|v_k\|.$$

In the analysis of SPA, it is not hard to get a lower bound for $d_{\max}$ in the first iteration. However, as the algorithm successively projects $\mathcal{S}$ into lower-dimensional subspaces, we need to keep track of this quantity for the projected simplex spanned by $V^\perp$. Lemma B.3 shows that the singular values of $V^\perp$ can be lower bounded. It motivates us to have a lemma that provides a lower bound of $d_{\max}$ in terms of the singular values of V.

Lemma B.4. *Fix $0 \le s \le K-2$. Suppose there are at least s indices, $\{k_1,\ldots,k_s\} \subset \{1,2,\ldots,K\}$, such that $\|v_k\| \le \delta$. If $\sigma^2_{K-1-s}(V) \ge 2(K-2)\delta^2$, then*

$$\max_{1\le k\le K} \|v_k\| \ge \frac{\sqrt{K-s-1}}{\sqrt{2(K-s)}}\,\sigma_{K-1-s}(V) \ge \frac{1}{2}\sigma_{K-1-s}(V). \tag{B.16}$$

B.2 The simplicial neighborhoods and a key lemma

We fix a simplex $\mathcal{S} \subset \mathbb{R}^d$ whose vertices are $v_1, v_2, \ldots, v_K$. Write $V = [v_1, v_2, \ldots, v_K] \in \mathbb{R}^{d\times K}$. Let $\mathcal{S}^*$ denote the standard probability simplex, and let $F : \mathcal{S}^* \to \mathcal{S}$ be the mapping in Lemma B.1. We introduce a local neighborhood for each vertex that has a "simplex shape":

Definition B.1. *Given $\epsilon \in (0,1)$, for each $1 \le k \le K$, the ϵ-simplicial-neighborhood of v_k inside the simplex $\mathcal{S}$ is defined by*

$$\mathcal{V}_k(\epsilon) := \{F(\pi) : \pi \in \mathcal{S}^*,\ \pi(k) \ge 1-\epsilon\}.$$

These simplicial neighborhoods are highlighted in blue in Figure 4.

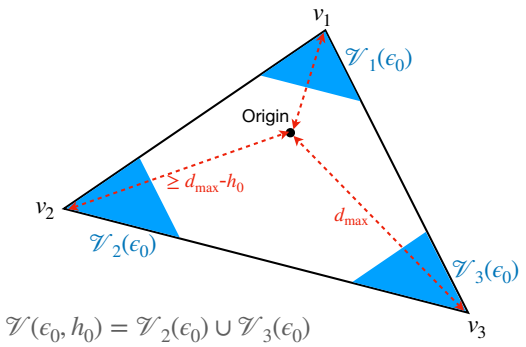

Figure 4: An illustration of the simplicial neighborhoods and $\mathcal{V}(\epsilon_0, h_0)$.

First, we verify that each $\mathcal{V}_k(\epsilon)$ is indeed a "neighborhood" in the sense each $x \in \mathcal{V}_k(\epsilon)$ is sufficiently close to v_k. Note that $v_k = F(e_k)$, where e_k is the kth standard basis vector of $\mathbb{R}^K$. For any $\pi \in \mathcal{S}^*$,

$$\|\pi - e_k\|_1 = 2[1-\pi(k)].$$

By Definition B.1, for any $x \in \mathcal{V}_k(\epsilon)$, its barycentric coordinate π satisfies $1-\pi(k) \le \epsilon$. It follows by Lemma B.1 that

$$\max_{x\in\mathcal{V}_k(\epsilon)} \|x - v_k\| = \max_{\pi\in\mathcal{S}^*:\pi(k)\le 1-\epsilon} \|F(\pi) - F(e_k)\| \le 2\gamma(V)\epsilon. \tag{B.17}$$

Hence, $\mathcal{V}_k(\epsilon)$ is within a ball centered at v_k with a radius of $2\gamma(V)\epsilon$. However, we opt to utilize these simplex-shaped neighborhoods instead of standard balls, as this choice greatly simplifies proofs.

Next, we show that as long as $\epsilon < 1/2$, the K neighborhoods $\mathcal{V}_1(\epsilon), \ldots, \mathcal{V}_K(\epsilon)$ are non-overlapping. By Lemma B.1,

$$\|v_k - v_\ell\| \geq \sigma_{K-1}(V)\|e_k - e_\ell\| \geq \sqrt{2}\sigma_{K-1}(V), \qquad \text{for } 1 \leq k \neq \ell \leq K. \tag{B.18}$$

When $x \in \mathcal{V}_k(\epsilon)$, the kth entry of $\pi := F^{-1}(x)$ is at least $1 - \epsilon > 1/2$. Since each $\pi \in \mathcal{S}^*$ cannot have two entries larger than $1/2$, these neighborhoods are disjoint:

$$\mathcal{V}_k(\epsilon) \cap \mathcal{V}_\ell(\epsilon) = \emptyset, \qquad \text{for any } 1 \leq k \neq \ell \leq K. \tag{B.19}$$

An intuitive explanation of our proof ideas for Theorem B.1: We outline our proof strategy using the example in Figure 4. The first step of SPA finds

$$i_1 = \text{argmax}_{1 \leq i \leq n} \|X_i\|.$$

The population counterpart of X_{i_1} is denoted by r_{i_1}. We will explore the region of the simplex that r_{i_1} falls into. In the noiseless case, $X_i = r_i$ for all $1 \leq i \leq n$. Since the maximum Euclidean norm over a simplex can only be attained at vertex, r_{i_1} must equal to one of the vertices. In Figure 4, the vertex v_3 has the largest Euclidean norm, hence, $r_{i_1} = v_3$ in the noiseless case. In the noisy case, the index i that maximizes $\|X_i\|$ may not maximize $\|r_i\|$; i.e., r_{i_1} may not have the largest Euclidean norm among r_i's. Noticing that $\|v_3\| > \|v_2\| > \|v_1\|$, we expect to see two possible cases:

- Possibility 1: r_{i_1} is in the ϵ-simplicial-neighborhood of v_3, for a small $\epsilon > 0$.
- Possibility 2 (when $\|v_2\|$ is close to $\|v_3\|$): r_{i_1} is in the ϵ-simplicial-neighborhood of v_2.

The focus of our proof will be showing that r_{i_1} falls into $\mathcal{V}_2(\epsilon) \cup \mathcal{V}_3(\epsilon)$. No matter $r_i \in \mathcal{V}_2(\epsilon)$ holds or $r_i \in \mathcal{V}_3(\epsilon)$ holds, the corresponding $\hat{v}_1 = X_{i_1}$ is close to one of the vertices.

Formalization of the above insights, and a key lemma: Introduce the notation

$$\mathcal{K}^* = \{k : \|v_k\| = d_{\max}\}, \qquad \text{where} \quad d_{\max} := \max_{x \in \mathcal{S}} \|x\| = \max_k \|v_k\|. \tag{B.20}$$

Given any $h_0 > 0$ and $\epsilon_0 \in (0, 1/2)$, let $\mathcal{V}_k(\epsilon_0)$ be the same as in Definition B.1, and we define an index set $\mathcal{K}(h_0)$ and a region $\mathcal{V}(\epsilon_0, h_0) \subset \mathcal{S}$ as follows:

$$\mathcal{K}(h_0) = \{k : \|v_k\| \geq d_{\max} - h_0\}, \qquad \mathcal{V}(\epsilon_0, h_0) = \cup_{k \in \mathcal{K}(h_0)} \mathcal{V}_k(\epsilon_0), \tag{B.21}$$

For the example in Figure 4, $\mathcal{K}^* = \{3\}$, $\mathcal{K}(h_0) = \{2, 3\}$, and $\mathcal{V}(\epsilon_0, h_0) = \mathcal{V}_2(\epsilon_0) \cup \mathcal{V}_3(\epsilon_0)$.

In the proof of Theorem B.1, we will repeatedly use the following key lemma, which states that the Euclidean norm of any point in $\mathcal{S} \setminus \mathcal{V}(\epsilon_0, h_0)$ is strictly smaller than $d_{\max}$ by a certain amount:

Lemma B.5. *Fix a simplex $\mathcal{S} \subset \mathbb{R}^d$ with vertices $v_1, v_2, \ldots, v_K$. Write $d_{\max} = \max_{1 \leq k \leq K} \|v_k\|$. Suppose there exists $\sigma_* > 0$ such that*

$$d_{\max} \geq \sigma_*/2, \qquad \textit{and} \qquad \min_{1 \leq k \neq \ell \leq K} \|v_k - v_\ell\| \geq \sqrt{2}\sigma_*. \tag{B.22}$$

Let $\mathcal{K}(h_0)$ and $\mathcal{V}(\epsilon_0, h_0)$ be as defined in (B.21). Given any $t > 0$ such that $\max\{1, d_{\max}/\sigma_*\}t < 3\sigma_*$, if we set (h_0, ϵ_0) such that

$$h_0 = \sigma_*/3, \qquad \textit{and} \qquad 1/2 > \epsilon_0 \geq 6\sigma_*^{-1} \max\{1, d_{\max}/\sigma_*\}t, \tag{B.23}$$

then

$$\|x\| \leq d_{\max} - t, \qquad \textit{for all } x \in \mathcal{S} \setminus \mathcal{V}(\epsilon_0, h_0). \tag{B.24}$$

Lemma B.5 will be proved in Section B.4.5, where we invoke Lemma B.2 to prove the claim here.

B.3 Proof of Theorem B.1 (Theorem 1 in the main paper)

The proof consists of three steps. In Step 1, we study the first iteration of SPA and show that $\hat{v}_1$ falls in the neighborhood of a true vertex. In Steps 2-3, we recursively study the remaining iterations and show that, if $\hat{v}_1, \ldots, \hat{v}_{s-1}$ fall into the neighborhoods of $(s-1)$ true vertices, one per each, then $\hat{v}_k$ will also fall into the neighborhood of another true vertex. For clarity, we first study the second

iteration in Step 2 (for which the notations are simpler), and then study the sth iteration for a general s in Step 3.

Let's denote for brevity:

$$\gamma = \gamma(V), \qquad d_{\max} = d_{\max}(V), \qquad \sigma_* = \sigma_{K-1}(V), \qquad \beta = \beta(X, V).$$

Write $J_k = \{1 \le i \le n : \pi_i(k) = 1\}$, for $1 \le k \le K$. From the definition of $\beta(X, V)$,

$$\max_{1\le i\le n} \mathrm{Dist}(X_i, \mathcal{S}) \le \beta, \qquad \max_{1\le k\le K} \min_{i\in J_k} \|X_i - v_k\| \le \beta. \tag{B.25}$$

Step 1: Analysis of the first iteration of SPA.

Applying Lemma B.4 with $s = 0$, we have $d_{\max} \ge \sigma_*/2$. We then apply Lemma B.5. Let $\mathcal{V}(\epsilon_0, h_0)$ be as in (B.21), with

$$h_0 = \sigma_*/3, \qquad \text{and} \qquad \epsilon_0 = 15 \max\{\sigma_*, \sigma_*^{-2} d_{\max}\}\beta. \tag{B.26}$$

Our assumptions yield $\epsilon_0 < 1/2$. Additionally, when $t = 7\beta/3$, $\epsilon_0 \ge 6\sigma_*^{-1} \max\{1, d_{\max}/\sigma^*\}t$, which satisfies (B.23). We apply Lemma B.5 with $t = 7\beta/3$. It yields

$$\max_{x\in\mathcal{S}\setminus\mathcal{V}(\epsilon_0,h_0)} \|x\| \le d_{\max} - 7\beta/3. \tag{B.27}$$

At the same time, let $\mathcal{K}^*$ be the same as in (B.20). For any $k \in \mathcal{K}^*$, it follows by (B.25) that

$$\text{there exists at least one } i^* \in J_k \text{ such that } \|X_{i^*} - v_k\| \le \beta.$$

Note that $\|v_k\| = d_{\max}$ for $k \in \mathcal{K}^*$. It follows by the triangle inequality that

$$\|X_{i^*}\| \ge \|v_k\| - \beta \ge d_{\max} - \beta.$$

Since $\|X_{i_1}\| = \max_i \|X_i\|$, we immediately have:

$$\|X_{i_1}\| \ge \|X_{i^*}\| \ge d_{\max} - \beta. \tag{B.28}$$

Combining (B.27) and (B.28), we conclude that $X_{i_1} \notin \mathcal{S} \setminus \mathcal{V}(\epsilon_0, h_0)$; in other words,

$$X_{i_1} \text{ can only be inside } \mathcal{V}(\epsilon_0, h_0) \text{ or outside } \mathcal{S}. \tag{B.29}$$

Suppose X_{i_1} is outside $\mathcal{S}$. Let $\mathrm{proj}_{\mathcal{S}}(X_{i_1}) \in \mathbb{R}^d$ be the point in the simplex that is closest to X_{i_1}. In other words, $\|X_{i_1} - \mathrm{proj}_{\mathcal{S}}(X_{i_1})\| = \min_{x\in\mathcal{S}} \|X_{i_1} - x\| = \mathrm{Dist}(X_{i_1}, \mathcal{S})$. Using the first inequality in (B.25), we have

$$\|X_{i_1} - \mathrm{proj}_{\mathcal{S}}(X_{i_1})\| \le \beta. \tag{B.30}$$

It follows by the triangle inequality and (B.28) that

$$\|\mathrm{proj}_{\mathcal{S}}(X_{i_1})\| \ge \|X_{i_1}\| - \beta \ge d_{\max} - 2\beta.$$

Combining it with (B.27), we conclude that $\mathrm{proj}_{\mathcal{S}}(X_{i_1})$ cannot be in $\mathcal{S} \setminus \mathcal{V}(\epsilon_0, h_0)$. So far, we have shown that one of the following cases must happen:

$$\begin{aligned} &\text{Case 1: } X_{i_1} \in \mathcal{V}(\epsilon_0, h_0), \\ &\text{Case 2: } X_{i_1} \notin \mathcal{S}, \text{ and } \mathrm{proj}_{\mathcal{S}}(X_{i_1}) \in \mathcal{V}(\epsilon_0, h_0). \end{aligned} \tag{B.31}$$

In Case 1, since $\mathcal{V}_1(\epsilon_0), \ldots, \mathcal{V}_K(\epsilon_0)$ are disjoint, there exists only one $k_1 \in \mathcal{K}(h_0)$ such that $X_{i_1} \in \mathcal{V}_{k_1}(\epsilon_0)$. It follows by (B.17) that

$$\|X_{i_1} - v_{k_1}\| \le 2\gamma\epsilon_0, \qquad \text{in Case 1.} \tag{B.32}$$

In Case 2, similarly, there is only one $k_1 \in \mathcal{K}(h_0)$ such that $\mathrm{proj}_{\mathcal{S}}(X_{i_1}) \in \mathcal{V}_{k_1}(\epsilon_0)$. It follows by (B.17) again that

$$\|\mathrm{proj}_{\mathcal{S}}(X_{i_1}) - v_{k_1}\| \le 2\gamma\epsilon_0.$$

Combining it with (B.30) gives

$$\begin{aligned} \|X_{i_1} - v_{k_1}\| &\le \|X_{i_1} - \mathrm{proj}_{\mathcal{S}}(X_{i_1})\| + \|\mathrm{proj}_{\mathcal{S}}(X_{i_1}) - v_{k_1}\| \\ &\le 2\gamma\epsilon_0 + \beta, \qquad \text{in Case 2.} \end{aligned} \tag{B.33}$$

We put (B.32) and (B.33) together and plug in the value of ϵ_0 in (B.26). It yields:

$$\begin{aligned}\|X_{i_1} - v_{k_1}\| &\le \beta + 2\gamma\epsilon_0 \\ &\le \Big(1 + \frac{30\gamma}{\sigma_*}\max\big\{1, \frac{d_{\max}}{\sigma_*}\big\}\Big)\beta, \qquad \text{for some } k_1. \end{aligned} \tag{B.34}$$

Step 2: Analysis of the second iteration of SPA.

Let $H_1 = I_d - \frac{1}{\|X_{i_1}\|^2} X_{i_1}X_{i_1}'$ and $\widetilde{X}_i = H_1 X_i$, for $1 \le i \le n$. The second iteration operates on the data points $\widetilde{X}_1, \ldots, \widetilde{X}_n \in \mathbb{R}^d$. Write

$$\tilde{r}_i = H_1 r_i, \qquad \tilde{\epsilon}_i = H_1\epsilon_i, \qquad \tilde{v}_k = H_1 v_k, \qquad \widetilde{V} = [\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_K].$$

It follows that

$$\widetilde{X}_i = \widetilde{V}\pi_i + \tilde{\epsilon}_i, \qquad 1 \le i \le n. \tag{B.35}$$

Let $\widetilde{S} \subset \mathbb{R}^d$ denote the projected simplex, whose vertices are $\tilde{v}_1, \ldots, \tilde{v}_K$. Let $\widetilde{F}$ denote the mapping from the standard probability simplex $\mathcal{S}^*$ to the projected simplex $\widetilde{S}$ (note that $\widetilde{F}$ is not necessarily a one-to-one mapping). We consider the neighborhoods of $\widetilde{S}$ using Definition B.1

$$\widetilde{\mathcal{V}}_k(\epsilon) = \big\{\widetilde{F}(\pi) : \pi \in \mathcal{S}^*,\, \pi_i(k) \ge 1 - \epsilon\big\} \subset \mathbb{R}^d, \qquad 1 \le k \le K. \tag{B.36}$$

Let k_1 be as in (B.34). Let $\tilde{d}_{\max} := \max_{x\in\widetilde{S}} \|x\|$. The maximum distance $\tilde{d}_{\max}$ is attained at one or multiple vertices. Same as before, let $\widetilde{\mathcal{K}}^*$ be the index set of k at which $\|\tilde{v}_k\| = \tilde{d}_{\max}$. We similarly define

$$\widetilde{\mathcal{K}}(h_0) = \{k : \|\tilde{v}_k\| \ge \tilde{d}_{\max} - h_0\}, \qquad \widetilde{\mathcal{V}}(\epsilon_0, h_0) = \cup_{k\in\widetilde{\mathcal{K}}(h_0)}\widetilde{\mathcal{V}}_k(\epsilon_0). \tag{B.37}$$

At the same time, let $\tilde{\beta} = \beta(\widetilde{X}, \widetilde{V})$. It is easy to see that for any points x and y, $\|H_1 x - H_1 y\| \le \|x - y\|$. Hence, $\tilde{\beta} \le \beta$. It follows that

$$\max_{1\le i\le n} \mathrm{Dist}(\widetilde{X}_i, \widetilde{\mathcal{S}}) \le \beta, \qquad \max_{1\le k\le K}\min_{i\in J_k} \|\widetilde{X}_i - \tilde{v}_k\| \le \beta. \tag{B.38}$$

Additionally, we have the following lemma:

Lemma B.6. *Under the conditions of Theorem B.1, for $\sigma_* = \sigma_{K-1}(V)$, the following claims are true:*

$$\tilde{d}_{\max} \ge \sigma_*/2, \quad \min_{\substack{(k,\ell):k\ne k_1,\\ \ell\ne k_1, k\ne\ell}} \|\tilde{v}_k - \tilde{v}_\ell\| \ge \sqrt{2}\sigma_*, \quad \textit{and} \quad k_1 \notin \widetilde{\mathcal{K}}(h_0). \tag{B.39}$$

Given (B.35)-(B.39), we now apply Lemma B.5 to study the projected simplex $\widetilde{S}$. Similarly as how we obtain (B.27), by choosing

$$h_0 = \sigma_*/3, \qquad \text{and} \qquad \epsilon_1 = 15\max\{\sigma_*, \sigma_*^{-2}\tilde{d}_{\max}\},$$

we get $\max_{x\in\widetilde{\mathcal{S}}\setminus\widetilde{\mathcal{V}}(\epsilon_1,h_0)} \|x\| \le \tilde{d}_{\max} - 7\beta/3$. Note that $\epsilon_1 \le \epsilon_0$, and the set $\widetilde{S} \setminus \widetilde{V}(\epsilon, h_0)$ becomes smaller as ϵ increases. We immediately have

$$\max_{x\in\widetilde{\mathcal{S}}\setminus\widetilde{\mathcal{V}}(\epsilon_0,h_0)} \|x\| \le \tilde{d}_{\max} - 7\beta/3. \tag{B.40}$$

At the same time, by (B.38) and (B.39), it is easy to get (similar to how we obtained (B.28))

$$\|\tilde{X}_{i_2}\| \ge \tilde{d}_{\max} - \beta.$$

We can mimic the analysis between (B.28) and (B.31) to show that one of the two cases happens:

$$\begin{aligned}&\text{Case 1: } \widetilde{X}_{i_2} \in \widetilde{\mathcal{V}}(\epsilon_0, h_0), \\ &\text{Case 2: } \widetilde{X}_{i_2} \notin \widetilde{\mathcal{S}}, \text{ and } \mathrm{proj}_{\widetilde{\mathcal{S}}}(\widetilde{X}_{i_2}) \in \widetilde{\mathcal{V}}(\epsilon_0, h_0).\end{aligned} \tag{B.41}$$

Consider Case 1. Since H_1 is a linear projector, $\widetilde{X}_i \in \widetilde{\mathcal{V}}_k(\epsilon_0)$ if and only if $X_i \in \mathcal{V}_k(\epsilon_0)$. Hence,

$$X_{i_2} \in \big(\cup_{k\in\widetilde{\mathcal{K}}(h_0)}\mathcal{V}_k(\epsilon_0)\big).$$

There exists a unique $k_2 \in \widetilde{\mathcal{K}}(h_0)$ such that $X_{i_2} \in \mathcal{V}_{k_2}(\epsilon_0)$. It follows by (B.17) that

$$\|X_{i_2} - v_{k_2}\| \le 2\gamma\epsilon_0, \qquad \text{in Case 1.}$$

Consider Case 2. Write $\tilde{x} = \mathrm{proj}_{\widetilde{\mathcal{S}}}(\widetilde{X}_{i_2})$ for short, and let $M = \{x \in \mathcal{S} : H_1 x = \tilde{x}\}$. For any k, $\tilde{x} \in \widetilde{\mathcal{V}}_k(\epsilon_0)$ implies that $x \in \mathcal{V}_k(\epsilon_0)$ for every $x \in M$. Additionally, $\widetilde{X}_i \in \widetilde{\mathcal{S}}$ if and only if $X_i \in \mathcal{S}$. Hence, it holds in Case 2 that

$$X_{i_2} \notin \mathcal{S}, \text{ and } x \in \big(\cup_{k\in\widetilde{\mathcal{K}}(h_0)}\mathcal{V}_k(\epsilon_0)\big), \text{ for every } x \in M.$$

We pick one $x \in M$. There exists a unique $k_2 \in \widetilde{\mathcal{K}}(h_0)$ such that $x \in \mathcal{V}_{k_2}(\epsilon_0)$. By mimicking the derivation of (B.33), we obtain that

$$\|X_{i_2} - v_{k_2}\| \le 2\gamma\epsilon_0 + \beta, \qquad \text{in Case 2.}$$

Combining the two cases and using the value of ϵ_0 in (B.26), we have the conclusion as

$$\|X_{i_2} - v_{k_2}\| \le \Big(1 + \frac{30\gamma}{\sigma_*}\max\big\{1, \frac{d_{\max}}{\sigma_*}\big\}\Big)\beta, \qquad \text{for some } k_2 \neq k_1. \tag{B.42}$$

Step 3: Analysis of the remaining iterations of SPA.

Fix $3 \le s \le K-1$. We now study the sth iteration. Let $i_1, \ldots, i_K$ denote the sequentially selected indices in SPA. We aim to show that there exist distinct $k_1, k_2, \ldots, k_s \in \{1, 2, \ldots, K\}$ such that

$$\|X_{i_s} - v_{k_s}\| \le \Big(1 + \frac{30\gamma}{\sigma_*}\max\big\{1, \frac{d_{\max}}{\sigma_*}\big\}\Big)\beta. \tag{B.43}$$

Let's denote $\mathcal{M}_{s-1} := \{k_1, \ldots, k_{s-1}\}$ for brevity. Suppose we have already shown (B.43) for every index $1, 2, \ldots, s-1$. Our goal is showing that (B.43) continues to hold for s and some $k_s \notin \mathcal{M}_{s-1}$.

Let $X_i^{(1)} = X_i$ and H_1 be the same as in Step 1 of this proof. We define $X_i^{(s)}$ and H_s recursively to describe the iterations in SPA:

$$\hat{y}_{s-1} = \frac{X_{i_{s-1}}^{(s-1)}}{\|X_{i_{s-1}}^{(s-1)}\|}, \qquad H_s = (I_d - \hat{y}_{s-1}\hat{y}_{s-1})H_{s-1}, \qquad X_i^{(s)} = H_s X_i^{(s-1)}. \tag{B.44}$$

It is seen that $H_{s-1} = \prod_{m=1}^{s-1}(I_d - \hat{y}_m\hat{y}_m')$. Note that each $\hat{y}_m$ is orthogonal to $\hat{y}_1, \ldots, \hat{y}_{m-1}$. As a result, H_{s-1} is a projection matrix with rank $(s-1)$. We apply Lemma B.3 to obtain that

$$\sigma_{K-s}(H_{s-1}V) \ge \sigma_{K-1}(V) \ge \sigma_*, \qquad \text{for } 3 \le s \le K-1. \tag{B.45}$$

Write $V^{(s-1)} = H_{s-1}V$ and $V^{(s)} = H_s V$. Using the notations in (B.44), we have

$$X_i^{(s)} = (I_d - \hat{y}_s\hat{y}_s')X_i^{(s-1)}, \qquad V^{(s)} = (I_d - \hat{y}_s\hat{y}_s')V^{(s-1)}.$$

Here, $\Gamma_s := I_d - \hat{y}_s\hat{y}_s'$ is a projection matrix. We observe:

$$\begin{aligned}&\text{The relationship between } (X_i^{(s-1)}, V^{(s-1)}) \text{ and } (X_i^{(s)}, V^{(s)}) \text{ is similar to the one}\\ &\text{between } (X_i, V) \text{ and } (\widetilde{X}_i, \widetilde{V}) \text{ in Step 2, except that } H_1 \text{ is replaced with } \Gamma_s.\end{aligned} \tag{B.46}$$

We aim to show that (B.35)-(B.38) still hold when those quantities are defined through $(X_i^{(s)}, V^{(s)})$. Recall that the proofs in Step 2 are inductive, where we actually showed that if (B.35)-(B.38) hold for the corresponding quantities defined through (X_i, V), then they also hold for the same quantities defined through $(\widetilde{X}_i, \widetilde{V})$. Given (B.46), the same is true here.

It remains to develop a counterpart of Lemma B.6. The following lemma will be in Section B.4.7. It is also an inductive proof, relying on that (B.43) already holds for $1, 2, \ldots, s-1$. .

Lemma B.6. *Under the conditions of Theorem B.1, write $\sigma_* = \sigma_{K-1}(V)$. Let $\tilde{v}_k = V^{(s)}e_k$, $\tilde{d}_{\max} = \max_k \|\tilde{v}_k\|$, and $\widetilde{\mathcal{K}}(h_0) = \{k : \|\tilde{v}_k\| \ge \tilde{d}_{\max} - h_0\}$. The following claims are true:*

$$\tilde{d}_{\max} \ge \sigma_*/2, \qquad \min_{\substack{\{k,\ell\}\cap\mathcal{M}_{s-1}=\emptyset,\\ k\neq\ell}} \|\tilde{v}_k - \tilde{v}_\ell\| \ge \sqrt{2}\sigma_*, \quad \textit{and} \quad \mathcal{M}_{s-1} \cap \widetilde{\mathcal{K}}(h_0) = \emptyset. \tag{B.47}$$

In Step 2, we have carefully shown how to use (B.35)-(B.39) to get (B.42). Using similar analyses, we can use the counterparts of (B.35)-(B.38), which are defined through $(X_i^{(s)}, V^{(s)})$, and the claim of Lemma B.6, to obtain (B.43). This completes the proof.

B.4 PROOF OF THE SUPPLEMENTARY LEMMAS

B.4.1 PROOF OF LEMMA B.1

By definition, $F(\pi) = \sum_{k=1}^K \pi(k) v_k$. Since $\sum_{k=1}^K \pi(k) = 1$, for any $v_0 \in \mathbb{R}^d$, we can re-express $F(\pi)$ as $F(\pi) = v_0 + \sum_{k=1}^K \pi(k)(v_k - v_0)$. It follows immediately that

$$\|F(\pi) - F(\tilde{\pi})\| = \Big\| \sum_{k=1}^K [\pi(k) - \tilde{\pi}(k)](v_k - v_0) \Big\| \leq \|\pi - \tilde{\pi}\|_1 \cdot \max_k \|v_k - v_0\|.$$

At the same time, since $\mathbf{1}_K'(\pi - \tilde{\pi}) = 0$, the vector $\pi - \tilde{\pi}$ is an $(K-1)$-dimensional linear subspace. It follows by basic properties of singular values that

$$\|F(\pi) - F(\tilde{\pi})\| = \|V(\pi - \tilde{\pi})\| \geq \sigma_{K-1}(V) \cdot \|\pi - \tilde{\pi}\|.$$

Combining the above gives (B.12).

Suppose there are $1 \leq k_1 < k_2 < \ldots < k_s \leq K$ such that $\pi(k_j) = \tilde{\pi}(k_j)$, for $1 \leq j \leq s$. Then, the vector $\delta = \pi - \tilde{\pi}$ satisfies $(s+1)$ constraints: $\mathbf{1}_K'\delta = 0$, $\delta(k_j) = 0$, for $1 \leq j \leq s$. In other words, δ lives in a $(K-1-s)$-dimensional linear space. It follows by properties of singular values that

$$\|F(\pi) - F(\tilde{\pi})\| = \|V(\pi - \tilde{\pi})\| \geq \sigma_{K-1-s}(V) \cdot \|\pi - \tilde{\pi}\|.$$

This proves (B.13).

B.4.2 PROOF OF LEMMA B.2

Write for short $x = \sum_{i=1}^m \pi_i x_i \in \mathbb{R}^d$ and $L = \sum_{i=1}^m w_i \|x_i\|$. By the triangle inequality,

$$\|x\| \leq L.$$

In this lemma, we would like to get a lower bound for $L - \|x\|$. By definition,

$$\|x\|^2 = \sum_i w_i^2 \|x_i\|^2 + \sum_{i \neq j} w_i w_j x_i' x_j. \tag{B.48}$$

For any vectors $u, v \in \mathbb{R}^d$, we have a universal equality: $2u'v = 2\|u\|\|v\| + (\|u\| - \|v\|)^2 - \|u - v\|^2$. By our assumption, $\|x_i - x_j\| \geq a$ and $(\|x_i\| - \|x_j\|)^2 \leq b^2$, for all $i \neq j$. It follows that

$$x_i' x_j \leq \|x_i\|\|x_j\| - (a^2 - b^2)/2, \qquad 1 \leq i \neq j \leq m. \tag{B.49}$$

We plug (B.49) into (B.48) to get

$$\begin{aligned} \|x\|^2 &\leq \sum_i w_i^2 \|x_i\|^2 + \sum_{i \neq j} w_i w_j \|x_i\|\|x_j\| - \frac{1}{2}(a^2 - b^2) \sum_{i \neq j} w_i w_j \\ &= L^2 - \frac{1}{2}(a^2 - b^2) \sum_{i \neq j} w_i w_j. \end{aligned} \tag{B.50}$$

Note that $\sum_{i \neq j} w_i w_j = \sum_i \sum_{j: i \neq j} w_j = \sum_i w_i(1 - w_i)$. Combining it with (B.50) gives

$$\|x\|^2 \leq L^2 - \frac{1}{2}(a^2 - b^2) \sum_i w_i(1 - w_i). \tag{B.51}$$

At the same time, $L + \|x\| \leq 2L$. It follows that

$$L - \|x\| = \frac{L^2 - \|x\|^2}{L + \|x\|} \geq \frac{L^2 - \|x\|^2}{2L} \geq \frac{a^2 - b^2}{4L} \sum_i w_i(1 - w_i). \tag{B.52}$$

This proves the claim.

B.4.3 PROOF OF LEMMA B.3

Since H is a projection matrix, there exists $Q_1 \in \mathbb{R}^s$ and $Q_2 \in \mathbb{R}^{d-s}$ such that $Q = [Q_1, Q_2]$ is an orthogonal matrix, $H = Q_1 Q_1'$, and $I_d - H = Q_2 Q_2'$. It follows that

$$(I_d - H)VV'(I_d - H) = Q_2(Q_2'VV'Q_2)Q_2'.$$

Since Q_2 has orthonormal columns, for any symmetric matrix $M \in \mathbb{R}^{(d-s)\times(d-s)}$, M and $Q_2 M Q_2'$ have the same set of nonzero eigenvalues. Hence,

$$\sigma_{K-1-s}^2((I_d - H)V) = \lambda_{K-1-s}(Q_2'VV'Q_2).$$

We note that $Q_2'VV'Q_2 \in \mathbb{R}^{(d-s)\times(d-s)}$ is a principal submatrix of $Q'VV'Q \in \mathbb{R}^{d\times d}$. Using the eigenvalue interlacing theorem (Horn & Johnson, 1985, Theorem 4.3.28),

$$\lambda_{K-1-s}(Q_2'VV'Q_2) \geq \lambda_{K-1}(Q'VV'Q).$$

The claim follows immediately by noting that $\lambda_{K-1}(Q'VV'Q) = \lambda_{K-1}(VV') = \sigma_{K-1}^2(V)$.

B.4.4 PROOF OF LEMMA B.4

Write $\ell_{\max} = \max_{1\leq k\leq K} \|v_k\|$. We target to show

$$\ell_{\max}^2 \geq \frac{K-s-1}{2(K-s)}\sigma_*^2, \qquad \text{with } \sigma_* := \sigma_{K-1-s}(V). \tag{B.53}$$

The right hand side of (B.53) is minimized at $s = K-2$, at which $\ell_{\max}^2 \geq \sigma_*^2/4$. We now show (B.53). When $s = 0$, it is seen that

$$K\ell_{\max}^2 \geq \sum_k \|v_k\|^2 = \mathrm{trace}(V'V) \geq (K-1)\sigma_{K-1}^2(V).$$

Therefore, $\ell_{\max}^2 \geq \frac{K-1}{K}\sigma_*^2$, which implies (B.16) for $s = 0$. When $1 \leq s \leq K-2$, since $\|v_k\| \leq \delta$ for at least s of the vertices,

$$s\delta^2 + (K-s)\ell_{\max}^2 \geq \sum_k \|v_k\|^2 = \mathrm{trace}(V'V) \geq (K-1-s)\sigma_{K-1-s}^2(V).$$

As a result, for $\sigma_* = \sigma_{K-1-s}(V)$,

$$\ell_{\max}^2 \geq \frac{(K-s-1)\sigma_*^2 - s\delta^2}{K-s}. \tag{B.54}$$

Note that $\frac{s}{K-s-1}$ is a monotone increasing function of s. Hence, $\frac{s}{K-s-1} \leq K-2$. The assumption of $2(K-2)\delta^2 \leq \sigma_*^2$ implies that $\frac{2s}{K-s-1}\delta^2 \leq \sigma_*^2$, or equivalently, $s\delta^2 \leq \frac{K-s-1}{2}\sigma_*^2$. We plug it into (B.54) to get $\ell_{\max}^2 \geq \frac{K-s-1}{2(K-s)}\sigma_*^2$. This proves (B.16) for $1 \leq s \leq K-2$.

B.4.5 PROOF OF LEMMA B.5

Write $\mathcal{K} = \mathcal{K}(h_0)$, $\mathcal{V}_k = \mathcal{V}_k(\epsilon_0)$, and $\mathcal{V} = \mathcal{V}(\epsilon_0, h_0)$ for short. By definition of $\mathcal{K}$,

$$d_{\max} - h_0 \leq \|v_k\| \leq d_{\max}, \text{ for } k \in \mathcal{K}, \quad \|v_k\| \leq d_{\max} - h_0, \text{ for } k \notin \mathcal{K}. \tag{B.55}$$

We shall fix a point $x \in \mathcal{S} \setminus \mathcal{V}$ and derive an upper bound for $\|x\|$.

First, we need some preparation, let F be the mapping in Lemma B.1. It follows that $\pi = F^{-1}(x)$ is the barycentric coordinate of x in the simplex. By definition of $\mathcal{V}$,

$$\max_{k\in\mathcal{K}} \pi(k) \leq 1 - \epsilon_0, \qquad \text{whenever } x := F(\pi) \text{ is in } \mathcal{S} \setminus \mathcal{V}. \tag{B.56}$$

The K vertices are naturally divided into two groups: those in $\mathcal{K}$ and those not in $\mathcal{K}$. Define

$$\rho := \sum_{k\in\mathcal{K}} \pi(k), \qquad \eta := \begin{cases} \rho^{-1}\sum_{k\in\mathcal{K}} \pi(k)v_k, & \text{if } \rho \neq 0, \\ \mathbf{0}_d, & \text{otherwise.} \end{cases} \tag{B.57}$$

Here, ρ is the total weight π puts on those vertices in $\mathcal{K}$, and we can re-write x as

$$x = \rho\eta + \sum_{k\notin\mathcal{K}} \pi(k) v_k.$$

By the triangle inequality,

$$\begin{aligned}\|x\| = \Big\|\rho\eta + \sum_{k\notin\mathcal{K}} \pi(k) v_k\Big\| &\le \rho\|\eta\| + \sum_{k\notin\mathcal{K}} \pi(k)\|v_k\| \\ &\le \rho\|\eta\| + (1-\rho)(d_{\max} - h_0). \end{aligned} \tag{B.58}$$

Next, we proceed with showing the claim. We consider two cases:

$$1-\rho \ge \epsilon_0/2 \ \text{ (Case 1)}, \qquad \text{and} \qquad 1-\rho < \epsilon_0/2 \ \text{ (Case 2)}.$$

In Case 1, the total weight that π_i puts on those vertices not in $\mathcal{K}$ is at least $\epsilon_0/2$. Since each vertex satisfies that $\|v_k\| \le d_{\max} - h_0$ (see (B.56)) and $\|\eta\| \le d_{\max}$, it follows from (B.58) that

$$\|x\| \le d_{\max} - (1-\rho)h_0 \le d_{\max} - \frac{h_0\epsilon_0}{2}, \qquad \text{in Case 1.} \tag{B.59}$$

In Case 2, if $\mathcal{K} = \{k^*\}$ is a singleton, then $\rho = \pi(k^*)$. By (B.56), $\pi(k^*) \le 1-\epsilon_0$, which leads to $1-\rho = 1-\pi(k^*) \ge \epsilon_0$. This yields a contradiction to $1-\rho < \epsilon_0/2$. Hence, it must hold that

$$|\mathcal{K}| \ge 2. \tag{B.60}$$

Now, η is a convex combination of more than one point in $\{v_k : k \in \mathcal{K}\}$, for which we hope to apply Lemma B.2. By (B.55), for each $k \in \mathcal{K}$, $\|v_k\|$ is in the interval $[d_{\max} - h_0, d_{\max}]$. Hence, we can take $b = h_0$ in Lemma B.2. In addition, from the assumption (B.22), $\|v_k - v_\ell\| \ge \sqrt{2}\sigma_*$ for any $k \ne \ell$. Hence, we set $a = \sqrt{2}\sigma_*$ in Lemma B.2. We apply this lemma to the vector η in (B.57). It yields

$$\|\eta\| \le L - \frac{(2\sigma_*^2 - h_0^2)}{4L} \sum_{k\in\mathcal{K}} \frac{\pi(k)[\rho - \pi(k)]}{\rho^2}, \qquad \text{with} \quad L := \sum_{k\in\mathcal{K}} \frac{\pi(k)}{\rho}\|v_k\|. \tag{B.61}$$

Since $L \le d_{\max}$, it follows from (B.61) that

$$\|\eta\| \le d_{\max} - \frac{2\sigma_*^2 - h_0^2}{4\rho d_{\max}} \sum_{k\in\mathcal{K}} \pi(k)[1 - \rho^{-1}\pi(k)].$$

Additionally, noticing that $\pi(k) \le 1-\epsilon_0$ for each $k \in \mathcal{K}$, we have the following inequality:

$$1 - \rho^{-1}\pi(k) = \rho^{-1}[1-\pi(k)] - \rho^{-1}(1-\rho) \ge \rho^{-1}[\epsilon_0 - (1-\rho)].$$

Combining these arguments and using the fact that $\sum_{k\in\mathcal{K}} \pi(k) = \rho$, we have

$$\begin{aligned}\|\eta\| &\le d_{\max} - \frac{(2\sigma_*^2 - h_0^2)[\epsilon_0 - (1-\rho)]}{4\rho^2 d_{\max}} \sum_{k\in\mathcal{K}} \pi(k) \\ &\le d_{\max} - \frac{(2\sigma_*^2 - h_0^2)[\epsilon_0 - (1-\rho)]}{4\rho d_{\max}}. \end{aligned} \tag{B.62}$$

Since $1-\rho \le \epsilon_0/2$, we immediately have $\|\eta\| \le d_{\max} - \frac{2\sigma_*^2 - h_0^2}{8\rho d_{\max}}$. We plug it into (B.58) to get

$$\begin{aligned}\|x\| &\le \rho\Big(d_{\max} - \frac{2\sigma_*^2 - h_0^2}{8\rho d_{\max}}\Big) + (1-\rho)(d_{\max} - h_0) \\ &\le \rho\Big(d_{\max} - \frac{2\sigma_*^2 - h_0^2}{8\rho d_{\max}}\Big) + (1-\rho) d_{\max} \\ &\le d_{\max} - \frac{(2\sigma_*^2 - h_0^2)\epsilon_0}{8 d_{\max}}, \qquad \text{in Case 2.} \end{aligned} \tag{B.63}$$

We now combine (B.59) for Case 1 and (B.63) for Case 2. By setting $h_0 = \sigma_*/3$, we have a unified expression:

$$\|x\| \le d_{\max} - \min\Big\{\frac{\sigma_*}{6},\ \frac{2\sigma_*^2}{9 d_{\max}}\Big\}\epsilon_0.$$

Consequently, a sufficient condition for $\|x\| \le d_{\max} - t$ to hold is

$$\min\Big\{\frac{\sigma_*}{6},\ \frac{\sigma_*^2}{6 d_{\max}}\Big\}\epsilon_0 \le t \qquad \Longleftrightarrow \qquad \epsilon_0 \ge \frac{6}{\sigma^*} \max\Big\{1,\ \frac{d_{\max}}{\sigma_*}\Big\} t.$$

This proves the claim.

B.4.6 PROOF OF LEMMA B.6

Without loss of generality, we assume $k_1 = 1$.

By definition, $\widetilde{V} = H_1 V$, where H_1 is a rank-1 projection matrix. It follows by Lemma B.3 that

$$\sigma_{K-2}(\widetilde{V}) \geq \sigma_{K-1}(V) = \sigma_*. \tag{B.64}$$

Note that $\tilde{d}_{\max} \geq \max_{k\neq 1} \|\tilde{v}_k\|$ and $\|\tilde{v}_1\| = 0$. We apply Lemma B.4 with $s = 1$ and $\delta = 0$ to get

$$\tilde{d}_{\max} \geq \frac{1}{2}\sigma_{K-2}(\widetilde{V}) \geq \frac{1}{2}\sigma_*.$$

This proves the first claim in (B.39). Note that $\tilde{v}_k = \widetilde{V}e_k$, where $e_k \in \mathbb{R}^K$ is a standard basis vector. For any $2 \leq k \neq \ell \leq K$, e_k and e_ℓ both have a zero at the first coordinate; and we apply Lemma B.1 with $s = 1$ to get

$$\|v_k - v_\ell\| \geq \sigma_{K-2}(\widetilde{V})\|e_k - e_\ell\| \geq \sqrt{2}\sigma_*.$$

This proves the second claim in (B.39).

Finally, we show the third claim. Note that

$$\tilde{v}_1 = H_1 v_1 = v_1 - \frac{v_1' X_{i_1}}{\|X_{i_1}\|^2} X_{i_1} = \frac{X_{i_1}'(X_{i_1} - v_1)}{\|X_{i_1}\|^2} v_1 - \frac{v_1' X_{i_1}}{\|X_{i_1}\|^2}(X_{i_1} - v_1). \tag{B.65}$$

Here, $\|v_1\| \leq d_{\max}$, and by (B.28), $\|X_{i_1}\| \geq d_{\max} - \beta$. Since $|X_{i_1}'(X_{i_1} - v_1)| \leq \|X_{i_1}\| \cdot \|X_{i_1} - v_1\|$, we have

$$\frac{|X_{i_1}'(X_{i_1} - v_1)|}{\|X_{i_1}\|^2}\|v_1\| \leq \frac{\|v_1\|}{\|X_{i_1}\|}\|X_{i_1} - v_1\| \leq \frac{d_{\max}}{d_{\max} - \beta}\|X_{i_1} - v_1\|,$$

and

$$\frac{v_1' X_{i_1}}{\|X_{i_1}\|^2} \leq \frac{\|v_1\|}{\|X_{i_1}\|} \leq \frac{d_{\max}}{d_{\max} - \beta}.$$

Plugging these inequalities into (B.65) and applying (B.34), we obtain:

$$\begin{aligned}\|\tilde{v}_1\| &\leq \frac{2d_{\max}}{d_{\max} - \beta}\|X_{i_1} - r_{i_1}\| \\ &\leq \frac{2d_{\max}}{d_{\max} - \beta}\Big(\beta + \frac{30\gamma}{\sigma_*}\max\big\{1, \frac{d_{\max}}{\sigma_*}\big\}\beta\Big).\end{aligned} \tag{B.66}$$

By our assumption, $\frac{30 d_{\max}}{\sigma_*}\max\big\{1, \frac{d_{\max}}{\sigma_*}\big\}\beta \leq \sigma_*/15$. Moreover, we have shown $d_{\max} \geq \tilde{d}_{\max} \geq \sigma_*/2$. It further implies $\beta \leq \frac{\sigma_*^2}{450 d_{\max}} \leq \frac{1}{225}\sigma_* \leq \frac{1}{100}\tilde{d}_{\max}$. As a result,

$$\|\tilde{v}_1\| \leq \frac{200}{99}(\beta + \frac{\sigma_*}{15}) \leq \frac{3}{10}\tilde{d}_{\max} \leq \tilde{d}_{\max} - \frac{7}{20}\sigma_*. \tag{B.67}$$

At the same time, $h_0 = \sigma_*/3$. Hence,

$$\|\tilde{v}_1\| < \widetilde{d}_{\max} - h_0 \qquad \Longrightarrow \qquad 1 \notin \widetilde{\mathcal{K}}(h_0).$$

This proves the third claim in (B.39).

B.4.7 PROOF OF LEMMA B.6

Suppose we have already obtained (B.47) and (B.43) for each $1 \leq j \leq s-1$, and we would like to show (B.47) for s.

First, consider the second claim in (B.47). For each $k \notin \mathcal{M}_{s-1}$, it has $(s-1)$ zeros in its barycentric coordinate (corresponding to those indices in $\mathcal{M}_{s-1}$). We apply Lemma B.1 to obtain:

$$\|\tilde{v}_k - \tilde{v}_\ell\| \geq \sqrt{2}\sigma_{K-s}(\widetilde{V}) \geq \sqrt{2}\sigma_*, \qquad \text{for all } k \neq \ell \text{ in } \{1, \ldots, K\} \setminus \mathcal{M}_{s-1},$$

where the first inequality is from (B.13) and the second inequality is from (B.45).

Next, consider the third claim in (B.47). Note that $\mathcal{M}_{s-1} = \{k_1, k_2, \ldots, k_{s-1}\}$. For each $1 \le j \le s-1$, by definition, $\tilde{v}_{k_j} = \big[\prod_{m \ge j}(I_d - \hat{y}_m \hat{y}_m')\big] \cdot (I_d - \hat{y}_j \hat{y}_j) H_{j-1} v_{k_j}$. It follows that

$$\|\tilde{v}_{k_j}\| \le \|(I_d - \hat{y}_j \hat{y}_j) H_{j-1} v_{k_j}\|, \qquad \text{where} \quad \hat{y}_j = \frac{H_{j-1} X_{i_j}}{\|H_{j-1} X_{i_j}\|}. \tag{B.68}$$

Here, $\|H_{j-1} X_{i_j}\|$ is the maximum Euclidean distance attained in the $(j-1)$th iteration. Since we have already established (B.47) for j, we immediately have

$$\|H_{j-1} X_{i_j}\| \ge \sigma_*/2, \qquad \text{for } 1 \le j \le s-1.$$

In addition, we have shown (B.42) for $1 \le j \le s-1$, which implies that

$$\|H_{j-1} X_{i_j} - H_{j-1} v_{k_j}\| \le \Big(1 + \frac{30\gamma}{\sigma_*} \max\big\{1, \frac{d_{\max}}{\sigma_*}\big\}\Big)\beta.$$

Using the above ineqaulities, we can mimic the proof of (B.66) to show that

$$\|(I_d - \hat{y}_j \hat{y}_j) H_{j-1} v_{k_j}\| \le \Big(1 + \frac{30\gamma}{\sigma_*} \max\big\{1, \frac{d_{\max}}{\sigma_*}\big\}\Big)\beta. \tag{B.69}$$

Write $\Gamma_j = I_d - \hat{y}_j \hat{y}_j'$. It is seen that

$$\|\tilde{v}_{k_j}\| = \Big\| \prod_{\ell=j+1}^{s} \Gamma_j H_{j-1} v_{k_j} \Big\| \le \|\Gamma_j H_{j-1} v_{k_j}\| \le \|(I_d - \hat{y}_j \hat{y}_j) H_{j-1} v_{k_j}\|.$$

Therefore, for $1 \le j \le s-1$,

$$\|\tilde{v}_{k_j}\| \le \Big(1 + \frac{30\gamma}{\sigma_*} \max\big\{1, \frac{d_{\max}}{\sigma_*}\big\}\Big)\beta. \tag{B.70}$$

We further mimic the argument in (B.67) to obtain:

$$\|\tilde{v}_{k_j}\| \le \tilde{\beta}_{\max} - 7\sigma_*/20 < \tilde{\beta} - h_0, \qquad \text{for all } 1 \le j \le s-1.$$

This implies that

$$k_j \notin \widetilde{\mathcal{K}}(h_0) \ \text{ for } 1 \le j \le s-1 \quad \Longrightarrow \quad \mathcal{M}_{s-1} \cap \widetilde{\mathcal{K}}(h_0) = \emptyset. \tag{B.71}$$

Last, consider the first claim in (B.47). Let Δ denote the right hand side of (B.70) for brevity. We have shown $\|\tilde{v}_k\| \le \Delta$, for all $k \in \mathcal{M}_{s-1}$. By our assumption, we can easily conclude that $\sigma_*^2 \ge 2(K-2)\Delta$. We then apply Lemma B.4 with $s-1$ and $\delta = \Delta$ to get

$$\tilde{d}_{\max} \ge \frac{1}{2}\sigma_{K-s}(\widetilde{V}) \ge \sigma_*/2, \tag{B.72}$$

where the last inequality is from (B.45).

C PROOF OF THE MAIN THEOREMS

We recall our pp-SPA procedure. On the hyperplane, we obtained the projected points

$$\tilde{X}_i := H(X_i - \bar{X}) + \bar{X} = (I_d - H)\bar{X} + H r_i + H\epsilon_i$$

after rotation by U, they become $Y_i = U'\tilde{X}_i = U' r_i + U'\epsilon_i = U' X_i \in \mathbb{R}^{K-1}$. Denote $\tilde{Y}_i = U_0' X_i = U_0' r_i + U_0'\epsilon_i \in \mathbb{R}^{K-1}$. In particular, $U_0'\epsilon_i \sim N(0, \sigma^2 I_{K-1})$. Then, without loss of generality, the vertex hunting analysis on $\tilde{Y}_i$ is equivalent to that of $X_i = r_i + \epsilon_i \in \mathbb{R}^p$, where $\epsilon_i \sim N(0, \sigma^2 I_p)$ with $p = K-1$. We provide the following theorems for the rate by applying D-SPA on the aforementioned low dimension $p = K-1$ space. The proof of these two theorems are postponed to Section C.2.

Theorem C.1. *Consider $X_i = r_i + \epsilon_i \in \mathbb{R}^p$, where $\epsilon_i \sim N(0, \sigma^2 I_p)$ for $1 \le i \le n$. Suppose $m \ge c_1 n$ for a constant $c_1 > 0$ and $p \ll \log(n)/\log\log(n)$. Let $p/\log(n) \ll \delta_n \ll 1$. Let $c_2^* = 0.9(2e^2)^{-1/p}\sqrt{(2/p)}(\Gamma(p/2+1))^{1/p}$. Then, $c_2^* \to 0.9e^{-1/2}$ as $p \to \infty$. . We apply D-SPA to $X_1, X_2, \ldots, X_n$ and output $X_1^*, \cdots, X_n^*$ where some X_i^* may be NA owing to the pruning. If we choose $N = \log(n)$ and*

$$\Delta = c_3\sigma\sqrt{p}\Big(\frac{\log(n)}{n^{1-\delta_n}}\Big)^{1/p} \text{ for a constant } c_3 \le c_2^*,$$

Then,

$$\beta_{new}(X^*) \le \sqrt{\delta_n} \cdot \sigma \cdot \sqrt{2\log(n)}$$

If the last inequality of (8) and (9) hold, then up to a permutation in the columns,

$$\max_{1\le k\le K} \|\hat{v}_k - v_k\| \le g_{new}(V) \cdot \sqrt{\delta_n} \cdot \sigma \cdot \sqrt{2\log(n)}.$$

The second theorem discuss the case there a fewer pure nodes.

Theorem C.2. *Consider $X_i = r_i + \epsilon_i \in \mathbb{R}^p$, where $\epsilon_i \sim N(0, \sigma^2 I_p)$ for $1 \le i \le n$. Fix $0 < c_0 < 1$ and assume that $m \ge n^{1-c_0+\delta}$ for a sufficiently small constant $0 < \delta < c_0$. Suppose $p \ll \log(n)/\log\log(n)$. Let $c_2^* = 0.9(2e^{2-c_0})^{-1/p}\sqrt{(2/p)}(\Gamma(p/2+1))^{1/p}$. Then $c_2^* \to 0.9e^{-1/2}$ as $p \to \infty$. Suppose we apply D-SPA to $X_1, X_2, \ldots, X_n$ and output $X_1^*, \cdots, X_n^*$ where some X_i^* may be NA owing to the pruning. If we choose $N = \log(n)$ and*

$$\Delta = c_3\sigma\sqrt{p}\Big(\frac{\log(n)}{n^{1-c_0}}\Big)^{1/p} \text{ for a constant } c_3 \le c_2^*.$$

Then,

$$\beta_{new}(X^*) \le \sqrt{c_0} \cdot \sigma \cdot \sqrt{2\log(n)}$$

If the last inequality of (8) and (9) hold, then up to a permutation in the columns,

$$\max_{1\le k\le K} \|\hat{v}_k - v_k\| \le g_{new}(V) \cdot \sqrt{c_0} \cdot \sigma\sqrt{2\log(n)}.$$

for any arbitrary small constant $\delta < 0$.

Based on the above two theorem, we have the results on $\{\tilde{Y}_i\}'s$. However, what we really care about is on $\{Y_i\}'s$ which differ from $\{\tilde{Y}_i\}'s$ by the rotation matrix. To bridge the gap, we need the following Lemma.

Lemma C.7. *Suppose that $s_{K-1}^2(R) \gg \max\{\sqrt{\sigma^2 d/n}, \sigma^2 d/n\}$ and $\sigma = O(1)$. Then, with probability $1 - o(1)$,*

$$\|U - U_0\| \asymp \|H - H_0\| \le \frac{C}{s_{K-1}^2(R)} \max\{\sqrt{\sigma^2 d/n}, \sigma^2 d/n\} \tag{C.73}$$

C.1 Proof of Theorems 2 and 3

With the help of Theorems C.1, C.2 and Lemma C.7, we now prove Theorems 2 and 3. We will present the detailed proof for Theorem 3. The proof of Theorem 2 is nearly identical to that of Theorem 3 with the only difference in employing Theorem C.1, and we refrain ourselves from repeated details.

Proof of Theorem 3. Recall that $Y_i = U'X_i = U'r_i + U'\epsilon_i$ and $\tilde{Y}_i = U_0'r_i + U_0'\epsilon_i$. Theorem C.2 indicates that applying D-SPA on $\bar{Y}_i$ improves the rate to $\sigma(1+o(1))\sqrt{2c_0\log(n)}$. Note that $\|r_i\| \le 1$. Also, by Lemma 5, $\|\epsilon_i\| \le (1+o(1))\sigma(\sqrt{\max\{d, 2\log(n)\}})$ simultaneously for all i, with high probability. Under the assumption $\alpha_n = o(1)$ for both cases and $s_{K-1}^2(R) \asymp s_{K-1}^2(\tilde{V})$ by Lemma 4, the first condition in Lemma C.7 is valid. By the last inequality in (8), we have the norm of r_i should be upper bounded for all $1 \le i \le n$ and therefore $s_{K-1}(\tilde{V}) \le C\max_{k\ne l}\|\tilde{v}_k - \tilde{v}_\ell\| \le$

C. Further with the condition (9), we obtain that $\sigma = O(1)$. Therefore, the conditions in Lemma C.7 are both valid. Then by employing Lemma C.7, we can derive that

$$\|Y_i - \tilde{Y}_i\| = O_{\mathbb{P}}\left(\frac{\sigma\sqrt{d}}{\sqrt{n}s^2_{K-1}(R)}(1 + \sigma\sqrt{\max\{d, 2\log(n)\}}\,)\right) = O_{\mathbb{P}}(\sigma\alpha_n)$$

where the last step is due to Lemma 4 under the condition (8).

Consider the first case that $\alpha_n \ll t_n^*$. We choose $\Delta = c_3 t_n^* \sigma$. It is seen that $\sigma\alpha_n \ll \Delta$. We will prove by contradiction that applying pp-SPA with $(\Delta, \log(n))$ on $\{Y_i\}$, the denoise step can remove outlying points whose distance to the underlying simplex larger than $\sigma[\sqrt{2c_0\log(n)} + C\alpha_n]$ for some $C > 0$.

First, suppose that with probability c for a small constant $c > 0$, there is one point Y_{i_0} away from the underlying simplex by a distance larger than $\sigma[\sqrt{2c_0\log(n)} + C\alpha_n]$ and it is not pruned out. Since $\sigma\alpha_n \ll \Delta$, we see that $\tilde{Y}_{i_0}$ is faraway to the simplex with distance $\sigma\sqrt{2c_0\log(n)}$ for certain large C and it cannot be pruned out by $(1.5\Delta, \log(n))$. Otherwise if it can be pruned out, $\mathcal{B}(Y_{i_0}, \Delta) \subset \mathcal{B}(\tilde{Y}_{i_0}, 1.5\Delta)$ and hence $N(\mathcal{B}(Y_{i_0}, \Delta)) \geq \log(n)$, which means that we can prune out Y_{i_0} with $(\Delta, \log(n))$. This is a contradiction. However, by employing Theorem C.2 on $\{\tilde{Y}_i\}$ with $p = K - 1$ and noticing $c_2^* = 1.8c_2$ with c_2 defined in the manuscript, we should be able to prune out $\tilde{Y}_{i_0}$ with high proability. This leads to a contradiction.

Second, suppose that with probability c for a small constant $c > 0$, all outliers can be removed but a vertex v_1 is also removed (which means all points near it are removed). Then, $N(\mathcal{B}(v_1, \Delta)) < \log(n)$. For the corresponding vertex for $\{\tilde{Y}_i\}$, denoted by $\tilde{v}_1$, it holds that $N(\mathcal{B}(\tilde{v}_i, \Delta/2)) < \log(n)$ which means the vertex $\tilde{v}_1$ for $\{\tilde{Y}_i\}$ is also pruned. However, again by Theorem C.2, this can only happen with probability $o(1)$. This leads to another contradiction.

Let us denote by $\beta(Y^*, U_0'V)$ the maximal distance of points in Y^* to the simplex formed by $U_0'V$. By the above two contradictions, we conclude that with high probability,

$$\beta(Y^*, U_0'V) \leq \sigma[\sqrt{2c_0\log(n)} + C\alpha_n].$$

where $U_0'V$ is the underlying simplex of $\{\tilde{Y}_i\}$. It is worth noting that $\alpha_n = o(1)$. Then, under the assumptions of the theorem, we can apply Theorem B.1 (Theorem 1 in the manuscript). It gives that

$$\max_{1\leq k\leq K} \|\hat{v}_k^* - U_0'v_k\| \leq \sigma g_{new}(V)[\sqrt{2c_0\log(n)} + C\alpha_n]$$

where we use $(\hat{v}_1^*, \cdots, \hat{v}_K^*)$ to denote the output vertices by applying SP on $\{Y_i\}$. Eventually, we output each vertex $\hat{v}_k = (I_K - UU')\bar{X} + U\hat{v}_k^*$. It follows that up to a permutation of the K vectors,

$$\begin{aligned}\max_{1\leq k\leq K} \|\hat{v}_k - v_k\| &\leq \max_{1\leq k\leq K} \|U\hat{v}_k^* - v_k\| + \|(I_d - UU')\bar{X} - (I_d - U_0U_0')\bar{r}\| \\ &\leq \max_{1\leq k\leq K} \|\hat{v}_k^* - U_0'v_k\| + \|U - U_0\| + \|(I_d - UU')\bar{X} - (I_d - U_0U_0')\bar{r}\|\end{aligned}$$

Further we can derive

$$\begin{aligned}\|(I_d - UU')\bar{X} - (I_d - U_0U_0')\bar{r}\| &\leq \|H - H_0\| + \|\bar{X} - \bar{r}\| \\ &\leq \sigma\alpha_n + \|\bar{\epsilon}\| \\ &\leq \sigma\alpha_n + \frac{2\sigma\sqrt{\max\{d, 2\log(n)\}}}{\sqrt{n}}\end{aligned}$$

this together with Lemma C.7, give rise to

$$\max_{1\leq k\leq K} \|\hat{v}_k - v_k\| \leq \sigma g_{new}(V)[\sqrt{2c_0\log(n)} + C\alpha_n] + \frac{2\sigma\sqrt{\max\{d, 2\log(n)\}}}{\sqrt{n}}\,.$$

Consider the second case that $\alpha_n \gg t_n^*$ where we choose $\Delta = \sigma\alpha_n$. By Lemma 5, it is observed that with high probability, $\max_{1\leq i\leq n} d(\tilde{Y}_i, \mathcal{S}) < (1 + o(1))\sigma\sqrt{2\log(n)}$. Notice that $\|Y_i - \tilde{Y}_i\| \leq C\sigma\alpha_n$ with high probability. For $\tilde{Y}_i$, if its distance to the underlying simplex is larger than $\sigma[(1 +$

$o(1))\sqrt{2\log(n)}+C_1\alpha_n]$ for a sufficiently large $C_1 > 3C+1$, then $d(\tilde{Y}_i, \mathcal{S}) \geq d(Y_i, \mathcal{S}) - C\sigma\alpha_n > \sigma[(1+o(1))\sqrt{2\log(n)} + (2C+1)\alpha_n]$. Hence, $\mathbb{B}(\tilde{Y}_i, (2C+1)\Delta))$ is away from the simplex by a distance larger than $\sigma(1+o(1))\sqrt{2\log(n)}$. It follows that $N(\mathbb{B}(Y_i, \Delta)) \leq N(\mathbb{B}(\tilde{Y}_i, (2C+1)\Delta)) < \log(n)$. This is equivalent to say that we prune out the points there. Consequently, with high probability,

$$\beta(Y^*, U_0'V) \leq \sigma[(1+o_{\mathbb{P}}(1))\sqrt{2\log(n)} + C_1\alpha_n]$$

and further by Theorem B.1 (Theorem 1 in the manuscript),

$$\max_{1\leq k\leq K} \|\hat{v}_k^* - U_0'v_k\| \leq \sigma g_{new}(V)[\sqrt{2\log(n)} + C_1\alpha_n]$$

Next, replicate the proof for $\max_{1\leq k\leq K}\|\hat{v}_k - v_k\|$ in the former case, we can conclude that

$$\begin{aligned}\max_{1\leq k\leq K} \|\hat{v}_k - v_k\| &\leq \sigma g_{new}(V)[(1+o_{\mathbb{P}}(1))\sqrt{2\log(n)} + C_1\alpha_n] + \frac{2\sigma\sqrt{\max\{d, 2\log(n)\}}}{\sqrt{n}} \\ &= \sigma g_{new}(V)(1+o_{\mathbb{P}}(1))\sqrt{2\log(n)}.\end{aligned}$$

This concludes our proof.

□

C.2 Proof of Theorems C.1 and C.2.

In the subsection, we provide the proofs of Theorems C.1 and C.2. We show the proof of Theorem C.2 in detail and briefly present the proof of Theorems C.1 as it is similar to that of Theorem C.2.

Proof of Theorem C.2. We first claim the limit of $c_2^* = 0.9(2e^{2-c_0})^{-1/p}\sqrt{(2/p)}(\Gamma(p/2+1))^{1/p}$. Note that $\Gamma(p/2+1) = (p/2)!$ if p is even and $\Gamma(p/2+1) = \sqrt{\pi}(p+1)!/(2^{p+1}(\frac{p+1}{2})!)$ if p is odd. Using Stirling's approximation, it is elementary to deduce that

$$c_2^* = e^{O(1/p)-(1-\log(p+1))(p+1)/2p-\log(p)/2} \to e^{-1/2}.$$

Define the radius $\Delta \equiv \Delta_n = c_3\sigma\sqrt{p}\Big(\frac{\log(n)}{n^{1-c_0}}\Big)^{1/p}$ for a constant $c_3 \leq c_2$. In the sequel, we will prove that applying D-SPA to $X_1, \cdots, X_n$ with (Δ, N), we can prune out the points whose distance to the underlying true simplex are larger than the rate in the theorem, while the points around vertices are captured.

Denote $d(x, \mathcal{S})$, the distance of x to the simplex $\mathcal{S}$. Let

$$\mathcal{R}_f := \{x \in \mathbb{R}^p : d(x, \mathcal{S}) \geq 2\sigma\sqrt{\log(n)}\,\}$$

We first claim that the number of points in $\mathcal{R}_f$, denoted by $N(\mathcal{R}_f)$, is bounded with probability $1-o(1)$. By definition, we deduce

$$N(\mathcal{R}_f) = \sum_{i=1}^n \mathbf{1}(x_i \in \mathcal{R}_f) \leq \sum_{i=1}^n \mathbf{1}(\|\varepsilon_i\| \geq 2\sigma\sqrt{\log n}\,)$$

The mean on the RHS is given by $n\mathbb{P}(\|\varepsilon_i\| \geq 2\sigma\sqrt{\log n}) = n\mathbb{P}(\chi_p^2 \geq 4\log n) \leq ne^{-1.5\log(n)} = n^{-1/2}$. By similar computations, the order of the variance is again $n^{-1/2}$. By Chebyshev's inequality, we conclude that $N(\mathcal{R}_f) = o_{\mathbb{P}}(1)$.

In the sequel, we use the notation $\mathbb{B}(x, r)$ to represent a ball centered at x with radius r and denote $N(\mathbb{B}(x, r))$ the number of points falling into this ball. And we also denote $\mathcal{S}$ the true underlying simplex.

Based on these notation, we introduce

$$P := \mathbb{P}(\exists\ X_i \text{ satisfying } \sigma\sqrt{2c_0\log(n)} \leq d(X_i, \mathcal{S}) \leq 2\sigma\sqrt{\log(n)} \text{ cannot be pruned out })$$

We aim to show that $P = o(1)$. To see this, we first derive

$$\begin{aligned}
P &= \binom{n}{N} N \cdot \mathbb{P}(X_1, \cdots X_N \in B(X_1, \Delta) \text{ s.t. } \sigma\sqrt{2c_0 \log(n)} \le d(X_1, \mathcal{S}) \le 2\sigma\sqrt{\log(n)}\,) \\
&\le \binom{n}{N} N \cdot \int_{a_n \le d(x,\mathcal{S}) \le b_n} f_{X_1}(x) \mathbb{P}(X_2, \cdots, X_N \in \mathcal{B}(x, \Delta)) \mathrm{d}x \\
&\le \binom{n}{N} N \cdot \int_{a_n \le d(x,\mathcal{S}) \le b_n} f_{X_1}(x) \prod_{t=2}^{N} \mathbb{P}(X_t \in \mathcal{B}(x, \Delta)) \mathrm{d}x
\end{aligned}$$

where $a_n := \sigma\sqrt{2c_0 \log(n)}$ and $b_n := 2\sigma\sqrt{\log(n)}$ for simplicity. We can compute that for any $2 \le t \le N$,

$$\begin{aligned}
\mathbb{P}(X_t \in \mathcal{B}(x, \Delta)) &= (2\pi\sigma^2)^{-\frac{p}{2}} \int_{\|y-x\| \le \Delta} \exp\{-\|y - r_t\|^2/2\sigma^2\} \mathrm{d}y \\
&\le \frac{(\Delta/\sigma)^p}{2^{p/2}\Gamma(p/2+1)} \exp\Big\{ - \frac{(\|x - r_t\| - \Delta)^2}{2\sigma^2} \Big\} \\
&\le (\Delta/\sigma)^p C_p \exp\Big\{ - \frac{\|x - r_t\|^2}{2(1+\tau_n)\sigma^2} \Big\} \qquad \text{(C.74)}
\end{aligned}$$

where $\tau_n := C\Delta/\sigma\sqrt{2c_0 \log(n)}$ for a large $C > 0$;and we write $C_p := 2^{1-p/2}/\Gamma(p/2+1)$. Here to obtain the last inequality, we used the definition of Δ and the derivation

$$\frac{\Delta}{\|x - r_t\|} \le \frac{\Delta}{\sigma\sqrt{2c_0 \log(n)}} \le C\tau_n \le C\sqrt{p}(\log(n))^{1/p - 1/2}/n^{(1-c_0)/p} = o(1)$$

so that

$$(1 - \Delta/\|x - r_t\|)^2 \le (1 + \tau_n)^{-1}$$

by choosing appropriate C in the definition of τ_n. Further, under the condition that $p \ll \log(n)/\log\log(n)$, one can verify that

$$\tau_n \ll 1/\log(n) = o(1)\,.$$

(C.74), together with

$$f_{X_1}(x) = (2\pi\sigma^2)^{-\frac{p}{2}} \exp\{-\|x - r_1\|^2/(2\sigma^2)\} \le (2\pi\sigma^2)^{-\frac{p}{2}} \exp\{-\|x - r_1\|^2/(2(1+\tau_n)\sigma^2)\},$$

leads to

$$P \le \binom{n}{N} N C_p^{N-1} (\Delta/\sigma)^{p(k-1)} \cdot \int_{a_n \le d(x,\mathcal{S}) \le b_n} (2\pi\sigma^2)^{-\frac{p}{2}} \exp\Big\{ - \frac{\sum_{t=1}^{N} \|x - r_t\|^2}{2(1+\tau_n)\sigma^2} \Big\} \mathrm{d}x$$

Also, notice that $\sum_{t=1}^{N} \|x - r_t\|^2 \ge N\|x - \bar{r}\|^2$ where $\bar{r} = N^{-1}\sum_{t=1}^{N} r_t$. Then,

$$\begin{aligned}
P &\le \binom{n}{N} N C_p^{N-1} (\Delta/\sigma)^{p(N-1)} \cdot \int_{a_n \le d(x,\mathcal{S}) \le b_n} (2\pi\sigma^2)^{-\frac{p}{2}} \exp\Big\{ - \frac{N\|x - \bar{r}\|^2}{2(1+\tau_n)\sigma^2} \Big\} \mathrm{d}x \\
&\le \binom{n}{N} N C_p^{N-1} (\Delta/\sigma)^{p(N-1)} \int_{\|x - \bar{r}\| \ge a_n} (2\pi\sigma^2)^{-\frac{p}{2}} \exp\Big\{ - \frac{N\|x - \bar{r}\|^2}{2(1+\tau_n)\sigma^2} \Big\} \mathrm{d}x \\
&\le \binom{n}{N} N C_p^{N-1} (\Delta/\sigma)^{p(N-1)} N^{-p/2} (1+\tau_n)^{p/2} \cdot \mathbb{P}(\chi_p^2 \ge 2Nc_0 \log n/(1+\tau_n))
\end{aligned}$$

where we used the fact that $\|x - \bar{r}\| \ge d(x, \mathcal{S})$ in the second step and we did change of variables so that the integral reduces to the tail probability of χ_p^2 distribution. By Mills ratio, the tail probability of χ_p^2 is given by

$$\mathbb{P}(\chi_p^2 \ge 2Nc_0 \log n/(1+\tau_n)) \le C n^{-Nc_0/(1+\tau_n)} \big(2Nc_0 \log n/(1+\tau_n)\big)^{p/2-1},$$

we obtain

$$P \le C \binom{n}{N} N C_p^{N-1} (\Delta/\sigma)^{p(N-1)} N^{-p/2} n^{-Nc_0/(1+\tau_n)} (2Nc_0 \log n)^{p/2-1}\,.$$

Using the approximation $\binom{n}{k} \le C(en/k)^k$, we deduce that

$$P \le C\left[e(2Nc_0\log n)^{(p-2)/(2N)}C_p^{1-1/N}N^{(1-p/2)/N}\cdot\frac{n^{1-c_0/(1+\tau_n)}(\Delta/\sigma)^{p(1-1/N)}}{N}\right]^N$$
$$=: C\Big[A(n,p,N)\cdot\frac{n^{1-c_0/(1+\tau_n)}(\Delta/\sigma)^{p(1-1/N)}}{N}\Big]^N$$

Now we plug in $N=\log(n)$ and $\Delta = c_3\sigma\sqrt{p}\Big(\frac{\log(n)}{n^{1-c_0}}\Big)^{1/p}$ for a constant $c_3 \le c_2$ where $c_2 = 0.9(2e^{2-c_0})^{-1/p}\sqrt{(2/p)}(\Gamma(p/2+1))^{1/p} = 0.9e^{-(2-c_0)/p}C_p^{-1/p}/\sqrt{p}$ with $C_p = 2^{1-p/2}/\Gamma(p/2+1)$. It is straightforward to compute that

$$A(n,p,N)\cdot\frac{n^{1-c_0/(1+\tau_n)}(\Delta/\sigma)^{p(1-1/N)}}{N}$$
$$\le e^{1-(2-c_0)(1-1/\log(n))}2^{\frac{p-2}{2\log(n)}}(c_0\log(n))^{\frac{p-2}{2\log(n)}}(0.9)^{p(1-1/\log(n))}n^{\tau_n c_0/(1+\tau_n)}\Big(\frac{n^{1-c_0}}{\log(n)}\Big)^{1/\log(n)}$$
$$\le e^{o(1)}(0.9)^p < 1.01\cdot 0.9 < 1$$

under the condition that $p \ll \log(n)/\log\log(n)$, which also give rise to $\tau_n\log(n) = o(1)$. This implies $P \le C(0.909)^{\log(n)} = o(1)$.

In the mean time, for each vertex v_k, recall that $J_k = \{i : r_i = v_k\}$,

$$N(\mathcal{B}(v_k,\Delta/2)) \ge \sum_{i\in J_k}\mathbf{1}(x_i\in\mathcal{B}(v_k,\Delta/2)) = \sum_{i\in J_k}\mathbf{1}(\|\varepsilon_i\|\le\Delta/2) \ge mp_\Delta - C\sqrt{mp_\Delta\log\log(n)}.$$

with probability $1-o(1)$, and

$$p_\Delta := \mathbb{P}(\|\varepsilon_i\|\le\Delta/2) = \mathbb{P}(\chi_p^2 \le 4^{-1}(\Delta/\sigma)^2) \ge \frac{e^{-(\Delta/\sigma)^2/8}2^{-p}}{2^{p/2}\Gamma(p/2+1)}(\Delta/\sigma)^p$$

Recall the condition that $m \ge n^\delta n^{1-c_0}$. It follows that

$$mp_\Delta \ge n^\delta\frac{e^{-(\Delta/\sigma)^2/8}2^{-p}}{2^{p/2}\Gamma(p/2+1)}n^{1-c_0}(\Delta/\sigma)^p = n^\delta\frac{e^{-(\Delta/\sigma)^2/8}}{2^{p/2}\Gamma(p/2+1)}\cdot\frac{c\log(n)}{C_p}2^{-p}(c_3/c_2)^p$$
$$\ge cn^\delta 2^{-p}(c_3/c_2)^p\log(n) \gg \log(n)$$

where $c>0$ is some small constant. The last step is due to the fact that $n^\delta 2^{-p}(c_3/c_2)^p = e^{\delta\log(n)-p\log(2c_2/c_3)} \gg 1$ as $2c_2/c_3 \ge 2$ is a constant and $p \ll \log(n)/\log\log(n)$. Thus, with probability $1-o(1)$, $N(\mathcal{B}(v_k,\Delta/2)) \gg \log(n)$. Under this event, for any point $X_{i_0}\in\mathcal{B}(v_k,\Delta/2)$, immediately $\mathcal{B}(v_k,\Delta/2)\subset\mathcal{B}(X_{i_0},\Delta)$ and further $N(\mathcal{B}(X_{i_0},\Delta)) \gg \log(n)$. Combining this, with $P=o(1)$ and $N(\mathcal{R}_f) = o_{\mathbb{P}}(1)$, we conclude that we can prune out all points with a distance to the simplex larger than $\sigma\sqrt{2c_0\log(n)}$ while preserve those points near vertices, with high probability. Thus we finish the claim for $\beta_{new}(X^*)$.

The last claim follows directly from Theorem B.1 (Theorem 1 in the manuscript) under condition (9). We therefore conclude the proof.

□

We briefly present the proof of Theorem C.1 below.

Proof. The proof strategy is roughly the same as that of Theorem C.2 When $m > c_1 n$, we take $\Delta = c_3\sigma\sqrt{p}\Big(\frac{\log(n)}{n^{1-\delta_n}}\Big)^{1/p}$ where $p/\log(n) \ll \delta_n \ll 1$ and $c_3 \le c_2$, then similarly we can derive that $N(\mathcal{B}(v_k,\Delta/2)) \ge c\log(n)n^{\delta_n}a^p = c\log(n)e^{\delta_n\log(n)-p\log(1/a)} \gg \log(n)$ where $c>0$ is a small constant and $0<a\le 1$. This gives rise to the conclusion that with high probability,

$N(\mathcal{B}(X_{i_0}, \Delta)) \gg \log(n)$ for any $X_{i_0} \in N(\mathcal{B}(v_k, \Delta/2))$.Moreover, in the same manner to the above derivations, replacing c_0 by δ_n, we can claim again that $N(\mathcal{R}_f) = o_{\mathbb{P}}(1)$ and

$$P \leq C\left(A(n,p,\log(n)) \cdot \frac{n^{1-\delta_n/(1+\tau_n)}(\Delta/\sigma)^{p(1-1/\log(n))}}{\log(n)}\right)^{\log(n)} = o(1).$$

Consequently, all the claims follow from the same reasoning as the proof of Theorem C.2. We therefore omit the details and conclude the proof . □

C.3 Proof of Lemma C.7

Recall that $R = n^{-1/2}[r_1 - \bar{r}, \ldots, r_n - \bar{r}]$. Let $R = U_0 D_0 V_0$ be its singular value decomposition and let $H_0 = U_0 U_0'$. Denote $\epsilon = [\epsilon_1, \ldots, \epsilon_n] \in \mathbb{R}^{d,n}$. We start by analyzing the convergence rate of $\|ZZ' - nRR' - n\sigma^2 I_d\|$. Recall that $\bar{X} = \bar{r} + \bar{\epsilon}$, where $\bar{\epsilon} = n^{-1}\sum_{i=1}^n \epsilon_i$. We obtain

$$Z = X_i - \bar{X} = r_i + \epsilon_i - \bar{r} - \bar{\epsilon}, \qquad Z = \sqrt{n}R + \epsilon - \bar{\epsilon}1_n'. \tag{C.75}$$

Observing the fact that $R1_n = 0$, we deduce

$$\begin{aligned} ZZ' - nRR' - n\sigma^2 I_d &= (\sqrt{n}R + \epsilon - \bar{\epsilon}1_n')(\sqrt{n}R + \epsilon - \bar{\epsilon}1_n')' - nRR' - n\sigma^2 I_d \\ &= \sqrt{n}(\epsilon - \bar{\epsilon}1_n')R' + \sqrt{n}R(\epsilon - 1_n\bar{\epsilon}')' + (\epsilon - \bar{\epsilon}1_n')(\epsilon - \bar{\epsilon}1_n')' - n\sigma^2 I_d \\ &= \sqrt{n}\epsilon R' + \sqrt{n}R\epsilon' + (\epsilon\epsilon' - n\sigma^2 I_d) - n\bar{\epsilon}\bar{\epsilon}'. \end{aligned} \tag{C.76}$$

The above equation implies that

$$\|ZZ' - nRR' - n\sigma^2 I_d\| \leq 2\sqrt{n}\|\epsilon R'\| + \|\epsilon\epsilon' - n\sigma^2 I_d\| + n\|\bar{\epsilon}\|^2. \tag{C.77}$$

We proceed to bound the three terms $\|\epsilon R'\|$, $\|\epsilon\epsilon' - n\sigma^2 I_d\|$ and $n\|\bar{\epsilon}\|^2$ respectively. First, notice that $\epsilon R' \in \mathbb{R}^{d\times d}$ is a Gaussian random matrix with independent rows which follow $N(0, RR')$. By Theorem 5.39 and Remark 5.40 in Vershynin (2010), we can deduce that with probability $1 - o(1)$,

$$n\|R\epsilon'\epsilon R'\| \leq Cnd\sigma^2 s_1^2(R).$$

This, together with the fact that $s_1(R) \leq c$ gives that

$$\sqrt{n}\|\epsilon R' + R\epsilon'\| \leq C\sigma\sqrt{nd}. \tag{C.78}$$

Second, by Bai-Yin law (Bai & Yin (2008)), we can estimate the bound of $\|\mathcal{E}\mathcal{E}' - n\sigma^2 I_d\|$ as follows.

$$\|\epsilon\epsilon' - n\sigma^2 I_d\| \leq n\sigma^2(2\sqrt{d/n} + d/n) \leq \sigma^2(2\sqrt{nd} + d), \tag{C.79}$$

with probability $1 - o(1)$. Third, observe that $\bar{\epsilon} \sim N(0, \sigma^2/nI_d)$. We therefore obtain that with probability $1 - o(1)$,

$$n\|\bar{\epsilon}\|^2 \leq \sigma^2[d + C\sqrt{d\log(n)}].$$

By applying the condition that $\sigma = O(1)$, combining the above equation with (C.77), (C.78) and (C.79) yields that, with probability at least $1 - o(1)$,

$$\begin{aligned} \|ZZ' - nRR' - n\sigma^2 I_d\| &\leq 2\sigma\sqrt{nd} + \sigma^2[d + C\sqrt{d\log(n)}] + \sigma^2(2\sqrt{nd} + d) \\ &\leq C(\sigma\sqrt{nd} + \sigma^2 d). \end{aligned} \tag{C.80}$$

Now, we compute the bound for $\|\widehat{H} - H_0\|$. Let $U^\perp, U_0^\perp \in \mathbb{R}^{d,d-K+1}$ such that their columns are the last $(d - K + 1)$ columns of U and U_0, respectively. It follows from direct calculations that

$$\begin{aligned} \|\widehat{H} - H_0\| &= \|U_0U_0' - UU'\| \leq \|U_0^\perp(U_0^\perp)'(U_0U_0' - UU')\| + \|U_0U_0'(U_0U_0' - UU')\| \\ &= \|U_0^\perp(U_0^\perp)'UU'\| + \|U_0U_0'U^\perp(U^\perp)'\| \leq \|(U_0^\perp)'U\| + \|U_0'U^\perp\| = 2\|\sin\Theta(U_0, U)\|. \end{aligned}$$

Notably, $U, U^\perp$ is also the eigen-space of $ZZ' - n\sigma^2 I_d$. By Weyl's inequality (see, for example, Horn & Johnson (1985)),

$$\max_{1\leq i\leq d}\left|\lambda_i(ZZ' - n\sigma^2 I_d) - \lambda_i(nRR')\right| \leq C\|ZZ' - n\sigma^2 I_d - nRR'\|$$

Under the condition that $s^2_{K-1}(R) \gg \max\{\sqrt{\sigma^2 d/n}, \sigma^2 d/n\}$, by Davis-Kahan Theorem (Davis & Kahan (1970)), we deduce that, with probability at least $1 - o(1)$,

$$\|\widehat{H} - H_0\| \leq 2\|\sin\Theta(U_0, U)\| \leq \frac{2\|ZZ' - nRR' - n\sigma^2 I_d\|}{\lambda_{K-1}(nRR')}$$
$$\leq C\frac{\max\{\sqrt{\sigma^2 d/n}, \sigma^2 d/n\}}{s^2_{K-1}(R)}. \tag{C.81}$$

The proof is complete.

D NUMERICAL SIMULATION FOR THEOREM 1

In this short section, we want to provide a better sense of our bound derived in Theorem 1 and how it compares with the one from the orthodox SPA. To make it easier for the reader to see the difference between the two bounds, we consider toy example where we fix $(K, d) = (3, 3)$ and

$$\widetilde{V} = \{(20, 20, 0), (20, 30, 0), (30, 20, 0)\}$$

while we let

$$V = \widetilde{V} + a \cdot (0, 0, 1).$$

We consider 50 different values for a ranging from 10 to 1000. It is not surprising to see that when a is close to 0 the bound of the orthodox SPA goes to infinity whereas as the simplex is bounded far away from the origin, the K^{th} singular value will be bounded away from 0. However, our bound still outperforms the traditional SPA bound even for very large values of a. Looking at two specific values of a we have the following. For $a = 10$,

$$\beta_{new} = 0.03, \qquad \beta(V) = 0.05$$

Moreover, as a changes, the Figure 5 below illustrate how much the ratio of

$$\frac{\text{our whole bound}}{\text{Gillis bound}}$$

changes as the parameter a changes. For example, when $a = 10$.

$$\frac{g_{new}(V)}{g(V)} = 0.015,$$

and so

$$\frac{\text{our whole bound}}{\text{Gillis bound}} = 0.009$$

so we reduce the bound by 111 . Similarly, when $a = 1000$,

$$\frac{g_{new}(V)}{g(V)} = 0.19, \qquad \frac{\text{our whole bound}}{\text{Gillis bound}} = 0.105,$$

so we have reduced the bound by 9.5.

REFERENCES

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, 2008.

M. C. U. Araújo, T. C. B. Saldanha, and R. K. H. Galvao et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst.*, 57 (2):65–73, 2001.

Zhi-Dong Bai and Yong-Qua Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. In *Advances In Statistics*, pp. 108–127. World Scientific, 2008.

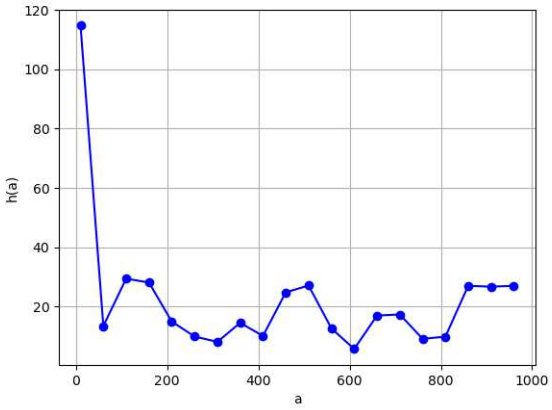

Figure 5: Factor of improvement of our bound over orthodox spa as the true simplex moves away from origin by a distance a.

Ainesh Bakshi, Chiranjib Bhattacharyya, Ravi Kannan, David P Woodruff, and Samson Zhou. Learning a latent simplex in input-sparsity time. *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–11, 2021.

Sohom Bhattacharya, Jianqing Fan, and Jikai Hou. Inferences on mixing probabilities and ranking in mixed-membership models. *arXiv:2308.14988*, 2023.

Chiranjib Bhattacharyya and Ravindran Kannan. Finding a latent k–simplex in o*(k· nnz (data)) time via subset smoothing. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 122–140. SIAM, 2020.

José M Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE journal of selected topics in applied earth observations and remote sensing*, 5(2):354–379, 2012.

Victor-Emmanuel Brunel. Adaptive estimation of convex and polytopal density support. *Probability Theory and Related Fields*, 164(1-2):1–16, 2016.

Maurice D Craig. Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, 32(3):542–552, 1994.

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.

Nicolas Gillis. Successive projection algorithm robust to outliers. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 331–335. IEEE, 2019.

Nicolas Gillis and Stephen A Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):698–714, 2013.

Nicolas Gillis and Stephen A Vavasis. Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization. *SIAM Journal on Optimization*, 25(1): 677–698, 2015.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer, 2nd edition, 2009.

Roger Horn and Charles Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

Sihan Huang, Jiajin Sun, and Yang Feng. Pcabm: Pairwise covariates-adjusted block model for community detection. *Journal of the American Statistical Association*, (just-accepted):1–26, 2023.

Hamid Javadi and Andrea Montanari. Nonnegative matrix factorization via archetypal analysis. *Journal of the American Statistical Association*, 115(530):896–907, 2020.

Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Mixed membership estimation for social networks. *J. Econom.*, https://doi.org/10.1016/j.jeconom.2022.12.003., 2023.

Zheng Tracy Ke and Jiashun Jin. The SCORE normalization, especially for heterogeneous network and text data. *Stat*, 12(1)(e545):https://doi.org/10.1002/sta4.545, 2023.

Zheng Tracy Ke and Minzhe Wang. Using SVD for topic modeling. *Journal of the American Statistical Association*, https://doi.org/10.1080/01621459.2022.2123813:1–16, 2022.

Tomohiko Mizutani and Mirai Tanaka. Efficient preconditioning for noisy separable nonnegative matrix factorization problems by successive projection based low-rank approximations. *Machine Learning*, 107:643–673, 2018.

Nicolas Nadisic, Nicolas Gillis, and Christophe Kervazo. Smoothed separable nonnegative matrix factorization. *Linear Algebra and its Applications*, 676:174–204, 2023.

Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E Priebe. A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1446–1473, 2022.

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

P Stein. A note on the volume of a simplex. *The American Mathematical Monthly*, 73(3):299–301, 1966.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *ArXiv.1011.3027*, 2010.

Michael E Winter. N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data. In *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, pp. 266–275, 1999.

Anru Zhang and Mengdi Wang. Spectral state compression of markov processes. *IEEE transactions on information theory*, 66(5):3202–3231, 2019.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Detecting overlapping communities in networks using spectral methods. *SIAM J. Math. Data Sci.*, 2(2):265–283, 2020.