Rigorous Restricted Isometry Property of Low-Dimensional Subspaces

Gen Li, Qinghua Liu, and Yuantao Gu

Tsinghua University, Beijing 100084, China

Abstract

Dimensionality reduction is in demand to reduce the complexity of solving large-scale problems with data lying in latent low-dimensional structures in machine learning and computer vision. Motivated by such need, in this work we study the Restricted Isometry Property (RIP) of Gaussian random projections for low-dimensional subspaces in $\mathbb{R}^N$, and rigorously prove that the projection Frobenius norm distance between any two subspaces spanned by the projected data in $\mathbb{R}^n$ ($n < N$) remain almost the same as the distance between the original subspaces with probability no less than $1 - \mathrm{e}^{-\mathcal{O}(n)}$. Previously the well-known Johnson-Lindenstrauss (JL) Lemma and RIP for sparse vectors have been the foundation of sparse signal processing including Compressed Sensing. As an analogy to JL Lemma and RIP for sparse vectors, this work allows the use of random projections to reduce the ambient dimension with the theoretical guarantee that the distance between subspaces after compression is well preserved.

Keywords: Restricted Isometry Property, Gaussian random matrix, random projection, low-dimensional subspaces, dimensionality reduction, subspace clustering

[1]The authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. This work was partially supported by National Natural Science Foundation of China (NSFC 61531166005, 61571263, 61371137) and the National Key Research and Development Program of China (Project No. 2016YFE0201900, 2017YFC0403600). The corresponding author of this work is Y. Gu (E-mail: gyt@tsinghua.edu.cn).

1. Introduction

This paper studies the Restricted Isometry Property (RIP) of random projections for subspaces. It reveals that the distance between two low-dimensional subspaces remain almost unchanged after being projected by a Gaussian random matrix with overwhelming probability, when the ambient dimension after projection is sufficiently large in comparison with the dimension of subspaces.

1.1. Motivation

In the era of data deluge, labeling huge amount of large-scale data can be time-consuming, costly, and even intractable, so unsupervised learning has attracted increasing attention in recent years. One of such methods emerging recently, subspace clustering (SC) [1, 2, 3, 4], which depicts the latent structure of a variety of data as a union of subspaces, has been shown to be powerful in a wide range of applications, including motion segmentation, face clustering, and anomaly detection. It also shows great potential to some previously less explored datasets, such as network data, gene series, and medical images.

Traditional subspace clustering methods, however, suffer from the deficiency in similarity representation, so it can be computationally expensive to adapt them to large-scale datasets. In order to alleviate the high computational burden, a variety of works have been done to address the crucial problem of how to efficiently handle large-scale datasets. Compressed Subspace Clustering (CSC) [5] also known as Dimensionality-reduced Subspace Clustering [6] is a method that performs SC on randomly compressed data points. Because the random compression reduces the dimension of the ambient space, the computational cost of finding the self-representation in SC can be efficiently reduced. Based on the concept of subspace affinity, which characterizes the similarity between two subspaces, and the mathematical tools introduced in [2], the conditions under which several popular algorithms can successfully cluster the compressed data have been theoretically studied and numerically verified [7, 8].

Because the data points are randomly projected from a high-dimensional ambient space $\mathbb{R}^N$ to a new medium-dimensional ambient space $\mathbb{R}^n$, a worry

is that the similarity between any two low-dimensional subspaces increases and the SC algorithms are less likely to perform well. Inspired by the well-known Johnson-Lindenstrauss (JL) Lemma [9, 10] and the Restricted Isometry Property (RIP) [11, 12, 13], which allows the use of random projection to reduce the space dimension while keeping the Euclidean distance between any two data points and leads to the boom of sparse signal processing including Compressed Sensing (CS) [14, 15, 16, 17, 18, 19, 20], one may speculate whether the similarity (or distance) between any two given subspaces can remain almost unchanged, if the dimension of the latent subspace that the data lie in is small compared with that of the ambient space after projection n. It should be highlighted that this conjecture is not confined to the SC problem, so we believe that it may benefit future studies on other subspace related topics.

Motivated by the conjecture about whether the similarity between any two given subspaces can remain almost unchanged after random projection, we study the RIP of Gaussian random projections for a finite set of subspaces. In order to give more solid guarantees and more precise insight into the law of magnitude of the dimensions for CSC and other subspace related problems, we derive an optimum probability bound of the RIP of Gaussian random compressions for subspaces in this paper. Compared with our previous work [21], the probability bound has been improve from $1-\mathcal{O}(1/n)$ to $1-\mathrm{e}^{-\mathcal{O}(n)}$, which is optimum when we consider the state-of-the-art statistical probability theories for Gaussian random matrix.

1.2. Main Results

The projection Frobenius norm (F-norm for short) distance is adopted in this work to measure the distance between two subspaces. It should be noted that we slightly generalize the definition in [22] to the situation where the dimensions of the two subspaces are different.

Definition 1 ([21]) (Projection Frobenius norm distance between subspaces) *The generalized projection F-norm distance between two subspaces* $\mathcal{X}_1$

and $\mathcal{X}_2$ is defined as

$$D(\mathcal{X}_1, \mathcal{X}_2) := \frac{1}{\sqrt{2}} \|\mathbf{U}_1\mathbf{U}_1^{\mathrm{T}} - \mathbf{U}_2\mathbf{U}_2^{\mathrm{T}}\|_{\mathrm{F}},$$

where $\mathbf{U}_i$ denotes an arbitrary orthonormal basis matrix for subspace $\mathcal{X}_i, i = 1, 2$.

We will focus on the change of the distance between any two low-dimensional subspaces after being randomly projected from $\mathbb{R}^N$ to $\mathbb{R}^n$ ($n < N$). The projection of a low-dimensional subspace by using a Gaussian random matrix is defined as below.

Definition 2 (Gaussian random projection for subspace) *The Gaussian random projection of a d-dimensional subspace $\mathcal{X} \subset \mathbb{R}^N$ onto $\mathbb{R}^n$ ($d < n < N$) is defined as below,*

$$\mathcal{X} \xrightarrow{\mathbf{\Phi}} \mathcal{Y} = \{\mathbf{y} | \mathbf{y} = \mathbf{\Phi}\mathbf{x}, \forall \mathbf{x} \in \mathcal{X}\},$$

where the projection matrix $\mathbf{\Phi} \in \mathbb{R}^{n \times N}$ is composed of entries independently drawn from Gaussian distribution $\mathcal{N}(0, 1/n)$.

One may notice that the dimensions of subspaces remain unchanged after random projection with probability one.

Based on the definitions above, the main theoretical result of this work is stated as follows.

Theorem 1 *Suppose $\mathcal{X}_1, \ldots, \mathcal{X}_L \subset \mathbb{R}^N$ are L subspaces with dimension less than or equal to d. After random projection by using a Gaussian random matrix $\mathbf{\Phi} \in \mathbb{R}^{n \times N}$, $\mathcal{X}_i \xrightarrow{\mathbf{\Phi}} \mathcal{Y}_i \subset \mathbb{R}^n, i = 1, \cdots, L, n < N$. There exist constants $c_1(\varepsilon), c_2(\varepsilon) > 0$ depending only on ε such that for any two subspaces $\mathcal{X}_i$ and $\mathcal{X}_j$, for any $n > c_1(\varepsilon) \max\{d, \ln L\}$,*

$$(1 - \varepsilon)\, D^2(\mathcal{X}_i, \mathcal{X}_j) < D^2(\mathcal{Y}_i, \mathcal{Y}_j) < (1 + \varepsilon)\, D^2(\mathcal{X}_i, \mathcal{X}_j) \tag{1}$$

holds with probability at least $1 - \mathrm{e}^{-c_2(\varepsilon)n}$.

Theorem 1 reveals that the distance between two subspaces remains almost unchanged after random projection with overwhelming probability, when the ambient dimension after projection n is sufficiently large.

1.3. Our Contribution

In this paper, we study the RIP of Gaussian random matrices for projecting a finite set of subspaces. The problem is challenging as random projections neither preserve orthogonality nor normalize the vectors defining orthonormal bases of the subspaces. In order to measure the change in subspace distance induced by random projections, both effects have to be carefully quantified. Based on building a metric space of subspaces with the projection F-norm distance, which is closely connected with subspace affinity, we start from verifying that the affinity between two subspaces concentrates on its estimate with overwhelming probability after Gaussian random projection. Then we successfully reach the RIP of two subspaces and generalize it to the situation of a finite set of subspaces, as stated in Theorem 1.

The main contribution of this work is to provide a mathematical tool, which can shed light on many problems including CSC. As a direct result of Theorem 1, when solving the SC problem at a large scale, one may conduct SC on randomly compressed samples to alleviate the high computational burden and still have theoretical performance guarantee. Because the distance between subspaces almost remains unchanged after projection, the clustering error rate of any SC algorithm may keep as small as that conducting in the original space. Considering that our theory is independent of SC algorithms, this may benefit future studies on other subspace related topics.

Except our previous work [21] that will be compared with in Section 6, as far as we know, there is no work that study the distance preserving property between low-dimensional subspaces after random projection.

1.3.1. Comparison with JL Lemma and RIP for Sparse Signals

The famous Johnson-Lindenstrauss Lemma illustrates that there exists a map from a higher-dimensional space into a lower-dimensional space such that the distance between a finite set of data points will change little after being mapped.

Lemma 1 (JL Lemma) *[9, 10] For any set $\mathcal{V}$ of L points in $\mathbb{R}^N$, there exists a map $f : \mathbb{R}^N \to \mathbb{R}^n, n < N$, such that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{V}$,*

$$(1-\varepsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \leq \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2$$

if n is a positive integer satisfying $n \geq 4\ln L/(\varepsilon^2/2 - \varepsilon^3/3)$, where $0 < \varepsilon < 1$ is a constant.

The RIP of random matrix illustrates that the distance between two sparse vectors will change little with high probability after random projection.

Definition 3 *[11, 12, 13] The projection matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times N}, n < N$ satisfies RIP of order k if there exists a $\delta_k \in (0, 1)$ such that*

$$(1-\delta_k)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}_1 - \mathbf{\Phi}\mathbf{x}_2\|_2^2 \leq (1+\delta_k)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2$$

holds for any two k-sparse vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^N$.

Theorem 2 *[13] A Gaussian random matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times N}, n < N$ has the RIP of order k for $n \geq c_1 k \ln\left(\frac{N}{k}\right)$ with probability $1 - \mathrm{e}^{-c_2 n}$, where $c_1, c_2 > 0$ are constants depending only on δ_k, the smallest nonnegative constant satisfying Definition 3.*

In summary, the above works focus on the change of the distance between points after determinate mapping or random projection. In comparison, our work views a subspace as a whole and studies the distance between subspaces, which to the best of our knowledge has never been studied before. Moreover, the above works study the points in Euclidean space with l_2-norm, while our work study the subspaces on the Grassmannian manifold with F-norm metric, which is highly nonlinear and more complex. A detailed comparison to explain the differences between our work and related works is presented in Table 1.

1.3.2. Comparison with RIP for the Angels of Sparse Signals

There are literatures studying the angle preserving properties of embedding sparse vectors. The authors [23] show that within a sparse signals set that

Table 1: Comparison with other dimension-reduction theories including JL Lemma, RIP for sparse signals, and our previous results [21]

	JL Lemma	RIP for sparse signals	RIP for low-dimensional subspaces	
			[21]	this work
object	any set of L points in $\mathbb{R}^N$	all k-sparse signals in $\mathbb{R}^N$	any set of L d-dimensional subspaces in $\mathbb{R}^N$	
metric	Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|_2$		projection F-norm distance $\frac{1}{\sqrt{2}}\|\mathbf{P}_i - \mathbf{P}_j\|_{\mathrm{F}}$	
compression method	some map f	Gaussian random matrix		
error bound	$(1-\varepsilon, 1+\varepsilon)$	$(1-\delta_k, 1+\delta_k)$	$(1-\varepsilon, 1+\varepsilon)$	
condition	$n \geq \frac{4\ln L}{\varepsilon^2/2 - \varepsilon^3/3}$	$n \geq c_1 k \ln\left(\frac{N}{k}\right)$	n large enough	$n > c_1 \max\{d, \ln L\}$
success probability	1	$1 - \mathrm{e}^{-c_2 n}$	$1 - \frac{2dL(L-1)}{(\varepsilon - d/n)^2 n}$	$1 - \mathrm{e}^{-c_2 n}$

is embedded with an RIP guarantee, all pairwise angles between vectors are preserved, which corresponds to the case $d = 1$ in our work. In addition, in Proposition 5 of [24], the authors also provide a similar result on preserving inner products between vectors.

Theorem 3 *[23] Suppose a matrix $\mathbf{\Phi}$ satisfies RIP for k-sparse vectors with $\delta_k \in [0, \frac{1}{3}]$. Then, for any vectors having sparsity at most k, supported on the*

same set of indices and separated by an angle $\alpha \in [0, \frac{\pi}{2}]$*, the angle* α_p *between the projected vectors obeys the bound*

$$(1 - \sqrt{3\delta_k})\alpha \leq \alpha_p \leq (1 + 3\delta_k)\alpha.$$

1.3.3. Comparison with RIP for Signals in UoS

There are literatures studying the distance preserving properties of compressed data points, which may be sparse on specific basis or lie in a couple of subspaces or surfaces [25, 26, 27, 28, 29].

The authors of [25] extended the RIP to signals that are sparse or compressible with respect to a certain basis $\mathbf{\Psi}$, i.e., $\mathbf{x} = \mathbf{\Psi}\boldsymbol{\alpha}$, where $\mathbf{\Psi}$ is represented as a unitary $N \times N$ matrix and $\boldsymbol{\alpha}$ is a k-sparse vector. The work of [26] proves that with high probability the random projection matrix $\mathbf{\Phi}$ can preserve the distance between two signals belonging to a Union of Subspaces (UoS). A unified RIP theory for Euclidean dimensionality reduction is proposed in [27]. In [28], it is shown that random projection preserves the structure of surfaces. Given a collection of L surfaces of linearization dimension d, if they are embedded into a space of $\mathcal{O}(d\delta^2 \log(Ld/\delta))$ dimension, the surfaces are preserved in the sense that for any pair of points on these surfaces the distance between them are preserved. The main contribution of [29] is stated as follows. If S is an n point subset of $\mathbb{R}^N$, $0 < \delta < \frac{1}{3}$ and $n = 256d \log n(\max\{d, 1/\delta\})^2$, there is a mapping of $\mathbb{R}^N$ into $\mathbb{R}^n$ under which volumes of sets of size at most d do not change by more than a factor of $1 + \delta$, and the distance of points from affine hulls of sets of size at most $k - 1$ is preserved within a relative error of δ.

According to above survey, those works study embedding of Euclidean distances between points in subspaces, while we discuss embedding of a finite set of subspaces in terms of the projection F-norm distance. In both the related works and this paper, the same mathematical tool of concentration inequalities and random matrix theory are adopted to derive the RIP for two different objects, i.e., data points in Euclidean space and subspaces in Euclidean space (or points on Grassmann manifold), respectively. In comparison, both Euclidean space

and random projection are linear, but Grassmannian is not linear, let along the projection on it, so the new problem is much more difficult than the existing one, and a core contribution of this work is dealing with the above challenges with a brand-new geometric proof, the technique in which has hardly been used previously to derive the RIP for data points.

1.4. Organization

The rest of this paper is organized as follows. Based on the introduction of principal angles, affinity, and its connection with the projection F-norm distance, we study the RIP for subspaces in the top level in Section 2. The main result of Theorem 1 is proved by using two core propositions of Lemma 4 and Theorem 4. In Section 3, we focus on the probability and concentration inequalities of Gaussian random matrix to prepare necessary mathematical tools that will be used through this work. In Section 4, we prove the first core proposition of Lemma 4, which states that the affinity between a line and a subspace will concentrate on its estimate with overwhelming probability after random projection. In Section 5, we prove the second core proposition of Theorem 4, which provides a general theory that the affinity between two subspaces with arbitrary dimensions demonstrates concentration after random projection. In Section 6, we compare those theories with our previous results and highlight the novelty. We conclude this work in Section 7. Most proofs of lemmas and remarks are included in the Appendix 8.

1.5. Notations

Vectors and matrices are denoted by lower-case and upper-case letter, respectively, both in boldface. $\mathbf{A}^{\mathrm{T}}$ denotes matrix transposition. $\|\mathbf{a}\|$ and $\|\mathbf{A}\|_{\mathrm{F}}$ denote ℓ_2 norm of vector $\mathbf{a}$ and Frobenius norm of matrix $\mathbf{A}$. $s_{\max}(\mathbf{A})$ and $s_{\min}(\mathbf{A})$ denote the largest and smallest singular value of matrix $\mathbf{A}$, respectively. Subspaces are denoted by $\mathcal{X}, \mathcal{Y}$, and $\mathcal{S}$. $\mathcal{C}(\mathbf{A})$ denotes the column space of matrix $\mathbf{A}$. We use $\mathcal{S}^{\perp}$ to denote the orthonormal complement space of $\mathcal{S}$. $\mathrm{P}_{\mathcal{S}}(\mathbf{v})$ denotes the projection of vector $\mathbf{v}$ onto subspace $\mathcal{S}$.

2. RIP of Gaussian Random Projection for Subspaces

2.1. Preliminary

Before starting the theoretical analysis, we first introduce the definition of principal angles and affinity. These two concepts have been widely adopted to describe the relative position and to measure the similarity between two subspaces. Our theoretical analysis will first focus on the estimation of these quantities before and after random projection. Then using the connection between affinity and projection F-norm distance derived in [21], we can readily derive the result in Theorem 1.

The principal angles (or canonical angles) between two subspaces provide a robust way to characterize the relative subspace positions [30, 31].

Definition 4 *The principal angles $\theta_1, \cdots, \theta_{d_1}$ between two subspaces $\mathcal{X}_1$ and $\mathcal{X}_2$ of dimensions $d_1 \leq d_2$, are recursively defined as*

$$\cos\theta_k = \max_{\mathbf{x}_1 \in \mathcal{X}_1} \max_{\mathbf{x}_2 \in \mathcal{X}_2} \frac{\mathbf{x}_1^{\mathrm{T}} \mathbf{x}_2}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|} =: \frac{\mathbf{x}_{1k}^{\mathrm{T}} \mathbf{x}_{2k}}{\|\mathbf{x}_{1k}\| \|\mathbf{x}_{2k}\|},$$

with the orthogonality constraints $\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{il} = 0, l = 1, \cdots, k-1, i = 1, 2$.

Beside definition, an alternative way of computing principal angles is to use the singular value decomposition [32].

Lemma 2 *Let the columns of $\mathbf{U}_i$ be orthonormal bases for subspace $\mathcal{X}_i$ of dimension $d_i, i = 1, 2$ and suppose $d_1 \leq d_2$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d_1} \geq 0$ be the singular values of $\mathbf{U}_1^{\mathrm{T}} \mathbf{U}_2$, then $\cos\theta_k = \lambda_k, k = 1, \cdots, d_1$.*

Based on principle angles, affinity is defined to measure the similarity between subspaces [2].

Definition 5 *The affinity between two subspaces $\mathcal{X}_1$ and $\mathcal{X}_2$ of dimension $d_1 \leq d_2$ is defined as*

$$\mathrm{aff}\,(\mathcal{X}_1, \mathcal{X}_2) := \left(\sum_{k=1}^{d_1} \cos^2\theta_k \right)^{1/2} = \|\mathbf{U}_1^{\mathrm{T}} \mathbf{U}_2\|_{\mathrm{F}},$$

where the columns of $\mathbf{U}_i$ are orthonormal bases of $\mathcal{X}_i, i = 1, 2$.

The relationship between distance and affinity is revealed in Lemma 3. Because of the concise definition and easy computation of affinity, we will start the theoretical analysis with affinity, and then present the results with distance by using Lemma 3.

Lemma 3 *[21] The* distance *and* affinity *between two subspaces $\mathcal{X}_1$ and $\mathcal{X}_2$ of dimension d_1, d_2, are connected by*

$$D^2(\mathcal{X}_1, \mathcal{X}_2) = \frac{d_1 + d_2}{2} - \mathrm{aff}^2(\mathcal{X}_1, \mathcal{X}_2).$$

2.2. Theoretical Results

In this section, we will present the main theoretical results about the affinity and distance between subspaces. Before that, let us introduce some basic notations to be used. We denote the random projection of subspaces of $\mathcal{X}_1$ and $\mathcal{X}_2$ as $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively. We denote $D_{\mathcal{X}} = D(\mathcal{X}_1, \mathcal{X}_2)$ and $D_{\mathcal{Y}} = D(\mathcal{Y}_1, \mathcal{Y}_2)$ as the distances before and after random projection. Similarly, we use $\mathrm{aff}_{\mathcal{X}} = \mathrm{aff}(\mathcal{X}_1, \mathcal{X}_2)$ and $\mathrm{aff}_{\mathcal{Y}} = \mathrm{aff}(\mathcal{Y}_1, \mathcal{Y}_2)$ to denote the affinities before and after projection. Without loss of generality, we always suppose that $d_1 \leq d_2$. For simplicity, we refer the affinity (distance) after random projection as *projected affinity* (*projected distance*).

To begin with, we focus on a special case that one subspace is degenerated to a line (one-dimensional subspace). The following lemma provides an estimation of the affinity between a line and a subspace after Gaussian random projection. When the dimensionality of the new ambient space is large enough, the real projected affinity will highly concentrate around this estimation with overwhelming probability.

Lemma 4 *Suppose $\mathcal{X}_1, \mathcal{X}_2 \subset \mathbb{R}^N$ are a line and a d-dimension subspace, $d \geq 1$, respectively. Let $\lambda = \mathrm{aff}_{\mathcal{X}}$ denote their affinity. If they are projected onto $\mathbb{R}^n, n < N$, by a Gaussian random matrix $\mathbf{\Phi} \in \mathbb{R}^{n \times N}$, $\mathcal{X}_i \xrightarrow{\mathbf{\Phi}} \mathcal{Y}_i, i = 1, 2$, then the projected affinity, $\mathrm{aff}_{\mathcal{Y}}$, can be estimated by*

$$\overline{\mathrm{aff}}_{\mathcal{Y}}^2 = \lambda^2 + \frac{d}{n}\left(1 - \lambda^2\right), \tag{2}$$

and there exist constants $c_1(\varepsilon), c_2(\varepsilon) > 0$ *depending only on* ε *such that for any* $n > c_1(\varepsilon)d$,

$$\left|\mathrm{aff}_{\mathcal{Y}}^2 - \overline{\mathrm{aff}}_{\mathcal{Y}}^2\right| < (1-\lambda^2)\varepsilon \tag{3}$$

holds with probability at least $1 - \mathrm{e}^{-c_2(\varepsilon)n}$.

Then, we study the general case of projecting two subspaces of arbitrary dimensions. As mentioned in the last subsection, we will begin with the estimation of affinity and then restate the result in terms of distance.

The following theorem reveals the concentration of affinity between two arbitrary subspaces after random projection.

Theorem 4 *Suppose* $\mathcal{X}_1, \mathcal{X}_2 \subset \mathbb{R}^N$ *are two subspaces with dimension* $d_1 \leq d_2$, *respectively. Take*

$$\overline{\mathrm{aff}}_{\mathcal{Y}}^2 = \mathrm{aff}_{\mathcal{X}}^2 + \frac{d_2}{n}(d_1 - \mathrm{aff}_{\mathcal{X}}^2) \tag{4}$$

as an estimate of the affinity between two subspaces after random projection, $\mathcal{X}_i \xrightarrow{\mathbf{\Phi}} \mathcal{Y}_i, i = 1, 2$. *Then there exist constants* $c_1(\varepsilon), c_2(\varepsilon) > 0$ *depending only on* ε *such that for any* $n > c_1(\varepsilon)d_2$,

$$\left|\mathrm{aff}_{\mathcal{Y}}^2 - \overline{\mathrm{aff}}_{\mathcal{Y}}^2\right| < (d_1 - \mathrm{aff}_{\mathcal{X}}^2)\varepsilon \tag{5}$$

holds with probability at least $1 - \mathrm{e}^{-c_2(\varepsilon)n}$.

Because of its concision when evaluating the relative position in Definition 5, we present the concentration by using affinity in Lemma 4 and Theorem 4, which play essential role in RIP for subspaces. Their proofs, which unfurl main text of this work, are postponed to Section 4 and Section 5, respectively.

Using Lemma 3 and Theorem 4, we derive an estimation of the projected distance. Similarly we prove that the true projected distance will highly concentrate around this estimate with overwhelming probability.

Corollary 1 *Suppose* $\mathcal{X}_1, \mathcal{X}_2 \subset \mathbb{R}^N$ *are two subspaces with dimension* $d_1 \leq d_2$, *respectively. We use*

$$\overline{D}_{\mathcal{Y}}^2 = D_{\mathcal{X}}^2 - \frac{d_2}{n}\left(D_{\mathcal{X}}^2 - \frac{d_2 - d_1}{2}\right) \tag{6}$$

as an estimation of the distance between two subspaces after random projection, $\mathcal{X}_i \xrightarrow{\mathbf{\Phi}} \mathcal{Y}_i, i = 1, 2$. *Then there exist constants* $c_1(\varepsilon), c_2(\varepsilon) > 0$ *depending only on* ε *such that for any* $n > c_1(\varepsilon)d_2$,

$$\left| D_{\mathcal{Y}}^2 - \overline{D}_{\mathcal{Y}}^2 \right| < \left(D_{\mathcal{X}}^2 - \frac{d_2 - d_1}{2} \right) \varepsilon \tag{7}$$

holds with probability at least $1 - \mathrm{e}^{-c_2(\varepsilon)n}$.

Proof Combining (4) and (6) by using Lemma 3, we readily get that $|D_{\mathcal{Y}}^2 - \overline{D}_{\mathcal{Y}}^2| = |\mathrm{aff}_{\mathcal{Y}}^2 - \overline{\mathrm{aff}}_{\mathcal{Y}}^2|$. Using Lemma 3 again, we have $D_{\mathcal{X}}^2 - \frac{d_2 - d_1}{2} = d_1 - \mathrm{aff}_{\mathcal{X}}^2$. Therefore, (7) is identical to (5). ∎

2.3. Proof of Theorem 1

Now we are ready to prove the RIP of Gaussian random matrix for projecting a finite set of subspaces using the results above.

Without loss of generality, we assume that $d_i \leq d_j \leq d$. According to Corollary 1, there exist constants $c_{1,1}$, $c_{2,1} > 0$ depending only on ε such that for any $n > c_{1,1}d_j$,

$$\begin{aligned} D^2(\mathcal{X}_i, \mathcal{X}_j) - \left(\frac{d_j}{n} + \frac{\varepsilon}{2} \right) \left(D^2(\mathcal{X}_i, \mathcal{X}_j) - \frac{d_j - d_i}{2} \right) &< D^2(\mathcal{Y}_i, \mathcal{Y}_j) \\ &< D^2(\mathcal{X}_i, \mathcal{X}_j) + \left(-\frac{d_j}{n} + \frac{\varepsilon}{2} \right) \left(D^2(\mathcal{X}_i, \mathcal{X}_j) - \frac{d_j - d_i}{2} \right). \end{aligned}$$

holds with probability at least $1 - \mathrm{e}^{-c_{2,1}n}$. When $n > 2d/\varepsilon$, we have $d_j/n \leq d/n < \varepsilon/2$. In this case, we have both

$$\begin{aligned} D^2(\mathcal{Y}_i, \mathcal{Y}_j) &> D^2(\mathcal{X}_i, \mathcal{X}_j) - \left(\frac{d_j}{n} + \frac{\varepsilon}{2} \right) D^2(\mathcal{X}_i, \mathcal{X}_j) \\ &= \left(1 - \frac{d_j}{n} - \frac{\varepsilon}{2} \right) D^2(\mathcal{X}_i, \mathcal{X}_j) > (1 - \varepsilon)\, D^2(\mathcal{X}_i, \mathcal{X}_j), \qquad (8) \\ D^2(\mathcal{Y}_i, \mathcal{Y}_j) &< D^2(\mathcal{X}_i, \mathcal{X}_j) + \left(-\frac{d_j}{n} + \frac{\varepsilon}{2} \right) D^2(\mathcal{X}_i, \mathcal{X}_j) \\ &= \left(1 - \frac{d_j}{n} + \frac{\varepsilon}{2} \right) D^2(\mathcal{X}_i, \mathcal{X}_j) < (1 + \varepsilon)\, D^2(\mathcal{X}_i, \mathcal{X}_j), \qquad (9) \end{aligned}$$

hold with probability at least $1 - \mathrm{e}^{-c_{2,1}n}$. Note that (8) and (9) hold for any $1 \leq i < j \leq L$. Then the probability is at least $1 - \frac{L(L-1)}{2}\mathrm{e}^{-c_{2,1}n}$. If

$n > \frac{1}{c_{2,1}} \ln \frac{L(L-1)}{2}$, there exists constant c_2 depending only on ε, such that $\frac{L(L-1)}{2} \mathrm{e}^{-c_{2,1} n} < \mathrm{e}^{-c_2 n}$. Take $c_1 := \max\{c_{1,1}, \frac{2}{\varepsilon}, \frac{2}{c_{2,1}}\}$, then when $n > c_1 \max\{d, \ln L\}$, conditions $n > c_{1,1} d$, $n > 2d/\varepsilon$, and $n > \frac{1}{c_{2,1}} \ln \frac{L(L-1)}{2}$ that are required above are all satisfied, the probability is at least $1 - \mathrm{e}^{-c_2 n}$ with $c_2 > 0$. Then we reach the final conclusion.

3. Concentration Inequalities for Gaussian Distribution

Before proving the main results, we first introduce some useful concentration inequalities for Gaussian distribution. Most of them are proved using the following lemma, which provides a strict estimation of the singular values of Gaussian random matrix.

Lemma 5 *[33] Let* $\mathbf{A}$ *be an* $N \times n$ *matrix whose elements* a_{ij} *are independent Gaussian random variables. Then for every* $t \geq 0$*, one has*

$$\mathbb{P}\left(s_{\max}(\mathbf{A}) \geq \sqrt{N} + \sqrt{n} + t\right) \leq \mathrm{e}^{-\frac{t^2}{2}}, \tag{10}$$

and

$$\mathbb{P}\left(s_{\min}(\mathbf{A}) \leq \sqrt{N} - \sqrt{n} - t\right) \leq \mathrm{e}^{-\frac{t^2}{2}}. \tag{11}$$

Based on Lemma 5, we are ready to prove some useful lemmas that will be directly used to prove our theories on RIP of random projection for subspaces. Before doing that, we first define standard Gaussian random matrix and verify that the function satisfying certain condition can be written as a single exponential function.

Definition 6 *A Gaussian random matrix (or vector) has* i.i.d. *zero-mean Gaussian random entries. A standard Gaussian random matrix* $\mathbf{A} \in \mathbb{R}^{n \times N}$ *has* i.i.d. *zero-mean Gaussian random entries with variance* $1/n$*. Each column of* $\mathbf{A}$ *is a standard Gaussian random vector.*

Lemma 6 *Given*

$$f(\varepsilon, n, \tau) = \frac{1}{K} \sum_{k=1}^{K} a_k(\varepsilon, n) \mathrm{e}^{-g_k(\varepsilon, n, \tau)}, \tag{12}$$

if for all k, it holds that

$$h_k(\varepsilon) := \lim_{\tau \to 0} \lim_{n \to \infty} \frac{g_k(\varepsilon, n, \tau)}{n} > 0, \tag{13}$$

$$b_k(\varepsilon) := \lim_{n \to \infty} \frac{\ln a_k(\varepsilon, n)}{n} < h_k(\varepsilon), \tag{14}$$

then there exist universal constants $n_0, c_1 > 0$, and $c_2 > 0$ depending only on ε, such that when $n > n_0, \tau < c_1$, it satisfies that $f(\varepsilon, n, \tau) < \mathrm{e}^{-c_2 n}$.

PROOF The proof is postponed to Appendix 8.1. ∎

Remark 1 *Lemma 6 illustrates that the summation of finite multiple exponential decay functions can always be bounded by a single exponential function.*

The following lemma illustrates that the norm of standard Gaussian random vector concentrates around 1 with high probability, especially when the dimensionality is high.

Lemma 7 *Assume that $\mathbf{a} \in \mathbb{R}^n$ is a standard Gaussian random vector. For any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\left|\|\mathbf{a}\|^2 - 1\right| > \varepsilon\right) < \mathrm{e}^{-c(\varepsilon) n} \tag{15}$$

hold for $n > n_0$, where n_0 and c are constants dependent on ε.

PROOF The proof is postponed to Appendix 8.2. ∎

Furthermore, Corollary 2 generalizes Lemma 7 to case when we project standard Gaussian random vector by orthonormal matrix.

Corollary 2 *Let $\mathbf{a} \in \mathbb{R}^n$ be a standard Gaussian random vector. For any given orthonormal matrix $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_d] \in \mathbb{R}^{n \times d}$ and $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\left|\|\mathbf{V}^{\mathrm{T}}\mathbf{a}\|^2 - \frac{d}{n}\right| > \varepsilon\right) < \mathrm{e}^{-c_2(\varepsilon) n} \tag{16}$$

hold for $n > c_1 d$, where c_1, c_2 are constants dependent on ε.

PROOF The proof is postponed to Appendix 8.3. ∎

Corollary 3 extends Lemma 5 to column-normalized standard Gaussian random matrix. It reveals that a column-normalized Gaussian random matrix is a high-quality approximation to an orthonormal matrix, because all its singular values are very close to 1.

Corollary 3 *Let* $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_k] \in \mathbb{R}^{n\times k}$ *be a standard Gaussian random matrix. Each column of* $\bar{\mathbf{A}}$ *is normalized from the corresponding column of* $\mathbf{A}$*, that is*

$$\bar{\mathbf{A}} = \left[\frac{\mathbf{a}_1}{\|\mathbf{a}_1\|}, \cdots, \frac{\mathbf{a}_k}{\|\mathbf{a}_k\|}\right].$$

Then we can get the bound of the minimum and maximum of the singular value of $\bar{\mathbf{A}}$ *as below*

$$\mathbb{P}\left(s_{\min}^2\left(\bar{\mathbf{A}}\right) < 1-\varepsilon\right) < \mathrm{e}^{-c_{2,1}(\varepsilon)n}, \quad \forall n > c_{1,1}k, \tag{17}$$

$$\mathbb{P}\left(s_{\max}^2\left(\bar{\mathbf{A}}\right) > 1+\varepsilon\right) < \mathrm{e}^{-c_{2,2}(\varepsilon)n}, \quad \forall n > c_{1,2}k, \tag{18}$$

where $c_{1,1}$ *and* $c_{2,1}$*,* $c_{1,2}$ *and* $c_{2,2}$ *are constants dependent on* ε *in* (17) *and* (18)*, respectively.*

PROOF The proof is postponed to Appendix 8.4. ∎

Remark 2 *For* $\mathbf{A} \in \mathbb{R}^{(n-d_0)\times k}$ *be a standard Gaussian random matrix, we have* (18) *hold for* $n > c_{1,2}\max\{k, d_0\}$*, where* $c_{1,2}$ *and* $c_{2,2}$ *are constants dependent on* ε*.*

The following lemma studies a property of Gaussian random projection. Intuitively, it illustrates that if a line and a subspace are perpendicular to each other, they will still be almost perpendicular after Gaussian random projection.

Lemma 8 *Assume* $\mathbf{u}_1 \in \mathbb{R}^N$ *is a unit vector,* $\mathbf{U}_2 \in \mathbb{R}^{N\times d}$ *is an orthonormal matrix, and* $\mathbf{u}_1$ *is perpendicular to* $\mathbf{U}_2$*. Let* $\mathbf{\Phi} \in \mathbb{R}^{n\times N}$ *be a standard Gaussian random matrix. We use* $\mathbf{a}_1 = \mathbf{\Phi}\mathbf{u}_1$ *and* $\mathbf{A}_2 = \mathbf{\Phi}\mathbf{U}_2$ *to denote the projection of* $\mathbf{u}_1$ *and* $\mathbf{U}_2$ *by using* $\mathbf{\Phi}$*. If* $\mathbf{V}_2$ *is an arbitrary orthonormal basis of* $\mathcal{C}(\mathbf{A}_2)$*, then for* $\varepsilon > 0$*, we have*

$$\mathbb{P}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1\right\|^2 > \varepsilon\right) \le \mathrm{e}^{-c_2(\varepsilon)n} \tag{19}$$

hold for $n > c_1(\varepsilon)d$*, where* c_1*,* c_2 *are constants determined by* ε*.*

PROOF The proof is postponed to Appendix 8.5. ∎

Corollary 4 *In Lemma 8, if we further define* $\bar{\mathbf{a}}_1 = \mathbf{a}_1/\|\mathbf{a}_1\|$ *as the normalized projection of* $\mathbf{u}_1$*, then for* $\varepsilon > 0$*, we have*

$$\mathbb{P}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_1\right\|^2 > \varepsilon\right) \leq \mathrm{e}^{-c_2(\varepsilon)n} \tag{20}$$

hold for $n > c_1(\varepsilon)d$*, where* c_1*,* c_2 *are constants determined by* ε*.*

PROOF The proof is postponed to Appendix 8.6. ∎

Remark 3 *Using the same notations in Corollary 4, if we take* $\mathbf{\Phi} \in \mathbb{R}^{(n-d_0)\times N}$*,* $\bar{\mathbf{a}}_1 \in \mathbb{R}^{n-d_0}$ *and* $\mathbf{V}_2 \in \mathbb{R}^{(n-d_0)\times d}$*, then we have still have*

$$\mathbb{P}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_1\right\|^2 > \varepsilon\right) \leq \mathrm{e}^{-c_2(\varepsilon)n}$$

for $n > c_1(\varepsilon)\max\{d, d_0\}$*, where* c_1 *and* c_2 *are constants dependent on* ε*. This can by readily verified by replacing* n *with* $n - d_0$ *in* (20) *and then applying Lemma 6. The detailed proof is postponed to Appendix 8.7.*

Finally we state the independence of random vectors, matrices, and their column-spanned subspaces.

Definition 7 *Two random vectors are independent, if and only if the distribution of any one of them does not have influence on that of the other. Two random matrices are independent, if and only if any two columns of them are independent. Furthermore, we introduce the independence between a random matrix and a subspace, which holds true if and only if the subspace is spanned by the columns of another random matrix that is independent of the first one. Finally, two subspaces are independent, if and only if they are spanned by the columns of two independent random matrices, respectively.*

Lemma 9 *Assume* $\mathbf{U}$ *and* $\mathbf{V}$ *are two matrices satisfying* $\mathbf{U}^{\mathrm{T}}\mathbf{V} = \mathbf{0}$*, and* $\mathbf{\Phi}$ *is a Gaussian random matrix. Then* $\mathbf{\Phi U}$ *and* $\mathbf{\Phi V}$ *are independent. This can be readily verified by calculating the correlation between any two entries in* $\mathbf{\Phi U}$ *and* $\mathbf{\Phi V}$*, respectively. They are all zero.*

4. Proof of Lemma 4

The proof of Lemma 4 is made up of two steps. At first, we derive an accurate expression of the error about estimating $\text{aff}_{\mathcal{Y}}$. Then, the estimate error is bounded by utilizing the concentration inequalities for Gaussian distribution that we derived in the previous section.

Step 1) Let us begin from choosing the bases for the line $\mathcal{X}_1$ and the subspace $\mathcal{X}_2$ and then calculating the affinity after projection.

According to the definition of affinity, $\lambda = \cos\theta$, where θ is the only principal angle between $\mathcal{X}_1$ and $\mathcal{X}_2$. We use $\mathbf{u}$ and $\mathbf{u}_1$ to denote, respectively, the basis of $\mathcal{X}_1$ and a unit vector in $\mathcal{X}_2$, which constructs the principal angle with $\mathbf{u}$. Therefore, $\mathbf{u}$ can be rewritten into the following form

$$\mathbf{u} = \lambda\mathbf{u}_1 + \sqrt{1-\lambda^2}\mathbf{u}_0, \tag{21}$$

where $\mathbf{u}_0$ denotes some unit vector orthogonal to $\mathcal{X}_2$. Based on the above definition, we can choose $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_d]$ as the basis of $\mathcal{X}_2$. Notice that $\{\mathbf{u}_2, \cdots, \mathbf{u}_d\}$ could be freely chosen as long as the orthonormality is satisfied.

After projecting $\mathcal{X}_1$ by random Gaussian matrix, we get subspace $\mathcal{Y}_1$, whose basis vector is

$$\begin{aligned}\mathbf{a} = \mathbf{\Phi u} &= \lambda\mathbf{\Phi u}_1 + \sqrt{1-\lambda^2}\mathbf{\Phi u}_0 \\ &= \lambda\mathbf{a}_1 + \sqrt{1-\lambda^2}\mathbf{a}_0,\end{aligned} \tag{22}$$

where $\mathbf{a}_1 := \mathbf{\Phi u}_1$ and $\mathbf{a}_0 := \mathbf{\Phi u}_0$ are not orthogonal to each other. As for $\mathcal{Y}_2$, considering that $\mathbf{\Phi U}$ is not a set of orthonormal basis, we do orthogonalization by using Gram-Schmidt process. Denote the orthonormalized matrix as $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_d]$. By the definition of Gram-Schmidt process, the first column of $\mathbf{V}$ should be

$$\mathbf{v}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|}, \tag{23}$$

which does not change its direction after the orthogonalization.

Remark 4 *Consider the affinity between two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$, with dimension 1 and $d \geq 1$. Let $\mathbf{v}_1$ and $\mathbf{V}_2$ be the orthonormal basis for $\mathcal{S}_1$ and $\mathcal{S}_2$,*

respectively. Then the affinity equals the norm of the projection of $\mathbf{v}_1$ *onto* $\mathcal{S}_2$*, i.e.,* $\lambda = \|\mathrm{P}_{\mathcal{S}_2}(\mathbf{v}_1)\| = \left\|\mathbf{V}_2^{\mathrm{T}}\mathbf{v}_1\right\|$.

By the definition of affinity and (22), we can calculate the affinity between $\mathcal{Y}_1$ and $\mathcal{Y}_2$ as

$$\begin{aligned}\mathrm{aff}_{\mathcal{Y}}^2 &= \left\|\mathbf{V}^{\mathrm{T}}\frac{\mathbf{a}}{\|\mathbf{a}\|}\right\|^2 \\ &= \frac{1}{\|\mathbf{a}\|^2}\left\|\lambda\mathbf{V}^{\mathrm{T}}\mathbf{a}_1 + \sqrt{1-\lambda^2}\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\right\|^2 \\ &= \frac{1}{\|\mathbf{a}\|^2}\left(\lambda^2\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_1\|^2 + (1-\lambda^2)\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\|^2 + \lambda\sqrt{1-\lambda^2}\|\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_0\|\right). \end{aligned} \tag{24}$$

Because $\mathbf{a}_1$ lies in $\mathcal{Y}_2$, we have

$$\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_1\| = \|\mathbf{a}_1\|. \tag{25}$$

By taking the norm on both sides of (22), we write

$$\|\mathbf{a}\|^2 = \lambda^2\|\mathbf{a}_1\|^2 + (1-\lambda^2)\|\mathbf{a}_0\|^2 + 2\lambda\sqrt{1-\lambda^2}\|\mathbf{a}_0\|\|\mathbf{a}_1\|. \tag{26}$$

Eliminating $\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_1\|$ and $\|\mathbf{a}_1\|$ by inserting (25) and (26) into (24), we get

$$\begin{aligned}\mathrm{aff}_{\mathcal{Y}}^2 &= \frac{1}{\|\mathbf{a}\|^2}\left(\|\mathbf{a}\|^2 - (1-\lambda^2)\|\mathbf{a}_0\|^2 + (1-\lambda^2)\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\|^2\right) \\ &= 1 - (1-\lambda^2)\left(\frac{\|\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2} - \frac{\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2}\right). \end{aligned} \tag{27}$$

Recalling (2) and inserting the estimation into (27), the estimate error is deduced as

$$\begin{aligned}\left|\mathrm{aff}_{\mathcal{Y}}^2 - \overline{\mathrm{aff}}_{\mathcal{Y}}^2\right| &= \left|\mathrm{aff}_{\mathcal{Y}}^2 - \left(\lambda^2 + \frac{d}{n}(1-\lambda^2)\right)\right| \\ &= \left|1-(1-\lambda^2)\left(\frac{\|\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2} - \frac{\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2}\right) - \left(1-(1-\lambda^2)\left(1-\frac{d}{n}\right)\right)\right| \\ &= (1-\lambda^2)\left|\left(1-\frac{d}{n}\right) - \left(\frac{\|\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2} - \frac{\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2}\right)\right|. \end{aligned} \tag{28}$$

Step 2) Before bounding the estimate error by concentration inequalities,

we first split the RHS of (28) into three parts using triangle inequality.

$$\begin{aligned}
&\left|\left(1-\frac{d}{n}\right)-\left(\frac{\|\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2}-\frac{\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2}\right)\right| \\
\leq&\left|\frac{\|\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2}-1\right|+\left|\frac{\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2}-\frac{d}{n}\right| \\
\leq&\left|\frac{\|\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2}-1\right|+\frac{1}{\|\mathbf{a}\|^2}\left|\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\|^2-\frac{d}{n}\right|+\frac{d}{n}\left|\frac{1}{\|\mathbf{a}\|^2}-1\right|. \qquad (29)
\end{aligned}$$

Since both $\mathbf{a}$ and $\mathbf{a}_0$ are standard Gaussian random vectors, by Lemma 7, for $n > n_0$, with probability at least $1-2\mathrm{e}^{-c_{2,1}(\varepsilon_1)n}$, we have

$$\left|\|\mathbf{a}\|^2-1\right|<\varepsilon_1 \qquad (30)$$

and

$$\left|\|\mathbf{a}_0\|^2-1\right|<\varepsilon_1. \qquad (31)$$

Since $\mathbf{u}_0$ is orthogonal to $\mathbf{u}_i$, $i=1,\cdots,d$, by Corollary 2, for $n > c_{1,2}(\varepsilon_2)d$, with probability at least $1-\mathrm{e}^{-c_{2,2}(\varepsilon_2)n}$, we have

$$\left|\left\|\mathbf{V}^{\mathrm{T}}\mathbf{a}_0\right\|^2-\frac{d}{n}\right|<\varepsilon_2. \qquad (32)$$

Using (30), (31), and (32) in (29), for $n > \max\{n_0, c_{1,2}\}d$, with probability at least

$$1-2\mathrm{e}^{-c_{2,1}(\varepsilon_1)n}-\mathrm{e}^{-c_{2,2}(\varepsilon_2)n}, \qquad (33)$$

we have

$$\begin{aligned}
\left|\left(1-\frac{d}{n}\right)-\frac{\|\mathbf{a}_0\|^2}{\|\mathbf{a}\|^2}\left(1-\sum_{i=1}^{d}\cos^2\theta_i\right)\right| \leq&\frac{2\varepsilon_1}{1-\varepsilon_1}+\frac{\varepsilon_2}{1-\varepsilon_1}+\frac{d}{n}\frac{\varepsilon_1}{1-\varepsilon_1} \\
\leq&\left(\frac{3d}{2n}+3\right)\varepsilon_1+\frac{3\varepsilon_2}{2}, \qquad (34)
\end{aligned}$$

where the last inequality holds for $\varepsilon_1 < 1/3$.

To complete the proof, we need to formulate (34) and (33) into the shape of (3) and a single exponential function, respectively. Letting $\varepsilon_1 = \varepsilon_2 =: \varepsilon/6$ and

inserting (34) into (28), we have

$$\begin{aligned}\left|\mathrm{aff}_{\mathcal{Y}}^2-\overline{\mathrm{aff}}_{\mathcal{Y}}^2\right| &\leq \left(1-\lambda^2\right)\left(\left(\frac{3d}{2n}+3\right)\varepsilon_1+\frac{3\varepsilon_2}{2}\right)\\ &\leq \left(1-\lambda^2\right)\left(\left(\frac{3}{2}+3\right)\varepsilon_1+\frac{3\varepsilon_2}{2}\right)\\ &= \left(1-\lambda^2\right)\varepsilon.\end{aligned}$$

By Remark 1, we know that there exist constants c_1, c_2, such that (33) is greater than $1-\mathrm{e}^{-c_2(\varepsilon)n}$ for any $n > c_1(\varepsilon)d$ and close the proof.

5. Proof of Theorem 4

The proof of Theorem 4 is divided into two parts. In subsection 5.1, we will complete the main body of the proof by using an important lemma, which will be proved in subsection 5.2. Before we start, let us introduce some auxiliary variables.

Remark 5 *Assume there are two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$, with dimension $d_1 \leq d_2$. Let $\tilde{\mathbf{U}}_i = [\tilde{\mathbf{u}}_{i,1}, \cdots, \tilde{\mathbf{u}}_{i,d_i}]$ denote any orthonormal matrix for subspace $\mathcal{S}_i, i = 1, 2$. One may do singular value decomposition as $\tilde{\mathbf{U}}_2^{\mathrm{T}}\tilde{\mathbf{U}}_1 = \mathbf{Q}_2\mathbf{\Lambda}\mathbf{Q}_1^{\mathrm{T}}$, where the singular values $\lambda_k = \cos\theta_k, 1 \leq k \leq d_1$ are located on the diagonal of $\mathbf{\Lambda}$, and $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{d_1}$ denote the principal angles between $\mathcal{S}_1$ and $\mathcal{S}_2$. After reshaping, we have*

$$\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1 := \left(\tilde{\mathbf{U}}_2\mathbf{Q}_2\right)^{\mathrm{T}}\tilde{\mathbf{U}}_1\mathbf{Q}_1 = \mathbf{\Lambda} = \left[\begin{array}{ccc} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{d_1} \\ \hline & \mathbf{0} & \end{array}\right],$$

where

$$\mathbf{U}_i := \tilde{\mathbf{U}}_i\mathbf{Q}_i = [\mathbf{u}_{i,1}, \cdots, \mathbf{u}_{i,d_i}], \quad i = 1, 2$$

are the orthonormal basis, which have the closest connection with the affinity between these two subspaces.

Definition 8 (*principal* orthonormal bases) *We refer* $\mathbf{U}_1$ *and* $\mathbf{U}_2$ *as* principal *orthonormal bases for* $\mathcal{S}_1$ *and* $\mathcal{S}_2$*, if they are derived by using the method in Remark 5.*

According to Remark 5, for subspaces $\mathcal{X}_1$ and $\mathcal{X}_2$, we can get their *principal* orthonormal bases $\mathbf{U}_1$ and $\mathbf{U}_2$, respectively. After projection by multiplying a standard Gaussian random matrix $\mathbf{\Phi}$, the original basis matrix changes to $\mathbf{A}_i = \mathbf{\Phi}\mathbf{U}_i = [\mathbf{a}_{i,1}, \cdots, \mathbf{a}_{i,d_i}]$, whose columns are no longer unitary and orthogonal to each other. Then we normalize each columns as $\bar{\mathbf{A}}_i = [\bar{\mathbf{a}}_{i,1}, \cdots, \bar{\mathbf{a}}_{i,d_i}] = \left[\frac{\mathbf{a}_{i,1}}{\|\mathbf{a}_{i,1}\|}, \cdots, \frac{\mathbf{a}_{i,d_i}}{\|\mathbf{a}_{i,d_i}\|}\right]$, whose columns are now unitary but still not orthogonal to each other. However, by Corollary 3, we know that $\bar{\mathbf{A}}_i$ can be used as a good approximation for the orthonormal basis of $\mathcal{Y}_i$. We will see that $\bar{\mathbf{A}}_i$ plays an important role in estimating the affinity after projection.

We also need to define an accurate orthonormal basis for the projected subspace. One efficient way is to process $\bar{\mathbf{A}}_i$ by using Gram-Schmidt orthogonalization, whose result is defined as $\mathbf{V}_i = [\mathbf{v}_{i,1}, \cdots, \mathbf{v}_{i,d_i}], i = 1, 2$.

5.1. Main Body

In order to prove Theorem 4, we need to calculate $\mathrm{aff}_{\mathcal{Y}}^2$ and estimate its bias from $\overline{\mathrm{aff}}_{\mathcal{Y}}^2$. Because $\bar{\mathbf{A}}_i$ is very close to an orthonormal matrix, we may use $\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{A}}_1\|$ to estimate the affinity after projection. By using triangle inequality, we have

$$\left|\mathrm{aff}_{\mathcal{Y}}^2 - \overline{\mathrm{aff}}_{\mathcal{Y}}^2\right| \leq \left|\mathrm{aff}_{\mathcal{Y}}^2 - \left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{A}}_1\right\|_{\mathrm{F}}^2\right| + \left|\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{A}}_1\right\|_{\mathrm{F}}^2 - \overline{\mathrm{aff}}_{\mathcal{Y}}^2\right|. \tag{35}$$

Therefore, the following proof can be divided into three steps. The first step is to bound the error caused by using $\bar{\mathbf{A}}_1$ as an approximation of $\mathbf{V}_1$ to compute the affinity. To do that, we will introduce an important lemma, which is the essence of the proof. The second step is to bound the difference between the approximated affinity and our estimate, which can be derived by using Lemma 4. Finally, we combine these two bounds and complete the proof.

Step 1) For the first item in the RHS of (35), according to the definition of affinity, we have

$$\begin{aligned}\left|\mathrm{aff}_{\mathcal{Y}}^2-\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{A}}_1\right\|_{\mathrm{F}}^2\right| &= \left|\|\mathbf{V}_2^{\mathrm{T}}\mathbf{V}_1\|_{\mathrm{F}}^2-\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{A}}_1\right\|_{\mathrm{F}}^2\right| \\ &= \left|\sum_{k=1}^{d_1}\left(\|\mathbf{V}_2^{\mathrm{T}}\mathbf{v}_{1,k}\|^2-\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\right\|^2\right)\right| \\ &\leq \sum_{k=1}^{d_1}\left|\|\mathbf{V}_2^{\mathrm{T}}\mathbf{v}_{1,k}\|^2-\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\right\|^2\right|. \end{aligned} \tag{36}$$

Lemma 10 *There exist constants $c_{1,1}(\varepsilon_1)$ and $c_{2,1}(\varepsilon_1)>0$ depending only on ε_1, such that for any $n>c_{1,1}(\varepsilon_1)d_2$, we have*

$$\left|\|\mathbf{V}_2^{\mathrm{T}}\mathbf{v}_{1,k}\|^2-\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\right\|^2\right|\leq\left(1-\lambda_k^2\right)\varepsilon_1,\quad\forall k=1,\cdots,d_1 \tag{37}$$

hold with probability at least $1-\mathrm{e}^{-c_{2,1}(\varepsilon_1)n}$.

Proof The proof is postponed to Section 5.2. ∎

Plugging (37) into (36), we have

$$\left|\mathrm{aff}_{\mathcal{Y}}^2-\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{A}}_1\right\|_{\mathrm{F}}^2\right|\leq\varepsilon_1\sum_{k=1}^{d_1}\left(1-\lambda_k^2\right) \tag{38}$$

hold with probability at least $1-d_1\mathrm{e}^{-c_{2,1}(\varepsilon_1)n}$ for any $n>c_{1,1}(\varepsilon_1)d_2$.

Step 2) For the second estimation error in the RHS of (35), we can convert this problem about one subspace $\mathcal{Y}_1$ with dimension d_1 into d_1 subproblems, each of which is about 1-dimensional subspace, and then use Lemma 4 to estimate the error. Denote $\mathcal{X}_{1,k}:=\mathcal{C}\{\mathbf{u}_{1,k}\}$, $\mathcal{Y}_{1,k}:=\mathcal{C}\{\mathbf{a}_{1,k}\}$, $1\leq k\leq d_1$. According to the definition of affinity and Remark 5, we have that the affinity between $\mathcal{X}_{1,k}$ and $\mathcal{X}_2$ is $\|\mathbf{U}_2^{\mathrm{T}}\mathbf{u}_{1,k}\|=\lambda_k$, and the affinity between $\mathcal{Y}_{1,k}$ and $\mathcal{Y}_2$ is equal to $\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\right\|$. Using Lemma 4, where $\mathcal{X}_{1,k}$ and $\mathcal{X}_2$ are the original subspaces and $\mathcal{Y}_{1,k}$ and $\mathcal{Y}_2$ are the projected subspaces, we have

$$\left|\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\right\|^2-\overline{\mathrm{aff}}_{\mathcal{Y}_k}^2\right|\leq\left(1-\lambda_k^2\right)\varepsilon_2,\quad k=1,\cdots,d_1 \tag{39}$$

hold with probability at least $1-d_1\mathrm{e}^{-c_{2,2}(\varepsilon_2)n}$ for any $n>c_{1,2}(\varepsilon_2)d_2$, where

$$\overline{\mathrm{aff}}_{\mathcal{Y}_k}^2:=\lambda_k^2+\frac{d_2}{n}\left(1-\lambda_k^2\right). \tag{40}$$

Plugging the definition of $\overline{\text{aff}}^2_{\mathcal{Y}}$ in (4), (39), and (40) into the second estimation error, we have

$$\begin{aligned}
\left|\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{A}}_1\right\|_{\mathrm{F}}^2 - \overline{\text{aff}}^2_{\mathcal{Y}}\right| &= \left|\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{A}}_1\right\|_{\mathrm{F}}^2 - \left(\text{aff}^2_{\mathcal{X}} + \frac{d_2}{n}(d_1 - \text{aff}^2_{\mathcal{X}})\right)\right| \\
&= \left|\sum_{k=1}^{d_1}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\right\|^2 - \left(\lambda_k^2 + \frac{d_2}{n}(1-\lambda_k^2)\right)\right)\right| \\
&\leq \sum_{k=1}^{d_1}\left|\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\right\|^2 - \overline{\text{aff}}^2_{\mathcal{Y}_k}\right| \\
&\leq \varepsilon_2 \sum_{k=1}^{d_1}(1-\lambda_k^2). \qquad (41)
\end{aligned}$$

Step 3) Combining (38) and (41) together into (35), we have that for any $n > \max\{c_{1,1}(\varepsilon_1), c_{1,2}(\varepsilon_2)\}d_2$,

$$\begin{aligned}
\left|\text{aff}^2_{\mathcal{Y}} - \overline{\text{aff}}^2_{\mathcal{Y}}\right| &\leq (\varepsilon_1 + \varepsilon_2)\sum_{k=1}^{d_1}\left(1-\lambda_k^2\right) \\
&= (\varepsilon_1 + \varepsilon_2)\left(d_1 - \text{aff}^2_{\mathcal{X}}\right) \qquad (42)
\end{aligned}$$

hold with probability at least $1 - d_1\mathrm{e}^{-c_{2,1}(\varepsilon_1)n} - d_1\mathrm{e}^{-c_{2,2}(\varepsilon_2)n}$. Let $\varepsilon_1 = \varepsilon_2 =: \varepsilon/2$, according to Remark 1, one can easily verify that there exist constants c_1, c_2 depending only on ε, when $n > c_1(\varepsilon)d_2$,

$$\left|\text{aff}^2_{\mathcal{Y}} - \overline{\text{aff}}^2_{\mathcal{Y}}\right| \leq \left(d_1 - \text{aff}^2_{\mathcal{X}}\right)\varepsilon$$

holds with probability at least $1 - \mathrm{e}^{-c_2(\varepsilon)n}$. Then we complete the proof.

5.2. Proof of Lemma 10

In order to improve the readability of the proof, we define intensively all the variables required in advance. Not that some variables defined before are also summarized here to make this part self-contained.

We use $\mathcal{X}_1$ and $\mathcal{X}_2$ to denote the subspaces before projection, with dimensions $d_1 \leq d_2$. The *principal* orthonormal bases for $\mathcal{X}_1$ and $\mathcal{X}_2$ are denoted as $\mathbf{U}_1$ and $\mathbf{U}_2$, respectively. The kth column of $\mathbf{U}_i$ is denoted as $\mathbf{u}_{i,k}$, which spans a 1-dimensional subspace denoted as $\mathcal{X}_{i,k}, k = 1, \cdots, d_i, i = 1, 2$. In addition,

we define $\mathbf{U}_{i,1:k}$ as the matrix composed of the first k columns of $\mathbf{U}_i$. That is $\mathbf{U}_{i,1:k} = [\mathbf{u}_{i,1}, \cdots, \mathbf{u}_{i,k}], 1 \leq k \leq d_i, i = 1, 2$. The subspace spanned by the columns of $\mathbf{U}_{i,1:k}$ is denoted as $\mathcal{X}_{i,1:k} = \mathcal{C}(\mathbf{U}_{i,1:k})$.

We use $\mathcal{Y}_1$ and $\mathcal{Y}_2$ to denote the subspaces after projection, respectively, from $\mathcal{X}_1$ and $\mathcal{X}_2$ by using a standard Gaussian random matrix $\mathbf{\Phi}$. The dimensions of $\mathcal{Y}_1$ and $\mathcal{Y}_2$ stay to be d_1 and d_2 with probability 1. $\mathbf{A}_i = \mathbf{\Phi}\mathbf{U}_i$ is a basis for $\mathcal{Y}_i$ and its kth column is denoted as $\mathbf{a}_{i,k}$, which spans a 1-dimensional subspace denoted as $\mathcal{Y}_{i,k} = \mathcal{C}(\mathbf{a}_{i,k})$. We define $\mathbf{A}_{i,1:k} = [\mathbf{a}_{i,1}, \cdots, \mathbf{a}_{i,k}]$ as the composition of the first k columns of $\mathbf{A}_i$. The subspace spanned by the columns in $\mathbf{A}_{i,1:k}$ is denoted as $\mathcal{Y}_{i,1:k} = \mathcal{C}(\mathbf{A}_{i,1:k})$.

We use $\bar{\mathbf{A}}_1$ and $\bar{\mathbf{A}}_2$ to denote the column-normalized result of $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively. $\mathbf{V}_1$ and $\mathbf{V}_2$ are defined as the orthonormalized result of $\bar{\mathbf{A}}_1$ and $\bar{\mathbf{A}}_2$, respectively, by using Gram-Schmidt orthogonalization. As a consequence, $\mathbf{V}_i$ provides an orthonormal basis for $\mathcal{Y}_i$. Similarly, $\mathbf{V}_{i,1:k}$ denotes the matrix composed of the first k columns of $\mathbf{V}_i$.

Let's start the proof of Lemma 10 from the LHS of (37). According to the definition of $\mathbf{V}_2, \mathbf{v}_{1,k}, \bar{\mathbf{a}}_{1,k}$, and Remark 4, we have

$$\|\mathbf{V}_2^{\mathrm{T}}\mathbf{v}_{1,k}\|^2 = \|\mathrm{P}_{\mathcal{Y}_2}(\mathbf{v}_{1,k})\|^2 = 1 - \left\|\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{v}_{1,k})\right\|^2, \tag{43}$$

$$\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\|^2 = \|\mathrm{P}_{\mathcal{Y}_2}(\bar{\mathbf{a}}_{1,k})\|^2 = 1 - \left\|\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\bar{\mathbf{a}}_{1,k})\right\|^2. \tag{44}$$

As a consequence, the LHS of (37) is derived as the difference of squared norm of the projection of $\mathbf{v}_{1,k}$ and $\bar{\mathbf{a}}_{1,k}$ onto to the orthogonal complement of $\mathcal{Y}_2$, i.e.

$$\left|\|\mathbf{V}_2^{\mathrm{T}}\mathbf{v}_{1,k}\|^2 - \|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\|^2\right| = \left|\left\|\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{v}_{1,k})\right\|^2 - \left\|\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\bar{\mathbf{a}}_{1,k})\right\|^2\right|. \tag{45}$$

In order to analyze $\mathbf{v}_{1,k}$, we take a close look at the Gram-Schmidt orthogonalization process. We introduce

$$\alpha_k := \|\mathrm{P}_{\mathcal{Y}_{1,1:k-1}}(\bar{\mathbf{a}}_{1,k})\| = \|\mathbf{V}_{1,1:k-1}^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\| \tag{46}$$

as the cosine of the only principal angle between $\mathcal{Y}_{1,k}$ and $\mathcal{Y}_{1,1:k-1}$, and

$$\mathbf{b}_k := \frac{1}{\alpha_k}\mathrm{P}_{\mathcal{Y}_{1,1:k-1}}(\bar{\mathbf{a}}_{1,k}) \tag{47}$$

as a unit vector along the direction of the projection of $\mathbf{a}_{1,k}$ onto $\mathcal{Y}_{1,1:k-1}$. As a consequence, the Gram-Schmidt orthogonalization process is represented by

$$\begin{aligned}\bar{\mathbf{a}}_{1,k} &= \mathrm{P}_{\mathcal{Y}_{1,1:k-1}}(\bar{\mathbf{a}}_{1,k}) + \mathrm{P}_{\mathcal{Y}_{1,1:k-1}^{\perp}}(\bar{\mathbf{a}}_{1,k}) \\ &= \alpha_k \mathbf{b}_k + \sqrt{1-\alpha_k^2}\mathbf{v}_{1,k}. \end{aligned} \tag{48}$$

Then we introduce

$$\hat{\lambda}_k := \|\mathrm{P}_{\mathcal{Y}_2}(\bar{\mathbf{a}}_{1,k})\| = \|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\|, \tag{49}$$

$$\beta_k := \|\mathrm{P}_{\mathcal{Y}_2}(\mathbf{b}_k)\| = \|\mathbf{V}_2^{\mathrm{T}}\mathbf{b}_k\| \tag{50}$$

to denote, respectively, the cosine of the only principal angle between $\mathcal{Y}_{1,k}$ and $\mathcal{Y}_2$ and that between $\mathcal{C}\{\mathbf{b}_k\}$ and $\mathcal{Y}_2$. Now projecting both side of (48) on the orthogonal complement of $\mathcal{Y}_2$, we have

$$\sqrt{1-\hat{\lambda}_k^2}\bar{\mathbf{a}}_{1,k}^{\perp} = \alpha_k\sqrt{1-\beta_k^2}\mathbf{b}_k^{\perp} + \sqrt{1-\alpha_k^2}\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{v}_{1,k}), \tag{51}$$

where $\bar{\mathbf{a}}_{1,k}^{\perp}$ and $\mathbf{b}_k^{\perp}$ denotes, respectively, the unit vectors along $\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\bar{\mathbf{a}}_{1,k})$ and $\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{b}_k)$.

Moving the first item in the RHS of (51) to the LHS and then taking norm on both sides, we get

$$(1-\alpha_k^2)\|\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{v}_{1,k})\|^2 = 1-\hat{\lambda}_k^2 + \alpha_k^2\left(1-\beta_k^2\right) - 2\alpha_k\sqrt{1-\hat{\lambda}_k^2}\sqrt{1-\beta_k^2}\langle\bar{\mathbf{a}}_{1,k}^{\perp}, \mathbf{b}_k^{\perp}\rangle. \tag{52}$$

In addition, the norm of the projection of $\bar{\mathbf{a}}_{1,k}$ onto to $\mathcal{Y}_2^{\perp}$ could be directly represented by using $\hat{\lambda}_k$ as

$$\|\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\bar{\mathbf{a}}_{1,k})\|^2 = 1-\hat{\lambda}_k^2. \tag{53}$$

Inserting both (52) and (53) into (45), we write

$$\begin{aligned} &\left|\|\mathbf{V}_2^{\mathrm{T}}\mathbf{v}_{1,k}\|^2 - \|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\|^2\right| \\ =&\left|\frac{1-\hat{\lambda}_k^2 + \alpha_k^2\left(1-\beta_k^2\right) - 2\alpha_k\sqrt{1-\hat{\lambda}_k^2}\sqrt{1-\beta_k^2}\langle\bar{\mathbf{a}}_{1,k}^{\perp}, \mathbf{b}_k^{\perp}\rangle}{1-\alpha_k^2} - (1-\hat{\lambda}_k^2)\right| \\ \leq&\frac{\alpha_k^2\left(1-\hat{\lambda}_k^2+1-\beta_k^2\right) + 2\left|\alpha_k\sqrt{1-\hat{\lambda}_k^2}\sqrt{1-\beta_k^2}\langle\bar{\mathbf{a}}_{1,k}^{\perp}, \mathbf{b}_k^{\perp}\rangle\right|}{1-\alpha_k^2}. \end{aligned} \tag{54}$$

Using the fact that geometric mean is no more than arithmetic mean and $\alpha_k^2 < 1/3$, which will be verified soon that α_k^2 is a small quantity, we further reshape (54) as

$$\left| \|\mathbf{V}_2^{\mathrm{T}} \mathbf{v}_{1,k}\|^2 - \left\|\mathbf{V}_2^{\mathrm{T}} \bar{\mathbf{a}}_{1,k}\right\|^2 \right| \leq \frac{3}{2} \left(1 - \hat{\lambda}_k^2 + 1 - \beta_k^2\right) \left(\alpha_k^2 + \left|\alpha_k \langle \bar{\mathbf{a}}_{1,k}^{\perp}, \mathbf{b}_k^{\perp} \rangle\right|\right). \quad (55)$$

In the following, we will estimate the four quantities of $1 - \hat{\lambda}_k^2$, α_k^2, $1 - \beta_k^2$, and $\langle \bar{\mathbf{a}}_{1,k}^{\perp}, \mathbf{b}_k^{\perp} \rangle$ separately.

Let us first consider $1 - \hat{\lambda}_k^2$. Recalling its definition in (49), we have already estimate $\hat{\lambda}_k^2$ in (39). Inserting (40) into (39), with probability at least $1 - \mathrm{e}^{-c_{2,1}(\varepsilon_1)n}$, we have for any $n > c_{1,1}(\varepsilon_1)d_2$

$$\lambda_k^2 + \frac{d_2}{n}\left(1 - \lambda_k^2\right) - \hat{\lambda}_k^2 \leq \left(1 - \lambda_k^2\right)\varepsilon_1, \quad (56)$$

Using basic algebra, (56) is reshaped to

$$\begin{aligned} 1 - \hat{\lambda}_k^2 \leq & 1 - \lambda_k^2 - \frac{d_2}{n}\left(1 - \lambda_k^2\right) + \left(1 - \lambda_k^2\right)\varepsilon_1 \\ = & \left(1 - \lambda_k^2\right)\left(1 - \frac{d_2}{n} + \varepsilon_1\right) \\ < & \left(1 - \lambda_k^2\right)\left(1 + \varepsilon_1\right), \quad k = 1, \cdots, d_1. \end{aligned} \quad (57)$$

The bound of (57) looks direct because $\hat{\lambda}_k$ denotes the affinity compressed from λ_k. According to Lemma 4, the former can be estimated by the latter.

Second let us check $\alpha_k = \|\mathbf{V}_{1,1:k-1}^{\mathrm{T}} \bar{\mathbf{a}}_{1,k}\|$. Intuitively, because $\mathcal{X}_{1,1:k-1}$ is orthogonal to $\mathcal{X}_{1,k}$, the new subspaces $\mathcal{Y}_{1,1:k-1}$ (projected from $\mathcal{X}_{1,1:k-1}$) and $\mathcal{Y}_{1,k}$ (projected from $\mathcal{X}_{1,k}$) are approximately orthogonal to each other. Actually, $\mathbf{V}_{1,1:k-1}$ and $\bar{\mathbf{a}}_{1,k}$ satisfy all conditions in Corollary 4. As a consequence, there exist constants $c_{1,2}(\varepsilon_2)$, $c_{2,2}(\varepsilon_2)$, such that for any $\varepsilon_2 < \frac{1}{3}$ and $n > c_{1,2}(\varepsilon_2)d_1 > c_{1,2}(\varepsilon_2)(k-1)$, we have $\alpha_k^2 < \varepsilon_2$ hold with probability at least $1 - \mathrm{e}^{-c_{2,2}(\varepsilon_2)n}$.

Next we consider $1 - \beta_k^2$, which is also bounded by $1 - \lambda_k^2$, Notice that $\mathbf{b}_k$ lies in $\mathcal{Y}_{1,1:k-1} \subset \mathcal{Y}_{1,1:k}$ and β_k is the norm of the projection of $\mathbf{b}_k$ onto $\mathcal{Y}_2$. Because the minimum of the norm of the projection of a unit vector in $\mathcal{Y}_{1,1:k}$ onto $\mathcal{Y}_2$ approximates λ_k, $1 - \beta_k^2$ should be very close to $1 - \lambda_k^2$. The difference between them can be bounded by the following lemma.

Lemma 11 *For any $n > c_{1,3}(\varepsilon_3)d_2$, we have*

$$1-\beta_k^2 \le \left(1-\lambda_k^2\right)(1+\varepsilon_3) \tag{58}$$

holds with probability at least $1-\mathrm{e}^{-c_{2,3}(\varepsilon_3)n}$.

PROOF The proof is postponed to Appendix 8.8. ∎

Finally, as for the last term to be estimated, $\langle \bar{\mathbf{a}}_{1,k}^{\perp}, \mathbf{b}_k^{\perp}\rangle$ is proved to be a small quantity in Lemma 12. Intuitively, $\bar{\mathbf{a}}_{1,k}^{\perp}$ and $\mathbf{b}_k^{\perp}$ is unit projections of $\bar{\mathbf{a}}_{1,k} \in \mathcal{Y}_{1,k}$ and $\mathbf{b}_k \in \mathcal{Y}_{1,1:k-1}$, respectively, onto $\mathcal{Y}_2^{\perp}$. Consequently, the inner product between $\bar{\mathbf{a}}_{1,k}^{\perp}$ and $\mathbf{b}_k^{\perp}$ should be very small if $\mathcal{Y}_{1,k}$ and $\mathcal{Y}_{1,1:k-1}$, which are independent with each other, are both independent with $\mathcal{Y}_2^{\perp}$.

Lemma 12 *There exist constants $c_{1,4}(\varepsilon_4)$, $c_{2,4}(\varepsilon_4)$, such that for any $n > c_{1,4}(\varepsilon_4)d_2$, we have $\left|\langle \bar{\mathbf{a}}_{1,k}^{\perp}, \mathbf{b}_k^{\perp}\rangle\right|^2 < \varepsilon_4$ holds with probability at least $1-\mathrm{e}^{-c_{2,4}(\varepsilon_4)n}$.*

PROOF The proof is postponed to Appendix 8.9. ∎

Now, we are ready to complete the proof by using the concentration properties derived above. Plugging (57), (58), $\alpha_k^2 < \varepsilon_2$, and $\left|\langle \bar{\mathbf{a}}_{1,k}^{\perp}, \mathbf{b}_k^{\perp}\rangle\right|^2 < \varepsilon_4$ into (55), we have for any $n > \max\{c_{1,l}(\varepsilon_l)\}d_2$, $l = 1,2,3,4$,

$$\left|\|\mathbf{V}_2^{\mathrm{T}}\mathbf{v}_{1,k}\|^2 - \left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\right\|^2\right| \le \frac{3}{2}\left(1-\lambda_k^2\right)(2+\varepsilon_1+\varepsilon_3)\left(\varepsilon_2+\sqrt{\varepsilon_2\varepsilon_4}\right)$$

hold with probability at least $1-\sum_{l=1}^{4}\mathrm{e}^{-c_{2,l}(\varepsilon_l)n}$. Let $\varepsilon_1 < 1$ and $\varepsilon_3 < 1$, $\varepsilon_2 = \varepsilon_4 =: \varepsilon/12$, then we have

$$\left|\|\mathbf{V}_2^{\mathrm{T}}\mathbf{v}_{1,k}\|^2 - \left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_{1,k}\right\|^2\right| \le \left(1-\lambda_k^2\right)\varepsilon. \tag{59}$$

According to Remark 1, we claim that there exist constants c_1, c_2, such that for any $n > c_1(\varepsilon)d_2$, (59) holds with probability at least $1-\mathrm{e}^{-c_2(\varepsilon)n}$.

6. Related Works

Our earlier results on the RIP of subspaces in [21] are cited below.

Theorem 5 *Suppose $\mathcal{X}_1, \mathcal{X}_2 \subset \mathbb{R}^N$ are two subspaces with dimension $d_1 \leq d_2$, respectively. If $\mathcal{X}_1$ and $\mathcal{X}_2$ are projected into $\mathbb{R}^n$ by a Gaussian random matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times N}$, $\mathcal{X}_k \xrightarrow{\mathbf{\Phi}} \mathcal{Y}_k, k = 1, 2$, then we have*

$$(1-\varepsilon)D_{\mathcal{X}}^2 \leq D_{\mathcal{Y}}^2 \leq (1+\varepsilon)D_{\mathcal{X}}^2,$$

with probability at least

$$1 - \frac{4d_1}{(\varepsilon - d_2/n)^2 n},$$

when n is large enough.

Theorem 6 *For any set composed by L subspaces $\mathcal{X}_1, \cdots, \mathcal{X}_L \in \mathbb{R}^N$ of dimension no more than d, if they are projected into $\mathbb{R}^n$ by a Gaussian random matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times N}$, $\mathcal{X}_k \xrightarrow{\mathbf{\Phi}} \mathcal{Y}_k, k = 1, \cdots, L$, and $d \ll n < N$, then we have*

$$(1-\varepsilon)D^2(\mathcal{X}_i, \mathcal{X}_j) \leq D^2(\mathcal{Y}_i, \mathcal{Y}_j) \leq (1+\varepsilon)D^2(\mathcal{X}_i, \mathcal{X}_j), \quad \forall i, j$$

with probability at least

$$1 - \frac{2dL(L-1)}{(\varepsilon - d/n)^2 n},$$

when n is large enough.

Compared with the our previous results, this paper has the following two main improvements. First, because we use more advanced random matrix theories and deal with the error more skillfully, the probability bound $1 - \mathrm{e}^{-\mathcal{O}(n)}$ derived in this paper is much tighter than the $1 - \mathcal{O}(1/n)$ in the previous work, where we used Chebyshev inequality. Such improvement provides a more accurate law of magnitude of the dimensions in this random projection problem, and the improved probability bound is optimum, if one compares it with the analogical conclusions in the theory of Compressed Sensing. Second, Theorem 6 requires $n \gg d$, but it does not specify how large n should be or the connection between ε, d, L, and the lower bound of n. In comparison, Theorem 1 in this paper rigorously clarifies that the conclusion will hold as long as n is larger than $c_1(\varepsilon)\max\{d, \ln L\}$.

Specifically, in the proof of Theorem 3 in [21], we need $\frac{d}{n} = o(1)$ to drop P_2 and $\frac{d}{n}$ in P_1 to derive

$$\mathbb{P}\left(\left|\mathrm{aff}_{\mathcal{Y}}^2 - \overline{\mathrm{aff}}^2\right| > (1-\lambda^2)\left(1-\frac{d}{n}\right)\lambda^2\varepsilon\right) \lesssim \frac{4}{\varepsilon^2 n},$$

so n cannot scale linearly with d. In this work, we are mainly concerned with the case $d \sim n$, and derive an exponential decay bound. Hence, except some geometry bases, the analysis in this work and that in [21] is completely different. Moreover, in this work, we only need $n > c \ln L$, while in [21], n needs to be at least $O(L^2)$ to make the failure probability less than 1. Hence, besides the improvement on the failure probability, this paper also improves the connection between n and d, L.

7. Conclusion

In this paper, we utilize the random matrix theory to rigorously prove the RIP of Gaussian random compressions for low-dimensional subspaces. Mathematically, we demonstrate that as long as the dimension after compression n is larger than $c_1(\varepsilon)\max\{d, \ln L\}$, with probability no less than $1-\mathrm{e}^{-c_2(\varepsilon)n}$, the distance between any two subspaces after compression remains almost unchanged. The probability bound $1-\mathrm{e}^{-\mathcal{O}(n)}$ is optimum in the asymptotic sense, in comparison with the analogical optimum theoretical result of RIP in Compressed Sensing. Our work can provide a solid theoretical foundation for Compressed Subspace Clustering and other low-dimensional subspace related problems.

8. Appendix

8.1. Proof of Lemma 6

We first prove the special case that $K = 1$, i.e., $f = a\mathrm{e}^{-g}$ for short. According to (13), (14), and the definition of limitation, there exist constants n_0 and $c_1 > 0$ depending only on ε. When $n > n_0$, $\tau < c_1$, we have $\frac{g}{n} > h - \frac{h-b}{3}$, and

$\frac{\ln a}{n} < b + \frac{h-b}{3}$. Let $c_2 := \frac{h-b}{3} > 0$ depending only on ε, we can have

$$\begin{aligned} f = a\mathrm{e}^{-g} &= \mathrm{e}^{-(g-\ln a)} = \exp\left(-n\left(\frac{g}{n} - \frac{\ln a}{n}\right)\right) \\ &\leq \exp\left(-n\left(h - \frac{h-b}{3} - b - \frac{h-b}{3}\right)\right) \\ &= \exp\left(-\frac{h-b}{3}n\right) = \mathrm{e}^{-c_2 n}. \end{aligned}$$

Now we consider the general case of arbitrary K. According to the above analysis, we have that, for each term of f, there exist constants $n_{0,k}, c_{1,k} > 0$, and $c_{2,k} > 0$ depending only on ε. When $n > n_{0,k}$, $\tau < c_{1,k}$, it satisfies that $a_k \mathrm{e}^{-g_k} < \mathrm{e}^{-c_{2,k} n}$. Let $n_0 := \max_k n_{0,k}$, $c_1 := \min_k c_{1,k} > 0$, $c_2 := \min_k c_{2,k} > 0$. Then when $n > n_0$, $\tau < c_1$, we have that

$$f = \frac{1}{K}\sum_{k=1}^{K} a_k \mathrm{e}^{-g_k} < \frac{1}{K}\sum_{k=1}^{K} \mathrm{e}^{-n c_{2,k}} \leq \mathrm{e}^{-c_2 n},$$

and complete the proof.

8.2. Proof of Lemma 7

Regarding $\sqrt{n}\mathbf{a}$ as a matrix belonging to $\mathbb{R}^{n\times 1}$, and using Lemma 5, we have that with probability at least $1 - 2\mathrm{e}^{-\frac{t^2}{2}}$,

$$\sqrt{n} - 1 - t \leq s_{\min}(\sqrt{n}\mathbf{a}) = \sqrt{n}\|\mathbf{a}\| = s_{\max}(\sqrt{n}\mathbf{a}) \leq \sqrt{n} + 1 + t.$$

Taking square and subtracting n from both sides, we have

$$-\left(2\sqrt{n}(1+t) + (1+t)^2\right) \leq -\left(2\sqrt{n}(1+t) - (1+t)^2\right) \leq n\|\mathbf{a}\|^2 - n \leq 2\sqrt{n}(1+t) + (1+t)^2,$$

with probability at least $1 - 2\mathrm{e}^{-\frac{t^2}{2}}$.

By choosing ε satisfying $n\varepsilon = 2\sqrt{n}(1+t) + (1+t)^2$, we can get

$$t = \sqrt{n}\left(\sqrt{1+\varepsilon} - 1 - 1/\sqrt{n}\right). \tag{60}$$

When $n > \left(\frac{1}{\sqrt{\varepsilon+1}-1}\right)^2 =: n_{0,1}$, we have $t > 0$. Substituting this equation into the expression of probability, we have

$$\mathbb{P}\left(\left|\|\mathbf{a}\|^2 - 1\right| > \varepsilon\right) < 2\exp\left(-n\left(\sqrt{1+\varepsilon} - 1 - 1/\sqrt{n}\right)^2/2\right). \tag{61}$$

According to Lemma 6, there exist constants $n_{0,2}$ and c dependent on ε, such that the RHS of (61) is smaller than e^{-cn}. Taking $n_0 = \max\{n_{0,1}, n_{0,2}\}$, we complete the proof.

8.3. Proof of Corollary 2

Proof Notice that $\sqrt{\frac{n}{d}}\mathbf{V}^{\mathrm{T}}\mathbf{a} \in \mathbb{R}^d$ is a standard Gaussian random vector. As a consequence, according to (61) in the proof of Lemma 7, by replacing ε with $\frac{n\varepsilon}{d}$, we have

$$\mathbb{P}\left(\left|\left\|\sqrt{\frac{n}{d}}\mathbf{V}^{\mathrm{T}}\mathbf{a}\right\|^2 - 1\right| > \frac{n\varepsilon}{d}\right) < 2\exp\left(-d\left(\sqrt{1+\frac{n}{d}\varepsilon} - 1 - 1/\sqrt{d}\right)^2/2\right), \tag{62}$$

where $d > \left(\frac{1}{\sqrt{\frac{n}{d}\varepsilon+1}-1}\right)^2$ is required. In order to satisfy this requirement, i.e., $\left(\frac{1}{\sqrt{\frac{n}{d}\varepsilon+1}-1}\right)^2 < 1 \leq d$, we need $n > \frac{3d}{\varepsilon} =: c_{1,1}d$. According to Lemma 6, there exist constants $c_{1,2}$, c_2 dependent on ε, such that the RHS of (62) is smaller than $\mathrm{e}^{-c_2 n}$. Taking $c_1 := \max\{c_{1,1}, c_{1,2}\}$ and dividing both sides of the expression in $\mathbb{P}(\cdot)$ in (62) by n/d, we complete the proof. ∎

8.4. Proof of Corollary 3

In order to bound $s^2_{\min}\left(\bar{\mathbf{A}}\right)$, noticing that

$$s^2_{\min}\left(\bar{\mathbf{A}}\right) \geq \frac{s^2_{\min}\left(\mathbf{A}\right)}{\max\limits_i \left\|\mathbf{a}_i\right\|^2}, \tag{63}$$

we may turn to estimate $s^2_{\min}\left(\mathbf{A}\right)$ and $\max_i \left\|\mathbf{a}_i\right\|^2$ separately.

We begin from estimating $s^2_{\min}\left(\mathbf{A}\right)$. According to Lemma 5, with probability at least $1 - \mathrm{e}^{-t^2/2}$, we have

$$s^2_{\min}\left(\mathbf{A}\right) \geq \frac{1}{n}\left(\sqrt{n} - \sqrt{k} - t\right)^2 = \left(1 - \sqrt{k/n} - t/\sqrt{n}\right)^2. \tag{64}$$

Let $1 - \varepsilon_1$ be the RHS of (64) then we have

$$t = \sqrt{n}\left(1 - \sqrt{1-\varepsilon_1} - \sqrt{k/n}\right). \tag{65}$$

When $n > \frac{k}{\left(1-\sqrt{1-\varepsilon_1}\right)^2} =: \hat{c}_{0,1}k$, we have $t > 0$. Plugging (65) into $\mathrm{e}^{-t^2/2}$, we can get the probability that (65) violates as $\exp\left(-n\left(1-\sqrt{k/n}-\sqrt{1-\varepsilon_1}\right)^2/2\right)$. According to Lemma 6, the above probability can be bounded by $\mathrm{e}^{-\hat{c}_{2,1}(\varepsilon_1)n}$ for $n > \hat{c}_{1,1}k$. Then we have

$$s_{\min}^2\left(\mathbf{A}\right) \geq 1-\varepsilon_1 \tag{66}$$

hold for $n > \max\{\hat{c}_{1,1}, \hat{c}_{0,1}\}k =: \hat{c}_3 k$ with probability at least $1-\mathrm{e}^{-\hat{c}_{2,1}(\varepsilon_1)n}$.

Next we estimate $\max_i \|\mathbf{a}_i\|^2$. According to Lemma 7, with probability at least $1-k\mathrm{e}^{-\hat{c}_{2,2}(\varepsilon_2)n}$, for $n > n_{0,1}$, we have

$$\max_i \|\mathbf{a}_i\|^2 \leq 1+\varepsilon_2. \tag{67}$$

Plugging (66) and (67) into (63), we have

$$s_{\min}^2\left(\bar{\mathbf{A}}\right) \geq \frac{1-\varepsilon_1}{1+\varepsilon_2} = 1-\frac{\varepsilon_1+\varepsilon_2}{1+\varepsilon_2}.$$

Let $\varepsilon_1 = \varepsilon_2 =: \varepsilon/2$, with probability at least $1-\mathrm{e}^{-\hat{c}_{2,1}(\varepsilon/2)n} - k\mathrm{e}^{-\hat{c}_{2,2}(\varepsilon/2)n}$, we have

$$s_{\min}^2\left(\bar{\mathbf{A}}\right) \geq 1-(\varepsilon/2+\varepsilon/2) = 1-\varepsilon. \tag{68}$$

Take $c_{1,1} := \max\{\hat{c}_3, n_{0,1}\}$ and we prove the first part of this corollary.

In order to bound $s_{\max}^2\left(\bar{\mathbf{A}}\right)$, following the same approach, we could derive step by step the counterparts of (63), (66), and (67), respectively, as

$$s_{\max}^2\left(\bar{\mathbf{A}}\right) \leq \frac{s_{\max}^2\left(\mathbf{A}\right)}{\min_i \|\mathbf{a}_i\|^2}, \tag{69}$$

$$s_{\max}^2\left(\mathbf{A}\right) \leq 1+\varepsilon_1 \tag{70}$$

for $n > \hat{c}_4 k$ with probability at least

$$1-\exp\left(-n\left(1+\sqrt{k/n}-\sqrt{1+\varepsilon_1}\right)^2/2\right) > 1-\mathrm{e}^{-\hat{c}_{2,3}n},$$

and

$$\min_i \|\mathbf{a}_i\|^2 \geq 1-\varepsilon_2 \tag{71}$$

for $n > n_{0,2}$ with probability at least $1 - k\mathrm{e}^{-\hat{c}_{2,2}(\varepsilon_2)n} - \mathrm{e}^{-\hat{c}_{2,3}(\varepsilon_1)n}$. Then we have

$$s_{\max}^2\left(\bar{\mathbf{A}}\right) \leq \frac{1+\varepsilon_1}{1-\varepsilon_2} = 1 + \frac{\varepsilon_1+\varepsilon_2}{1-\varepsilon_2}. \tag{72}$$

Similarly reshaping (72) and letting $\varepsilon_2 \leq 1/2$, $\varepsilon_1 = \varepsilon_2 =: \varepsilon/4$, taking $c_{1,2} := \max\{\hat{c}_4, n_{0,2}\}$, we prove the second part of the corollary.

8.5. Proof of Lemma 8

Using the orthogonality between $\mathbf{u}_1$ and $\mathbf{U}_2$, $\mathcal{C}(\mathbf{A}_2)$ is independent with $\mathbf{a}_1$. As an orthonormal basis of such subspace, $\mathbf{V}_2$ is also independent with $\mathbf{a}_1$. Then, according to Definition 6, $\mathbf{a}_1$ conditioned on $\mathbf{V}_2$ is still a standard Gaussian random vector. As a consequence, $\sqrt{n}\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1 \in \mathbb{R}^{d\times 1}$, the entries of which are independent standard Gaussian random variables, satisfies the condition in Lemma 5. With probability no more than $\mathrm{e}^{-\frac{t^2}{2}}$, we have

$$\left\|\sqrt{n}\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1\right\|^2 = s_{\max}^2\left(\sqrt{n}\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1\right) \geq \left(\sqrt{d}+1+t\right)^2. \tag{73}$$

Let $\varepsilon := \left(\sqrt{d}+1+t\right)^2/n$, we can get

$$t = \sqrt{n\varepsilon} - \sqrt{d} - 1. \tag{74}$$

When $n > \frac{4d}{\varepsilon} =: c_{1,1}d$, we have $t > \sqrt{d} - 1 \geq 0$. Plugging (74) into $\mathrm{e}^{-\frac{t^2}{2}}$, the probability of (73) holding is at least $2\exp\left(-\left(\sqrt{n\varepsilon} - \sqrt{d} - 1\right)^2/2\right)$. According to Lemma 6, there exist constants $c_{1,2}$, c_2, such that when $n > c_{1,2}d$, this probability is smaller than $\mathrm{e}^{-c_2 n}$. Taking $c_1 := \max\{c_{1,1}, c_{1,2}\}$ and dividing both sides of (73) by n, we conclude the lemma.

8.6. Proof of Corollary 4

According to the definition of $\bar{\mathbf{a}}_1$ and basic probability, we have

$$\begin{aligned}
\mathbb{P}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_1\right\|^2 > \varepsilon\right) =& \mathbb{P}\left(\frac{\left\|\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1\right\|^2}{\left\|\mathbf{a}_1\right\|^2} > \varepsilon\right) \\
=& 1 - \mathbb{P}\left(\frac{\left\|\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1\right\|^2}{\left\|\mathbf{a}_1\right\|^2} < \varepsilon\right) \\
\leq& 1 - \mathbb{P}\left(\left\|\mathbf{a}_1\right\|^2 > 1-\varepsilon \text{ and } \left\|\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1\right\|^2 < \varepsilon\left(1-\varepsilon\right)\right) \\
=& \mathbb{P}\left(\left\|\mathbf{a}_1\right\|^2 < 1-\varepsilon \text{ or } \left\|\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1\right\|^2 > \varepsilon\left(1-\varepsilon\right)\right) \\
\leq& \mathbb{P}\left(\left\|\mathbf{a}_1\right\|^2 < 1-\varepsilon\right) + \mathbb{P}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1\right\|^2 > \varepsilon\left(1-\varepsilon\right)\right). \qquad (75)
\end{aligned}$$

Now we may estimate the two items in the RHS of (75), separately. By using Lemma 7, for $n > n_0$, we have

$$\mathbb{P}(\left\|\mathbf{a}_1\right\|^2 < 1-\varepsilon) < \mathrm{e}^{-c_{2,1}(\varepsilon)n}. \qquad (76)$$

By using Lemma 8, for $\varepsilon < \frac{1}{3}$ and $n > c_{1,2}d$, we have

$$\mathbb{P}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\mathbf{a}_1\right\|^2 > \varepsilon(1-\varepsilon)\right) < \mathrm{e}^{-c_{2,2}(2\varepsilon/3)n}. \qquad (77)$$

Plugging (76) and (77) into (75), we readily get

$$\mathbb{P}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_1\right\|^2 > \varepsilon\right) < \mathrm{e}^{-c_{2,1}(\varepsilon)n} + \mathrm{e}^{-c_{2,2}(2\varepsilon/3)n}.$$

According to Remark 1, we claim that there exist constants c_1, c_2, such that for any $n > c_1(\varepsilon)d$, we have $\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_1\right\|^2 > \varepsilon$ with probability no more than $\mathrm{e}^{-c_2(\varepsilon)n}$.

8.7. Proof of Remark 3

Replacing n with $n - d_0$ in (20), we can get

$$\mathbb{P}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_1\right\|^2 > \varepsilon\right) \leq \mathrm{e}^{-\hat{c}_2(\varepsilon)(n-d_0)}. \qquad (78)$$

We only need to prove that there exist constants c_1, c_2, when $n > c_1 \max\{d, d_0\}$, we have $\mathbb{P}\left(\left\|\mathbf{V}_2^{\mathrm{T}}\bar{\mathbf{a}}_1\right\|^2 > \varepsilon\right) \leq \mathrm{e}^{-\hat{c}_2(\varepsilon)(n-d_0)} \leq \mathrm{e}^{-c_2 n}$. Let $\tau := \frac{d_0}{n}$, according to

Lemma 6, we have

$$h := \lim_{\tau\to 0}\lim_{n\to\infty}\frac{\hat{c}_2(\varepsilon)\,(n-d_0)}{n} = \lim_{\tau\to 0}\lim_{n\to\infty}\hat{c}_2\,(1-\tau) = \hat{c}_2 > 0,$$
$$b := \lim_{n\to\infty}\frac{\ln 1}{n} = 0 < h.$$

Then when $n > n_0$, $\tau = \frac{d_0}{n} \le \tau_0$, there exists constant c_2, such that $\mathrm{e}^{-\hat{c}_2(\varepsilon)(n-d_0)} \le \mathrm{e}^{-c_2 n}$. By choosing $c_1 := \max\{n_0, \frac{1}{\tau_0}, \hat{c}_1\}$, when $n > c_1 d_0$, we have $n > n_0$, $\tau = \frac{d_0}{n} \le \tau_0$. The condition of (78) holding, that is $n > \hat{c}_1 d$, is also satisfied.

8.8. Proof of Lemma 11

Using the definition of β_k in (50) and the fact that $\mathcal{Y}_{1,1:k-1} \subset \mathcal{Y}_{1,1:k}$, we have

$$\beta_k^2 = \left\|\mathbf{V}_2^{\mathrm{T}}\mathbf{b}_k\right\|^2 = \min_{\substack{\|\mathbf{b}\|=1\\ \mathbf{b}\in\mathcal{Y}_{1,1:k-1}}} \|\mathbf{V}_2^{\mathrm{T}}\mathbf{b}\|^2 \ge \min_{\substack{\|\mathbf{b}\|=1\\ \mathbf{b}\in\mathcal{Y}_{1,1:k}}} \|\mathbf{V}_2^{\mathrm{T}}\mathbf{b}\|^2. \tag{79}$$

Removing both side of (79) from one, we write

$$1-\beta_k^2 \le 1 - \min_{\substack{\|\mathbf{b}\|=1\\ \mathbf{b}\in\mathcal{Y}_{1,1:k}}} \|\mathbf{V}_2^{\mathrm{T}}\mathbf{b}\|^2 = 1 - \min_{\substack{\|\mathbf{b}\|=1\\ \mathbf{b}\in\mathcal{Y}_{1,1:k}}} \|\mathrm{P}_{\mathcal{Y}_2}(\mathbf{b})\|^2 = \max_{\substack{\|\mathbf{b}\|=1\\ \mathbf{b}\in\mathcal{Y}_{1,1:k}}} \|\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{b})\|^2. \tag{80}$$

Then we will loose the condition and rewrite the expression of this maximization problem step by step, and finally convert it to a problem about the extreme singular value of random matrix.

For any vector $\mathbf{b}$ in $\mathcal{Y}_{1,1:k}$, it can be spanned by the columns of $\bar{\mathbf{A}}_{1,1:k}$ as

$$\mathbf{b} = \bar{\mathbf{A}}_{1,1:k}\mathbf{x} = \sum_{j=1}^{k} x_j \bar{\mathbf{a}}_{1,j}, \tag{81}$$

where $\mathbf{x} = [x_1, \cdots, x_k]^{\mathrm{T}}$ denotes the weight vector. Consequently, the condition of $\|\mathbf{b}\| = 1$ can be loosen to the condition on $\mathbf{x}$, i.e.,

$$\|\mathbf{x}\|^2 \le \frac{\left\|\bar{\mathbf{A}}_{1,1:k}\mathbf{x}\right\|^2}{s_{\min}^2\left(\bar{\mathbf{A}}_{1,1:k}\right)} = \frac{\|\mathbf{b}\|^2}{s_{\min}^2\left(\bar{\mathbf{A}}_{1,1:k}\right)} = \frac{1}{s_{\min}^2\left(\bar{\mathbf{A}}_{1,1:k}\right)} =: x_{\mathrm{u}}. \tag{82}$$

Inserting (81) and (82) in (80), we have

$$\begin{aligned}1-\beta_k^2 &\leq \max_{\|\mathbf{x}\|^2\leq x_{\mathrm{u}}} \left\| \mathrm{P}_{\mathcal{Y}_2^{\perp}} \left(\sum_{j=1}^{k} x_j \bar{\mathbf{a}}_{1,j} \right) \right\|^2 \\ &= \max_{\|\mathbf{x}\|^2\leq x_{\mathrm{u}}} \left\| \sum_{j=1}^{k} x_j \mathrm{P}_{\mathcal{Y}_2^{\perp}} \left(\bar{\mathbf{a}}_{1,j} \right) \right\|^2 \\ &= \max_{\|\mathbf{x}\|^2\leq x_{\mathrm{u}}} \left\| \sum_{j=1}^{k} x_j \sqrt{1-\hat{\lambda}_j^2} \bar{\mathbf{a}}_{1,j}^{\perp} \right\|^2, \end{aligned} \tag{83}$$

where $\hat{\lambda}_j$ is defined in (49).

According to (57) and the decreasing order of $\lambda_1 \geq \cdots \geq \lambda_{d_1}$, we have

$$1-\hat{\lambda}_j^2 \leq (1-\lambda_j^2)(1+\varepsilon) \leq (1-\lambda_k^2)(1+\varepsilon), \quad \forall j = 1, \cdots, k < d_1, \tag{84}$$

hold with probability at least $1-\mathrm{e}^{-c_{2,1}(\varepsilon_1)n}$ for any $n > c_{1,1}(\varepsilon_1)d_2$. Inserting (84) in (83), we have

$$\begin{aligned}1-\beta_k^2 &\leq (1-\lambda_k^2)(1+\varepsilon_1) \max_{\|\mathbf{x}\|^2\leq x_{\mathrm{u}}} \left\| \sum_{j=1}^{k} x_j \bar{\mathbf{a}}_{1,j}^{\perp} \right\|^2 \\ &\leq (1-\lambda_k^2)(1+\varepsilon_1) s_{\max}^2 \left(\bar{\mathbf{A}}_{1,1:k}^{\perp} \right) x_{\mathrm{u}} \\ &= (1-\lambda_k^2)(1+\varepsilon_1) \frac{s_{\max}^2 \left(\bar{\mathbf{A}}_{1,1:k}^{\perp} \right)}{s_{\min}^2 \left(\bar{\mathbf{A}}_{1,1:k} \right)}, \end{aligned} \tag{85}$$

where $\bar{\mathbf{A}}_{1,1:k}^{\perp} = \left[\bar{\mathbf{a}}_{1,1}^{\perp}, \cdots, \bar{\mathbf{a}}_{1,k}^{\perp} \right]$.

Now we need to bound the denominator and numerator in the RHS of (85), separately. As to the denominator, according to Corollary 3, we have for $n > c_{1,2}(\varepsilon_2)d_1$,

$$\mathbb{P}\left(s_{\min}^2 \left(\bar{\mathbf{A}}_{1,1:k} \right) > 1-\varepsilon_2 \right) > 1-\mathrm{e}^{-c_{2,2}(\varepsilon_2)n}. \tag{86}$$

As for estimating the numerator, since that $\mathbf{A}_{1,1:k}$ is correlated with $\mathcal{Y}_2$, we can not directly apply the available lemmas about the concentration inequalities of independent Gaussian random matrix. However, by using the following techniques, we could manage to convert the problem of estimating $s_{\max}^2 \left(\bar{\mathbf{A}}_{1,1:k}^{\perp} \right)$ to

a problem about the singular value of a normalized random matrix satisfying the independence condition.

Remark 6 *Recalling Remark 5 about the characteristics of* principal *orthonormal bases* $\mathbf{U}_1$, $\mathbf{U}_2$ *for subspaces* $\mathcal{X}_1, \mathcal{X}_2$, *and following the decomposition way in the proof of Lemma 4, we can decompose each column of* $\mathbf{U}_1$ *as the projections onto* $\mathcal{X}_2$ *and its orthogonal complement and get*

$$\mathbf{U}_1 = \mathbf{U}_2\mathbf{\Lambda} + \mathbf{U}_0\mathbf{\Lambda}^{\perp}, \tag{87}$$

where

$$\mathbf{\Lambda}^{\perp} = \begin{bmatrix} \sqrt{1-\lambda_1^2} & & \\ & \ddots & \\ & & \sqrt{1-\lambda_{d_1}^2} \end{bmatrix},$$

and $\mathcal{C}(\mathbf{U}_0) \in \mathbb{R}^{N\times d_1}$ *is a subspace of* $\mathcal{X}_2^{\perp}$, *that is* $\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_0 = \mathbf{0}$.

After random projection, the decomposition in Remark 6 changes to

$$\mathbf{A}_1 = \mathbf{A}_2\mathbf{\Lambda} + \mathbf{A}_0\mathbf{\Lambda}^{\perp}, \tag{88}$$

where $\mathbf{A}_0 = \mathbf{\Phi}\mathbf{U_0}$. Projecting both side of (88) onto the orthogonal complement of $\mathcal{Y}_2$, we have

$$\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{A}_1) = \mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{A}_0)\mathbf{\Lambda}^{\perp},$$

which means that the normalized column of $\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{A}_1)$, i.e., $\bar{\mathbf{a}}_{1,k}^{\perp}$ defined in (51) are exactly identical to the normalized column of $\mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{A}_0)$, which is denoted as $\bar{\mathbf{a}}_{0,k}^{\perp}, k = 1, \cdots, d_1$. That is

$$\bar{\mathbf{A}}_{1:1:k}^{\perp} = \left[\bar{\mathbf{a}}_{0,1}^{\perp}, \cdots, \bar{\mathbf{a}}_{0,k}^{\perp}\right] =: \bar{\mathbf{A}}_{0,1:k}^{\perp}. \tag{89}$$

Considering its property of isotropy, a Gaussian random vector remains Gaussian distribution when it is projected to an independent subspace. This is demonstrated in Remark 7.

Remark 7 *Let* $\mathbf{A}_1 \in \mathbb{R}^{n\times d_1}$ *and* $\mathbf{A}_2 \in \mathbb{R}^{n\times d_2}, d_1 \le d_2$ *be two Gaussian random matrices. We denote* $\mathbf{V}_2$ *as an orthonormal basis of* $\mathcal{C}(\mathbf{A}_2)$. *The projection*

of $\mathbf{A}_1 = [\mathbf{a}_{1,1}, \cdots, \mathbf{a}_{1,d_1}]$ *onto* $\mathcal{C}(\mathbf{A}_2)$ *is denoted by* $\mathbf{B}_1 = [\mathbf{b}_{1,1}, \cdots, \mathbf{b}_{1,d_1}]$*, i.e.,* $\mathbf{b}_{1,k} = \mathrm{P}_{\mathcal{C}(\mathbf{A}_2)}(\mathbf{a}_{1,k})$*. If* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are independent, we have* $\mathbf{B}_1 = \mathbf{V}_2\mathbf{\Omega}$*, where* $\mathbf{\Omega} \in \mathbb{R}^{d_2 \times d_1}$ *is a Gaussian random matrix.*

This can be readily verified by using the fact that $\mathbf{B}_1 = \mathbf{V}_2\mathbf{V}_2^{\mathrm{T}}\mathbf{A}_1 =: \mathbf{V}_2\mathbf{\Omega}$*, where* $\mathbf{\Omega} = (\omega_{i,j}) := \mathbf{V}_2^{\mathrm{T}}\mathbf{A}_1$*. Because* $\mathbf{A}_1$ *is independent with* $\mathcal{C}(\mathbf{A}_2)$*, as well as its orthonormal basis* $\mathbf{V}_2$*, the distribution of* $\mathbf{A}_1$ *is not influenced, if we first condition* $\mathbf{V}_2$ *and regard it as a given matrix. Consequently, we can readily check that* $\omega_{i,j}$ *are* i.i.d. *zero mean Gaussian random variables.*

Recalling that $\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_2 = \mathbf{0}$, which means $\mathbf{u}_{0,i}^{\mathrm{T}}\mathbf{u}_{2,j} = 0$, $1 \le i \le d_1$, $1 \le j \le d_2$. According to Lemma 9, we have that $\mathbf{a}_{0,i}$ and $\mathbf{a}_{2,j}$ are independent. Moreover, $\mathbf{a}_{0,i}$ is independent with $\mathcal{Y}_2$ and thus independent with its orthogonal complement, $\mathcal{Y}_2^{\perp}$. Then according to Remark 7, the projection of $\mathbf{A}_{0,1:k} \in \mathbb{R}^{n\times k}$ onto $\mathcal{Y}_2^{\perp}$ can be written as

$$\mathbf{A}_{0,1:k}^{\perp} := \left[\mathbf{a}_{0,1}^{\perp}, \cdots, \mathbf{a}_{0,k}^{\perp}\right] = \mathbf{V}_2^{\perp}\mathbf{\Omega}_{1:k}, \tag{90}$$

where $\mathbf{a}_{0,j}^{\perp} := \mathrm{P}_{\mathcal{Y}_2^{\perp}}(\mathbf{a}_{0,j})$, $\mathbf{V}_2^{\perp} \in \mathbb{R}^{n\times(n-d_2)}$ is an arbitrary orthonormal basis of $\mathcal{Y}_2^{\perp}$, and $\mathbf{\Omega}_{1:k} \in \mathbb{R}^{(n-d_2)\times k}$ is a Gaussian random matrix. According to the orthonormality of $\mathbf{V}_2^{\perp}$, we normalize both sides of (90) as

$$\bar{\mathbf{A}}_{0,1:k}^{\perp} = \mathbf{V}_2^{\perp}\bar{\mathbf{\Omega}}_{1:k}, \tag{91}$$

where $\bar{\mathbf{\Omega}}_{1:k}$ denotes the column-normalized $\mathbf{\Omega}_{1:k}$. Because left multiplying an orthonormal matrix does not change its singular value, we have

$$s_{\max}\left(\bar{\mathbf{A}}_{0,1:k}^{\perp}\right) = s_{\max}\left(\bar{\mathbf{\Omega}}_{1:k}\right). \tag{92}$$

Combining (89) and (92), and using Remark 2, we have

$$\begin{aligned}\mathbb{P}\left(s_{\max}^2\left(\bar{\mathbf{A}}_{1,1:k}^{\perp}\right) < 1+\varepsilon_3\right) &= \mathbb{P}\left(s_{\max}^2\left(\bar{\mathbf{\Omega}}_{1:k}\right) < 1+\varepsilon_3\right)\\ &> 1 - \mathrm{e}^{-c_{2,3}(\varepsilon_3)n}\end{aligned} \tag{93}$$

hold when $n > c_{1,3}(\varepsilon_3)d_2$.

Plugging both bounds of denominator and numerator, i.e., (86) and (93), into (85), we can get

$$\begin{aligned} 1-\beta_k^2 &\le \left(1-\lambda_k^2\right)(1+\varepsilon_1)\frac{1+\varepsilon_3}{1-\varepsilon_2} \\ &= \left(1-\lambda_k^2\right)\left(1+\frac{\varepsilon_2+\varepsilon_3}{1-\varepsilon_2}+\varepsilon_1\frac{1+\varepsilon_3}{1-\varepsilon_2}\right), \end{aligned} \tag{94}$$

with probability at least $1-\sum_{l=1}^{3}\mathrm{e}^{-c_{2,l}(\varepsilon_l)n}$ for any $n>\max\{c_{1,l}(\varepsilon_l)\}d_2$. Let $\varepsilon_2\le 1/2$, $\varepsilon_3\le 1/2$, $\varepsilon_1=\varepsilon_2=\varepsilon_3=:\varepsilon/7$, we have $\frac{\varepsilon_2+\varepsilon_3}{1-\varepsilon_2}+\varepsilon_1\frac{1+\varepsilon_3}{1-\varepsilon_2}\le 2\left(\varepsilon_2+\varepsilon_3\right)+3\varepsilon_1=\varepsilon$. By using Remark 1 to reshape the probability, we readily complete the proof.

8.9. Proof of Lemma 12

We will calculate the inner product between $\bar{\mathbf{a}}_{1,k}^{\perp}$ and $\mathbf{b}_k^{\perp}$. Recalling that $\mathbf{b}_k$ is the projection of $\mathbf{a}_{1,k}$ onto $\mathcal{C}(\mathbf{A}_{1,1:k-1})$, it is not obvious whether $\bar{\mathbf{a}}_{1,k}^{\perp}$ and $\mathbf{b}_k^{\perp}$ are independent, and this aggravate the problem to estimate their inner product directly. In order to solve this, therefore, we have to find the relationship between product and projection and then convert the problem to the situation described in Remark 3.

Recalling the previous result that $\bar{\mathbf{a}}_{1,k}^{\perp}=\bar{\mathbf{a}}_{0,k}^{\perp}$ in (89) and using the fact of $\mathbf{b}_k^{\perp}\in\mathcal{C}(\bar{\mathbf{A}}_{1,1:k-1}^{\perp})$, we write

$$\begin{aligned} \left|\langle\bar{\mathbf{a}}_{1,k}^{\perp},\mathbf{b}_k^{\perp}\rangle\right| &= \left|\langle\bar{\mathbf{a}}_{0,k}^{\perp},\mathbf{b}_k^{\perp}\rangle\right| \\ &\le \left\|\mathrm{P}_{\mathcal{C}(\bar{\mathbf{A}}_{1,1:k-1}^{\perp})}(\bar{\mathbf{a}}_{0,k}^{\perp})\right\| \\ &= \left\|\mathrm{P}_{\mathcal{C}(\bar{\mathbf{A}}_{0,1:k-1}^{\perp})}(\bar{\mathbf{a}}_{0,k}^{\perp})\right\|. \end{aligned} \tag{95}$$

Now we need to construct an orthonormal basis for $\mathcal{C}(\bar{\mathbf{A}}_{0,1:k-1}^{\perp})$ and build its connection with $\bar{\mathbf{a}}_{0,k}^{\perp}$. Recalling Remark 7 and the deduction in the proof of Lemma 11, we reshape (91) as

$$\left[\bar{\mathbf{A}}_{0,1:k-1}^{\perp},\bar{\mathbf{a}}_{0,k}\right]=\mathbf{V}_2^{\perp}\left[\bar{\mathbf{\Omega}}_{1:k-1},\bar{\boldsymbol{\omega}}_k\right],$$

where $\bar{\boldsymbol{\omega}}_k$ denotes the last column of $\bar{\mathbf{\Omega}}_{1:k}\in\mathbb{R}^{(n-d_2)\times k}$. We next apply Gram-Schimidt orthogonalization to $\mathbf{\Omega}_{1:k-1}$ and get $\mathbf{W}_{1:k-1}$, which is an orthonormal

basis for $\mathcal{C}(\bar{\mathbf{\Omega}}_{1:k-1})$. Because of the orthonormality of $\mathbf{V}_2^{\perp}$, $\mathbf{V}_2^{\perp}\mathbf{W}_{1:k-1}$ is an orthonormal basis for $\mathcal{C}(\bar{\mathbf{A}}_{0,1:k-1}^{\perp})$. As a consequence, we are able to calculate the RHS of (95) as

$$\begin{aligned}\left\|\mathrm{P}_{\mathcal{C}(\bar{\mathbf{A}}_{0,1:k-1}^{\perp})}(\bar{\mathbf{a}}_{0,k}^{\perp})\right\| &= \left\|\left(\mathbf{V}_2^{\perp}\mathbf{W}_{1:k-1}\right)^{\mathrm{T}}\bar{\mathbf{a}}_{0,k}^{\perp}\right\| \\ &= \left\|\left(\mathbf{V}_2^{\perp}\mathbf{W}_{1:k-1}\right)^{\mathrm{T}}\mathbf{V}_2^{\perp}\bar{\boldsymbol{\omega}}_k\right\| \\ &= \left\|\mathbf{W}_{1:k-1}^{\mathrm{T}}\bar{\boldsymbol{\omega}}_k\right\|. \end{aligned} \tag{96}$$

Recalling that $\bar{\mathbf{\Omega}}_{1:k}$ is a column-normalized Gaussian random matrix, $\bar{\boldsymbol{\omega}}_k$ should be independent with each column of $\bar{\mathbf{\Omega}}_{1:k-1}$, and thus independent with $\mathcal{C}(\bar{\mathbf{\Omega}}_{1:k-1}) = \mathcal{C}(\mathbf{W}_{1:k-1})$. Combining (95) and (96), and using Remark 3, we have

$$\mathbb{P}\left(\left|\langle\bar{\mathbf{a}}_{1,k}^{\perp},\mathbf{b}_k^{\perp}\rangle\right|^2 > \varepsilon_4\right) \leq \mathbb{P}\left(\left\|\mathbf{W}_{1:k-1}^{\mathrm{T}}\bar{\boldsymbol{\omega}}_k\right\|^2 > \varepsilon_4\right) < \mathrm{e}^{-c_{2,4}(\varepsilon_4)n} \tag{97}$$

for all $n \geq c_{1,4}d_2$. The proof is completed.

References

[1] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 2790–2797.

[2] M. Soltanolkotabi, E. J. Candes, A geometric analysis of subspace clustering with outliers, The Annals of Statistics 40 (4) (2012) 2195–2238.

[3] E. Elhamifar, R. Vidal, Sparse subspace clustering: Algorithm, theory, and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11) (2013) 2765–2781.

[4] R. Heckel, H. Bölcskei, Robust subspace clustering via thresholding, IEEE Transactions on Information Theory 61 (11) (2015) 6320–6342.

[5] X. Mao, Y. Gu, Compressed subspace clustering: A case study, in: IEEE Global Conference on Signal and Information Processing, 2014, pp. 453 – 457.

[6] R. Heckel, M. Tschannen, H. Bolcskei, Subspace clustering of dimensionality-reduced data, in: IEEE International Symposium on Information Theory, 2014, pp. 2997–3001.

[7] R. Heckel, M. Tschannen, H. Bölcskei, Dimensionality-reduced subspace clustering, arXiv preprint arXiv:1507.07105.

[8] Y. Wang, Y.-X. Wang, A. Singh, A theoretical analysis of noisy sparse subspace clustering on dimensionality-reduced data, arXiv preprint arXiv:1610.07650.

[9] W. B. Johnson, J. Lindenstrauss, Extensions of lipschitz maps into a hilbert space, Contemporary mathematics 26 (1984) 189–206.

[10] S. Dasgupta, A. Gupta, An elementary proof of the johnson-lindenstrauss lemma, International Computer Science Institute, Technical Report (1999) 99–006.

[11] E. J. Candes, T. Tao, Decoding by linear programming, IEEE Transactions on Information Theory 51 (12) (2005) 4203–4215.

[12] E. J. Candès, The restricted isometry property and its implications for compressed sensing, Comptes Rendus Mathematique 346 (9–10) (2008) 589–592.

[13] R. Baraniuk, M. Davenport, R. Devore, M. Wakin, A simple proof of the restricted isometry property for random matrices, Constructive Approximation 28 (28) (2015) 253–263.

[14] D. L. Donoho, Compressed sensing, IEEE Transactions on Information Theory 52 (4) (2006) 1289–1306.

[15] E. J. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory 52 (2) (2006) 489–509.

[16] S. Aeron, V. Saligrama, M. Zhao, Information theoretic bounds for compressed sensing, IEEE Transactions on Information Theory 56 (10) (2010) 5111–5130.

[17] E. Candes, J. Romberg, Sparsity and incoherence in compressive sampling, Inverse problems 23 (3) (2007) 969.

[18] Y. C. Eldar, G. Kutyniok, Compressed sensing: theory and applications, Cambridge University Press, 2012.

[19] A. Eftekhari, M. B. Wakin, New analysis of manifold embeddings and signal recovery from compressive measurements, Applied and Computational Harmonic Analysis 39 (1) (2015) 67–109.

[20] G. Kutyniok, A. Pezeshki, R. Calderbank, T. Liu, Robust dimension reduction, fusion frames, and grassmannian packings, Applied and Computational Harmonic Analysis 26 (1) (2009) 64–76.

[21] G. Li, Y. Gu, Restricted isometry property of gaussian random projection for finite set of subspaces, IEEE Transactions on Signal Processing 66 (7) (2018) 1705–1720.

[22] A. Edelman, T. A. Arias, S. T. Smith, The geometry of algorithms with orthogonality constraints, SIAM Journal on Matrix Analysis and Applications 20 (2) (1998) 303–353.

[23] J. Haupt, R. Nowak, A generalized restricted isometry property, University of Wisconsin–Madison, Tech. Rep. ECE-07-1.

[24] A. Eftekhari, M. B. Wakin, What happens to a manifold under a bi-lipschitz map?, Discrete & Computational Geometry 57 (3) (2017) 641–673.

[25] M. A. Davenport, P. T. Boufounos, M. B. Wakin, R. G. Baraniuk, Signal processing with compressive measurements, IEEE Journal of Selected Topics in Signal Processing 4 (2) (2010) 445–460.

[26] T. Blumensath, M. E. Davies, Sampling theorems for signals from the union of finite-dimensional linear subspaces, IEEE Transactions on Information Theory 55 (4) (2009) 1872–1882.

[27] S. Dirksen, Dimensionality reduction with subgaussian matrices: a unified theory, Foundations of Computational Mathematics 16 (5) (2016) 1367–1396.

[28] P. K. Agarwal, S. Har-Peled, H. Yu, Embeddings of surfaces, curves, and moving points in euclidean space, in: Proceedings of the twenty-third annual symposium on Computational geometry, ACM, 2007, pp. 381–389.

[29] A. Magen, Dimensionality reductions that preserve volumes and distance to affine spaces, and their algorithmic applications, Randomization and approximation techniques in computer science (2002) 953–953.

[30] C. Jordan, Essai sur la géométrie à n dimensions, Bulletin de la Société mathématique de France 3 (1875) 103–174.

[31] A. Galántai, H. C. J., Jordan's principal angles in complex vector spaces, Nunerical Linear Algebra with Applications 13 (2006) 589–598.

[32] A. Björck, G. H. Golub, Numerical methods for computing the angles between linear subspaces, Mathematics of Computation 27 (1973) 579–594.

[33] K. R. Davidson, S. J. Szarek, Local operator theory, random matrices and banach spaces, Handbook in Banach Spaces (2001) 317–366.