Unraveling the Veil of Subspace RIP Through Near-Isometry on Subspaces

Xingyu Xu, Gen Li, and Yuantao Gu*

Abstract

Dimensionality reduction is a popular approach to tackle high-dimensional data with low-dimensional nature. Subspace Restricted Isometry Property, a newly-proposed concept, has proved to be a useful tool in analyzing the effect of dimensionality reduction algorithms on subspaces. In this paper, we provide a characterization of subspace Restricted Isometry Property, asserting that matrices which act as a near-isometry on low-dimensional subspaces possess subspace Restricted Isometry Property. This points out a unified approach to discuss subspace Restricted Isometry Property. Its power is further demonstrated by the possibility to prove with this result the subspace RIP for a large variety of random matrices encountered in theory and practice, including subgaussian matrices, partial Fourier matrices, partial Hadamard matrices, partial circulant/Toeplitz matrices, matrices with independent strongly regular rows (for instance, matrices with independent entries having uniformly bounded $4 + \epsilon$ moments), and log-concave ensembles. Thus our result could extend the applicability of random projections in subspace-based machine learning algorithms including subspace clustering and allow for the application of some useful random matrices which are easier to implement on hardware or are more efficient to compute.

Keywords: Restricted Isometry Property, dimensionality reduction, random matrix, heavy-tailed distribution, subspace clustering

*The authors are with Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. The corresponding author of this paper is Y. Gu (gyt@tsinghua.edu.cn).

1 Introduction

In this paper we investigate the subspace Restricted Isometry Property (RIP) of random projections and try to capture the root of subspace RIP. In more intuitive language, given two linear subspaces in an ambient space, we ask that for which type of random projections, the "distance" of these two subspaces, when defined properly, is almost invariant after being projected. The precise meaning of these terms will be presented later in this section. Before that, we ground our results with some preliminaries.

1.1 Background

High-dimensional signals can be computationally expensive, or even intractable to analyze. Fortunately, many real-world high-dimensional signals are of low-dimensional nature. In this vein, numerous low dimensional models have been proposed and have remarkably fascinated researches in signal processing [1–3]. Union of Subspaces (UoS) is a powerful low dimensional model which subsumes many classical models including sparse representation and has been used extensively in the recent decade [4]. Briefly speaking, UoS model assumes that in a dataset with high ambient dimension, the data points actually lie on a few low dimensional linear subspaces, and these subspaces characterize the intrinsic structure of the dataset.

Subspace clustering [5–8] is one of the various successful applications of the UoS model that has achieved impressive performance in tasks such as motion segmentation, face clustering, and anomaly detection. Moreover, the performance of subspace clustering is theoretically guaranteed under fairly general conditions, a fact proved in [5] based on the concept of affinity, c.f. Definition 1. However, for traditional subspace clustering algorithms there is a high computational cost in building the so-called similarity representation when the dataset is of high dimension. This defect can be overcome by random compression, as was done in Compressed Subspace Clustering (CSC) [9, 10]. While random compression can significantly reduce the computational burden, it raised a new concern that the affinity between two subspaces may not be preserved after random compression, hence it is not clear whether there is a theoretical guarantee for CSC.

Part of the above concern was addressed in [11–13], which provided theoretical analyses for several popular CSC algorithms. However, these analyses are done per algorithm and do not focus on the concept of affinity. A theorem on "invariance property" of affinity

under random projections would constitute a more universal framework to analyze the performance of CSC. Such a theorem was given in [14, 15], which basically states that the change of affinity between two subspaces is small with high probability under Gaussian random projections. Since affinity is closely related to the notion of projection Frobenius-norm distance between subspaces, this implies that the projection Frobenius-norm distance between subspaces is approximately preserved by a Gaussian random projection, a property named by *subspace Restricted Isometry Property* (subspace RIP) resembling the classical RIP for sparse vectors.

This paper is devoted to a thorough investigation of subspace RIP. Our first aim is to answer the question: what should be the proper abstract setting to study subspace RIP, or more precisely, what is the essential property of a matrix which would lead to subspace RIP? We will prove that such essential property is that the matrix acts as a near-isometry on any low-dimensional subspace. This is not obvious and requires involved analysis. In fact, a naive argument using near-isometry will lose a factor of the dimension of the subspaces, hence will be far from optimal. This fundamental result will be used to prove the subspace RIP for a large variety of random matrices, including subgaussian matrices and other random matrices with exponential Johnson-Lindenstrauss property, partial Fourier/Hadamard matrices and other randomly sampled Bounded Orthonormal Systems (BOS) [16], partial circulant/Toeplitz matrices [17], and also some typical heavy-tailed matrices, e.g. those with independent strongly regular rows [18] or log-concave ensembles [19, 20]. These results provide a universal framework to analyze the subspace RIP of random matrices and their effects on subspace related tasks, which requires rather weak assumptions on the random matrix but yields universal performance guarantee that are not constrained to specific algorithms.

1.2   Our Contribution

In this paper we proved that the essential property of a matrix that leads to subspace RIP is that the matrix acts as a near-isometry on any low-dimensional subspace. This accounts for the root of subspace RIP and provides the proper abstract setting, or a unified approach, to discuss subspace RIP. Both the statement and the proof of this result are deterministic, thus apparently differ from previous work on subspace RIP [14,15] which relied heavily on delicate probabilistic analysis of Gaussian matrices and cannot be decoupled into deterministic and

probabilistic parts in an obvious way; it is even not clear how the proof there generalizes to subgaussian matrices. More discussions on this difference are carried out after sufficient technical preparation, in Section 6.

With this result, we are able to provide an easy proof of subspace RIP for random matrices with exponential Johnson-Lindenstrauss property, e.g. subgaussian matrices, which generalizes the conclusion of [14, 15]. Moreover, we will also prove that randomly sampled BOS, e.g. partial Fourier/Hadamard matrices, and partial circulant/Toeplitz matrices possess subspace RIP. These are matrices with fast matrix-vector multiplication algorithms that permit a wide application and could significantly accelerate computation in practice, and our results validate their use in subspace related tasks. Note that in [12] it was claimed that randomly sampled BOS could be used for random compression in subspace clustering meanwhile keeping the clustering performance, but the proof was based on the assertion that randomly sampled BOS satisfies exponential Johnson-Lindenstrauss property, which was not provided with a legitimate proof there. The proof strategy in [12] appears only feasible to show that randomly sampled BOS satisfies exponential Johnson-Lindenstrauss property with unreasonably small constants, which is not helpful in practice. As such, our results constitute a more effective guarantee for performance of partial Fourier/Hadamard matrices and partial circulant/Toeplitz matrices on subspace clustering.

Recently, there are rising interests on heavy-tailed random matrices. We will deal with two typical types of such random matrices, namely the ones with finite $4 + \epsilon$ moments and log-concave ensembles, and show that how a combination of our characterization of subspace RIP and well-known results in covariance estimation implies easily the subspace RIP of these random matrices.

From a practical point of view, our result holds for much more general random matrices compared with the subspace RIP for Gaussian random matrices in [15], hence allows the application of random matrices that are more useful in practice, for instance those matrices which are easier to generate and store on hardware, e.g. Bernoulli matrices, or those who arise in the physical world naturally and are more efficient to compute, e.g. partial Fourier/Hadamard matrices and partial circulant/Toeplitz matrices. Most of these matrices are inaccessible within the proof strategy in previous works on subspace RIP [14, 15]. As pointed out in [21], in applications such as compressive spectral imaging, typical projection matrices are not Gaussian. Instead, Bernoulli matrices can be used [22]. Our result demon-

strate more practical scenarios where techniques of random projections and in particular, CSC algorithms may apply.

1.3 Notations and Conventions

Throughout this paper, c and C denote two positive universal constants that may vary upon each appearance, while $\tilde{c}$ is the constant appearing in the definition of exponential Johnson-Lindenstrauss property, c.f. Definition 3. Bold upper case letters, e.g. $\mathbf{A}$, are used to denote a matrix, while bold lower case letters, e.g. $\mathbf{u}$, are used to denote a vector. $\boldsymbol{\Phi}$ will always be a random matrix. If $\mathcal{X}$ is a linear subspace of $\mathbb{R}^N$, $\mathcal{X}^\perp$ denotes its orthogonal complement. Orthogonal projections onto subspace $\mathcal{X}$ will be denoted by $\mathcal{P}_{\mathcal{X}}$. The maximal and minimal singular value of a matrix $\mathbf{A}$ will be denoted by $s_{\max}(\mathbf{A})$ and $s_{\min}(\mathbf{A})$. $\|\mathbf{v}\|$ is the Euclidean norm of the vector $\mathbf{v}$, and $\|\mathbf{A}\|_{\mathrm{F}}$ is the Frobenius norm of the matrix $\mathbf{A}$. The $(n-1)$-dimensional unit sphere in $\mathbb{R}^n$ is denoted by $\mathbb{S}^{n-1}$, i.e. $\mathbb{S}^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$. The affinity between subspaces $\mathcal{X}_1$, $\mathcal{X}_2$, defined in Definition 1, will be denoted by $\mathrm{aff}(\mathcal{X}_1, \mathcal{X}_2)$. Occasionally we will write $\mathrm{aff}_{\mathcal{X}}$ (resp. $\mathrm{aff}_{\mathcal{Y}}$) as an abbreviation of $\mathrm{aff}(\mathcal{X}_1, \mathcal{X}_2)$ (resp. $\mathrm{aff}(\mathcal{Y}_1, \mathcal{Y}_2)$). The probability of an event is denoted by $\mathbb{P}(\cdot)$. The expectation of a random variable/vector/matrix is denoted by $\mathbb{E}(\cdot)$.

We will be a bit blurry when using "infinitesimal" ε. That is, we will implicitly shrink the value of ε by a constant ratio when needed. For example, we will assert $\mathbb{P}(X > \varepsilon) < \mathrm{e}^{-c\varepsilon^2 n}$ while we actually proved $\mathbb{P}(X > 2\varepsilon) < \mathrm{e}^{-c\varepsilon^2 n}$. Such gaps are usually easy to fill and harmless to skip. In fact, the former statement can be easily derived from the latter by replacing ε with $\varepsilon/2$ and replacing c with $4c$.

1.4 Organization

The rest of this paper is organized as follows. In Section 2, definitions and basic properties of affinity are provided. In Section 3, we state our main theorem that a matrix acting as a near-isometry on a subspace $\mathcal{X}$ preserves the affinity and projection Frobenius-norm distance between any pair of subspaces in $\mathcal{X}$. Using this theorem, we analyze several important classes of random matrices in Section 4 and prove their subspace RIP. Section 5 is devoted to the proof of the main theorem. Section 6 provides some further comments on proof strategies and comparison with related works. Section 7 briefly introduces some examples among the various potential applications of our theory. Section 8 verifies our

results on a real-world dataset. Finally in Section 9 we conclude the paper.

2 Preliminaries

A key ingredient in the statement of our results is the *affinity* between two subspaces, defined as following [5, 12, 15]:

Definition 1. Let $\mathcal{X}_1$, $\mathcal{X}_2$ be subspaces of dimension d_1, d_2 in $\mathbb{R}^n$. Denote by $\mathcal{P}_{\mathcal{X}_1}$, $\mathcal{P}_{\mathcal{X}_2}$ the matrix of orthogonal projection onto $\mathcal{X}_1$ and $\mathcal{X}_2$. The affinity between $\mathcal{X}_1$ and $\mathcal{X}_2$ is

$$\mathrm{aff}(\mathcal{X}_1, \mathcal{X}_2) = \sqrt{\mathrm{tr}(\mathcal{P}_{\mathcal{X}_1} \mathcal{P}_{\mathcal{X}_2})}.$$

There are several alternative ways to compute the affinity which will be used interchangeably. They are summarized in the following lemma.

Lemma 1. Let $\mathcal{X}_1$, $\mathcal{X}_2$ be subspaces of dimension d_1, d_2 in $\mathbb{R}^n$. Denote by $\mathcal{P}_{\mathcal{X}_1}$, $\mathcal{P}_{\mathcal{X}_2}$ the orthogonal projection onto $\mathcal{X}_1$ and $\mathcal{X}_2$.

i) If $\mathbf{U}_1$, $\mathbf{U}_2$ are orthonormal bases of $\mathcal{X}_1$, $\mathcal{X}_2$, then

$$\mathrm{aff}(\mathcal{X}_1, \mathcal{X}_2) = \|\mathbf{U}_1^{\mathrm{T}} \mathbf{U}_2\|_{\mathrm{F}},$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

ii) If $\mathbf{U}_2$ are orthonormal bases of $\mathcal{X}_2$, then

$$\mathrm{aff}(\mathcal{X}_1, \mathcal{X}_2) = \|\mathcal{P}_{\mathcal{X}_1} \mathbf{U}_2\|_{\mathrm{F}}.$$

iii) There exists orthonormal bases $\mathbf{U}_1$, $\mathbf{U}_2$ of $\mathcal{X}_1$, $\mathcal{X}_2$ and nonnegative real numbers $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{\min(d_1, d_2)}$, such that

$$\langle \mathbf{u}_{1,i}, \mathbf{u}_{2,j} \rangle = \begin{cases} \lambda_i, & i = j; \\ 0, & i \neq j, \end{cases}$$

where $\mathbf{u}_{1,i}, \mathbf{u}_{2,j}$ denotes the i-th column of $\mathbf{U}_1$ and the j-th column of $\mathbf{U}_2$ respectively. Such $\mathbf{U}_1$, $\mathbf{U}_2$ are called *principal orthonormal bases* of $\mathcal{X}_1$, $\mathcal{X}_2$. Furthermore,

$$\mathrm{aff}^2(\mathcal{X}_1, \mathcal{X}_2) = \sum_{i=1}^{\min(d_1, d_2)} \lambda_i^2.$$

As its name suggests, affinity measures how close two subspaces are to each other. A relevant notion is the *projection Frobenius-norm distance* of two subspaces [15].

Definition 2. The projection Frobenius-norm distance of two subspaces $\mathcal{X}_1, \mathcal{X}_2$ is defined as

$$D(\mathcal{X}_1, \mathcal{X}_2) = \frac{1}{\sqrt{2}} \| \mathcal{P}_{\mathcal{X}_1} - \mathcal{P}_{\mathcal{X}_2} \|_{\mathrm{F}},$$

where $\mathcal{P}_{\mathcal{X}_i}$ is the matrix of orthogonal projection onto $\mathcal{X}_i$, $i = 1, 2$.

Affinity and projection Frobenius-norm distance are related by

$$D^2(\mathcal{X}_1, \mathcal{X}_2) = \frac{d_1 + d_2}{2} - \mathrm{aff}^2(\mathcal{X}_1, \mathcal{X}_2). \tag{1}$$

Intuitively, this means that the closer (in affinity) two subspaces are to each other, the less distant (in projection Frobenius-norm) they are to each other, which sounds tautological. Our main results will be stated based on affinity, but they can be easily translated to statements on projection Frobenius-norm distance by (1).

We are now in the position to state the main result of [15].

Theorem 1. Assume $\mathbf{\Phi}$ is an $n \times N$ Gaussian matrix with i.i.d. entries sampled from $\mathcal{N}(0, 1/n)$. For any two subspaces $\mathcal{X}_1$, $\mathcal{X}_2$ of dimension d_1, d_2 in $\mathbb{R}^N$, assuming $d_1 \leq d_2$, denote by $\mathcal{Y}_1, \mathcal{Y}_2$ the image of $\mathcal{X}_1, \mathcal{X}_2$ under $\mathbf{\Phi}$. Then for any $0 < \varepsilon < 1/2$ there exists positive constants $c_1(\varepsilon)$, $c_2(\varepsilon)$, such that for $n > c_1(\varepsilon) d_2$, the following is true with probability exceeding $1 - \mathrm{e}^{c_2(\varepsilon) n}$.

$$\left| \mathrm{aff}^2(\mathcal{Y}_1, \mathcal{Y}_2) - \mathrm{aff}^2(\mathcal{X}_1, \mathcal{X}_2) \right| \leq \left(d_1 - \mathrm{aff}^2(\mathcal{X}_1, \mathcal{X}_2) \right) \varepsilon. \tag{2}$$

Remark 1. Using the notion of projection Frobenius-norm distance (2) has the following corollary in an easy-to-remember form:

$$\left| D^2(\mathcal{Y}_1, \mathcal{Y}_2) - D^2(\mathcal{X}_1, \mathcal{X}_2) \right| \leq \varepsilon D^2(\mathcal{X}_1, \mathcal{X}_2). \tag{3}$$

In other words, the distance of two subspaces only changes by a small portion after random projections with overwhelming probability. We thus call the "affinity preserving" property in (2) by *subspace Restricted Isometry property* (subspace RIP), a term resembling the classical Restricted Isometry Property for sparse vectors [23].

The aim of this paper is to illuminate the root of subspace RIP and to extend Theorem 1 to a much wider range of random matrices that are more useful in practice.

3 Main Theorem

Lying in the center of our theory is the following theorem:

Theorem 2. Let $\mathcal{X}$ be a d-dimensional subspace in $\mathbb{R}^N$. Let $\mathcal{X}_1$, $\mathcal{X}_2$ be subspaces in $\mathcal{X}$ whose dimensions are respectively d_1 and d_2, and (without loss of generality) assume that $d_1 \leq d_2$. Denote by $\mathbf{U}$ a matrix whose columns constitute an orthonormal basis of $\mathcal{X}$. Suppose $\mathbf{\Phi}$ is a $n \times N$ matrix satisfying for some $\delta \in (0, 1/4)$ that

$$1 - \delta < s_{\min}^2(\mathbf{\Phi U}) \leq s_{\max}^2(\mathbf{\Phi U}) < 1 + \delta. \tag{4}$$

Then with $\mathcal{Y}_i = \mathbf{\Phi}\mathcal{X}_i$, $\text{aff}_\mathcal{Y} = \text{aff}(\mathcal{Y}_1, \mathcal{Y}_2)$, $\text{aff}_\mathcal{X} = \text{aff}(\mathcal{X}_1, \mathcal{X}_2)$, we have

$$\left| \text{aff}_\mathcal{Y}^2 - \text{aff}_\mathcal{X}^2 \right| \leq C(d_1 - \text{aff}_\mathcal{X}^2)\delta, \tag{5}$$

where $C > 0$ is some universal constant.

Remark 2. Consequently, the projection Frobenius-norm distance of $\mathcal{X}_1, \mathcal{X}_2$ is preserved by $\mathbf{\Phi}$:

$$\left| D^2(\mathcal{Y}_1, \mathcal{Y}_2) - D^2(\mathcal{X}_1, \mathcal{X}_2) \right| \leq C\delta D^2(\mathcal{X}_1, \mathcal{X}_2), \tag{6}$$

which means that $\mathbf{\Phi}$ possesses *subspace RIP* for subspaces of $\mathcal{X}$.

Remark 3. The assumption (4) has a intimate connection with the concept of *subspace embedding* in numerical linear algebra [24]. The main difference is that subspace embedding in [24] asks (4) to hold with probability at least $1 - \varepsilon$, hence is a probabilistic assumption, while in our assumption (4) is deterministic and removes the need of probabilistic argument; we believe our assumption better captures the essence of the matter.

The assumption (4) is equivalent to saying that $\mathbf{\Phi}$ acts as a near-isometry on $\mathcal{X}$, i.e. $(1 - \delta)\|\mathbf{u}\|^2 < \|\mathbf{\Phi u}\|^2 < (1 + \delta)\|\mathbf{u}\|^2$ for any $\mathbf{u} \in \mathcal{X}$. Thus the essence of Theorem 2 is that a near-isometry on a subspace $\mathcal{X}$ preserves the pairwise distance of subspaces of $\mathcal{X}$. This is not an obvious fact, since affinity, hence subspace distance, is defined in a subtle way that involves orthonormal bases of both subspaces, but the latter is not preserved by a near-isometry. The overall effect of such structural degeneration makes the desired factor $(d_1 - \text{aff}_\mathcal{X}^2)$, which is crucial in establishing (6), out of immediate reach. One has to perform some careful analysis to obtain (5) and (6).

Before we present the proof (in Section 5), it is of interest to explain how Theorem 2 leads easily to a series of corollaries on subspace RIP for a wide variety of random matrices, which we will do in the next section.

4 Random Matrices and Near-Isometry on Subspaces

This section discusses in detail the near-isometry condition (4) and its connection with random matrices. Furthermore, this section examines various random matrices encountered in practice and shows that they satisfy the near-isometry condition, hence possess subspace RIP, which would validate their application in subspace-related tasks to accelerate computation.

The near-isometry condition (4) and Theorem 2 are best understood in the context of random matrices. In practice, it is useless to discuss a pair of low-dimensional subspaces $\mathcal{X}_1, \mathcal{X}_2$ contained in a *specific* subspace $\mathcal{X}$; one would often need (5) and (6) for *any* pair of such subspaces. For any pair of subspaces $\mathcal{X}_1$, $\mathcal{X}_2$, we consider their sum

$$\mathcal{X} = \{\mathbf{x}_1 + \mathbf{x}_2 : \mathbf{x}_1 \in \mathcal{X}_1, \mathbf{x}_2 \in \mathcal{X}_2\}.$$

This is a subspace of $\mathbb{R}^N$ of dimension at most $2d_2$ that contains both $\mathcal{X}_1$ and $\mathcal{X}_2$. If $\boldsymbol{\Phi}$ acts as a near-isometry on $\mathcal{X}$, then $\boldsymbol{\Phi}$ preserves the affinity and the distance between $\mathcal{X}_1$, $\mathcal{X}_2$. In order that $\boldsymbol{\Phi}$ preserves the affinity and the distance between any pair of $\mathcal{X}_1$, $\mathcal{X}_2$, one may impose that $\boldsymbol{\Phi}$ acts as a near-isometry on any subspace of dimension $2d_2$ in $\mathbb{R}^N$. This is, however, apparently impossible for deterministic $\boldsymbol{\Phi}$, and a standard way to resolve this is to use a random matrix $\boldsymbol{\Phi}$ instead.

It is clear from the above argument that, if $\boldsymbol{\Phi}$ is a random matrix which acts as a near-isometry with high probability on any subspace of dimension $2d_2$, then $\boldsymbol{\Phi}$ preserves the affinity and the distance between $\mathcal{X}_1$, $\mathcal{X}_2$ with high probability for any pair of $\mathcal{X}_1$, $\mathcal{X}_2$ of dimension respectively d_1, d_2, where $d_1 \leq d_2$.

As a consequence, analysis of subspace RIP now boils down to analysis of the singular values of $\boldsymbol{\Phi}\mathbf{U}$, where $\mathbf{U}$ is a matrix whose columns constitute an orthonormal basis for some subspace in $\mathbb{R}^N$. This will be carried out in the rest of this section.

Throughout this section, $\mathcal{X}_1, \ldots, \mathcal{X}_L$ always denote subspaces in $\mathbb{R}^N$ of dimension $d_1, \ldots, d_L$, and $d_* = \max\{d_1, \ldots, d_L\}$. The image of $\mathcal{X}_i$ under the random projection $\boldsymbol{\Phi}$ will be denoted by $\mathcal{Y}_i$. The subspace RIP of $\boldsymbol{\Phi}$ will be characterized by maximum discrepancy

$$\Delta = \max_{1 \leq i < j \leq L} \frac{|\mathrm{aff}^2(\mathcal{Y}_i, \mathcal{Y}_j) - \mathrm{aff}^2(\mathcal{X}_i, \mathcal{X}_j)|}{\max\{d_i, d_j\} - \mathrm{aff}^2(\mathcal{X}_i, \mathcal{X}_j)}.$$

Note that

$$\Delta \geq \max_{1 \leq i < j \leq L} \frac{|D^2(\mathcal{Y}_i, \mathcal{Y}_j) - D^2(\mathcal{X}_i, \mathcal{X}_j)|}{D^2(\mathcal{X}_i, \mathcal{X}_j)}.$$

4.1 Example: Exponential Johnson-Lindenstrauss Property

A class of random projections that deserves much emphasis is the ones with exponential Johnson-Lindenstrauss property[1], defined as following

Definition 3. A random matrix $\mathbf{A} \in \mathbb{R}^{n \times N}$ is said to satisfy *exponential Johnson-Lindenstrauss property*, if there exists some constant $\tilde{c} > 0$, such that for any $0 < \varepsilon < 1$ and for any $\mathbf{x} \in \mathbb{R}^N$,

$$\mathbb{P}\big(\big|\|\mathbf{A}\mathbf{x}\|^2 - \|\mathbf{x}\|^2\big| > \varepsilon\|\mathbf{x}\|^2\big) \leq 2\mathrm{e}^{-\tilde{c}\varepsilon^2 n}.$$

Examples of random matrices with exponential Johnson-Lindenstrauss property are pervasive in both theory and practice. Section 10 provides a non-comprehensive list of such examples (Gaussian matrices with independent columns, subgaussian matrices with independent rows, and partial Fourier/Hadamard matrices) and also a related theorem which asserts that classical RIP for sparse vectors with sufficiently small restricted isometry constants implies exponential Johnson-Lindenstrauss property.

Taking the route discussed at the beginning of this section, we have

Lemma 2. Let $\mathbf{U}$ be a matrix whose columns constitute an orthonormal basis for a d-dimensional subspace in $\mathbb{R}^N$. Assume the random matrix $\mathbf{\Phi}$ satisfies exponential Johnson-Lindenstrauss property. Then for any $0 < \varepsilon < 1$, we have

$$\mathbb{P}(1 - \varepsilon < s_{\min}^2(\mathbf{\Phi}\mathbf{U}) \leq s_{\max}^2(\mathbf{\Phi}\mathbf{U}) < 1 + \varepsilon) \geq 1 - \mathrm{e}^{-\tilde{c}\varepsilon^2 n + 3d}.$$

The proof is by a standard covering argument and is deferred to Section 10.

As a corollary, we have the following result on subspace RIP of random matrices with exponential Johnson-Lindenstrauss property, which generalizes the main result in [15].

Corollary 1. Assume the random matrix $\mathbf{\Phi}$ satisfies exponential Johnson-Lindenstrauss property. Then for some universal constant $c > 0$ and for any $0 < \varepsilon < 1$, we have

$$\Delta \leq \varepsilon$$

with probability at least $1 - L^2 \mathrm{e}^{-c\tilde{c}\varepsilon^2 n + 6d_*}$. In particular, whenever $n > 24c^{-1}\tilde{c}^{-1}\varepsilon^{-2}\max\{d_*, \log L\}$, the probability is at least $1 - \mathrm{e}^{-c\tilde{c}\varepsilon^2 n/2}$.

---

[1]In literature the same property is usually under the name "Johnson-Lindenstrauss property", without "exponential"; see for instance [16], Section 9.5.

Proof. By the argument at the beginning of this section, this follows from Lemma 2, Theorem 2 and union bound. $\qquad\square$

Note that how this simple proof supersedes, in both effectivity and generality, the complicated probabilistic analysis which spans tens of pages in [15], thanks to Theorem 2. We will discuss this difference in more detail in Section 6.

Corollary 1 permits to apply various matrices used in practice (e.g. Bernoulli or partial Fourier) to subspace related tasks. For subgaussian matrices, the constant $\tilde{c}$ in exponential Johnson-Lindenstrauss property depends only on the subgaussian norm and is inverse proportional to the square of the subgaussian norm, which is quite satisfying. However, for partial Fourier matrices the above analysis is a bit rough, as the constant $\tilde{c}$ is proportional to $\sqrt{N}$. This is problematic when N is large[2]. Moreover, this leaves out the commonly-used partial circulant matrices and partial Toeplitz matrices, which do not satisfy exponential Johnson-Lindenstrauss property with reasonable $\tilde{c}$. These matrices are endowed with fast matrix-vector multiplication algorithms that significantly accelerate the random compression procedure, hence are worth a more refined treatment, as shown in the next example.

4.2 Example: Some Random Matrices With Fast Algorithms

Some random matrices are particularly fascinating for practical use due to their advantages in computational efficiency and their natural emergence in signal processing tasks. Such examples include partial Fourier matrices and other randomly sampled Bounded Orthonormal Systems (BOS) [16], which correspond to random subsampling in frequency domain and other feature domains. Another important example is partial circulant/Toeplitz matrix, which corresponds to subsampling after a random convolution. These matrices allow for $O(N \log N)$-time multiplication-by-vector algorithms by virtue of Fast Fourier Transform (FFT), Fast Walsh-Hadamard Transform (FWHT), etc.

Proving subspace RIP of these matrices would legitimate their use in subspace related tasks, hence significantly improves the efficiency in handling such tasks. In fact, we will show in Section 8 how the application of random compression by these matrices boosts up

---

[2]In [12] it is claimed that $\tilde{c} = \Omega(\log^4 N)$, which was not legitimately proved there and is likely wrong. In fact, they argue that $\tilde{c} = \Omega(\log^4 N)$ follows from Theorem 6 and Theorem 7 presented in our Section 10, but to achieve the $\mathrm{e}^{-\Omega(n)}$ probability bound one has to take $s = \Omega(n)$ and $\zeta = \mathrm{e}^{-\Omega(n)}$, which requires that $n = \Omega(n^2)$ and is absurd.

subspace clustering on a real-world dataset. This motivates the following results.

Lemma 3. Let $\mathbf{U}$ be a matrix whose columns constitute an orthonormal basis for a d-dimensional subspace in $\mathbb{R}^N$. Let $\mathbf{A} \in \mathbb{C}^{n \times N}$ be the random sampling associated to a BOS[3] with constant $K \geq 1$. Let $\mathbf{D}_\epsilon$ be a diagonal matrix with i.i.d. Rademacher random variables on its diagonal. Let $\mathbf{\Phi} = \mathbf{A}\mathbf{D}_\epsilon$. Then there exists some constant $C > 1$ such that for any $\varepsilon \in (0,1)$ and for any $n > CK^3\varepsilon^{-3}\max\{d\log^3 d, \log^3 N\}$, we have

$$1 - \varepsilon < s_{\min}^2(\mathbf{\Phi}\mathbf{U}) \leq s_{\max}^2(\mathbf{\Phi}\mathbf{U}) < 1 + \varepsilon$$

with probability at least

$$1 - \exp\left(-C^{-1}\left(\sqrt{d^2 + K^{-2}\varepsilon^2 n} - d\right)\right).$$

The proof is by a careful application of well-known properties of randomly sample BOS and is deferred to Section 11. Note that the exponent 3 in K^3, ε^{-3}, $\log^3 d$ and $\log^3 N$ can be replaced by $2 + \epsilon$ for any $\epsilon > 0$, and we chose 3 only for typographical convenience.

Corollary 2. Let $\mathbf{\Phi}$ be as in Lemma 3. Then there exists some constant $C > 1$ such that for any $\varepsilon \in (0,1)$ and any $n > CK^3\varepsilon^{-3}\max\{d_*(\log^3 d_* + \log L), \log^2 L, \log^3 N\}$, we have

$$\Delta \leq \varepsilon$$

with probability at least

$$1 - \exp\left(-C^{-1}\left(\sqrt{d_*^2 + K^{-2}\varepsilon^2 n} - d_*\right)\right).$$

Proof. By the argument at the beginning of this section, this follows from Lemma 3, Theorem 2 and union bound. □

While Lemma 3 follows from standard results on randomly sampled BOS, for partial circulant/Toeplitz matrices the situation is more subtle. The following lemma will be proved in Section 11 via some modifications of the proof strategy in [25].

Lemma 4. Let $\mathbf{U}$ be a matrix whose columns constitute an orthonormal basis for a d-dimensional subspace in $\mathbb{R}^N$. Let $\mathbf{a}$ be a random vector in $\mathbb{R}^N$ with i.i.d. standard complex circular Gaussian entries. Let $\mathcal{C}(\mathbf{a})$ be the circulant matrix generated by $\mathbf{a}$, i.e. whose

---

[3]Partial Fourier matrices and partial Hadamard matrices are both randomly samplings associated to a BOS with constant $O(1)$.

first row is $\mathbf{a}$. Choose arbitrarily n rows of $\mathcal{C}(\mathbf{a})$ and form with these rows a new matrix $\mathbf{A} \in \mathbb{R}^{n \times N}$. Let $\mathbf{\Phi} = \frac{1}{\sqrt{n}} \mathbf{A}$. Then there exists some constant $C > 1$ such that for any $\varepsilon \in (0,1)$ and any $n > C\varepsilon^{-2} \max\{d, \log N\}^2$, we have

$$1 - \varepsilon < s_{\min}^2(\mathbf{\Phi}\mathbf{U}) \leq s_{\max}^2(\mathbf{\Phi}\mathbf{U}) < 1 + \varepsilon$$

with probability at least $1 - \mathrm{e}^{-c\varepsilon\sqrt{n}}$.

Remark 4. The above lemma requires that $n = O(d_*^2)$. With a substantial amount of work (using some modern results on generic chaining bound of suprema of order-2 chaos process, e.g. [26]), it is possible to attain a sub-optimal scaling similar to the one in Corollary 2, as well as a similar probability bound as in Corollary 2. An easier way to improve the scaling is to utilize the result in [17], which allows to obtain $n = O(d_*^{3/2})$ with a trade-off in the probability bound that becomes $\mathrm{e}^{-O(n^{1/3})}$; this turns out to be a special case of the aforementioned "harder" treatment. We will not pursue these directions here to avoid unnecessary technicality.

Corollary 3. Let $\mathbf{\Phi}$ be as in Lemma 4. Then there exists some constant $C > 1$ such that for any $\varepsilon \in (0,1)$ and any $n > C\varepsilon^{-2} \max\{d_*, \log N, \log L\}^2$, we have

$$\Delta \leq \varepsilon$$

with probability at least $1 - \mathrm{e}^{-c\varepsilon\sqrt{n}}$.

Proof. By the argument at the beginning of this section, this follows from Lemma 4, Theorem 2 and union bound. $\square$

Remark 5. The same is true for Toeplitz matrix since a Toeplitz matrix can be embedded into a circulant matrix with twice dimension.

Remark 6. In fact, we will prove the Lemma 4 for random vector $\mathbf{a}$ with independent complex uniformly-subgaussian entries (see Section 10, Definition 4). This involves modifying and generalizing the proof in [25].

4.3 Example: Heavy-Tailed Distributions

In practice one may also have to deal with heavy-tailed random matrices. Subspace RIP of heavy-tailed random matrices is now handy by Theorem 2 and standard results in

13

covariance estimation. Before presenting these results, we need to set up some customary assumptions on the rows of $\boldsymbol{\Phi}$. Denote the rows of $\boldsymbol{\Phi}$ by $\frac{1}{\sqrt{n}}\mathbf{x}_1^{\mathrm{T}}, \ldots, \frac{1}{\sqrt{n}}\mathbf{x}_n^{\mathrm{T}}$.

1. $\mathbf{x}_i$'s are centered, i.e. $\mathbb{E}\mathbf{x}_i = \mathbf{0}$.

2. $\mathbf{x}_i$'s are isotropic, i.e. $\mathbb{E}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \mathbf{I}$.

3. $\mathbf{x}_i$'s are independent.

These assumptions (centered, isotropic and independent rows) are quite natural and often serve as the default setting in compressed sensing and non-asymptotic random matrix theory, especially in context of Bai-Yin law; see [27] for examples and further discussions. Here we briefly mention that the a row of $\boldsymbol{\Phi}$ can be regarded as a linear functional that observes the data vector, and independence of rows means that different observations are independent, which is reasonable in many applications. On the other hand, centeredness and isotropy can be fulfilled by a preprocessing step before observing the data, thus are not really restrictive. Some results in random matrix theory are also valid for weakly-independent rows, but such results usually bear considerable technicality imposed by the difficulty to quantitively define weak dependence, hence are not discussed here.

We will deal with two important types of heavy-tailed distributions: those with finite moments, and log-concave ensembles.

Finite moments Let $\eta > 1$, $C' \geq 1$ be constants and $\mathbf{x}$ be an N-dimensional random vectors which is centered and isotropic. The random vector $\mathbf{x}$ is said to satisfy the strong regularity condition [18] if

$$\mathbb{P}(\|\mathcal{P}\mathbf{x}\|^2 > t) \leq C't^{-\eta}, \quad \text{for } t > C' \operatorname{rank}\mathcal{P}. \tag{7}$$

for every orthogonal projection $\mathcal{P}$ of rank at most d in $\mathbb{R}^N$, where C is some universal constant. This condition is satisfied, for example, by those $\mathbf{x}$ whose entries are independent and have uniformly bounded $(4+\varepsilon)$-moments. We will discuss the meaning of this condition later.

Lemma 5. Let $\mathbf{U}$ be a matrix whose columns constitute an orthonormal basis for a d-dimensional subspace in $\mathbb{R}^N$. Assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent centered isotropic random vectors in $\mathbb{R}^N$ that satisfy the strong regularity condition (7). Let $\boldsymbol{\Phi}$ be a $n \times N$ random matrix whose rows are $\frac{1}{\sqrt{n}}\mathbf{x}_1^{\mathrm{T}}, \ldots, \frac{1}{\sqrt{n}}\mathbf{x}_n^{\mathrm{T}}$. Then there exists a polynomial function

poly$(\cdot)$ whose coefficients depend only on η and C', such that whenever $\varepsilon \in (0,1)$ and $n > \text{poly}(\varepsilon^{-1})d$, we have

$$1 - \varepsilon < s_{\min}^2(\mathbf{\Phi U}) \leq s_{\max}^2(\mathbf{\Phi U}) < 1 + \varepsilon$$

with probability at least $1 - \varepsilon$.

Sketch. Set $\mathbf{y}_i = \mathbf{U}^{\mathrm{T}}\mathbf{x}_i$, then it is easy to check that $\mathbf{y}_i$'s are independent, centered and isotropic, and

$$\|\mathbf{U}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi U} - \mathbf{I}\| = \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}} - \mathbf{I} \right\|.$$

One may verify that $\mathbf{y}_i$ satisfies strong regularity condition if $\mathbf{x}_i$ satisfies strong regularity condition, and the main theorem in [18] yields the desired results. Details are postponed to Section 12. □

Corollary 4. Let $\mathbf{\Phi}$ be as in Lemma 5. Then there exists a polynomial function poly$(\cdot)$ whose coefficients depend only on η and C', such that whenever $\varepsilon \in (0,1)$ and $n > \text{poly}(\varepsilon^{-1})Ld_*$, we have

$$\Delta \leq \varepsilon$$

with probability at least $1 - \varepsilon$.

Proof. Note that this does *not* follow from union bound. Instead, we take $\mathcal{X}$ to be the sum of all $\mathcal{X}_i$, then $\dim \mathcal{X} \leq Ld_*$. Now the conclusion follows from Lemma 5 (applied to an orthonormal basis of $\mathcal{X}$) and Theorem 2. □

Remark 7. A recent result [28] of some of the authors show that the partial strong regularity condition can be further relaxed to that (7) holds for any $\mathcal{P}$ with rank $\lceil Cd_* \rceil$. Though this assumption is logically weaker, it does not seem to yield sensible improvement in our case.

An intriguing feature of these results is that strong regularity condition, which in its original context seems to be an artifact due to the spectral sparsfier method adopted in [18], arises very naturally in our case. In fact, to control the singular values of $\mathbf{\Phi U}$ in Lemma 5, it is necessary to have a reasonably fast tail-decay for $\|\mathbf{y}_i\|$ (see [18, 29, 30] for discussions), but $\|\mathbf{y}_i\|$ is just the norm of the orthogonal projection of $\mathbf{x}_i$ onto the arbitrary d-dimensional subspace spanned by the columns of $\mathbf{U}$; thus it is necessary to assume that the tails of all orthogonal projections of $\mathbf{x}_i$ decay fast enough, which is exactly what is meant by strong regularity condition.

If $\mathbf{x}$ has independent entries, (7) is satisfied when its entries have uniformly bounded 4η-th moments. (For instance, see Proposition 1.3 in [18]). Hence one may regard strong regularity condition as a finite $4 + \epsilon$ moment assumption. Note that it has been known for long that finite fourth moments are necessary for covariance estimation [29].

Log-concave ensembles If the row vectors of $\mathbf{\Phi}$ have better tail behavior, stronger probability bounds can be obtained. A class of distributions with heavy, but not too heavy tails that plays a role in geometric functional analysis is log-concave distribution. A probability distribution $\mathbb{P}$ on $\mathbb{R}^N$ is said to be log-concave if

$$\mathbb{P}(\theta A + (1-\theta)B) \geq \mathbb{P}(A)^\theta \mathbb{P}(B)^{1-\theta}$$

for any measurable sets A, B in $\mathbb{R}^N$ and any $\theta \in [0,1]$. It follows from definition that the marginal of a log-concave distribution is again log-concave. Typical examples of log-concave distributions include the uniform distribution on a convex body (e.g. a Euclidean ball) or more generally, the distribution with density $C \exp(-f(\mathbf{x}))$, where f is a convex function[4]. This subsumes Laplace, Gaussian, Gamma, Beta, Weibull and Logistic distributions (in suitable region of parameters).

Lemma 6. Let $\mathbf{U}$ be a matrix whose columns constitute an orthonormal basis for a d-dimensional subspace in $\mathbb{R}^N$. Assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent random vectors in $\mathbb{R}^N$ that are centered, isotropic and log-concave. Let $\mathbf{\Phi}$ be a $n \times N$ random matrix whose rows are $\frac{1}{\sqrt{n}}\mathbf{x}_1^{\mathrm{T}}, \ldots, \frac{1}{\sqrt{n}}\mathbf{x}_n^{\mathrm{T}}$. Then there exists some universal constants $c > 0$ and $C > 1$ such that for any $\varepsilon \in (0,1)$ and any $n > C\varepsilon^{-2}d$, we have

$$1 - \varepsilon < s_{\min}^2(\mathbf{\Phi U}) \leq s_{\max}^2(\mathbf{\Phi U}) < 1 + \varepsilon$$

with probability at least $1 - \mathrm{e}^{-c\varepsilon\sqrt{n}}$.

This follows from a tricky application of a famous theorem from [19, 20]. Details are postponed to Section 12.

Corollary 5. Let $\mathbf{\Phi}$ be as in Lemma 6 Then there exists some universal constants $c > 0$ and $C > 0$ such that whenever $\varepsilon \in (0, 1/2)$ and $n > C\varepsilon^{-2}\max\{d_*, \log^2 L\}$, we have

$$\Delta \leq \varepsilon \tag{8}$$

---

[4]By setting f to be the indicator function of a convex body we recover the case of uniform distribution on a convex body.

with probability at least $1 - \mathrm{e}^{-c\varepsilon\sqrt{n}}$.

Proof. By the argument at the beginning of this section, this follows from Lemma 6, Theorem 2 and union bound. $\qquad\square$

Remark 8. The $e^{-\Omega(\sqrt{n})}$ probability bound is optimal due to thin shell probability of log-concave ensembles, see [31].

5 Proof of Theorem 2

5.1 Some First Consequences of (4)

Recall that (4) means $\boldsymbol{\Phi}$ acts as a near-isometry on $\mathcal{X}$. An immediate consequence of this assumption is:

Proposition 1. Let $\mathcal{X}$ be a subspace of $\mathbb{R}^N$ and $\mathbf{U}$ be a matrix whose columns constitute an orthonormal basis of $\mathcal{X}$. Let $\boldsymbol{\Phi}$ be a $n \times N$ matrix such that (4) holds for some $\delta \in (0,1)$. Then we have

$$\sqrt{1-\delta}\|\mathbf{u}\| \leq \|\boldsymbol{\Phi}\mathbf{u}\| \leq \sqrt{1+\delta}\|\mathbf{u}\|$$

for any $\mathbf{u} \in \mathcal{X}$. In particular, if $\delta \in (0, 1/4)$ we have $\sqrt{\frac{3}{4}}\|\mathbf{u}\| \leq \|\boldsymbol{\Phi}\mathbf{u}\| \leq \sqrt{\frac{5}{4}}\|\mathbf{u}\|$.

Moreover, assume $\mathcal{X}_1$, $\mathcal{X}_2$ are two subspaces of $\mathcal{X}$ which are orthogonal to each other, and let $\mathbf{U}_1$ (resp. $\mathbf{U}_2$) be a matrix whose columns constitute an orthonormal basis of $\mathcal{X}_1$ (resp. $\mathcal{X}_2$), then

$$\|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{U}_1\| \leq \delta.$$

Proof. For any $\mathbf{u} \in \mathcal{X}$, one may find a vector $\mathbf{x}$ in $\mathbb{R}^{\dim \mathcal{X}}$ such that $\mathbf{u} = \mathbf{U}\mathbf{x}$, hence $\|\mathbf{u}\| = \|\mathbf{x}\|$. Thus $\|\boldsymbol{\Phi}\mathbf{u}\| = \|\boldsymbol{\Phi}\mathbf{U}\mathbf{x}\|$ is at least $s_{\min}(\boldsymbol{\Phi}\mathbf{U})\|u\|$ and at most $s_{\max}(\boldsymbol{\Phi}\mathbf{U})\|\mathbf{u}\|$. This proves the first part of the proposition.

For the second part, note that

$$\|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{U}_1\| = \sup_{\substack{\mathbf{x}_1 \in \mathbb{R}^{\dim \mathcal{X}_1} \\ \mathbf{x}_2 \in \mathbb{R}^{\dim \mathcal{X}_2}}} \mathbf{x}_2^{\mathrm{T}}\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{U}_1\mathbf{x}_1$$

$$= \sup_{\substack{\mathbf{u}_1 \in \mathcal{X}_1, \|\mathbf{u}_1\|=1 \\ \mathbf{u}_2 \in \mathcal{X}_2, \|\mathbf{u}_2\|=1}} \mathbf{u}_2^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{u}_1.$$

But

$$\mathbf{u}_2^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{u}_1 = \frac{1}{4}(\|\boldsymbol{\Phi}(\mathbf{u}_1 + \mathbf{u}_2)\|^2 - \|\boldsymbol{\Phi}(\mathbf{u}_1 - \mathbf{u}_2)\|^2).$$

the conclusion follows from the first part and the fact that $\|\mathbf{u}_1 + \mathbf{u}_2\|^2 = \|\mathbf{u}_1 - \mathbf{u}_2\|^2 = 2$ (since $\mathcal{X}_1$, $\mathcal{X}_2$ are orthogonal to each other). $\qquad\square$

Lemma 7. Under the same setting as in Lemma 2, assume further that $\mathcal{X}_1$, $\mathcal{X}_2$ are orthogonal to each other. Then we have

$$\|\mathcal{P}_{\mathcal{Y}_2}\boldsymbol{\Phi}\mathbf{u}\| \leq \sqrt{\frac{4}{3}}\|\mathbf{u}\|\delta < \frac{4}{3}\|\boldsymbol{\Phi}\mathbf{u}\|\delta \tag{9}$$

for any $\mathbf{u} \in \mathcal{X}_1$.

Proof. Let $\mathbf{U}_1$ be an orthonormal basis for $\mathcal{X}_1$ and $\mathbf{U}_2$ be an orthonormal basis for $\mathcal{X}_2$, thus $\mathbf{U}_1$ (resp. $\mathbf{U}_2$) is a $N \times d_1$ (resp. $N \times d_2$) matrix with orthonormal columns. Set $\mathbf{V}_1 = \boldsymbol{\Phi}\mathbf{U}_1$, $\mathbf{V}_2 = \boldsymbol{\Phi}\mathbf{U}_2$. By (4), $\mathbf{V}_2$ is of full rank and all of its singular values are in $(\sqrt{1-\delta}, \sqrt{1+\delta})$. In this case we have

$$\mathcal{P}_{\mathcal{Y}_2} = \mathbf{V}_2(\mathbf{V}_2^{\mathrm{T}}\mathbf{V}_2)^{-1}\mathbf{V}_2^{\mathrm{T}}.$$

Thus

$$\begin{aligned}
\|\mathcal{P}_{\mathcal{Y}_2}\mathbf{V}_1\|^2 &= \mathbf{V}_1^{\mathrm{T}}\mathbf{V}_2(\mathbf{V}_2^{\mathrm{T}}\mathbf{V}_2)^{-1}\mathbf{V}_2^{\mathrm{T}}\mathbf{V}_1 \\
&\leq s_{\min}^{-2}(\mathbf{V}_2)\|\mathbf{V}_2^{\mathrm{T}}\mathbf{V}_1\|^2 \\
&\leq (1-\delta)^{-1}\|\mathbf{U}_2^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{U}_1\|^2 \\
&\leq (1-\delta)^{-1}\delta^2,
\end{aligned}$$

where the last inequality follows from Proposition 1. For any $\mathbf{u} \in \mathcal{X}_1$, it is possible to find some $\mathbf{x} \in \mathbb{R}_1^d$ such that $\mathbf{u} = \mathbf{U}_1\mathbf{x}$, hence $\|\mathbf{u}\| = \|\mathbf{x}\|$; we thus have

$$\|\mathcal{P}_{\mathcal{Y}_2}\boldsymbol{\Phi}\mathbf{u}\| = \|\mathcal{P}_{\mathcal{Y}_2}\boldsymbol{\Phi}\mathbf{U}_1\mathbf{x}\| = \|\mathcal{P}_{\mathcal{Y}_2}\mathbf{V}_1\mathbf{x}\| \leq \|\mathcal{P}_{\mathcal{Y}_2}\mathbf{V}_1\|\|\mathbf{u}\|.$$

Hence $\|\mathcal{P}_{\mathcal{Y}_2}\boldsymbol{\Phi}\mathbf{u}\| \leq (1-\delta)^{-1/2}\delta\|\mathbf{u}\| \leq \sqrt{\frac{4}{3}}\|\mathbf{u}\|\delta$. Along with Proposition 1, this proves (9). $\qquad\square$

5.2 One-dimensional Subspaces

Lemma 8. Under the same setting as in Theorem 2, we have

$$\left|\text{aff}^2(\mathcal{Y}_2, \boldsymbol{\Phi}\mathbf{u}) - \text{aff}^2(\mathcal{X}_2, \mathbf{u})\right| \leq C\left(1 - \text{aff}^2(\mathcal{X}_2, \mathbf{u})\right)\delta \tag{10}$$

for all $\mathbf{u} \in \mathcal{X}_1$, $\mathbf{u} \neq \mathbf{0}$.

Proof. The proof is by straightforward computation. It suffices to prove (10) for unit vectors $\mathbf{u}_1 \in \mathcal{X}_1$. For any such unit vector, there exists some unit vector $\mathbf{u}_2 \in \mathcal{X}_2$, such that $\lambda \overset{\triangle}{=} \langle \mathbf{u}_1, \mathbf{u}_2 \rangle = \text{aff}(\mathcal{X}_2, \mathbf{u})$ by Lemma 1). In fact, $\mathbf{u}_2$ is the direction vector of the projection of $\mathbf{u}$ onto $\mathcal{X}_2$. We thus have

$$\mathbf{u}_1 = \lambda \mathbf{u}_2 + \sqrt{1 - \lambda^2} \mathbf{u}_0, \tag{11}$$

where $\mathbf{u}_0 \in \mathcal{X}$ is some unit vector orthogonal to $\mathcal{X}_2$. Recall that the squared affinity of $\mathcal{Y}_1$ and $\mathcal{Y}_2$ is defined as

$$\text{aff}_{\mathcal{Y}}^2 = \frac{1}{\|\mathbf{\Phi u}_1\|^2} \|\mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_1\|^2.$$

In light of (11), we have

$$\|\mathbf{\Phi u}_1\|^2 = \lambda^2 \|\mathbf{\Phi u}_2\|^2 + (1 - \lambda^2) \|\mathbf{\Phi u}_0\|^2$$
$$+ 2\lambda \sqrt{1 - \lambda^2} \langle \mathbf{\Phi u}_2, \mathbf{\Phi u}_0 \rangle,$$
$$\|\mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_1\|^2 = \lambda^2 \|\mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_2\|^2 + (1 - \lambda^2) \|\mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_0\|^2$$
$$+ 2\lambda \sqrt{1 - \lambda^2} \langle \mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_2, \mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_0 \rangle. \tag{12}$$

Note that $\mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_2 = \mathbf{\Phi u}_2$ since $\mathbf{\Phi u}_2 \in \mathbf{\Phi} \mathcal{X}_2 = \mathcal{Y}_2$. Furthermore,

$$\langle \mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_2, \mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_0 \rangle = \langle \mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_2, \mathbf{\Phi u}_0 \rangle$$
$$= \langle \mathbf{\Phi u}_2, \mathbf{\Phi u}_0 \rangle,$$

where the first equality is elementary geometry. Thus

$$\|\mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_1\|^2 = \lambda^2 \|\mathbf{\Phi u}_2\|^2 + (1 - \lambda^2) \|\mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_0\|^2$$
$$+ 2\lambda \sqrt{1 - \lambda^2} \langle \mathbf{\Phi u}_2, \mathbf{\Phi u}_0 \rangle.$$

Combining these equations, we have

$$\left| \text{aff}_{\mathcal{Y}}^2 - \lambda^2 \right| = \frac{(1 - \lambda^2)}{\|\mathbf{\Phi u}_1\|^2} \left(\lambda^2 (\|\mathbf{\Phi u}_2\|^2 - \|\mathbf{\Phi u}_0\|^2) \right.$$
$$+ \|\mathcal{P}_{\mathcal{Y}_2} \mathbf{\Phi u}_0\|^2$$
$$\left. + 2\lambda \sqrt{1 - \lambda^2} \langle \mathbf{\Phi u}_2, \mathbf{\Phi u}_0 \rangle \right).$$

Since $\|\mathbf{u}_1\| = \|\mathbf{u}_2\| = \|\mathbf{u}_0\| = 1$ and that $\mathbf{u}_0$ is perpendicular to $\mathcal{X}_2$ (hence to $\mathbf{u}_2$), the above quantity is bounded by $C(1 - \lambda^2)\delta$ by Proposition 1 and Lemma 7. This completes the proof. $\qquad \square$

5.3 The General Case

Now we prove the full version of Theorem 2. Choose a principal orthonormal bases (Lemma 1) $\mathbf{U}_1$, $\mathbf{U}_2$ for $\mathcal{X}_1, \mathcal{X}_2$. In this proof we also borrow the notation of λ_k from Lemma 1. Denote $\mathbf{V}_1 = \boldsymbol{\Phi}\mathbf{U}_1$, $\mathbf{V}_2 = \boldsymbol{\Phi}\mathbf{U}_2$. The k-th column of $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{V}_1$ are respectively denoted by $\mathbf{u}_{1,k}$, $\mathbf{u}_{2,k}$ and $\mathbf{v}_{1,k}$. Note that $\mathbf{v}_{1,k} = \boldsymbol{\Phi}\mathbf{u}_{1,k}$ by definition.

By (4), $\mathbf{V}_1$, $\mathbf{V}_2$ are of full rank and all of their singular values lie in $(\sqrt{1-\delta}, \sqrt{1+\delta})$. We shall need two auxiliary matrices derived from $\mathbf{V}_1$. The first one is the column-normalized version of $\mathbf{V}_1$, defined as $\hat{\mathbf{V}}_1 = [\frac{\mathbf{v}_{1,1}}{\|\mathbf{v}_{1,1}\|}, \ldots, \frac{\mathbf{v}_{1,d_1}}{\|\mathbf{v}_{1,d_1}\|}]$. The second one is the orthogonal matrix obtained from Gram-Schmidt orthogonalization of columns of $\mathbf{V}_1$, which we denote by $\mathbf{Q}_1$. The k-th column of $\hat{\mathbf{V}}_1$ and $\mathbf{Q}_1$ are respectively denoted by $\hat{\mathbf{v}}_{1,k}$ and $\mathbf{q}_{1,k}$. Let $\mathcal{P}_{\mathcal{Y}_2}$ be the orthogonal projection onto $\mathcal{Y}_2$, i.e. the column space of $\mathbf{V}_2$. We have

$$
\begin{aligned}
\mathrm{aff}_{\mathcal{Y}}^2 - \mathrm{aff}_{\mathcal{X}}^2 = & \|\mathcal{P}_{\mathcal{Y}_2}\mathbf{Q}_1\|_{\mathrm{F}}^2 - \|\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1\|_{\mathrm{F}}^2 \\
= & (\|\mathcal{P}_{\mathcal{Y}_2}\mathbf{Q}_1\|_{\mathrm{F}}^2 - \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{V}}_1\|_{\mathrm{F}}^2) \\
& + (\|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{V}}_1\|_{\mathrm{F}}^2 - \|\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1\|_{\mathrm{F}}^2).
\end{aligned}
\tag{13}
$$

We estimate the last two quantities in (13) respectively.

Proposition 2. We have

$$
\left| \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{V}}_1\|_{\mathrm{F}}^2 - \|\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1\|_{\mathrm{F}}^2 \right| \le C(d_1 - \mathrm{aff}_{\mathcal{X}}^2)\delta.
$$

Proof. Note that

$$
\|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{V}}_1\|_{\mathrm{F}}^2 - \|\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1\|_{\mathrm{F}}^2 = \sum_{k=1}^{d_1} \left(\|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{v}}_{1,k}\|^2 - \|\mathbf{U}_2^{\mathrm{T}}\mathbf{u}_{1,k}\|^2 \right).
\tag{14}
$$

Observe that $\|\mathbf{U}_2^{\mathrm{T}}\mathbf{u}_{1,k}\|$ is the affinity between $\mathcal{X}_2$ and the one-dimensional subspace spanned by $\mathbf{u}_{1,k}$, while $\|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{v}}_{1,k}\|$ is the affinity between $\mathcal{Y}_2$ and the one-dimensional subspace spanned by $\boldsymbol{\Phi}\mathbf{u}_{1,k}$. Furthermore, $\|\mathbf{U}_2^{\mathrm{T}}\mathbf{u}_{1,k}\| = \lambda_k$ since $\mathbf{U}_1$, $\mathbf{U}_2$ are principal orthonormal bases. By Lemma 7, we have

$$
\left| \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{v}}_{1,k}\|^2 - \|\mathbf{U}_2^{\mathrm{T}}\mathbf{u}_{1,k}\|^2 \right| \le C(1 - \lambda_k^2)\delta.
\tag{15}
$$

Summing up, we obtain

$$
\begin{aligned}
\left| \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{V}}_1\|_{\mathrm{F}}^2 - \|\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1\|_{\mathrm{F}}^2 \right| & \le C(d_1 - \sum_{k=1}^{d_1} \lambda_k^2)\delta \\
& = C(d_1 - \mathrm{aff}_{\mathcal{X}}^2)\delta,
\end{aligned}
$$

as desired. □

Proposition 3. We have

$$\left| \|\mathcal{P}_{\mathcal{Y}_2}\mathbf{Q}_1\|_{\mathrm{F}}^2 - \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{V}}_1\|_{\mathrm{F}}^2 \right| \le C(d_1 - \mathrm{aff}_{\mathcal{X}}^2)\delta.$$

Proof. Similar to (14), we have

$$\|\mathcal{P}_{\mathcal{Y}_2}\mathbf{Q}_1\|_{\mathrm{F}}^2 - \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{V}}_1\|_{\mathrm{F}}^2 = \sum_{i=1}^{d_1}(\|\mathcal{P}_{\mathcal{Y}_2}\mathbf{q}_{1,k}\|^2 - \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{v}}_{1,k}\|^2).$$

Denote by $\mathcal{Z}_k$ the space spanned by $\mathbf{v}_{1,1}, \ldots, \mathbf{v}_{1,k}$. Then

$$\mathbf{q}_{1,k} = \frac{\hat{\mathbf{v}}_{1,k} - \mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}}{\|\hat{\mathbf{v}}_{1,k} - \mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|}. \tag{16}$$

Note that $\mathbf{q}_{1,k}$, $\hat{\mathbf{v}}_{1,k}$ are unit vectors, hence by Pythagorean theorem we have

$$\|\hat{\mathbf{v}}_{1,k} - \mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2 = 1 - \|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2, \tag{17}$$

and

$$
\begin{aligned}
&\|\mathcal{P}_{\mathcal{Y}_2}\mathbf{q}_{1,k}\|^2 - \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{v}}_{1,k}\|^2 \\
&= (1 - \|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{q}_{1,k}\|^2) - (1 - \|\mathcal{P}_{\mathcal{Y}_2^\perp}\hat{\mathbf{v}}_{1,k}\|^2) \\
&= \|\mathcal{P}_{\mathcal{Y}_2^\perp}\hat{\mathbf{v}}_{1,k}\|^2 - \|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{q}_{1,k}\|^2,
\end{aligned} \tag{18}
$$

where $\mathcal{Y}_2^\perp$ denotes the orthogonal complement of $\mathcal{Y}_2$. Combining (16) and (17) we obtain

$$
\begin{aligned}
\|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{q}_{1,k}\|^2 &= \frac{\|\mathcal{P}_{\mathcal{Y}_2^\perp}\hat{\mathbf{v}}_{1,k} - \mathcal{P}_{\mathcal{Y}_2^\perp}\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2}{1 - \|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2} \\
&= \frac{\|\mathcal{P}_{\mathcal{Y}_2^\perp}\hat{\mathbf{v}}_{1,k}\|^2}{1 - \|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2} \\
&\quad - \frac{2\langle\mathcal{P}_{\mathcal{Y}_2^\perp}\hat{\mathbf{v}}_{1,k}, \mathcal{P}_{\mathcal{Y}_2^\perp}\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\rangle}{1 - \|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2} \\
&\quad + \frac{\|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2}{1 - \|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2}.
\end{aligned}
$$

This together with (18) yields

$$
\begin{aligned}
&\|\mathcal{P}_{\mathcal{Y}_2}\mathbf{q}_{1,k}\|^2 - \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{v}}_{1,k}\|^2 \\
&= -\frac{\|\mathcal{P}_{\mathcal{Y}_2^\perp}\hat{\mathbf{v}}_{1,k}\|^2\|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2}{1 - \|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2} \\
&\quad + \frac{2\langle\mathcal{P}_{\mathcal{Y}_2^\perp}\hat{\mathbf{v}}_{1,k}, \mathcal{P}_{\mathcal{Y}_2^\perp}\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\rangle}{1 - \|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2} \\
&\quad - \frac{\|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2}{1 - \|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2}.
\end{aligned} \tag{19}
$$

21

Since $\mathbf{u}_{1,k}$ is perpendicular to the subspace spanned by $\mathbf{u}_{1,1}, \ldots, \mathbf{u}_{1,k-1}$, by Lemma 7, $\|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2 \leq \frac{16}{9}\delta^2 < \frac{4}{9}\delta < 1/2$. The proof would be complete once we show

$$\|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2 \leq C(1-\lambda_k^2)\delta^2, \tag{20}$$

$$\|\mathcal{P}_{\mathcal{Y}_2^\perp}\hat{\mathbf{v}}_{1,k}\|^2 \leq C(1-\lambda_k^2). \tag{21}$$

By (15) and the discussion prior to it (which says that $\|\mathbf{U}_2^{\mathrm{T}}\mathbf{u}_{1,k}\| = \lambda_k$), the following holds:

$$\begin{aligned}
\|\mathcal{P}_{\mathcal{Y}_2^\perp}\hat{\mathbf{v}}_{1,k}\|^2 &= 1 - \|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{v}}_{1,k}\|^2 \\
&= (1-\lambda_k^2) - (\|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{v}}_{1,k}\|^2 - \|\mathbf{U}_2^{\mathrm{T}}\mathbf{u}_{1,k}\|^2) \\
&\leq (1-\lambda_k^2)(1+C\delta) \\
&\leq C(1-\lambda_k^2),
\end{aligned}$$

where the first inequality follows from Lemma 8 applied to $\mathcal{X}_2$ and the 1-dimensional subspace spanned by $\mathbf{u}_{1,k}$, noting that $\|\mathcal{P}_{\mathcal{Y}_2}\hat{\mathbf{v}}_{1,k}\|$ is the affinity between $\mathcal{Y}_2$ and $\mathbf{\Phi}\mathbf{u}_{1,k}$. This establishes (21). For (20), however, some more work is required. Let $\mathbf{Z}_{k-1}$ be a orthonormal basis of $\mathcal{Z}_{k-1}$. Then $\mathcal{P}_{\mathcal{Z}_{k-1}} = \mathbf{Z}_{k-1}\mathbf{Z}_{k-1}^{\mathrm{T}}$, which implies

$$\begin{aligned}
\|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2 &= \|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{Z}_{k-1}\mathbf{Z}_{k-1}^{\mathrm{T}}\hat{\mathbf{v}}_{1,k}\|^2 \\
&\leq s_{\max}^2(\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{Z}_{k-1})\|\mathbf{Z}_{k-1}^{\mathrm{T}}\hat{\mathbf{v}}_{1,k}\|^2.
\end{aligned}$$

One recognizes at once that $\|\mathbf{Z}_{k-1}^{\mathrm{T}}\hat{\mathbf{v}}_{1,k}\|^2 = \|\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|^2$, which is bounded by $C\delta^2$ according to Lemma 7. It remains to prove

$$s_{\max}^2(\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{Z}_{k-1}) \leq C(1-\lambda_k^2). \tag{22}$$

Since $\mathbf{Z}_{k-1}$ has orthonormal columns, it follows that

$$\begin{aligned}
s_{\max}^2(\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{Z}_{k-1}) &= \sup_{\mathbf{x}\in\mathbb{S}^{k-2}} \|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{Z}_{k-1}\mathbf{x}\|^2 \\
&= \sup_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}\in\mathcal{Z}_{k-1}}} \|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{x}\|^2. \tag{23}
\end{aligned}$$

By definition, $\mathcal{Z}_{k-1}$ is spanned by the first $(k-1)$ columns of $\mathbf{V}_1 = \mathbf{\Phi}\mathbf{U}_1$. Denote by

$\mathbf{U}_{1,1:k-1}$ the first $(k-1)$ columns of $\mathbf{U}_1$. We have

$$\sup_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}\in\mathcal{Z}_{k-1}}} \|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{x}\|^2 = 1 - \inf_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}\in\mathcal{Z}_{k-1}}} \|\mathcal{P}_{\mathcal{Y}_2}\mathbf{x}\|^2$$

$$= 1 - \inf_{\mathbf{x}\in\mathbb{S}^{k-2}} \frac{\|\mathcal{P}_{\mathcal{Y}_2}\mathbf{\Phi}\mathbf{U}_{1,1:k-1}\mathbf{x}\|^2}{\|\mathbf{\Phi}\mathbf{U}_{1,1:k-1}\mathbf{x}\|^2}$$

Note that

$$\frac{\|\mathcal{P}_{\mathcal{Y}_2}\mathbf{\Phi}\mathbf{U}_{1,1:k-1}\mathbf{x}\|^2}{\|\mathbf{\Phi}\mathbf{U}_{1,1:k-1}\mathbf{x}\|^2} = \mathrm{aff}^2(\mathcal{Y}_2, \mathbf{\Phi}\mathbf{U}_{1,1:k-1}\mathbf{x}).$$

Applying Lemma 8 to $\mathcal{X}_2$ and the 1-dimensional subspace spanned by $\mathbf{U}_{1,1:k-1}\mathbf{x}$, we obtain

$$\sup_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}\in\mathcal{Z}_{k-1}}} \|\mathcal{P}_{\mathcal{Y}_2^\perp}\mathbf{x}\|^2 \leq (1 - \|\mathcal{P}_{\mathcal{X}_2}\mathbf{U}_{1,1:k-1}\mathbf{x}\|^2)(1 + C\delta)$$

$$\leq C(1 - \|\mathcal{P}_{\mathcal{X}_2}\mathbf{U}_{1,1:k-1}\mathbf{x}\|^2) \tag{24}$$

Finally we need to bound $\|\mathcal{P}_{\mathcal{X}_2}\mathbf{U}_{1,1:k-1}\mathbf{x}\|^2$. Since $\mathcal{P}_{\mathcal{X}_2}\mathbf{u}_{1,i} = \lambda_i\mathbf{u}_{2,i}$ by choice of principal orthonormal bases, we have

$$\|\mathcal{P}_{\mathcal{X}_2}\mathbf{U}_{1,1:k-1}\mathbf{x}\|^2 = \sum_{i=1}^{k-1}\lambda_i^2 x_i^2 \geq \sum_{i=1}^{k-1}\lambda_k^2 x_i^2$$

$$= \lambda_k^2. \tag{25}$$

Combining (23), (24) and (25), we obtain (22) as desired. $\qquad\square$

6 Discussions

This section discusses the proofs of Theorem 2 and other results in this paper, and their differences from previous works.

A notable feature of Theorem 2 is being purely deterministic, i.e. does not involve randomness of $\mathbf{\Phi}$. This cleans up the fog spanned across the mixture of probabilistic arguments and deterministic arguments in previous works on dimensionality-reduced subspace clustering [12,13] and on subspace RIP [15]. Note that the concept of subspace embedding (see Remark 4) that is somehow similar to (4) appears in the analysis in [13]. Theorem 2 is distinguished from theirs at least in two aspects: our analysis is deterministic and does not involve randomness, while the assumption in [13] is an probabilistic inequality; we proceed

23

to analyze the subspace RIP of $\boldsymbol{\Phi}$, which is capable of handling various subspace-related tasks and algorithms, while [13] considered only SSC, a special algorithm of the specific problem of subspace clustering. Also note that the result in [13] requires $n = \Omega(d^{9/2})$ for subgaussian matrices to ensure success of CSC, which is worse than the optimal scaling $n = O(d)$ that can be obtained by our theory, for instance, using the framework in [10].

Readers may find some similarity between the proof of Theorem 2 and the proof in [15]. Indeed, the routine computations in both proofs are the same, e.g. (12), (13)–(19). However, these routine computations are merely a non-substantial part of the proof, and the core difficulty is how to bound the quantities involved in these equations. To this end, our proof deviate significantly from that in [15]. In fact, [15] relied heavily on the fact that the Gaussian projection of two orthogonal vectors are independent to bound terms such as $\|\mathcal{P}_{\mathcal{Y}_2}\mathbf{V}_1\|$ in Lemma 7 and $\|\mathcal{P}_{\mathcal{Y}_2}\boldsymbol{\Phi}\mathbf{u}_0\|$ in (12), which cannot be generalized to even subgaussian matrices, let alone partial Fourier matrices, partial circulant matrices or log-concave ensembles, and the situation got more involved when the quantity of interest is complicated: it took over four pages in [15] (see Appendix 8.8 there) to obtain the sought-for conditional independence to bound $\|\mathcal{P}_{\mathcal{Y}_2^{\perp}}\mathcal{P}_{\mathcal{Z}_{k-1}}\hat{\mathbf{v}}_{1,k}\|$ in (20). In our proof, we propose the novel Lemma 8 (which was proved in [15] but only for one-dimensional $\mathcal{X}_1$) as a key intermediate step, and all the bounds needed follow easily from either assumption (4) or Lemma 8. This demonstrates the power of our abstract setting (4).

7 Applications

In this section we briefly mention some of the applications of our theory. As we mentioned before, our theory provides a universal framework to analyze the effects of random compression on subspace related tasks, and providing a list of such tasks would be out of scope of this paper. However, it is possible to describe the universal framework which works for any subspace-related algorithms that admits a theoretical guarantee via affinity, as follows.

Framework. *Assume that we have a theoretical guarantee for an algorithm A that succeeds on a collection of L subspaces $\mathcal{X}_1, \ldots, \mathcal{X}_L$ with probability at least $1 - \delta$. Let $\boldsymbol{\Phi}$ be the random matrix described in Corollary 1, 2, 3, 4 or 5. Let ε be a sufficiently small positive number that depends only on the relative position of subspaces. Then for n satisfying the restriction in the corresponding corollary, algorithm A succeeds on the data projected by $\boldsymbol{\Phi}$ with prob-*

ability at least $1 - \delta_{proj} - \delta$, where δ_{proj} is the error probability given in the corresponding corollary.

As examples, we roughly describe some useful consequences of our theory on two tasks: subspace clustering and active subspace detection [32].

Theorem 3 (Compressed subspace clustering)**.** Let $\boldsymbol{\Phi}$ be a partial Fourier matrix. Under some technical assumptions on the algorithm parameters (that is irrelevant of $\boldsymbol{\Phi}$, see [10]), the Threshold-based Subspace Clustering (TSC) algorithm succeeds (see [10]) on the dataset compressed by $\boldsymbol{\Phi}$ with probability at least

$$1 - \frac{10}{M} - \sum_{l=1}^{L} \mathrm{e}^{-c(N_l+1)} - \mathrm{e}^{-c(\sqrt{d_*^2 + K^{-2}\varepsilon^2 n} - d_*)}$$

given $n > C\varepsilon^{-3} \max\{d_*(\log^3 d_* + \log L), \log^2 L, \log^3 N\}$ and

$$\max_{k \neq l} \sqrt{\frac{\mathrm{aff}^2(\mathcal{X}_k, \mathcal{X}_l)(1 - \varepsilon) + d_*\varepsilon}{d_k \wedge d_l}}$$

$$+ 6\sqrt{\frac{d_*}{n}} \left(1 + \sqrt{\frac{6\log M}{d_*}}\right)^2$$

$$\leq \frac{1}{\sqrt{6M_{\max}\log M}},$$

where $c > 0$ is a constant, M denotes the number of data points and $M_{\max}$ denotes the maximal number of data points lying in the same subspace.

This follows from Corollary 3 the analysis scheme proposed in [10]. We have chosen partial Fourier matrix and TSC algorithm only for simplicity of presentation; similar results hold for subgaussian matrices, partial circulant matrices, etc., and for SSC, SSC-OMP algorithms, etc. It is perhaps worth mentioning that Theorem 3 is, to the best of our knowledge, the first effective performance analysis with sub-optimal scaling $n = O(d_*\mathrm{polylog}(d_*))$ for TSC compressed by partial Fourier matrices, since the analysis in [12] appears flawed as pointed out in a footnote of this paper.

We now turn to the task of active subspace detection.

Theorem 4 (Compressed active subspace detection)**.** Let $\boldsymbol{\Phi}$ be a partial Fourier matrix. The compressed maximum-likelihood detector (see [33]) for noiseless active subspace detection succeeds, under Gaussian assumption on the data distribution, with probability at least

$$1 - 4\sum_{i \neq j} \mathrm{e}^{-K(\mathrm{aff}_*(1-\varepsilon)+d\varepsilon)d_*} - \mathrm{e}^{-c(\sqrt{d_*^2 + K^{-2}\varepsilon^2 n} - d_*)},$$

where

$$K(x) := \frac{1}{8} \frac{(1 - x/d_* - 8/d_*)^2}{4 + (1 - x/d_* - 8/d_*)}$$

and aff_* denotes $\max_{k \neq l} \mathrm{aff}(\mathcal{X}_k, \mathcal{X}_l)$, given $\varepsilon \in (0, 1)$ and $n > C\varepsilon^{-3} \max\{d_*(\log^3 d_* + \log L), \log^2 L, \log^3 N\}$. For noisy case a similar conclusion holds.

This follows from Corollary 2 and the analysis scheme proposed in [33]. Note that in [33] the above theorem is proved for random matrices with exponential Johnson-Lindenstrauss property (Theorem 5 there) using results from the first version of this paper. Again, the appearance of partial Fourier matrices is arbitrary and can be replaced by any other random matrices discussed in this paper.

8 Simulations

We verify our results on Yale Face Database B [34] and test the performance of Sparse Subspace Clustering (SSC) after random projection by Gaussian matrix, partial Fourier/Hadamard matrix, partial circulant matrix, and matrix with i.i.d. Student-t distributed ($\nu = 5$) entries. Yale Face Database B is a database with ambient dimension $N = 32256$ that contains the face images of 10 human subjects. For convenience we randomly select 4 subjects whose face images are subsequently clustered. The matrices we chose are representatives of the three classes of random matrices which we have inspected: Gaussian matrix represents matrices with exponential Johnson-Lindenstrauss property, partial Fourier/Hadamard matrix and partial circulant matrix represent matrices with fast algorithms, and i.i.d. Student-t distributed matrices, which has infinite fifth moments, represents matrices with heavy tails.

Performance is evaluated in terms of clustering error rate [6], i.e. the rate that SSC algorithm clusters a randomly compressed image to the correct subject, see Fig. 1. We are also concerned with the boost-up in computational efficiency supplied by fast matrix-vector multiplication algorithms for partial Fourier/Hadamard matrices and partial circulant matrices, which will be evaluated in terms of average running time, i.e. the time it takes to compute the random projection of a high-dimensional vector, see Fig. 2.

Computational Complexity For unstructured $n \times N$ random matrices such as Gaussian matrices and Student-t matrices, it takes $\mathcal{O}(nN)$ time to compute the random projection of a

vector. For partial Fourier/Hadamard matrices and partial circulant matrices, $\mathcal{O}(N \log N)$-time algorithms exist, thanks to Fast Fourier Transform (FFT) and Fast Walsh-Hadamard Transform (FWHT). More precisely, one may compute the sign-randomized[5] version of a vector $\mathbf{x} \in \mathbb{R}^N$ in $\mathcal{O}(N)$ time, and then compute its fast Fourier/Walsh-Hadamard transform $\hat{\mathbf{x}}$ in $\mathcal{O}(N \log N)$ time. By randomly sampling n entries from $\hat{\mathbf{x}}$, which takes $\mathcal{O}(n)$ time, one finally obtains the randomly projected version of $\mathbf{x}$. For typical scenarios in practice we have $n \gg \log N$, hence random projections by partial Fourier/Hadamard matrices and partial circulant matrices are much efficient than random projections by unstructured matrices.

Discussions   As one may see in Fig. 1, the error rates of all types of random projections converge to the baseline, i.e. the error rate of SSC without random projection, as n tends to N. This is consistent with our theory that all these random matrices preserve affinities between subspaces. The running time of random projection shown in Fig. 2 coincides with our analysis above. Note that the running time is plotted in logarithmic scale. For partial Fourier/Hadamard matrix and partial circulant matrix, the running time is almost constant in n. Both the running times of Student-t matrix and Gaussian matrix grow linearly with respect to n; this is because that they both involve a $\mathcal{O}(nN)$-time matrix-by-vector multiplication. Note that Student-t matrix takes a somewhat longer time than Gaussian matrix, possibly due to higher complexity in its implementation, i.e. in generating Student-t distributed variables. The running time of partial Hadamard matrix is longer than that of partial Fourier matrix and partial circulant matrix, which may be caused by a less efficient implementation of FWHT than that of FFT. Except for very small n, partial Fourier/Hadamard matrix and partial circulant matrix is significantly faster than Gaussian matrix and Student-t matrix. For small n the running time of Gaussian/Student-t matrix is shorter than that of partial Hadamard matrix. However, as n grows large, for instance when $n > 10000$, partial Hadamard matrix becomes the better choice. Our analysis indicates that this advantage would become even more obvious when the ambient dimension N is larger and $n \gg \log N$.

---

[5]This means multiplying each entry of $\mathbf{x}$ by a Rademacher random variable. See Theorem 6 for details.
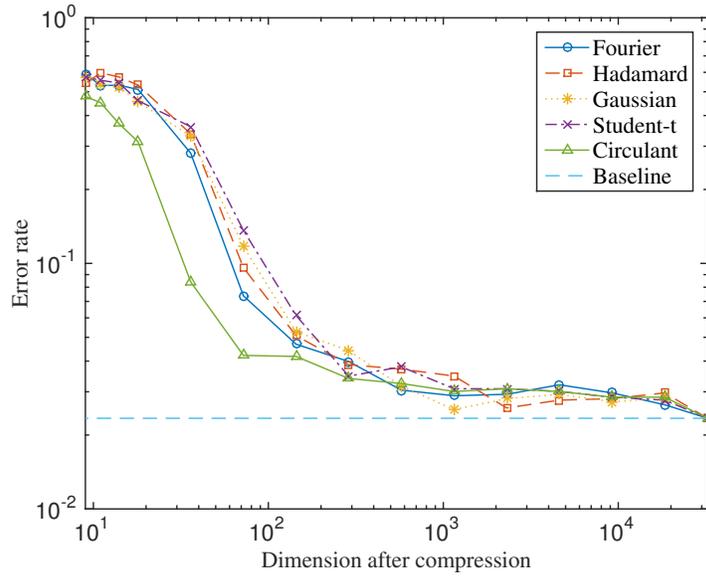
Figure 1: Clustering error rate vs. compressed dimension n for Yale Face Dataset B. The error rate of SSC without random compression is approximately 2.3%.
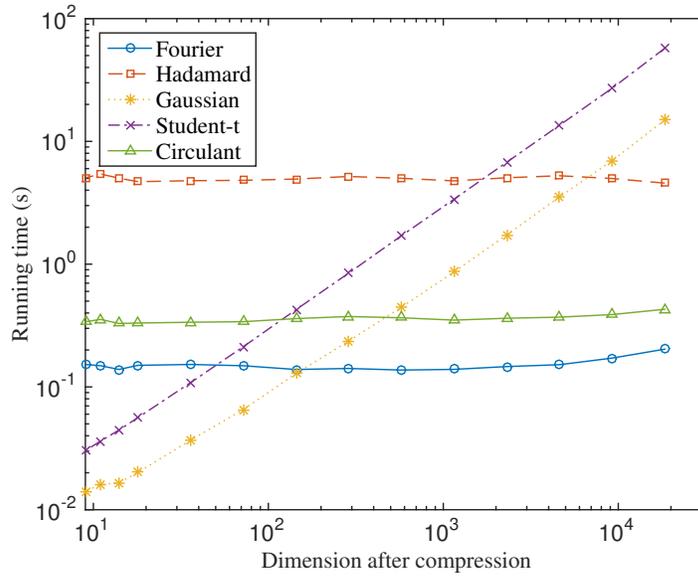

Figure 2: Average running time of random projecting 256 vectors vs. compressed dimension n. Note that for partial Fourier matrix and partial Hadamard matrix the running time is almost independent of n. Both x-axis and y-axis are drawn in logarithmic scale, so the running time of Gaussian matrix and Bernoulli matrix grows linearly in the figure.

9   Conclusion

In this paper we provided a deterministic characterization of subspace RIP in terms of near-isometry on subspaces. This result enables to analyze the subspace RIP of matrices with a unified approach. As examples, we prove with this result that a large variety of random matrices, including subgaussian matrices, partial Fourier/Hadamard matrices and partial circulant/Toeplitz matrices, random matrices with independent strongly regular rows, and log-concave ensembles. This significantly enlarges the collection of random matrices known to possess subpace RIP in literature, demonstrating the applicability of subspace RIP.

Subspace RIP, or in plain language, the almost-invariance of affinity under random projections, has played an important role in the analysis of Compressed Subspace Clustering algorithms. Hence our result demonstrates more scenarios where random projection and CSC may apply and has the potential to give better performance guarantee for CSC algorithms. Furthermore, since subspace RIP is a universal concept that does not depend on any specific algorithm, our result may find its application in various subspace-based machine learning algorithms, which we leave to future research.

10   Appendix: Exponential Johnson-Lindenstrauss Property

This appendix deals with random matrices with exponential Johnson-Lindenstrauss property and their subspace RIP. The appendix is divided into two parts. In the first part, we prove that subgaussian matrices and partial Fourier/Hadamard matrices satisfy exponential Johnson-Lindenstrauss property to illustrate the wide applicability of this concept. In the second part, we introduce the standard tool of covering arguments and use it to prove Lemma 2.

10.1   Examples of Random Matrices with Exponential JL Property

Here we provide a non-comprehensive list of common random matrices that fulfill exponential Johnson-Lindenstrauss property. Such matrices can be roughly divided into two categories, whose exponential Johnson-Lindenstrauss property respectively stems from subgaussian concentration property and Restricted Isometry Property for sparse vectors. The most important example in the first category is subgaussian random matrices, and that in the second category is randomly sampled Bounded Orthogonal Systems (BOS), a class of

random matrices including partial Fourier matrices and partial Hadamard matrices. We discuss these two categories respectively.

Subgaussian concentration   We begin by defining subgaussian random variables and subgaussian random vectors.

Definition 4 (Section 7.1, [16])**.** Fix a positive constant K. A K-*subgaussian random variable* is a random variable X satisfying

$$\mathbb{E}e^{tx} \leq e^{K^2 t^2/2},$$

for any $t \in \mathbb{R}$. A *subgaussian matrix* is a random matrix with each entry a K-subgaussian random variable.

Gaussian variables might be the most common examples of subgaussian random variables. Other examples include variables with Rademacher distribution or uniform distribution on $[-1, 1]$. In fact, any centered bounded random variable is subgaussian.

It is possible to generalized the above definition to a multi-dimensional setting.

Definition 5. Let $\mathbf{\Gamma}$ be a positive semidefinite matrix. An $\mathbf{\Gamma}$-*subgaussian random vector* is a random vector $\mathbf{u}$ taking value in $\mathbb{R}^n$ such that for any $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbb{E}e^{\langle \mathbf{u}, \mathbf{x} \rangle} \leq e^{\langle \mathbf{\Gamma}\mathbf{x}, \mathbf{x} \rangle/2}.$$

Such a random vector is said to satisfy *Bernstein condition*, if

$$\mathbb{E}\left| \|\mathbf{u}\|^2 - \mathbb{E}\|\mathbf{u}\|^2 \right|^k \leq Ck! \|\mathbf{\Gamma}\|_{\mathrm{op}}^{k-2} \|\mathbb{E}(\mathbf{u}\mathbf{u}^{\mathrm{T}})\|_{\mathrm{F}}^2$$

for some constant $C > 0$, where $\| \cdot \|_{\mathrm{op}}$ denotes the operator norm.

Theorem 5 ([35], Theorem 2.10)**.** If $\mathbf{A}$ is a random matrix with independent $\mathbf{\Gamma}$-subgaussian columns satisfying Bernstein condition. Furthermore, assume $\mathbb{E}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = \mathbf{I}$. Then $\mathbf{A}$ satisfies exponential Johnson-Lindenstrauss property.

Corollary 6. Assume $\mathbf{A}$ is a random matrix satisfying $\mathbb{E}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = \mathbf{I}$. If in addition $\mathbf{A}$ is in one of the following form, then $\mathbf{A}$ satisfies exponential Johnson-Lindenstrauss property:

 a) A random matrix with independent subgaussian rows;

 b) A Gaussian matrix with independent columns;

c) A product of a positive semidefinite matrix and a subgaussian matrix with independent entries.

Proof. a) is classical and can be found in [27]; b) and c) are proved in [35]. □

Restricted Isometry Property   Restricted isometry property for sparse vectors [23] has been a very powerful tool in analysis of compressed sensing and related algorithms. A vector is called s-sparse if at most s of its entries are non-zero. RIP for sparse vectors are defined in the following way:

Definition 6. A matrix $\mathbf{A}$ is said to possess the RIP if there exists a function $\delta(s) \geq 0$, such that for any positive integer s and any s-sparse vector $\mathbf{x}$,

$$(1 - \delta(s))\|\mathbf{x}\|^2 \leq \|\mathbf{A}\mathbf{x}\|^2 \leq (1 + \delta(s))\|\mathbf{x}\|^2.$$

The function $\delta(s)$ is called the *restricted isometry constant* of $\mathbf{A}$.

It is easy to see that exponential Johnson-Lindenstrauss property implies RIP, see for instance Theorem 5.2 in [36]). The converse is also true in some sense, as the following theorem shows.

Theorem 6 ([37, 38]). Assume $\mathbf{A}$ is an $n \times N$ matrix with RIP and restricted isometry constant $\delta(s)$. Fix some $\varepsilon \in (0, 1/2)$. Assume further that for some $s > 0$ we have $\delta(s) < \varepsilon/4$. Let $\mathbf{D}_\epsilon$ be a diagonal matrix with i.i.d. Rademacher random variables on its diagonal, then for any $\mathbf{x} \in \mathbb{R}^N$, we have

$$\mathbb{P}\big(\big|\|\mathbf{A}\mathbf{D}_\epsilon\mathbf{x}\|^2 - \|\mathbf{x}\|^2\big| > \varepsilon\|\mathbf{x}\|^2\big) \leq 2\mathrm{e}^{-\tilde{c}s}$$

for some universal constant $\tilde{c} > 0$.

Partial Fourier matrices and partial Hadamard matrices are both examples of a more general class of random matrices, namely random sampled Bounded Orthonormal Systems (BOS). For such matrices it was shown that their restricted isometry constants are sufficiently small:

Theorem 7 ([16, 37]). Let $\mathbf{A} \in \mathbb{C}^{n \times N}$ be the random sampling associated to a BOS with constant $K \geq 1$. For $\zeta, \eta_1, \eta_2 \in (0, 1)$, if

$$\frac{n}{\log(9n)} \geq C_1 \eta_1^{-2} K^2 s \log^2(4s) \log(8N),$$

$$n \geq C_2 \eta_2^{-2} K^2 s \log(\zeta^{-1}),$$

then with probability at least $1 - \zeta$ the restricted isometry constant $\delta(s)$ of $\frac{1}{\sqrt{n}}\mathbf{A}$ satisfies $\delta(s) \le \eta_1 + \eta_1^2 + \eta_2$. (Here C_1, C_2 are universal positive constants.)

One may combine Theorem 6 and 7 to obtain several modified versions of exponential Johnson-Lindenstrauss property that random sampled BOS satisfy. For example, taking $\eta_1 = \eta_2 = \varepsilon/4$, $s = \lceil C\varepsilon^2 n/\sqrt{N} \rceil$ and $\zeta = \mathrm{e}^{-C's}$, one may obtain that

$$\mathbb{P}(\big| \|\mathbf{A}\mathbf{D}_\epsilon \mathbf{x}\|^2 - \|\mathbf{x}\|^2 \big| > \varepsilon \|\mathbf{x}\|^2) \le 2\mathrm{e}^{-c\varepsilon^2 n/\sqrt{N}}.$$

10.2 Proof of Lemma 2

We will use some standard covering arguments, e.g. [36], to prove (4) for random matrices with exponential Johnson-Lindenstrauss property.

Definition 7. An ε-*net* in a subset X of a Euclidean space is a finite subset $\mathcal{N}$ of X such that for any $x \in X$ we have

$$\min_{z \in \mathcal{N}} \|x - z\| < \varepsilon.$$

The *metric entropy* of X is a function $N(X, \varepsilon)$ defined as the minimum cardinality of an ε-net of X.

For subsets of Euclidean space, the metric entropy can be easily bounded by a volume packing argument. For the Euclidean unit ball the corresponding result reads as following:

Lemma 9 (Proposition C.3, [16]). Let B_n be the unit ball in $\mathbb{R}^n$. Then

$$N(B_n, \varepsilon) \le \left(1 + \frac{2}{\varepsilon}\right)^n.$$

The usage of covering arguments is demonstrated by the following lemma:

Lemma 10 ([27], Lemma 5.3). Suppose $\mathcal{N}$ is a $\frac{1}{2}$-net of $\mathbb{S}^{n-1}$. Let $\mathbf{A}$ be a $n \times n$ matrix. Then

$$\|\mathbf{A}\| \le 2 \sup_{\mathbf{x} \in \mathcal{N}} \|\mathbf{A}\mathbf{x}\|.$$

Now we are ready to finish the proof of Lemma 2. Note that it suffices to show

$$\mathbb{P}\left(\max_{\mathbf{x} \in \mathbb{S}^{d-1}} \big| \|\mathbf{\Phi}\mathbf{U}\mathbf{x}\|^2 - 1 \big| > \varepsilon\right) \le \mathrm{e}^{-\tilde{c}\varepsilon^2 n + 3d}. \tag{26}$$

For any $\mathbf{x} \in \mathbb{S}^{d-1}$, exponential Johnson-Lindenstrauss property implies

$$\mathbb{P}(\big| \|\mathbf{\Phi}\mathbf{U}\mathbf{x}\|^2 - 1 \big| > \varepsilon) \le 2\mathrm{e}^{-\tilde{c}\varepsilon^2 n}. \tag{27}$$

The desired inequality (26) follows from (27) and a standard covering argument. By Lemma 9, one may find a set $\mathcal{N} \subseteq \mathbb{S}^{d-1}$ with cardinality 5^d such that

$$\max_{\mathbf{x} \in \mathbb{S}^{d-1}} \min_{\mathbf{z} \in \mathcal{N}} \|\mathbf{x} - \mathbf{z}\| \leq \frac{1}{2}.$$

Then by Lemma 10

$$\max_{\mathbf{x} \in \mathbb{S}^{d-1}} \left| \|\boldsymbol{\Phi}\mathbf{U}\mathbf{x}\|^2 - 1 \right| \leq 4 \max_{\mathbf{x} \in \mathcal{N}} \left| \|\boldsymbol{\Phi}\mathbf{U}\mathbf{x}\|^2 - 1 \right|. \tag{28}$$

By (27), (28) and union bound,

$$\mathbb{P}(\max_{\mathbf{x} \in \mathcal{N}} \left| \|\boldsymbol{\Phi}\mathbf{U}\mathbf{x}\|^2 - 1 \right| > \varepsilon) \leq 2 \cdot 5^d \mathrm{e}^{-\tilde{c}\varepsilon^2 n}.$$

The proof is completed once we note that $2 \cdot 5^d \leq e^{3d}$ for $d \geq 1$.

11 Appendix: Randomly Sampled BOS and Partial Circulant Matrices

The purpose of this appendix is to prove Lemma 3 and Lemma 4. We will need Theorem 6 and Theorem 7 as stated in Section 10 as well as the covering argument adapted there.

11.1 Randomly Sampled BOS

Proof of Lemma 3. By Theorem 7, the restricted isometry constants of $\boldsymbol{\Phi}$ satisfy

$$\mathbb{P}\left(\delta(s) \leq \frac{\varepsilon}{4} \right) \geq 1 - \exp\left(-C^{-1}\varepsilon^2 K^{-2} \frac{n}{s} \right), \tag{29}$$

given

$$n \geq C K^2 \varepsilon^{-2} s \log^2 s \log(K^2 \varepsilon^{-2} s \log N) \log N. \tag{30}$$

If $\delta(s) \leq \varepsilon/4$ holds, then by Theorem 6 and a standard covering argument (c.f. proof of Lemma 2), $1-\varepsilon < s_{\min}(\boldsymbol{\Phi}\mathbf{U}) \leq s_{\max}(\boldsymbol{\Phi}\mathbf{U}) < 1+\varepsilon$ holds with probability at least $1-\mathrm{e}^{-cs+3d}$. Thus by union bound,

$$1 - \varepsilon < s_{\min}(\boldsymbol{\Phi}\mathbf{U}) \leq s_{\max}(\boldsymbol{\Phi}\mathbf{U}) < 1 + \varepsilon$$

holds with probability at least

$$1 - \exp(-cs + 3d) - \exp\left(-C^{-1}\varepsilon^2 K^{-2} \frac{n}{s} \right)$$

if (30) holds.

Set

$$s = \left\lceil \frac{3d + \sqrt{9d^2 + 4cC^{-1}\varepsilon^2 K^{-2} n}}{2c} \right\rceil$$

Then for $n > C'\varepsilon^{-2}K^2$ the probability above is at least

$$1 - 2\exp\left(\frac{3}{4}d - \frac{1}{4}\sqrt{9d^2 + C'^{-1}\varepsilon^2 K^{-2} n}\right),$$

as desired. It remains to check that (30) holds. Note that $s \geq C'^{-1}\varepsilon K^{-1}\sqrt{n}$. For $n \geq C'\log N$ we have $K^2\varepsilon^{-2}\log N \leq s^2$, thus $\log(K^2\varepsilon^{-2}s\log N) \leq 3\log s$. It then suffices to show

$$n \geq CK^2\varepsilon^{-2}s\log^3 s\log N. \tag{31}$$

But $s \leq C'\max\{d, \varepsilon K^{-1}\sqrt{n}\}$, which implies (31) when $n \geq C'K^3\varepsilon^{-3}\max\{d\log^3 d, \log^3 N\}$. $\qquad\square$

11.2 Partial Circulant/Toeplitz Matrices

Proof of Lemma 4. The equations (3.3)-(3.7) in [25], together with (3.8) there replaced by Hanson-Wright inequality to control the norm of $\|\mathbf{\Sigma V^* a}\|^2$, imply that for any $\mathbf{x} \in \mathbb{S}^{N-1}$, we have $\|\|\mathbf{\Phi x}\|^2 - 1\| < \varepsilon$ with probability at least

$$1 - 4N\mathrm{e}^{-\frac{t}{4}} - 2\mathrm{e}^{-\frac{cn\varepsilon^2}{t}}$$

for any $t > 0$. Take $t = \varepsilon\sqrt{n}$, the above probability is at least $1 - 4N\mathrm{e}^{-c\varepsilon\sqrt{n}}$. By a standard covering argument (c.f. Proof of Corollary 1), $1 - \varepsilon < s_{\min}(\mathbf{\Phi U}) \leq s_{\max}(\mathbf{\Phi U}) < 1 + \varepsilon$ holds with probability at least $1 - 4N \cdot 5^d \cdot \mathrm{e}^{-c\varepsilon\sqrt{n}}$, which is greater than $1 - \mathrm{e}^{-c\varepsilon\sqrt{n}/2}$ when $n > C\varepsilon^{-2}(d + \log N)^2$. $\qquad\square$

12 Appendix: Random Matrices With Heavy-Tailed Distributions

The purpose of this appendix is to provide some material on heavy-tailed distributions, in particular, distribution with finite moments characterized by strong regularity condition and log-concave ensembles, and to provide a proof of Lemma 5 and Lemma 6.

12.1 Finite Moments

Our proof depends on a theorem from [18], which asserts that

Theorem 8 ([18]). Consider independent isotropic random vectors $\mathbf{x}_i$ valued in $\mathbb{R}^d$. Assume that $\mathbf{x}_i$ satisfies the strong regularity assumption: for some $C', \eta > 1$, one has

$$\mathbb{P}(\|\mathcal{P}\mathbf{x}_i\|^2 > t) \leq C' t^{-\eta}, \quad \text{for } t > C' \operatorname{rank} \mathcal{P}$$

for every orthogonal projection $\mathcal{P}$ in $\mathbb{R}^d$. Then there exists a polynomial function $\operatorname{poly}(\cdot)$ whose coefficients depend only on C' and η, such that for any $\varepsilon \in (0,1)$ and for $n > \operatorname{poly}(\varepsilon^{-1})d$, we have

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} - \mathbf{I} \right\| \leq \varepsilon.$$

Proof of Lemma 5. Let $d = \lceil \alpha n \rceil$ where $\alpha \in (0,1)$ is to be determined later (we shall choose some α that does not depend on n). Fix a d-dimensional subspace of $\mathbb{R}^N$ and denote by $\mathbf{U}$ any of its orthonormal basis; it follows that $\mathbf{U}$ is an $N \times d$ matrix. We shall prove (4) for some δ and ε using Theorem 8 and partial strong regularity condition (7). Denote the rows of $\mathbf{\Phi}$ by $\frac{1}{\sqrt{n}}\mathbf{x}_1^{\mathrm{T}}, \ldots, \frac{1}{\sqrt{n}}\mathbf{x}_n^{\mathrm{T}}$. Let $\mathbf{y}_i = \mathbf{U}^{\mathrm{T}}\mathbf{x}_i$. Then $\mathbf{y}_i$'s are independent, centered and isotropic. Furthermore, we have

$$\|\mathbf{U}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{U} - \mathbf{I}\| = \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}} - \mathbf{I} \right\|$$

Before applying Theorem 8, we need to show that $\mathbf{y}_i$ fulfills strong regularity condition. For any orthogonal projection $\mathcal{P}$ of rank k in $\mathbb{R}^d$, we consider the tail of $\|\mathcal{P}\mathbf{y}_i\| = \|\mathcal{P}\mathbf{U}^{\mathrm{T}}\mathbf{x}_i\|$. First we note that there exists some $d \times k$ matrix $\mathbf{V}$ with orthonormal columns such that $\mathcal{P} = \mathbf{V}\mathbf{V}^{\mathrm{T}}$. Thus $\|\mathcal{P}\mathbf{U}^{\mathrm{T}}\mathbf{x}_i\| = \|\mathbf{U}\mathcal{P}\mathbf{U}^{\mathrm{T}}\mathbf{x}_i\| = \|(\mathbf{U}\mathbf{V})(\mathbf{U}\mathbf{V})^{\mathrm{T}}\mathbf{x}_i\|$. But $\mathbf{U}\mathbf{V}$ is a matrix of rank$\leq k$ with orthonormal columns, since $\operatorname{rank}(\mathbf{U}\mathbf{V}) \leq \operatorname{rank}\mathbf{V}$ and $(\mathbf{U}\mathbf{V})^{\mathrm{T}}(\mathbf{U}\mathbf{V}) = \mathbf{V}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{V} = \mathbf{I}$. By (7) we have

$$\mathbb{P}(\|\mathcal{P}\mathbf{y}_i\|^2 > t) \leq C' t^{-\eta}, \quad \text{for } t > C' \operatorname{rank}(\mathbf{U}\mathbf{V}),$$

hence for $t > C'k$. This shows that $\mathbf{y}_i$ satisfies strong regularity condition. The corollary follows from Theorem 8 and Chebyshev inequality. $\qquad \square$

12.2 Log-Concave Ensembles

We will need the following well-known results[6] on covariance estimation with log-concave ensembles.

Theorem 9 ([19, 20]). Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent centered isotropic random vectors in $\mathbb{R}^d$ with log-concave distributions. Then there exists some universal constants $c \in (0, 1)$, $C > 0$ such that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} - \mathbf{I} \right\| \leq C \sqrt{\frac{d}{n}}$$

with probability at least $1 - 2 \exp(-c\sqrt{d})$.

By definition, it is easy to check that the low-dimensional marginal of a log-concave distribution is log-concave.

Proof of Lemma 6. Let $\mathbf{U}$ be an orthonormal basis for a d-dimensional subspace of $\mathbb{R}^N$, where $d = \lfloor C^{-2} \varepsilon^2 n \rfloor$. For $n > 2C^2 \varepsilon^{-2} d_2$ we have $d \geq 2d_2$, which suffices for our purpose. Set $\mathbf{y}_i = \mathbf{U}^{\mathrm{T}} \mathbf{x}_i$. Argue as in the proof of Lemma 5, we obtain that $\mathbf{y}_i$ is independent, centered, and isotropic, and that

$$\| \mathbf{U}^{\mathrm{T}} \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi} \mathbf{U} - \mathbf{I} \| = \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}} - \mathbf{I} \right\|. \tag{32}$$

The distribution of $\mathbf{y}_i$ is a d-dimensional marginal of $\mathbf{x}_i$, hence is log-concave. It follows from Theorem 9 and (32) that

$$1 - \varepsilon < s_{\min}(\mathbf{\Phi} \mathbf{U}) \leq s_{\max}(\mathbf{\Phi} \mathbf{U}) < 1 + \varepsilon$$

with probability at least $1 - 2 \exp(-c\sqrt{d})$. When $n > 10 c^{-2} C^2 \varepsilon^{-2} d_2$, we have $d > \frac{1}{2} C^{-2} \varepsilon^2 n + 3 c^{-2}$, thus the probability above is at least $1 - \exp(-c' \varepsilon \sqrt{n})$, as desired. $\qquad \square$

References

[1] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

---

[6]Sharper results are known in literature, e.g. [39], but this does not yield significant improvement in our case.

[2] R. G. Baraniuk and M. B. Wakin, "Random projections of smooth manifolds," *Foundations of computational mathematics*, vol. 9, no. 1, pp. 51–77, 2009.

[3] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Science & Business Media, 2010.

[4] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[5] M. Soltanolkotabi and E. J. Candès, "A geometric analysis of subspace clustering with outliers," *The Annals of Statistics*, vol. 40, no. 4, pp. 2195–2238, 2012.

[6] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[7] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès, "Robust subspace clustering," *The Annals of Statistics*, vol. 42, no. 2, pp. 669–699, 2014.

[8] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6320–6342, 2015.

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

[10] L. Meng, G. Li, J. Yan, and Y. Gu, "A general framework for understanding compressed subspace clustering algorithms," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1504 – 1519, 2018.

[11] R. Heckel, M. Tschannen, and H. Bölcskei, "Subspace clustering of dimensionality-reduced data," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 2997–3001.

[12] ——, "Dimensionality-reduced subspace clustering," *Information and Inference: A Journal of the IMA*, vol. 6, no. 3, pp. 246–283, 2017.

[13] Y. Wang, Y.-X. Wang, and A. Singh, "A theoretical analysis of noisy sparse subspace clustering on dimensionality-reduced data," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 685–706, 2019.

[14] G. Li and Y. Gu, "Restricted isometry property of gaussian random projection for finite set of subspaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1705–1720, 2018.

[15] G. Li, Q. Liu, and Y. Gu, "Rigorous restricted isometry property of low-dimensional subspaces," to appear on *Applied and Computational Harmonic Analysis, available at arXiv:1801.10058*, 2019.

[16] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing.* Springer Science & Business Media, 2013.

[17] H. Rauhut, J. Romberg, and J. A. Tropp, "Restricted isometries for partial random circulant matrices," *Applied and Computational Harmonic Analysis*, vol. 32, no. 2, pp. 242–254, 2012.

[18] N. Srivastava and R. Vershynin, "Covariance estimation for distributions with $2 + \varepsilon$ moments," *Ann. Probab.*, vol. 41, no. 5, pp. 3081–3111, 09 2013.

[19] R. Adamczak, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann, "Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles," *Journal of the American Mathematical Society*, vol. 23, no. 2, pp. 535–561, 2010.

[20] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann, "Sharp bounds on the rate of convergence of the empirical covariance matrix," *Comptes Rendus Mathematique*, vol. 349, no. 3, pp. 195 – 200, 2011.

[21] C. Hinojosa, J. Bacca, and H. Arguello, "Coded aperture design for compressive spectral subspace clustering," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1589–1600, 2018.

[22] G. Martín and J. M. Bioucas-Dias, "Hyperspectral blind reconstruction from random spectral projections," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 6, pp. 2390–2399, 2016.

[23] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[24] K. L. Clarkson and D. P. Woodruff, "Low-rank approximation and regression in input sparsity time," *Journal of the ACM (JACM)*, vol. 63, no. 6, p. 54, 2017.

[25] J. Vybíral, "A variant of the johnson–lindenstrauss lemma for circulant matrices," *Journal of Functional Analysis*, vol. 260, no. 4, pp. 1096–1105, 2011.

[26] S. Dirksen, "Tail bounds via generic chaining," *Electronic Journal of Probability*, vol. 20, 2015.

[27] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[28] X. Xu and Y. Gu, "Convergence of empirical covariance matrix without strong regularity assumption," *in preparation*.

[29] Z. D. Bai, J. W. Silverstein, and Y. Q. Yin, "A note on the largest eigenvalue of a large dimensional sample covariance matrix," *Journal of Multivariate Analysis*, vol. 26, no. 2, pp. 166–168, 1988.

[30] K. Tikhomirov, "Sample Covariance Matrices of Heavy-Tailed Distributions," *International Mathematics Research Notices*, vol. 2018, no. 20, pp. 6254–6289, 04 2017.

[31] O. Guédon, "Concentration phenomena in high dimensional geometry," in *ESAIM: Proceedings*, vol. 44. EDP Sciences, 2014, pp. 47–60.

[32] M. A. Lodhi and W. U. Bajwa, "Detection theory for union of subspaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6347–6362, 2018.

[33] Y. Jiao, G. Li, and Y. Gu, "Compressed subspace learning based on canonical angle preserving property," *arXiv preprint arXiv:1907.06166*, 2019.

[34] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 643–660, 2001.

[35] X. Chen and Y. Yang, "Hanson-wright inequality in hilbert spaces with application to k-means clustering for non-euclidean data," *arXiv preprint arXiv:1810.11180*, 2018.

[36] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[37] H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical foundations and numerical methods for sparse recovery*, vol. 9, pp. 1–92, 2010.

[38] F. Krahmer and R. Ward, "New and improved johnson–lindenstrauss embeddings via the restricted isometry property," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 3, pp. 1269–1281, 2011.

[39] S. Mendelson and G. Paouris, "On the singular values of random matrices," *Journal of the European Mathematical Society*, vol. 16, no. 4, pp. 823–834, 2014.