Exact and Stable Covariance Estimation From Quadratic Sampling via Convex Programming

Yuxin Chen, *Student Member, IEEE*, Yuejie Chi, *Member, IEEE*, and Andrea J. Goldsmith, *Fellow, IEEE*

***Abstract*— Statistical inference and information processing of high-dimensional data often require an efficient and accurate estimation of their second-order statistics. With rapidly changing data, limited processing power and storage at the acquisition devices, it is desirable to extract the covariance structure from a single pass over the data and a small number of stored measurements. In this paper, we explore a quadratic (or rank-one) measurement model which imposes minimal memory requirements and low computational complexity during the sampling process, and is shown to be optimal in preserving various low-dimensional covariance structures. Specifically, four popular structural assumptions of covariance matrices, namely, low rank, Toeplitz low rank, sparsity, jointly rank-one and sparse structure, are investigated, while recovery is achieved via convex relaxation paradigms for the respective structure. The proposed quadratic sampling framework has a variety of potential applications, including streaming data processing, high-frequency wireless communication, phase space tomography and phase retrieval in optics, and noncoherent subspace detection. Our method admits universally accurate covariance estimation in the absence of noise, as soon as the number of measurements exceeds the information theoretic limits. We also demonstrate the robustness of this approach against noise and imperfect structural assumptions. Our analysis is established upon a novel notion called the mixed-norm restricted isometry property (RIP-ℓ_2/ℓ_1), as well as the conventional RIP-ℓ_2/ℓ_2 for near-isotropic and bounded measurements. In addition, our results improve upon the best-known phase retrieval (including both dense and sparse signals) guarantees using PhaseLift with a significantly simpler approach.**

***Index Terms*— Quadratic measurements, rank-one measurements, covariance sketching, energy measurements, phase retrieval, phase tomography, RIP-ℓ_2/ℓ_1, Toeplitz, low rank, sparsity.**

I. Introduction

ACCURATE estimation of second-order statistics of stochastic processes and data streams is of ever-growing importance to various applications that exhibit high dimensionality. Covariance estimation is the cornerstone of modern statistical analysis and information processing, as the covariance matrix constitutes the sufficient statistics to many signal processing tasks, and is particularly crucial for extracting reduced-dimension representation of the objects of interest. For signals and data streams of high dimensionality, there might be limited memory and computation power available at the data acquisition devices to process the rapidly changing input, which requires the covariance estimation task to be performed with a single pass over the data stream, minimal storage, and low computational complexity. This is not possible unless appropriate structural assumptions are incorporated into the high-dimensional problems. Fortunately, a broad class of high-dimensional signals indeed possesses low-dimensional structures, and the intrinsic dimension of the covariance matrix is often far smaller than the ambient dimension. For different types of data, the covariance matrix may exhibit different structures; four of the most widely considered structures are listed below.

- *Low Rank:* The covariance matrix is (approximately) low-rank, which occurs when a small number of components accounts for most of the variability in the data. Low-rank covariance matrices arise in applications including traffic data monitoring, array signal processing, collaborative filtering, and metric learning.
- *Stationarity and Low Rank:* The covariance matrix is simultaneously low-rank and Toeplitz, which arises when the random process is generated by a few spectral spikes. Recovery of the stationary covariance matrix, often equivalent to spectral estimation, is crucial in many tasks in wireless communications (e.g. detecting spectral holes in cognitive radio networks), and array signal processing (e.g. direction-of-arrival analysis [3]).
- *Sparsity:* The covariance matrix can be approximated in a sparse form [4]. This arises when a large number of variables have small pairwise correlation, or when several variables are mutually exclusive. Sparse covariance matrices arise in finance, biology and spectrum estimation.
- *Joint Sparsity and Rank-One:* The covariance matrix can be approximated by a jointly sparse and rank-one matrix. This has received much attention in recent development of sparse PCA, and is closely related to sparse signal recovery from magnitude measurements (called *sparse phase retrieval*).

In this paper, we wish to reconstruct an unknown covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ with the above structure from a small number of rank-one measurements. In particular, we explore

Manuscript received December 20, 2013; revised December 18, 2014; accepted March 18, 2015. Date of publication May 4, 2015; date of current version June 12, 2015. Y. Chen and A. J. Goldsmith were supported in part by the NSF Center for Science of Information and in part by AFOSR through MURI under Grant FA9550-12-1-0215. Y. Chi was supported in part by NSF CCF-1422966, in part by AFOSR through the Young Investigator Program under Grant FA9550-15-1-0205, and in part by the Google Faculty Research Award. This paper was presented at the 2014 International Conference on Acoustics, Speech, and Signal Processing [1] and the 2014 International Symposium on Information Theory [2].

Y. Chen is with the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: yxchen@stanford.edu).

Y. Chi is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: chi.97@osu.edu).

A. J. Goldsmith is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: andrea@ee.stanford.edu).

Communicated by Y. Ma, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2015.2429594

0018-9448 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

sampling methods of the form

$$y_i = \boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i + \eta_i, \quad i = 1, \ldots, m, \tag{1}$$

where $\boldsymbol{y} := \{y_i\}_{i=1}^m$ denotes the measurements, $\boldsymbol{a}_i \in \mathbb{R}^n$ represents the sensing vector, $\boldsymbol{\eta} := \{\eta_i\}_{i=1}^m$ stands for the noise term, and m is the number of measurements. The noise-free measurements $\boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i$'s are henceforth referred to as *quadratic measurements* (or *rank-one measurements*). In practice, the number of measurements one can obtain is constrained by the storage requirement in data acquisition, which could be much smaller than the ambient dimension of $\boldsymbol{\Sigma}$. This sampling scheme finds applications in a wide spectrum of practical scenarios, admits optimal covariance estimation with tractable algorithms, and brings in computational and storage advantages in comparison with other types of measurements, as detailed in the rest of the paper.

A. Motivation

The quadratic measurements in the form of (1) are motivated by several application scenarios listed below, which illustrate the practicability and benefits of the proposed quadratic measurement scheme.

1) Covariance Sketching for Data Streams: A high-dimensional data stream model represents *real-time* data that arrives sequentially at a high rate, where each data instance is itself high-dimensional. In many resource-constrained applications, the available memory and processing power at the data acquisition devices are severely limited compared with the volume and rate of the data [5]. Therefore it is desirable to extract the covariance matrix of the data instances from inputs on the fly without storing the whole stream. Interestingly, the quadratic measurement strategy can be leveraged as an effective data stream processing method to extract the covariance information from real-time data, with limited memory and low computational complexity.

Specifically, consider an input stream $\{\boldsymbol{x}_t\}_{t=1}^\infty$ that arrives sequentially, where each $\boldsymbol{x}_t \in \mathbb{R}^n$ is a high-dimensional data instance generated at time t. The goal is to estimate the covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_t^\top] \in \mathbb{R}^{n\times n}$. The prohibitively high rate at which data is generated forces covariance extraction to function with as small a memory as possible. The scenario we consider is quite general, and we only impose that the covariance of a random substream of the original data stream converges to the true covariance $\boldsymbol{\Sigma}$. No prior information on the correlation statistics across consecutive instances is assumed to be known *a priori* (e.g. they are not necessarily independently drawn), and hence it is not feasible to exploit these statistics to enable lower sample complexity.

We propose to pool the data stream $\{\boldsymbol{x}_t\}_{t=1}^\infty$ into a small set of measurements in an easy-to-adapt fashion with a collection of sketching vectors $\{\boldsymbol{a}_i\}_{i=1}^m$. Our covariance sketching method, termed *quadratic sketching*, is outlined as follows:

1) At each time t, we randomly choose a sketching vector indexed by $\ell_t \in \{1, \ldots, m\}$, and obtain a single nonnegative *quadratic sketch* $(\boldsymbol{a}_{\ell_t}^\top \boldsymbol{x}_t)^2$.
2) All sketches employing the same sketching vector $\boldsymbol{a}_i$ are aggregated and normalized, which converge *rapidly* to a *measurement*[1]

$$y_i = \mathbb{E}\left[(\boldsymbol{a}_i^\top \boldsymbol{x}_t)^2\right] + \boldsymbol{\eta}_i = \boldsymbol{a}_i^\top \mathbb{E}\left[\boldsymbol{x}_t \boldsymbol{x}_t^\top\right] \boldsymbol{a}_i + \boldsymbol{\eta}_i$$
$$= \boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i + \boldsymbol{\eta}_i, \quad i = 1, \ldots, m, \tag{2}$$

where $\boldsymbol{\eta} := \{\eta_i\}_{i=1}^m$ denotes the error term.

There are several benefits of this covariance sketching method. First, the storage complexity m, as will be shown, can be much smaller than the ambient dimension of $\boldsymbol{\Sigma}$. The computational cost for sketching each instance is linear with respect to the dimension of the instance in the data stream. Unlike the uncompressed sketching methods where each instance one measures usually affects many stored measurements, our scheme allows each aggregate quadratic sketch to be composed by completely different instances, which allows sketching to be performed in a distributed and asynchronous manner. This arises since each randomized sketch is a compressive snapshot of the second-order statistics, while each uncompressed measurement itself is unable to capture the correlation information. As we will demonstrate later, this sketching scheme allows optimal covariance estimation with information theoretically minimal memory complexity at the data acquisition stage. One motivating application for this covariance sketching method is covariance estimation of ultra-wideband random processes, as is further elaborated in Section I-A2.

2) Noncoherent Energy Measurements in Communications and Signal Processing: When communication takes place in the high-frequency regime, *empirical* energy measurements are often more accurate and cheaper to obtain than phase measurements. For instance, energy measurements will be more reliable when communication systems are operating with extremely high carrier frequencies (e.g. 60GHz communication systems [6]).

- *Spectrum Estimation of Stochastic Processes from Energy Measurements:* Many wireless communication systems operating in stochastic environments rely on reliable estimation of the spectral characteristics of random processes [7], such as recovering the power spectrum of the ultra-wideband random process characterizing the spectrum occupancy in cognitive radio [8], [9]. Moreover, optimal signal transmissions are often based on the Karhunen–Loeve decomposition of a random process, which requires accurate covariance information [10]. If one employs a sensing vector $\boldsymbol{a}_i$, which is implementable using random demodulators [11], and observes the average energy measurements over N instances $\{\boldsymbol{x}_t\}_{1\leq t\leq N}$, then the energy measurements read

$$y_i = \frac{1}{N}\sum_{t=1}^{N}\left|\boldsymbol{a}_i^\top \boldsymbol{x}_t\right|^2 = \boldsymbol{a}_i^\top \boldsymbol{\Sigma}_N \boldsymbol{a}_i, \quad i = 1, \ldots, m \tag{3}$$

where $\boldsymbol{\Sigma}_N := \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^\top$ denotes the sample covariance matrix, leading to the quadratic-form observations.

[1]Note that we might only be able to obtain measurements for empirical covariance matrices instead of $\boldsymbol{\Sigma}$, but this inaccuracy can be absorbed into the noise term $\boldsymbol{\eta}$. In fact, for stationary data streams, y_i converges rapidly to $\boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i$ with *a few* instances $\boldsymbol{x}_t$.

- *Noncoherent Subspace Detection from Energy Measurements:* Matched subspace detection [12] spans many applications in wireless communication, radar, and pattern recognition when the transmitted signal is encoded by the subspaces. The problem can also be cast as recovering the principal subspace of a dataset $\{\boldsymbol{x}_t\}_{t=1}^{N}$, with an energy detector obtaining m measurements in the form of (3). Thus, the noncoherent subspace detection is subsumed by the formulation (1).

3) Phaseless Measurements in Physics: Optical imaging devices are incapable of acquiring phase measurements due to ultra-high frequencies associated with light. In many applications, measurements taking the form of (1) arise naturally.

- *Compressive Phase Space Tomography:* Phase Space Tomography [13] is an appealing method to measure the correlation function of a wave field in physics. However, tomography becomes challenging when the dimensionality of the correlation matrix becomes large. Recently, it was proposed experimentally in [14] to recover an approximately low-rank correlation matrix, which often holds in physics, by only taking a small number of measurements in the form of (1).
- *Phase Retrieval:* Due to the physical constraints, one can only measure amplitudes of the Fourier coefficients of an optical object. This gives rise to the problem of recovering a signal $\boldsymbol{x} \in \mathbb{R}^n$ from magnitude measurements, which is often referred to as phase retrieval. Several convex (see [15]–[17]) and nonconvex algorithms (see [18]–[21]) have been proposed that enable exact phase retrieval (i.e. recovers $\boldsymbol{x} \cdot \boldsymbol{x}^\top$) from random magnitude measurements. If we set $\boldsymbol{\Sigma} := \boldsymbol{x}\boldsymbol{x}^\top$, then our problem formulation (1) subsumes phase retrieval as a special case in the low-rank setting.

Apart from the preceding applications, we are aware that this rank-one measurement model naturally arises in the mixture of linear regression problem [22]. All in all, all of these applications require structured matrix recovery from *a small number* of rank-one measurements (1). The aim of this paper to develop tractable recovery algorithms that enjoy near-optimal performance guarantees.

B. Contributions

Our main contributions are three fold. First, we have developed convex optimization algorithms for covariance estimation from a set of quadratic measurements as given in (1) for a variety of structural assumptions including low-rank, Toeplitz low-rank, sparse, and sparse rank-one covariance matrices. The proposed algorithms exploit the presumed low-dimensional structures using convex relaxation tailored for respective structures. For a large class of sub-Gaussian sensing vectors, we derive theoretical performance guarantees (Theorems 1 – 4) from the following aspects:

1) *Exact and Universal Recovery:* once the sensing vectors are selected, then with high probability, all covariance matrices satisfying the presumed structure can be recovered;
2) *Stable Recovery:* the proposed algorithms allow reconstruction of the true covariance matrix to within high accuracy even under imperfect structural assumptions; additionally, if the measurements are corrupted by noise, possibly adversarial, the estimate deviates from the true covariance matrix by at most a constant multiple of the noise level;
3) *Near-minimal Measurements:* the proposed algorithms succeed as soon as the number of measurements is slightly above the information theoretic limits for most of the respective structure. For the special case of (sparse) rank-one matrices, our result recovers and strengthens the best-known reconstruction guarantees of (sparse) phase retrieval using PhaseLift [15], [23], [24] with a much simpler proof technique.

Secondly, to obtain some of the above theoretical guarantees (Theorems 1, 3, and 4), we have introduced a novel mixed-norm restricted isometry property, denoted by RIP-ℓ_2/ℓ_1. An operator is said to satisfy the RIP-ℓ_2/ℓ_1 if the strength of the signal class of interest before and after measurements are preserved when measured in the ℓ_2 norm and in the ℓ_1 norm, respectively. While the conventional RIP-ℓ_2/ℓ_2 does not hold for the quadratic sensing model for general low-rank structures as pointed out by [15], we have established that the sensing mechanism does satisfy the RIP-ℓ_2/ℓ_1 after a "debiasing" modification, under general low-rank, sparse, and simultaneously sparse and rank-one structural assumptions. This seemingly subtle change enables a significantly simpler analytical approach without resorting to complicated dual construction as in [15], [23], and [24].

On the other hand, we demonstrate, via the entropy method [25], that linear combinations of the quadratic measurements satisfy RIP-ℓ_2/ℓ_2 when restricted to *Toeplitz* low-rank covariance matrices. This leads to near-optimal recovery guarantees for Toeplitz low-rank covariance matrices (Theorem 2). Along the way, we have also established a RIP-ℓ_2/ℓ_2 for bounded and near-isometric operators (Theorem 5), which strengthens previous work [26], [27] by offering universal and stable recovery guarantees for a broader class of operators including Fourier-type measurements.

Last but not least, our measurement schemes and algorithms may be of independent interest to high-dimensional data processing. The measurements in (1) are rank-one measurements with respect to the covariance matrix, which are much easier to implement and bear a smaller computational cost than full-rank measurement matrices with i.i.d. entries. Moreover, the performance guarantees of the measurement scheme (1) is universal, which does not require any additional incoherence conditions on the covariance matrix as required in the standard matrix completion framework [26], [28], [29].

C. Related Work

In most existing work, the covariance matrix is estimated from a collection of *full* data samples, and fundamental guarantees have been derived on how many samples are sufficient to approximate the ground truth [4], [30]. In contrast, this paper is motivated by the success of Compressed Sensing (CS) [31], [32], which asserts that compression can be achieved at the same time as sensing without losing information. Efficient algorithms have been developed to estimate a *deterministic signal* from a much smaller number of linear measurements that is proportional to the complexity of the parsimonious signal model. As we will show in this paper, covariance estimation from compressive measurements can be highly robust.

When the covariance matrix is assumed to be approximately sparse, recent work [8], [33] explored reconstruction of second-order statistics of a cyclostationary signal from random linear measurements, by ℓ_1-minimization without performance guarantees. Other spectral prior information has been considered as well in [34] for stationary processes. These problem setups are quite different from (1) in the current work. Another work by Dasarathy *et al.* [35] proposed estimating an approximately sparse covariance matrix from measurements of the form $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ denotes the sketching matrix constructed from expander graphs. Nevertheless, this scheme cannot accommodate low-rank covariance matrix estimation.

Our covariance estimation method is inspired by recent developments in phase retrieval [15], [17], [20], [23], [36], [37], which is tantamount to recovering rank-one covariance matrices from quadratic measurements. In particular, our recovery algorithm coincides with *PhaseLift* [15], [23] when applied to low-rank matrices. In [23], it is shown that PhaseLift succeeds at reconstructing a signal of dimensionality n from $\Theta(n)$ phaseless *Gaussian* measurements, and stable recovery has also been established in the presence of noise. When specializing our result to this case, we have shown that the same type of theoretical guarantee holds for a much larger class of *sub-Gaussian* measurements, with a different proof technique that yields a much simpler proof. Moreover, when the signal is further assumed to be k-sparse, the pioneering work [24] showed that $O(k^2 \log n)$ *Gaussian* measurements suffice; this result is extended to accommodate sub-Gaussian measurements and approximately sparse signals by our framework with a much simpler proof. More details can be found in Section II-D.

We also put the proposed covariance sketching scheme in Section I-A1 into perspective. In a streaming setting, online principal component analysis (PCA) has been an active area of research for decades [38] using full data samples, where non-asymptotic convergence guarantees have only been recently developed [39]. Inspired by CS, subspace tracking from partial observations of a data stream [40], [41], which can be regarded as a variant of incremental PCA [42] in the presence of missing values, is also closely related. However, existing subspace tracking algorithms mainly aim to recover the data stream, which is not necessary if one only cares to extract the second-order statistics.

Finally, after we posted our work on Arxiv, Cai and Zhang made available their manuscript [43], an independent work that studies low-rank matrix recovery under rank-one measurements via the notion of restricted uniform boundedness. In comparison, our results accommodate a larger class of covariance structures including Toeplitz low-rank, sparse, and jointly low-rank and sparse matrices.

D. Organization

The rest of this paper is organized as follows. We first present the convex optimization based algorithms in Section II, and establish their theoretical guarantees. The analysis framework is based upon a novel mixed-norm restricted isometry property as well as conventional RIP for near-isotropic and bounded measurements, as elaborated in Sections III and IV. The proof of main theorems is deferred to the appendices. Numerical examples are provided in Sections V. Finally, Section VI concludes the paper with a summary of our findings and a discussion of future directions.

E. Notations

Before proceeding, we provide a brief summary of useful notations that will be used throughout this paper. A variety of matrix norms will be discussed; in particular, we denote by $\|\boldsymbol{X}\|$, $\|\boldsymbol{X}\|_{\mathrm{F}}$, and $\|\boldsymbol{X}\|_*$ the spectral norm, the Frobenius norm, and the nuclear norm (i.e. sum of all singular values) of $\boldsymbol{X}$, respectively. When $\boldsymbol{X}$ is a positive semidefinite (PSD) matrix, the nuclear norm coincides with the trace $\|\boldsymbol{X}\|_* = \mathrm{Tr}(\boldsymbol{X})$. We use $\|\boldsymbol{X}\|_1$ and $\|\boldsymbol{X}\|_0$ to denote the ℓ_1 norm and support size of the vectorized $\boldsymbol{X}$, respectively. The Euclidean inner product between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{Y})$. We will abuse the notation and let $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_k$ stand for the best rank-r approximation and the best k-term approximation of $\boldsymbol{\Sigma}$ respectively, i.e.

$$\boldsymbol{\Sigma}_r = \mathrm{argmin}_{\boldsymbol{M}:\mathrm{rank}(\boldsymbol{M})=r} \|\boldsymbol{\Sigma} - \boldsymbol{M}\|_{\mathrm{F}},$$

and

$$\boldsymbol{\Sigma}_k = \mathrm{argmin}_{\boldsymbol{M}:\|\boldsymbol{M}\|_0=k} \|\boldsymbol{\Sigma} - \boldsymbol{M}\|_{\mathrm{F}},$$

whenever clear from context. Besides, we denote by $\mathcal{T}$ the orthogonal projection operator onto Toeplitz matrices, and $\mathcal{T}^\perp$ its orthogonal complement. Some useful notations are summarized in Table I.

II. Convex Relaxation and Its Performance Guarantees

In general, recovering the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ from $m < n(n+1)/2$ measurements is ill-posed, unless the sampling mechanism can effectively exploit the low-dimensional covariance structure. Random sampling often preserves the information structure from minimal observations, and allows robust recovery from noisy measurements.

In this paper, we restrict our attention to the following random sampling model. We assume that the sensing vectors are composed of i.i.d. *sub-Gaussian* entries. In particular, we assume $\boldsymbol{a}_i$'s ($1 \le i \le m$) are i.i.d. copies of $\boldsymbol{z} = [z_1, \cdots, z_n]^\top$, where each z_i is i.i.d. drawn from a distribution with the following properties

$$\mathbb{E}[z_i] = 0, \quad \mathbb{E}[z_i^2] = 1, \quad \text{and} \quad \mu_4 := \mathbb{E}\left[z_i^4\right] > 1. \tag{4}$$

TABLE I
SUMMARY OF NOTATION AND PARAMETERS

$\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_r$, $\boldsymbol{\Sigma}_\mathrm{c}$	true covariance matrix, best rank-r approximation of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_\mathrm{c} := \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_r$
$\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{\Omega_0}$, $\boldsymbol{\Sigma}_{\Omega_0^\mathrm{c}}$	true covariance matrix, best k-sparse approximation of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_{\Omega_0^\mathrm{c}} := \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\Omega_0}$
$\mathcal{T}$, $\mathcal{T}^\perp$	orthogonal projection operator onto Toeplitz matrices, and its orthogonal complement.
$\boldsymbol{\eta}, \boldsymbol{y} \in \mathbb{R}^m$,	noise, quadratic measurements $\left\{\boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i + \eta_i\right\}_{1\leq i\leq m}$
$\boldsymbol{a}_i \in \mathbb{R}^n$, $\boldsymbol{A}_i \in \mathbb{R}^{n\times n}$	ith sensing vector, ith sensing matrix $\boldsymbol{A}_i := \boldsymbol{a}_i \cdot \boldsymbol{a}_i^\top$
$\boldsymbol{B}_i \in \mathbb{R}^{n\times n}$	auxiliary sensing matrix
$\mathcal{A}_i$,$\mathcal{A}$	linear transformation $\boldsymbol{X} \mapsto \boldsymbol{a}_i^\top \boldsymbol{X} \boldsymbol{a}_i$, linear mapping $\boldsymbol{X} \mapsto \left\{\boldsymbol{a}_i^\top \boldsymbol{X} \boldsymbol{a}_i\right\}_{1\leq i\leq m}$
$\mathcal{B}_i$,$\mathcal{B}$	linear transformation $\boldsymbol{X} \mapsto \langle \boldsymbol{B}_i, \boldsymbol{X}\rangle$, linear mapping $\boldsymbol{X} \mapsto \left\{\mathcal{B}_i(\boldsymbol{X})\right\}_{1\leq i\leq m}$

We assume that the error term $\boldsymbol{\eta} := [\eta_1, \cdots, \eta_m]^\top$ is bounded in either ℓ_1 norm or ℓ_2 norm as specified later in the theoretical guarantees. For notational simplicity, let $\boldsymbol{A}_i := \boldsymbol{a}_i\boldsymbol{a}_i^\top$ represent the equivalent sensing matrix, and hence the measurements $\boldsymbol{y} := [y_1, \cdots, y_m]^\top$ obeys $y_i := \langle \boldsymbol{A}_i, \boldsymbol{\Sigma}\rangle + \eta_i$. We also define the linear operator $\mathcal{A}(\boldsymbol{M}) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ that maps a matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ to $\{\langle \boldsymbol{M}, \boldsymbol{A}_i\rangle\}_{i=1}^m$. These notations allow us to express the measurements as

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{\Sigma}) + \boldsymbol{\eta}. \tag{5}$$

A. Recovery of Low-Rank Covariance Matrices

Suppose that $\boldsymbol{\Sigma}$ is approximately low-rank, a natural heuristic is to perform rank minimization to encourage the low-rank structure

$$\hat{\boldsymbol{\Sigma}} = \mathrm{argmin}_{\boldsymbol{M}}\, \mathrm{rank}(\boldsymbol{M}) \quad \text{subject to} \quad \boldsymbol{M} \succeq 0, \quad \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_1 \leq \epsilon_1, \tag{6}$$

where ϵ_1 is an upper bound on $\|\boldsymbol{\eta}\|_1$ and assumed known *a priori*. However, the rank minimization problem is in general NP-hard. Therefore, we replace it with trace minimization over all matrices compatible with the measurements

$$\hat{\boldsymbol{\Sigma}} = \mathrm{argmin}_{\boldsymbol{M}}\, \mathrm{Tr}(\boldsymbol{M}) \quad \text{subject to} \quad \boldsymbol{M} \succeq 0, \quad \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_1 \leq \epsilon_1. \tag{7}$$

Since $\boldsymbol{\Sigma}$ is PSD, the trace norm forms a convex surrogate for the rank function, which has proved successful in matrix completion and phase retrieval problems [15], [28], [44]. It turns out that this convex relaxation approach (7) admits stable and faithful estimates even when $\boldsymbol{\Sigma}$ is approximately low rank and/or when the measurements are corrupted by bounded noise. This is formally stated in the following theorem.

Theorem 1: Consider the sub-Gaussian sampling model in (4) *and assume that* $\|\boldsymbol{\eta}\|_1 \leq \epsilon_1$. *Then with probability exceeding* $1 - C_0 \exp(-c_0 m)$, *the solution* $\hat{\boldsymbol{\Sigma}}$ *to* (7) *satisfies*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\mathrm{F} \leq C_1 \frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_r\|_*}{\sqrt{r}} + C_2 \frac{\epsilon_1}{m} \tag{8}$$

simultaneously for all $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$, *provided that* $m > c_1 nr$. *Here,* $\boldsymbol{\Sigma}_r$ *represents the best rank-r approximation of* $\boldsymbol{\Sigma}$, *and* c_0, c_1, C_0, C_1 *and* C_2 *are some positive numerical constants.*

The main implications of Theorem 1 and its associated performance bound (8) are listed as follows.

1) *Exact Recovery From Noiseless Measurements:* Consider the case where $\mathrm{rank}\,(\boldsymbol{\Sigma}) = r$. In the absence of noise, one can see from (8) that the trace minimization program (7) (with $\epsilon_1 = 0$) allows perfect covariance recovery with exponentially high probability, provided that the number m of measurements exceeds the order of nr. Notice that each PSD matrix can be uniquely decomposed as $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ has orthogonal columns. That said, the the intrinsic degrees of freedom carried by PSD matrices is $\Theta(nr)$, indicating that our algorithm achieves order-wise optimal recovery.
2) *Near-Optimal Universal Recovery:* The trace minimization program (7) allows universal recovery, in the sense that once the sensing vectors are chosen, *all* low-rank covariance matrices can be perfectly recovered in the absence of noise. This highlights the power of convex programming, which allows universally accurate estimates as soon as the number of measurements exceeds the order of the information theoretic limit. In addition, the universality and optimality results hold for a large class of sub-Gaussian measurements beyond the Gaussian sampling model.
3) *Robust Recovery for Approximately Low-Rank Matrices:* In the absence of noise ($\epsilon_1 = 0$), if $\boldsymbol{\Sigma}$ is approximately low-rank, then by (8) the reconstruction inaccuracy is at most
$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\mathrm{F} \leq O\left(\frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_r\|_*}{\sqrt{r}}\right)$$
with probability at least $1 - \exp(-c_1 m)$, as soon as m is about the same order of nr. One can obtain a more intuitive understanding through the following *power-law* covariance model. Let λ_ℓ represent the ℓth largest singular value of $\boldsymbol{\Sigma}$, and suppose the decay of λ_ℓ obeys a power law, i.e. $\lambda_\ell \leq \frac{\alpha}{\ell^\beta}$ for some constant $\alpha > 0$ and decay rate exponent $\beta > 1$. Then simple computation reveals that
$$\frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_r\|_*}{\sqrt{r}} \leq \frac{1}{\sqrt{r}} \sum_{\ell=r+1}^{n} \frac{\alpha}{\ell^\beta} \leq \frac{\alpha}{(\beta-1)r^{\beta-\frac{1}{2}}},$$
which in turn implies
$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\mathrm{F} = O\left(\frac{1}{r^{\beta-\frac{1}{2}}}\right). \tag{9}$$
This asserts that (7) returns an almost accurate estimate of $\boldsymbol{\Sigma}$ in a manner which requires no prior knowledge on the signal (other than the power law decay that is natural for a broad class of data).
4) *Stable Recovery From Noisy Measurements:* When $\boldsymbol{\Sigma}$ is exactly of rank r and the noise is bounded $\|\boldsymbol{\eta}\|_1 \leq \epsilon_1$,

the reconstruction inaccuracy of (7) is bounded above by

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} = O\left(\frac{\epsilon_1}{m}\right) \tag{10}$$

with exponentially high probability, provided that m exceeds $\Theta(nr)$. This reveals that the algorithm (7) recovers an unknown object with an error at most proportional to the average *per-entry* noise level, which makes it practically appealing.

5) *Phase Retrieval With Sub-Gaussian Measurements:* The proposed algorithm (7) appears in the same form as the convex algorithm called *PhaseLift*, which was proposed in [15] for phase retrieval. It is equivalent to treating $\boldsymbol{\Sigma}$ as the rank-one lifted matrix $\boldsymbol{x}\boldsymbol{x}^{\top}$ from an unknown signal $\boldsymbol{x}$. It has been established in [23] that with high probability, it is feasible to recover $\boldsymbol{x}$ exactly from $\Theta(n)$ quadratic measurements, assuming that the sensing vectors are i.i.d. Gaussian. Our result immediately recovers all results of [15] and [23] including exact and stable recovery. In fact, our analysis framework yields a much simpler and shorter proof of all these results, and immediately extends to a broader class of sub-Gaussian sampling mechanisms. We will further discuss our improvement of sparse recovery from magnitude measurements [24], [45] in Section II-D.

Remark 1: A lower bound on the minimax risk has recently been established by Cai and Zhang [43, Th. 2.4]. Specifically, if the noise $\boldsymbol{\eta} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$ with $\sigma = \Theta\left(\frac{\epsilon_1}{m}\right)$, then for any estimator $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{y})$,

$$\inf_{\tilde{\boldsymbol{\Sigma}}(\cdot)} \sup_{\boldsymbol{\Sigma}:\, \mathrm{rank}(\boldsymbol{\Sigma})=r} \sqrt{\mathbb{E}_{\boldsymbol{\eta}}\left[\left\|\tilde{\boldsymbol{\Sigma}}(\boldsymbol{y}) - \boldsymbol{\Sigma}\right\|_{\mathrm{F}}^2\right]} \gtrsim \sigma = \Theta\left(\frac{\epsilon_1}{m}\right),$$

provided that $m = \Theta(nr)$. While our results are established for bounded (possibly adversarial) noise, it is straightforward to see that the above argument reveals the orderwise minimaxity of our stability bound.

B. Recovery of Low-Rank Covariance Matrices for Stationary Instances

Suppose that $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ is simultaneously low-rank and Toeplitz, which can represent the covariance matrix of a wide-sense stationary random process. Similar to recovery in the general low-rank model, we propose to seek a nuclear norm minimizer over all matrices compatible with the measurements as well as the Toeplitz constraint, which results in the following estimator:

$$\hat{\boldsymbol{\Sigma}} = \mathrm{argmin}_{\boldsymbol{M}}\ \mathrm{Tr}(\boldsymbol{M}) \quad \text{subject to} \quad \boldsymbol{M} \succeq 0, \qquad \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_2 \le \epsilon_2, \quad \boldsymbol{M} \text{ is Toeplitz}, \tag{11}$$

where ϵ_2 is an upper bound of $\|\boldsymbol{\eta}\|_2$.

Encouragingly, the PSD Toeplitz cone can be very pointy around many low-rank feasible points, as illustrated in Fig. 1. Therefore, the intersection between the PSD Toeplitz cone and a random hyperplane passing through $\boldsymbol{\Sigma}$ often contains only a single point. As a result, the semidefinite relaxation (11) is exact with high probability under noise-free measurements, as stated in the following theorem.

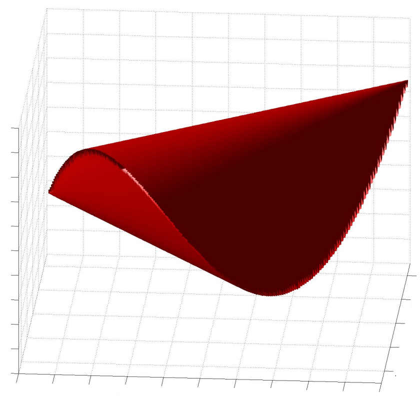


Fig. 1. Representation of the unit PSD Toeplitz ball consisting of all (x, y, z) such that $\boldsymbol{T} = \begin{bmatrix} 1/4 & x & y & z \\ x & 1/4 & x & y \\ y & x & 1/4 & x \\ z & y & x & 1/4 \end{bmatrix} \succeq \mathbf{0}$ and $\mathrm{Tr}(\boldsymbol{T}) = 1$.

Theorem 2: Consider the sub-Gaussian sampling model in (4)*, and assume that $\mu_4 \le 3$ and $\|\boldsymbol{\eta}\|_2 \le \epsilon_2$. Then with probability exceeding $1 - 1/n^2$,*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} \le C_2 \frac{\epsilon_2}{\sqrt{m}} \tag{12}$$

holds simultaneously for all Toeplitz covariance matrices $\boldsymbol{\Sigma}$ of rank at most r, provided that $m > c_0 r \log^{10} n$. Here, c_0 and C_2 are some universal constants.

We highlight some implications of Theorem 2 as follows.

1) *Exact Recovery Without Noise:* As any rank-r PSD Toeplitz matrix admits a unique rank-r Vandemonde decomposition that can be specified by $2r$ parameters, by Theorem 2, exact recovery of Toeplitz low-rank covariance matrices occurs as soon as m is slightly larger than the information theoretic limit $\boldsymbol{\Omega}(r)$ (modulo some poly-logarithmic factor). Note that this sampling requirement is much smaller than that for general low-rank matrices, and also much smaller than the degrees of freedom for general Toeplitz matrices (which is n).
2) *Stable and Universal Recovery From Noisy Measurements:* The proposed convex relaxation (11) returns faithful estimates in the presence of noise, as revealed by Theorem 2. This feature is universal: if $\mathcal{A}$ is randomly sampled and then fixed thereafter, then, with high probability, the error bounds (12) hold simultaneously for all Toeplitz low-rank matrices. Note that the error bound (12) is stated in terms of the ℓ_2 norm of $\boldsymbol{\eta}$. This is out of mathematical convenience for this special setup, which will be discussed later.

Remark 2: Two aspects of Theorem 2 are worth noting. First, Theorem 2 does not guarantee recovery with exponentially high probability as ensured in Theorem 1. This arises from our use of stochastic RIP, as will be seen in the analysis. Secondly, we are only able to provide theoretical guarantees when $\mu_4 \le 3$; roughly speaking, the tails of these distributions are typically not heavier than those of the Gaussian measure (e.g. $\mu_4 = 3$ for Gaussian distribution and $\mu_4 = 1$ for Bernoulli distribution). We conjecture that these two aspects can be improved via other proof techniques.

C. Recovery of Sparse Covariance Matrices

Assume that $\boldsymbol{\Sigma}$ is approximately sparse, we propose to seek a matrix with minimal support size that is compatible with observations:

$$\hat{\boldsymbol{\Sigma}} = \operatorname{argmin}_{\boldsymbol{M}} \|\boldsymbol{M}\|_0 \quad \text{subject to} \quad \boldsymbol{M} \succeq 0, \quad \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_1 \leq \epsilon_1, \tag{13}$$

where ϵ_1 is an upper bound on $\|\boldsymbol{\eta}\|_1$. However, the ℓ_0 minimization problem in (13) is also intractable, and one can instead solve a tractable convex relaxation of (13), given as

$$\hat{\boldsymbol{\Sigma}} = \operatorname{argmin}_{\boldsymbol{M}} \|\boldsymbol{M}\|_1 \quad \text{subject to} \quad \boldsymbol{M} \succeq 0, \quad \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_1 \leq \epsilon_1. \tag{14}$$

Here, the ℓ_1 norm is the convex relaxation of the support size, which has proved successful in many compressed sensing algorithms [32], [46]. It turns out that the convex relaxation (14) allows stable and reliable estimates even when $\boldsymbol{\Sigma}$ is only approximately sparse and the measurements are contaminated by noise, as stated in the following theorem.

Theorem 3: Consider the sub-Gaussian sampling model in (4) *and assume that* $\|\boldsymbol{\eta}\|_1 \leq \epsilon_1$. *Then with probability exceeding* $1 - C_0 \exp(-c_0 m)$, *the solution* $\hat{\boldsymbol{\Sigma}}$ *to* (14) *satisfies*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} \leq C_1 \frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_\Omega\|_1}{\sqrt{k}} + C_2 \frac{\epsilon_1}{m}, \tag{15}$$

simultaneously for all $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, *provided that* $m > c_1 k \log(n^2/k)$. *Here,* $\boldsymbol{\Sigma}_\Omega$ *denotes the best* k*-sparse approximation of* $\boldsymbol{\Sigma}$, *and* c_0, c_1, C_0, C_1 *and* C_2 *are positive universal constants.*

Theorem 3 leads to similar implications as those listed in Section II-A, which we briefly summarize as follows.

1) *Exact Recovery Without Noise:* When $\boldsymbol{\Sigma}$ is exactly k-sparse and no noise is present, by setting $\epsilon_1 = 0$, the solution to (14) is exactly equal to the ground truth with exponentially high probability, as soon as the number m of measurements is about the order of $k \log(n^2/k)$. Therefore our performance guarantee in (15) is *optimal* within a constant factor.
2) *Universal Recovery:* Our performance guarantee in (15) is universal in the sense that the same sensing mechanism simultaneously works for all sparse covariance matrices.
3) *Imperfect Structural Models:* The estimate (15) allows robust recovery for approximately sparse matrices (which appears in a similar form as that for CS [46]), indicating that quadratic measurements are order-wise at least as good as linear measurements.

D. Recovery of Jointly Sparse and Rank-One Matrices

If we set the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{x}\boldsymbol{x}^\top$ to be a rank-one matrix, then covariance estimation from quadratic measurements is equivalent to phase retrieval as studied in [15]. In addition to the general rank-one model, our approach allows simple analysis for recovering jointly sparse and rank-one covariance matrices or, equivalently, sparse signal recovery from magnitude measurements. Specifically, suppose that $\boldsymbol{x}$ is (approximately) sparse, and we collect a small number of phaseless measurements as

$$\boldsymbol{y} := \left\{ |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 + \eta_i \right\}_{1 \leq i \leq m}.$$

When $\boldsymbol{x}$ is sparse, the lifting matrix $\boldsymbol{x}\boldsymbol{x}^\top$ is *simultaneously* low rank and sparse, which motivates us to adapt the convex program proposed in [24] to accommodate bounded noise as follows

$$\begin{aligned} \hat{\boldsymbol{X}} = \operatorname{argmin}_{\boldsymbol{M}} \quad & \operatorname{Tr}(\boldsymbol{M}) + \lambda \|\boldsymbol{M}\|_1 \\ \text{subject to} \quad & \boldsymbol{M} \succeq 0, \\ & \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_1 \leq \epsilon_1. \end{aligned} \tag{16}$$

Here, λ is a regularization parameter that balances the two convex surrogates (i.e. trace norm and ℓ_1 norm) associated with the low-rank and sparse structural assumptions, respectively, and ϵ_1 is an upper bound of $\|\boldsymbol{\eta}\|_1$. Our analysis framework ensures stable recovery of an approximately sparse signal, as stated in the following theorem.

Theorem 4: Set $\lambda \in \left[\frac{1}{n}, \frac{1}{\sqrt{k}}\rho\right]$ *for some quantity* ρ. *Consider the sub-Gaussian sampling model in* (4) *and assume that* $\|\boldsymbol{\eta}\|_1 \leq \epsilon_1$. *Then with probability at least* $1 - C_0 \exp(-c_0 m)$, *the solution* $\hat{\boldsymbol{X}}$ *to (16) satisfies*

$$\left\|\hat{\boldsymbol{X}} - \boldsymbol{x}\boldsymbol{x}^\top\right\|_{\mathrm{F}} \leq C_1 \left\{ \left\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}_\Omega\boldsymbol{x}_\Omega^\top\right\|_* + \lambda \left\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}_\Omega\boldsymbol{x}_\Omega^\top\right\|_1 + \frac{\epsilon_1}{m} \right\} \tag{17}$$

simultaneously for all signals $\boldsymbol{x} \in \mathbb{R}^n$ *satisfying* $\frac{\|\boldsymbol{x}_\Omega\|_2}{\|\boldsymbol{x}_\Omega\|_1} \geq \rho$, *provided that* $m > \frac{C_2 \log n}{\lambda^2}$. *Here,* $\boldsymbol{x}_\Omega$ *denotes the best* k*-sparse approximation of* $\boldsymbol{x}$, *and* C_0, C_1, C_2 *and* c_0 *are positive universal constants.*

Theorem 4, depending on the choice of λ, provides universal recovery guarantees over a large class of signals obeying $\frac{\|\boldsymbol{x}_\Omega\|_2}{\|\boldsymbol{x}_\Omega\|_1} \geq \rho$. Some implications of Theorem 4 are as follows.

1) *Exact Recovery for Exactly Sparse Signals:* When $\boldsymbol{x}$ is an exactly k-sparse signal, we can set $\rho = \frac{1}{\sqrt{k}}$ and $\lambda = \frac{1}{k}$ in Theorem 4, which implies the algorithm (16) universally recovers all k-sparse signals $\boldsymbol{x}$ from $O(k^2 \log n)$ noise-free measurements, with exponentially high probability. This recovers the theoretical performance guarantees established in [24] for Gaussian sensing vectors, but extends it to a large class of sub-Gaussian sensing vectors, using a simpler proof.
2) *Near-Optimal Recovery for Power-Law Exactly Sparse Signals:* Somewhat surprisingly, if the *nonzero entries* of $\boldsymbol{x}$ are known to be decaying in a power-law fashion, then the algorithm (16) allows near-optimal recovery. Specifically, suppose that the *non-zero entries* of $\boldsymbol{x}$ satisfies the power-law decay such that the magnitude of the lth largest entry of $\boldsymbol{x}_\Omega / \|\boldsymbol{x}_\Omega\|_2$ is bounded above by c_{pl}/l^α for some constants c_{pl} and exponent $\alpha > 1$, then
$$\|\boldsymbol{x}_\Omega\|_2 / \|\boldsymbol{x}_\Omega\|_1 = O\left(1/\log k\right) := \rho.$$

TABLE II
SUMMARY OF MAIN RESULTS

Structure	Number of Measurements	Noise	RIP
rank-r	$O(nr)$	ℓ_1	ℓ_2/ℓ_1
Toeplitz rank-r	$O(r\text{polylog}n)$	ℓ_2	ℓ_2/ℓ_2
k-sparse	$O(k\log(n^2/k))$	ℓ_1	ℓ_2/ℓ_1
k-sparse and rank-one	$O(k^2\log n)$ (general sparse); $O(k\log^2 n)$ (power-law sparse)	ℓ_1	ℓ_2/ℓ_1

By setting $\lambda = \Theta((\sqrt{k}\log n)^{-1})$, one can obtain accurate recovery from $O\left(k\log^2 n\right)$ noiseless samples, which is only a logarithmic factor from the minimum sample complexity requirement.

3) *Stable and Universal Recovery for Imperfect Models and Noisy Samples:* When the sparsity assumption is inexact, or measurements are noisy, the estimate $\hat{\boldsymbol{X}}$ will not be exact, and we can recover the estimate of the signal $\hat{\boldsymbol{x}}$ as the top (normalized) eigenvector of $\hat{\boldsymbol{X}}$. Using the Davis-Kahan theorem in standard matrix perturbation theory [47], we have

$$\sin\angle(\hat{\boldsymbol{x}},\boldsymbol{x}) \leq \frac{1}{\|\boldsymbol{x}\|_2^2}\left\|\hat{\boldsymbol{X}}-\boldsymbol{x}\boldsymbol{x}^\top\right\|_{\mathrm{F}}$$

bounded by Theorem 4, where $\angle(\hat{\boldsymbol{x}},\boldsymbol{x})$ represents the angle between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$. The recovered signal $\hat{\boldsymbol{x}}$ is a highly accurate estimate if $\boldsymbol{x}_{\Omega^c}$ is small enough. The estimation inaccuracy due to noise corruption is also small, in the sense that it is at most proportional to the per-entry noise level. This generalizes prior work [24] to imperfect structural assumptions as well as noisy measurements.

E. Extension to General Matrices

Table II summarizes the main results of Theorems 1–4. We further remark that the main results hold even when $\boldsymbol{\Sigma}$ is not PSD but a symmetric matrix.[2] When $\boldsymbol{\Sigma}$ is not a covariance matrix but a general low-rank, Toeplitz low-rank, or sparse matrix, one can simply drop the PSD constraint in the proposed algorithms, and replace the trace norm objective by the nuclear norm in (7). As will be shown, the PSD constraint is never invoked in the proof, hence it is straightforward to extend all results to the more general cases where $\boldsymbol{\Sigma}$ is a general $n\times n$ low-rank, Toeplitz low-rank, or sparse matrix. Note that in this more general scenario, the measurements in (1) are no longer nonnegative.

III. APPROXIMATE ℓ_2/ℓ_1 ISOMETRY FOR LOW-RANK AND SPARSE MATRICES

In this section, we present a novel concept called the mixed-norm restricted isometry property (RIP-ℓ_2/ℓ_1) that allows us to establish Theorems 1, 3, and 4 concerning universal recovery of low-rank, sparse and sparse rank-one covariance matrices from quadratic measurements.

[2]The proposed framework and proof arguments can also be easily extended to handle asymmetric matrices without difficulty, using bilinear rank-one measurements.

Prevailing wisdom in CS asserts that perfect recovery from minimal samples is possible if the dimensionality reduction projection preserves the signal strength when acting on the class of matrices of interest [32], [44]. While there are various ways to define the restricted isometry properties (RIP), an appropriately chosen approximate isometry leads to a very simple yet powerful theoretical framework.

A. Mixed-Norm Restricted Isometry (RIP-ℓ_2/ℓ_1)

Recall that the RIP occurs if the sampling output preserves the input strength under certain metrics. The most commonly used one is RIP-ℓ_2/ℓ_2, for which the signal strength before and after the projection are both measured in terms of the Frobenius norm [44], [46]. This, however, fails to hold under rank-one measurements – see detailed arguments by Candes *et al.* in [15]. Another isometry concept called RIP-ℓ_1/ℓ_1 has also been investigated, for which the signal strength before and after the operation $\mathcal{A}$ are measured both in terms of the ℓ_1 norms.[3] This is initially developed to account for measurements from expander graphs [48], and has become a powerful metric when analyzing phase retrieval [15], [23], [24]. Nevertheless, when considering general low-rank matrices, RIP-ℓ_1/ℓ_1 no longer holds. To see this, consider two matrices

$$\boldsymbol{X}_1 = \mathrm{diag}\{\boldsymbol{I}_{r/2},\boldsymbol{I}_{r/2}\boldsymbol{0}\}$$
$$\boldsymbol{X}_2 = \mathrm{diag}\{\boldsymbol{I}_{r/2},-\boldsymbol{I}_{r/2}\boldsymbol{0}\}$$

enjoying the same nuclear norm. When $m = \Omega(nr)$, one can see from the Bernstein inequality (for sub-exponential variables) that

$$\frac{1}{m}\|\mathcal{A}(\boldsymbol{X}_1)\|_1 = \Theta(r),\quad \frac{1}{m}\|\mathcal{A}(\boldsymbol{X}_2)\|_1 = \Theta\left(\sqrt{r}\right),$$

precluding the existence of a small RIP-ℓ_1/ℓ_1 constant. Leaving out this matter, the proof based on RIP-ℓ_1/ℓ_1 typically relies on delicate construction of dual certificates [15], [23], [24], which is often mathematically complicated.

One of the key and novel ingredients in our analysis is a mixed-norm approximate isometry, which measures the signal strength before and after the projection with different metrics. Specifically, we introduce RIP-ℓ_2/ℓ_1, where the input and output are measured in terms of the Frobenius norm and the ℓ_1 norm, respectively. It turns out that as long as the input is measured with the Frobenius norm, the standard trick pioneered in [46] in treating linear measurements carry over to quadratic measurements with slight modifications and saves the need for dual construction. We make formal definitions of RIP-ℓ_2/ℓ_1 for low-rank/sparse matrices as follows.

Definition 1 (RIP-ℓ_2/ℓ_1 for Low-Rank Matrices): For the set of rank-r matrices, we define the RIP-ℓ_2/ℓ_1 constants δ_r^{lb} and δ_r^{ub} with respect to an operator $\mathcal{B}$ as the smallest numbers such that for all $\boldsymbol{X}$ of rank at most r:

$$\left(1-\delta_r^{\mathrm{lb}}\right)\|\boldsymbol{X}\|_{\mathrm{F}} \leq \frac{1}{m}\|\mathcal{B}(\boldsymbol{X})\|_1 \leq \left(1+\delta_r^{\mathrm{ub}}\right)\|\boldsymbol{X}\|_{\mathrm{F}}.$$

[3]Note that the nuclear norm is the ℓ_1-norm counterpart for matrices.

Definition 2 (RIP-ℓ_2/ℓ_1 for Sparse Matrices): For the set of k-sparse matrices, we define the RIP-ℓ_2/ℓ_1 constants γ_k^{lb} and γ_k^{ub} with respect to an operator $\mathcal{B}$ as the smallest numbers such that for all $\boldsymbol{X}$ of sparsity at most k:

$$\left(1-\gamma_k^{\text{lb}}\right)\|\boldsymbol{X}\|_{\text{F}} \leq \frac{1}{m}\|\mathcal{B}(\boldsymbol{X})\|_1 \leq \left(1+\gamma_k^{\text{ub}}\right)\|\boldsymbol{X}\|_{\text{F}}.$$

Definition 3 (RIP-ℓ_2/ℓ_1 for Low-Rank Plus Sparse Matrices): Consider the class of index sets

$$\mathcal{S}_k := \Big\{\Omega \in [n]\times[n] \;\Big|\; \exists \text{ an index set } \omega \in [n] \text{ of cardinality } k \text{ such that } \Omega = \omega\times\omega\Big\}.$$

For the set of matrices

$$\mathcal{M}_{k,r,l} = \Big\{\boldsymbol{X}_1+\boldsymbol{X}_2 \;\Big|\; \exists\Omega\in\mathcal{S}_k,\ \operatorname{rank}(\boldsymbol{X}_1)\leq r,\ \operatorname{supp}(\boldsymbol{X}_1)\in\Omega,\ \|\boldsymbol{X}_2\|_0\leq l\Big\}. \tag{18}$$

we define the RIP-ℓ_2/ℓ_1 constants $\delta_{k,r,l}^{\text{lb}}$ and $\delta_{k,r,l}^{\text{ub}}$ with respect to an operator $\mathcal{B}$ as the smallest numbers such that $\forall\boldsymbol{X}\in\mathcal{M}_{k,r,l}$:

$$\left(1-\delta_{k,r,l}^{\text{lb}}\right)\|\boldsymbol{X}\|_{\text{F}} \leq \frac{1}{m}\|\mathcal{B}(\boldsymbol{X})\|_1 \leq \left(1+\delta_{k,r,l}^{\text{ub}}\right)\|\boldsymbol{X}\|_{\text{F}}.$$

Remark 3: In short, any matrix within $\mathcal{M}_{k,r,l}$ can be decomposed into two components $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, where $\boldsymbol{X}_1$ is simultaneously low-rank and sparse, and $\boldsymbol{X}_2$ is sparse. This allows us to treat each matrix perturbation as a superposition of a collection of jointly low-rank and sparse matrices and a collection of general sparse matrices, where the rank-one measurements of each term can be well controlled under minimal sample complexity.

B. RIP-ℓ_2/ℓ_1 of Quadratic Measurements for Low-rank and Sparse Matrices

Unfortunately, the original sampling operator $\mathcal{A}$ does not satisfy RIP-ℓ_2/ℓ_1. This occurs primarily because each measurement matrix $\boldsymbol{A}_i$ has non-zero mean, which biases the output measurements. In order to get rid of this undesired bias effect, we introduce a set of "debiased" auxiliary measurement matrices as follows

$$\boldsymbol{B}_i := \boldsymbol{A}_{2i-1} - \boldsymbol{A}_{2i}. \tag{19}$$

Without loss of generality, denote $\mathcal{B}_i(\boldsymbol{X}) := \langle\boldsymbol{B}_i, \boldsymbol{X}\rangle$ for all $1\leq i\leq m$, and let $\mathcal{B}(\boldsymbol{X})$ represent the linear transformation that maps $\boldsymbol{X}$ to $\{\mathcal{B}_i(\boldsymbol{X})\}_{i=1}^m$. Note that by representing the sensing process using m rank-2 measurements $\mathcal{B}_i$, we have implicitly doubled the number of measurements for notational simplicity. This, however, will not change our order-wise results.

It turns out that the auxiliary operator $\mathcal{B}$ exhibits the RIP-ℓ_2/ℓ_1 in the presence of minimal measurements, which can be shown by combining the following proposition with a standard covering argument as applied in [49].

Proposition 1: Let $\mathcal{A}$ be sampled from the sub-Gaussian model in (4)*. For any matrix $\boldsymbol{X}$, there exist universal constants $c_1, c_2, c_3 > 0$ such that with probability exceeding $1-\exp(-c_3 m)$, one has*

$$c_1\|\boldsymbol{X}\|_{\text{F}} \leq \frac{1}{m}\|\mathcal{B}(\boldsymbol{X})\|_1 \leq c_2\|\boldsymbol{X}\|_{\text{F}}. \tag{20}$$

Proof: See Appendix A. □

Remark 4: This statement extends without difficulty to the bilinear rank-one measurement model where $\boldsymbol{y}_i = \boldsymbol{a}_i^\top\boldsymbol{\Sigma}\boldsymbol{b}_i$ for some independently generated sensing vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$. This indicates that all our results hold for this asymmetric sensing model as well.

An immediate consequence of Proposition 1 is the establishment of RIP-ℓ_2/ℓ_1 of the sampling operator $\mathcal{B}$ for either general low-rank or sparse matrices. The proof of the corollaries below follows immediately from a standard covering argument detailed in [49, Sec. III-B] and [50, Sec. 5]. We thus omit the details but refer interested readers to the above references for details.

Corollary 1 (RIP-ℓ_2/ℓ_1 for Low-Rank Matrices): Consider the sub-Gaussian sampling model in (4) *and the universal constants $c_1, c_2 > 0$ given in (20). There exist universal constants $c_3, c_4, C_3 > 0$ such that with probability exceeding $1 - C_3\exp(-c_3 m)$, $\mathcal{B}$ satisfies RIP-ℓ_2/ℓ_1 for all matrices $\boldsymbol{X}$ of rank at most r, and obeys*

$$1-\delta_r^{\text{lb}} \geq \frac{c_1}{2}, \quad 1+\delta_r^{\text{ub}} \leq 2c_2, \tag{21}$$

provided that $m > c_4 nr$.

Corollary 2 (RIP-ℓ_2/ℓ_1 for Sparse Matrices): Consider the sub-Gaussian sampling model in (4) *and the universal constants $c_1, c_2 > 0$ given in (20). Then with probability exceeding $1 - C_3\exp(-c_3 m)$, $\mathcal{B}$ satisfies the RIP-ℓ_2/ℓ_1 for all matrices $\boldsymbol{X}$ of sparsity at most k, and obeys*

$$1-\gamma_k^{\text{lb}} \geq \frac{c_1}{2}, \quad 1+\gamma_k^{\text{ub}} \leq 2c_2, \tag{22}$$

provided that $m > c_4 k\log(n^2/k)$, where $c_3, c_4, C_3 > 0$ are some universal constants.

Corollary 3 (RIP-ℓ_2/ℓ_1 for Low-Rank Plus Sparse Matrices): Consider the sub-Gaussian sampling model in (4) *and the universal constants $c_1, c_2 > 0$ given in (20). Then with probability exceeding $1 - C_3\exp(-c_3 m)$, $\mathcal{B}$ satisfies the RIP-ℓ_2/ℓ_1 with respect to $\mathcal{M}_{k,r,l}$ (defined in (18)), and obeys*

$$1-\delta_{k,r,l}^{\text{lb}} \geq \frac{c_1}{2}, \quad 1+\delta_{k,r,l}^{\text{ub}} \leq 2c_2, \tag{23}$$

provided that $m > c_4\max\left\{kr\log(n/k), l\log(n^2/l)\right\}$, where $c_3, c_4, C_3 > 0$ are some universal constants.

Remark 5: Recall that each matrix in $\mathcal{M}_{k,r,l}$ is a sum of some $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, where $\boldsymbol{X}_1$ is a rank-r matrix in a $k\times k$ subspace, while $\boldsymbol{X}_2$ is an l-sparse matrix. Consequently, if we let $\mathcal{C}_\epsilon(\mathcal{M})$ stand for the covering number of a set $\mathcal{M}$ (i.e. the the fewest number of points in any ϵ-net of $\mathcal{M}$), then $\mathcal{C}_\epsilon\left(\mathcal{M}_{k,r,l}\right)$ is apparently bounded above by the product of $\mathcal{C}_{\epsilon/2}(\mathcal{M}_r)$ and $\mathcal{C}_{\epsilon/2}(\mathcal{M}_l)$, where $\mathcal{M}_r$ and $\mathcal{M}_l$ denotes the rank-r manifold (with ambient dimension k) and the ℓ-sparse manifold (with ambient dimension n^2), respectively. Thus, $\log\mathcal{C}_\epsilon\left(\mathcal{M}_{k,r,l}\right)$ cannot exceed $kr\log(n/k) + l\log\left(n^2/l\right)$.

C. Proof of Theorems 1, 3 and 4 via RIP-ℓ_2/ℓ_1

Theorems 1 and 3 can thus be proved given that the auxiliary operator $\mathcal{B}$ satisfies RIP-ℓ_2/ℓ_1 with sufficiently small constants, as asserted in Corollaries 1 and 2. We first present Lemma 1 which in turn establishes Theorem 1.

Lemma 1: Consider any matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_{\mathrm{c}}$, where $\boldsymbol{\Sigma}_r$ is the best rank-r approximation of $\boldsymbol{\Sigma}$. If there exists a number $K_1 > 2r$ such that

$$\frac{1-\delta_{2r+K_1}^{\mathrm{lb}}}{\sqrt{2}} - \left(1+\delta_{K_1}^{\mathrm{ub}}\right)\sqrt{\frac{2r}{K_1}} \geq \beta_1 > 0 \tag{24}$$

holds for some numerical value β_1, then the minimizer $\hat{\boldsymbol{\Sigma}}$ to (7) *obeys*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} \leq \left(\frac{C_1}{\beta_1} + C_3\right)\frac{\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_*}{\sqrt{K_1}} + \frac{C_2}{\beta_1}\cdot\frac{\epsilon_1}{m} \tag{25}$$

for some positive universal constants C_1, C_2 and C_3 depending only on the RIP-ℓ_2/ℓ_1 constants.

Proof: See Appendix C. □

By choosing

$$K_1 = 8\left(\frac{4c_2}{c_1}\right)^2 r \geq 8\left(\frac{1+\delta_{K_1}^{\mathrm{ub}}}{1-\delta_{2r+K_1}^{\mathrm{ub}}}\right)^2 r$$

for the universal constants c_1, c_2 given in Corollary 1, we obtain (24) when $m > c_4\,(K_1 + 2r)\,n$ for some constant c_4. This establishes Theorem 1.

Theorem 3 is a direct consequence from the following lemma.

Lemma 2: Consider any matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\Omega + \boldsymbol{\Sigma}_{\Omega^c}$, where $\boldsymbol{\Sigma}_\Omega$ is the best k-term approximation of $\boldsymbol{\Sigma}$. If there exists a number $K_2 > 2k$ such that

$$\frac{(1-\gamma_{k+K_2}^{\mathrm{lb}})}{\sqrt{2}} - \left(1+\gamma_{K_2}^{\mathrm{ub}}\right)\sqrt{\frac{k}{K_2}} \geq \beta_2 > 0 \tag{26}$$

holds for some numerical value β_2, then the minimizer $\hat{\boldsymbol{\Sigma}}$ to (7) *obeys*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} \leq \left(\frac{C_1}{\beta_2} + C_3\right)\frac{\|\boldsymbol{\Sigma}_{\Omega^c}\|_1}{\sqrt{K_2}} + \frac{C_2}{\beta_2}\frac{\epsilon_1}{m} \tag{27}$$

for some positive universal constants C_1, C_2, C_3 depending only on the RIP-ℓ_2/ℓ_1 constants.

Proof: See Appendix D. □

By picking

$$K_2 = 4\left(\frac{4c_2}{c_1}\right)^2 k \geq 4\left(\frac{1+\gamma_{K_2}^{\mathrm{ub}}}{1-\gamma_{k+K_2}^{\mathrm{lb}}}\right)^2 k,$$

one obtains (26) as soon as $m > c_4\,(K_2 + 2k)\log(n^2/k)$ for the constant c_4 given in Corollary 2. This concludes the proof of Theorem 3.

Furthermore, the specialized RIP-ℓ_2/ℓ_1 concept allows us to prove Theorem 4 through the following lemma.

Lemma 3: Set λ to be any number within the interval $\left[\frac{1}{n}, \frac{1}{\sqrt{k}}\frac{\|\boldsymbol{x}_\Omega\|_2}{\|\boldsymbol{x}_\Omega\|_1}\right]$. Suppose that $\boldsymbol{x}_\Omega$ is the best k-term approximation of $\boldsymbol{x}$. If there exists a number K_1 such that

$$\begin{cases} \dfrac{\frac{1}{\sqrt{3}}\left(1-\delta^{\mathrm{lb}}_{k,2K_1,\frac{2K_1}{\lambda^2}}\right) - \frac{3}{\sqrt{K_1}}\left(1+\delta^{\mathrm{ub}}_{k,K_1,\frac{K_1}{\lambda^2}}\right)}{2\max\left\{\frac{1}{\sqrt{K_1}}\left(1+\delta^{\mathrm{ub}}_{k,K_1,\frac{K_1}{\lambda^2}}\right),1\right\}} \geq \beta_3 > 0, \\ \dfrac{1+\delta^{\mathrm{ub}}_{k,K_1,\frac{K_1}{\lambda^2}}}{\left(1-\delta^{\mathrm{lb}}_{k,K_1,\frac{K_1}{\lambda^2}}\right)\sqrt{K_1}} \leq \beta_4 \end{cases} \tag{28}$$

for some absolute constants β_3 and β_4, then the solution $\hat{\boldsymbol{X}}$ to (16) satisfies

$$\begin{aligned}\left\|\hat{\boldsymbol{X}} - \boldsymbol{x}\boldsymbol{x}^\top\right\|_{\mathrm{F}} \leq C\Big\{&\left\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}_\Omega\boldsymbol{x}_\Omega^\top\right\|_* \\ &+ \lambda\left\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}_\Omega\boldsymbol{x}_\Omega^\top\right\|_1 + \frac{\epsilon_1}{m}\Big\}\end{aligned} \tag{29}$$

for some constant C that depends only on β_3 and β_4.

Proof: See Appendix E. □

From Corollary 3, one can ensure small RIP-ℓ_2/ℓ_1 constants satisfying (28), provided that

$$m > c_4 \max\left\{kK_1\log n, \frac{K_1}{\lambda^2}\log n\right\} = c_4\frac{K_1}{\lambda^2}\log n.$$

This in turn establishes Theorem 4.

Finally, note that we have not discussed general Toeplitz low-rank matrices using RIP-ℓ_2/ℓ_2. We are unaware of a rigorous approach to prove exact recovery using RIP-ℓ_2/ℓ_1 for the Toeplitz case, partly due to the difficulty in characterizing the covering number for general low-rank Toeplitz matrices. Fortunately, the analysis for Toeplitz low-rank matrices can be performed by means of a different method, as detailed in the next section.

IV. Approximate ℓ_2/ℓ_2 Isometry for Toeplitz Low-Rank Matrices

While quadratic measurements in general do not exhibit RIP-ℓ_2/ℓ_2 (as introduced in [44]) with respect to the set of general low-rank matrices (as pointed out in [15]), a slight variant of them can indeed satisfy RIP-ℓ_2/ℓ_2 when restricted to *Toeplitz* low-rank matrices. In this section, we first provide a characterization of RIP-ℓ_2/ℓ_2 for the set of general low-rank matrices under bounded and near-isotropic measurements, and then convert quadratic measurements into equivalent isotropic measurements.

A. RIP-ℓ_2/ℓ_2 for Near-Isotropic and Bounded Measurements

Before proceeding to the Toeplitz low-rank matrices, we investigate near-isotropic and bounded operators for the set of general low-rank matrices as follows. For convenience of presentation, we repeat the definition of RIP-ℓ_2/ℓ_2 as follows, followed by a theorem characterizing RIP-ℓ_2/ℓ_2 for near-isotropic and bounded operators.

Definition 4 (RIP-ℓ_2/ℓ_2 for Low-Rank Matrices): For the set of rank-r matrices, we define the RIP-ℓ_2/ℓ_2 constants δ_r

w.r.t. an operator $\mathcal{B}$ as the smallest number such that for all $\boldsymbol{X}$ of rank at most r,

$$(1-\delta_r)\,\|\boldsymbol{X}\|_{\mathrm{F}} \le \frac{1}{m}\,\|\mathcal{B}(\boldsymbol{X})\|_2 \le (1+\delta_r)\,\|\boldsymbol{X}\|_{\mathrm{F}}.$$

Theorem 5: Suppose that for all $1 \le i \le m$,

$$\|\boldsymbol{B}_i\| \le K \quad \text{and} \quad \left\|\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]-\mathcal{I}\right\| \le \frac{c_5}{n} \tag{30}$$

hold for some quantity $K \le n^2$. For any small constant $\delta > 0$, if $m > c_0 r K^2 \log^7 n$, then with probability at least $1-1/n^2$, one has[4]

i) $\mathcal{B}$ satisfies RIP-ℓ_2/ℓ_2 w.r.t. all matrices of rank at most r and obeys $\delta_r \le \delta$;

ii) Suppose that $\mathcal{K}$ is some convex set. Then for all $\boldsymbol{\Sigma}$ of rank at most r and $\boldsymbol{\Sigma} \in \mathcal{K}$, if $\|\boldsymbol{y}-\mathcal{B}(\boldsymbol{\Sigma})\|_2 \le \epsilon_2$, the solution

$$\hat{\boldsymbol{\Sigma}} = \operatorname{argmin}_{\boldsymbol{M}} \|\boldsymbol{M}\|_* \quad \text{subject to} \quad \|\boldsymbol{y}-\mathcal{B}(\boldsymbol{M})\|_2 \le \epsilon_2, \quad \boldsymbol{M} \in \mathcal{K},$$

satisfies

$$\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_{\mathrm{F}} \le C_2\frac{\epsilon_2}{\sqrt{m}} \tag{31}$$

for some universal constants $c_0, C_2, c_5 > 0$.

Proof: See Appendix B. □

In fact, the bound on $\|\boldsymbol{B}_i\|$ can be as small as $\Theta\left(\sqrt{n}\right)$, and we say a measurement matrix $\boldsymbol{B}_i$ is *well-bounded* if $K = O\left(\sqrt{n}\mathrm{polylog}n\right)$. Simultaneously well-bounded and near-isotropic operators (i.e. those satisfying (30)) subsume the Fourier-type basis discussed in [26]. Theorem 5 strengthens the result in [26] to admit universal and stable recovery of low-rank matrices with random subsampling using Fourier-type basis, by justifying RIP-ℓ_2/ℓ_2 as soon as $m = \Omega(nr\mathrm{polylog}n)$.

Unfortunately, Theorem 5 cannot be directly applied to the class of Toeplitz low-rank matrices for the following reasons: i) The sampling operator $\mathcal{A}$ is neither isotropic nor well-bounded; ii) Theorem 5 requires $m = \Omega(nr\mathrm{polylog}n)$ measurements, which far exceeds the ambient dimension of a Toeplitz matrix, which is n. This motivates us to construct another set of equivalent sampling operators that satisfies the assumptions of Theorem 5, which is the focus of the following subsection.

B. Construction of RIP-ℓ_2/ℓ_2 Operators for Toeplitz Low-Rank Matrices

Note that the quadratic measurement matrices $\boldsymbol{A}_i = \boldsymbol{a}_i\boldsymbol{a}_i^\top$ are neither non-isotropic nor well bounded. For instance, when $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, simple calculation reveals that

$$\|\boldsymbol{A}_i\| = \Theta\left(\sqrt{n}\right), \quad \text{and} \quad \mathbb{E}\left[\boldsymbol{A}_i\left\langle\boldsymbol{A}_i,\boldsymbol{X}\right\rangle\right] = 2\boldsymbol{X}+\operatorname{tr}(\boldsymbol{X})\cdot\boldsymbol{I}, \tag{32}$$

precluding $\boldsymbol{A}_i$'s from being isotropic. In order to facilitate the use of Theorem 5, we generate a new set of measurement matrices $\tilde{\boldsymbol{B}}_i$ through the following procedure.

1) Define a set of matrices $\boldsymbol{B}_i$ of rank at most 3

$$\boldsymbol{B}_i := \begin{cases} \frac{1}{2}\left(\boldsymbol{A}_{2i-1}-\boldsymbol{A}_{2i}\right), & \text{if } \mu_4 = 3, \\ \alpha\boldsymbol{A}_{3i}+\beta\boldsymbol{A}_{3i-1}+\gamma\boldsymbol{A}_{3i-2}, & \text{if } \mu_4 < 3, \end{cases} \tag{33}$$

where α, β, γ are specified in Lemma 4.

2) Generate M (whose choice will be specified later) matrices independently such that

$$\hat{\boldsymbol{B}}_i = \begin{cases} \sqrt{n}\mathcal{T}\left(\boldsymbol{B}_i\right), & \text{with probability } \frac{1}{n}, \\ \sqrt{\frac{n}{n-1}}\mathcal{T}^\perp\left(\boldsymbol{G}_i\right), & \text{with probability } \frac{n-1}{n}, \end{cases} \tag{34}$$

where $\boldsymbol{G}_i$ is a random matrix with i.i.d. standard Gaussian entries.

3) Define a *truncated* version $\tilde{\boldsymbol{B}}_i$ of $\hat{\boldsymbol{B}}_i$ as follows

$$\tilde{\boldsymbol{B}}_i = \hat{\boldsymbol{B}}_i \cdot \mathbb{1}_{\left\{\left\|\hat{\boldsymbol{B}}_i\right\| \le c_{10}\log^{3/2}n\right\}}, \quad 1 \le i \le M. \tag{35}$$

We will demonstrate that the $\tilde{\boldsymbol{B}}_i$'s are nearly-isotropic and well-bounded, and hence by Theorem 5 the associated operator $\tilde{\mathcal{B}}$ enables exact and stable recovery for all rank-r matrices when M exceeds $O(nr\mathrm{polylog}n)$. This in turn establishes Theorem 2 through an equivalence argument, detailed later.

1) Isotropy Trick: While $\boldsymbol{A}_i$'s are in general non-isotropic, a linear combination of them can be made isotropic when restricted to Toeplitz matrices. This is stated in the following lemma.

Lemma 4: Consider the sub-Gaussian sampling model in (4).

1) When $\mu_4 = 3$, then for any $\boldsymbol{X}$, the matrix

$$\boldsymbol{B}_i = \frac{1}{2}\left(\boldsymbol{A}_{2i-1}-\boldsymbol{A}_{2i}\right) \tag{36}$$

satisfies

$$\mathbb{E}\left[\boldsymbol{B}_i\left\langle\boldsymbol{B}_i,\boldsymbol{X}\right\rangle\right] = \boldsymbol{X}. \tag{37}$$

2) When $\mu_4 < 3$, take any constant $\xi > 0$ obeying $\xi^2 > 1.5(3-\mu_4)$ and set

$$\boldsymbol{B}_i = \alpha\boldsymbol{A}_{3i}+\beta\boldsymbol{A}_{3i-1}+\gamma\boldsymbol{A}_{3i-2}, \tag{38}$$

with the choice of $\Delta := -\left(1-\frac{\xi}{n}\right)^2 - 2 + \frac{2\xi^2}{3-\mu_4}$,

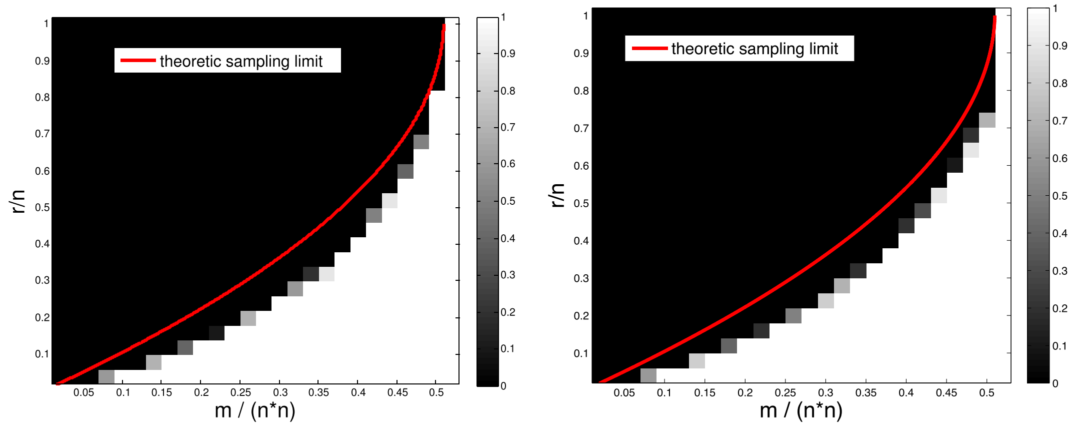
$$\begin{cases} \alpha = \sqrt{\dfrac{3-\mu_4}{2\xi^2}}, \\ \beta := \dfrac{-\left(1-\dfrac{\xi}{\sqrt{n}}\right)+\sqrt{\Delta}}{2}\alpha, \\ \gamma := \dfrac{-\left(1-\dfrac{\xi}{\sqrt{n}}\right)-\sqrt{\Delta}}{2}\alpha. \end{cases} \tag{39}$$

Then, for any norm $\|\cdot\|_{\mathrm{n}}$ and any $\boldsymbol{X}$ that satisfies $\boldsymbol{X}_{11} = \boldsymbol{X}_{22} = \cdots = \boldsymbol{X}_{nn}$, one has

$$\begin{cases} \mathbb{E}\left[\boldsymbol{B}_i\right] = \sqrt{\dfrac{3-\mu_4}{2n}}; \\ \mathbb{E}\left[\boldsymbol{B}_i\left\langle\boldsymbol{B}_i,\boldsymbol{X}\right\rangle\right] = \boldsymbol{X}; \\ \|\boldsymbol{B}_i\|_{\mathrm{n}} \le \sqrt{3}\max_{i:1\le i\le m}\|\boldsymbol{A}_i\|_{\mathrm{n}}. \end{cases} \tag{40}$$

Proof: See Appendix F. □

[4] The proof of Theorem 5 follows the entropy method introduced in [25], where $\log^7 n$ factor is a natural consequence, and might be refined a bit by generic chaining due to Talagrand [51] as employed in [52]. However, we are unaware of an approach that can get rid of the logarithmic factor.



(a) (b)

Fig. 2. Recovery of covariance matrices from quadratic measurements when $n = 50$. For each (m, r) pair, we repeated Monte Carlo trials 20 times. A PSD matrix $\boldsymbol{\Sigma}$ and m sensing vectors are selected at random. The colormap for each cell indicates the empirical probability of success, and the red line reflects the fundamental information theoretic limit. The results are shown for (a) Gaussian sensing vectors and (b) symmetric Bernoulli sensing vectors.

Lemma 4 asserts that a large class of measurement matrices can be made isotropic when restricted to the class of matrices with identical diagonal entries (e.g. Toeplitz matrices). This immediately implies that the operator $\hat{\mathcal{B}}$ associated with $\hat{\boldsymbol{B}}_i$'s (defined in (34)) is isotropic. Specifically, for any symmetric $\boldsymbol{X}$,

$$\begin{aligned}
&\mathbb{E}\left[\hat{\boldsymbol{B}}_i\left\langle\hat{\boldsymbol{B}}_i,\boldsymbol{X}\right\rangle\right]\\
&\quad=\mathbb{E}\left[\mathcal{T}(\boldsymbol{B}_i)\left\langle\boldsymbol{B}_i,\mathcal{T}(\boldsymbol{X})\right\rangle\right]+\mathbb{E}\left[\mathcal{T}^{\perp}(\boldsymbol{G}_i)\left\langle\boldsymbol{G}_i,\mathcal{T}^{\perp}(\boldsymbol{X})\right\rangle\right]\\
&\quad=\mathcal{T}\left(\mathbb{E}\left[\boldsymbol{B}_i\left\langle\boldsymbol{B}_i,\mathcal{T}(\boldsymbol{X})\right\rangle\right]\right)+\mathcal{T}^{\perp}\left(\mathbb{E}\left[\boldsymbol{G}_i\left\langle\boldsymbol{G}_i,\mathcal{T}^{\perp}(\boldsymbol{X})\right\rangle\right]\right)\\
&\quad=\mathcal{T}\left(\mathcal{T}(\boldsymbol{X})\right)+\mathcal{T}^{\perp}\left(\mathcal{T}^{\perp}(\boldsymbol{X})\right)=\boldsymbol{X},
\end{aligned}$$

which follows since $\boldsymbol{B}_i$ and $\boldsymbol{G}_i$ are both isotropic matrices, a consequence of Lemma 4.

2) Truncation of $\hat{\mathcal{B}}$ is Near-Isotropic: The operators associated with $\hat{\boldsymbol{B}}_i$'s are in general not well-bounded. Fortunately, $\hat{\boldsymbol{B}}_i$'s are well-bounded with high probability, which comes from the following lemma whose proof can be found in Appendix G.

Lemma 5: Consider a random vector $\boldsymbol{z}$ that follows the sub-Gaussian sampling model as described in (4). *There exists an absolute constant $c_{10} > 0$ such that*

$$\left\|\mathcal{T}\left(\boldsymbol{z}\boldsymbol{z}^{\top}\right)\right\| \leq c_{12}\log^{\frac{3}{2}} n \tag{41}$$

holds with probability exceeding $1 - n^{-10}$.

As $\|\boldsymbol{B}_i\|$ can be bounded above by $\max_{1\leq i\leq m}\|\boldsymbol{A}_i\|$ up to some constant factor, Lemma 5 reveals that $\|\mathcal{T}(\boldsymbol{B}_i)\|$ can be well controlled for sub-Gaussian vectors, i.e.

$$\|\mathcal{T}(\boldsymbol{B}_i)\| \leq c_{10}\log^{\frac{3}{2}} n, \quad 1 \leq i \leq m \tag{42}$$

with probability exceeding $1 - 3n^{-8}$. Similarly, classical results in random matrices (see [53]) assert that $\|\boldsymbol{G}_i\|$ can also be bounded above by $O\left(\sqrt{n}\log n\right)$ with overwhelming probability. These bounds taken collectively suggest that

$$\|\hat{\boldsymbol{B}}_i\| \leq K := c_{10}\sqrt{n}\log^{\frac{3}{2}} n, \quad 1 \leq i \leq m \tag{43}$$

for some constant $c_{10} > 0$ with probability exceeding $1 - n^{-7}$.

The above stochastic boundedness property motivates us to study the truncated version $\tilde{\boldsymbol{B}}_i$ of $\hat{\boldsymbol{B}}_i$ as defined in (35). Interestingly, $\tilde{\boldsymbol{B}}_i$ is near-isotropic, a consequence of the following lemma whose proof can be found in Appendix H.

Lemma 6: Suppose that the restriction of $\mathcal{B}_i$ to Toeplitz matrices is isotropic. Consider any event E obeying $\mathbb{P}(E) \geq 1 - \frac{1}{n^5}$. Then there is some constant $c_5 > 0$ such that

$$\left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_E\right] - \mathcal{T}\right\| \leq \frac{c_5}{n^2}. \tag{44}$$

The truncated version of $\boldsymbol{G}_i$ can be easily bounded as in [52], which we omit for simplicity of presentation. This combined with (44) indicates that

$$\begin{aligned}
&\left\|\mathbb{E}\left[\tilde{\mathcal{B}}_i^*\tilde{\mathcal{B}}_i\right] - \mathcal{I}\right\|\\
&\quad\leq \left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\right] - \mathcal{T}\right\| + \left\|\mathbb{E}\left[\mathcal{T}^{\perp}\mathcal{G}_i^*\mathcal{G}_i\mathcal{T}^{\perp}\right] - \mathcal{T}^{\perp}\right\|\\
&\quad\leq \frac{c_5}{n}.
\end{aligned}$$

C. Proof of Theorem 2

So far we have demonstrated that $\tilde{\boldsymbol{B}}_i$'s are near-isotropic and satisfy $\|\tilde{\boldsymbol{B}}_i\| = O\left(\sqrt{n}\log^{\frac{3}{2}} n\right)$. Suppose that $\|\boldsymbol{y} - \tilde{\mathcal{B}}(\boldsymbol{\Sigma})\|_2 \leq \tilde{\epsilon}_2$. Theorem 5 implies that if M exceeds $\Theta\left(nr\log^{10}(n)\right)$, then the solution to

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}} := \operatorname{argmin}_{\boldsymbol{M}} \|\boldsymbol{M}\|_* \quad &\text{subject to } \|\boldsymbol{y} - \tilde{\mathcal{B}}(\boldsymbol{M})\|_2 \leq \tilde{\epsilon}_2,\\
&\boldsymbol{M} \text{ is Toeplitz},
\end{aligned} \tag{45}$$

satisfies

$$\left\|\boldsymbol{\Sigma} - \tilde{\boldsymbol{\Sigma}}\right\|_{\mathrm{F}} \leq C_2\frac{\tilde{\epsilon}_2}{\sqrt{M}} \tag{46}$$

for the entire set of rank-r matrices $\boldsymbol{\Sigma}$. Apparently, such low-rank manifolds subsume all rank-r Toeplitz matrices as special cases. This claim in turn establishes Theorem 2 through the following argument:

1) From (34) and the Chernoff bound, $\tilde{\mathcal{B}}$ entails $\Theta\left(\frac{M}{n}\right) = \Theta\left(r\log^{10} n\right)$ independent copies of $\sqrt{n}\mathcal{T}(\boldsymbol{B}_i)$, and all other measurements are on the orthogonal complement of the Toeplitz space.

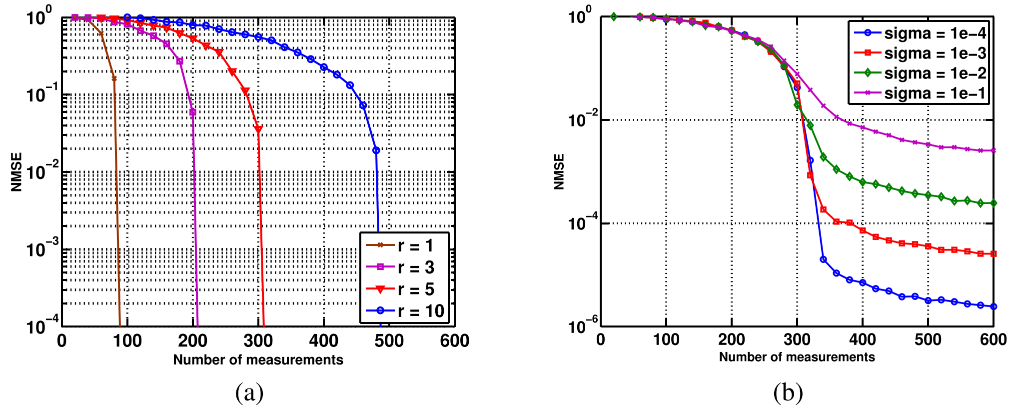


Fig. 3. The NMSE of the reconstructed covariance matrix via trace minimization vs. the number of measurements when $n = 40$: (a) for different ranks when no noise is present; (b) for different noise levels when $r = 5$.

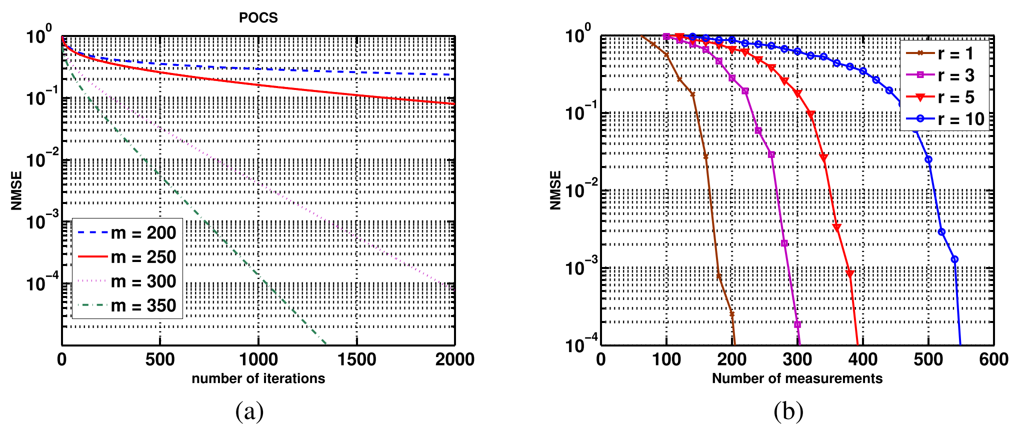


Fig. 4. The NMSE of the reconstructed covariance matrix via POCS for $n = 40$: (a) the NMSE vs. the number of iterations for different numbers of measurements when $r = 3$; (b) the NMSE vs. the number of measurements for different ranks when running 2000 iterations.

2) For any rank-r Toeplitz matrix $\boldsymbol{\Sigma}$, the original $\mathcal{A}$ entails $m/3 > \Theta\left(r \log^{10} n\right)$ measurement matrices of the form $\mathcal{T}(\boldsymbol{B}_i)$, and any non-Toeplitz component of $\boldsymbol{X}$ is perfectly known (i.e. equal to 0). This indicates that the convex program (11) is tighter than (45) when $\tilde{\epsilon}_2 = \Theta\left(\sqrt{n}\epsilon_2\right)$, i.e. one can construct (via coupling) a new probability space over which if the solution $\tilde{\boldsymbol{\Sigma}}$ to (45) is exact and unique, then it will be the unique solution to (11) as well. This combined with the universal bound (46) establishes Theorem 2.

V. Numerical Examples

To demonstrate the practical applicability of the proposed convex relaxation under quadratic sensing, in this section we present a variety of numerical examples for low-rank or sparse covariance matrix estimation.

A. Recovery of Low-Rank Covariance Matrices

We conduct a series of Monte Carlo trials for various parameters. Specifically, we choose $n = 50$, and for each (m, r) pair, we repeat the following experiments 20 times. We generate $\boldsymbol{\Sigma}$, an $n \times n$ PSD matrix via $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a randomly generated $n \times r$ matrix with independent

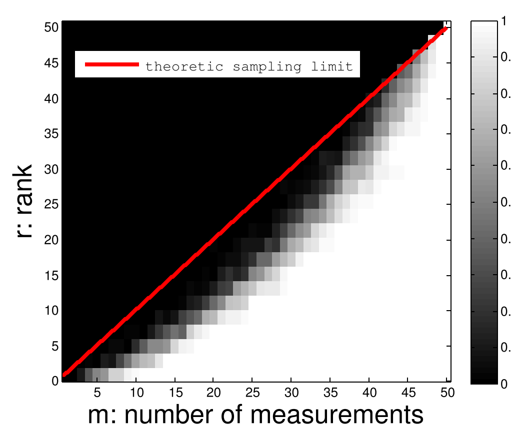
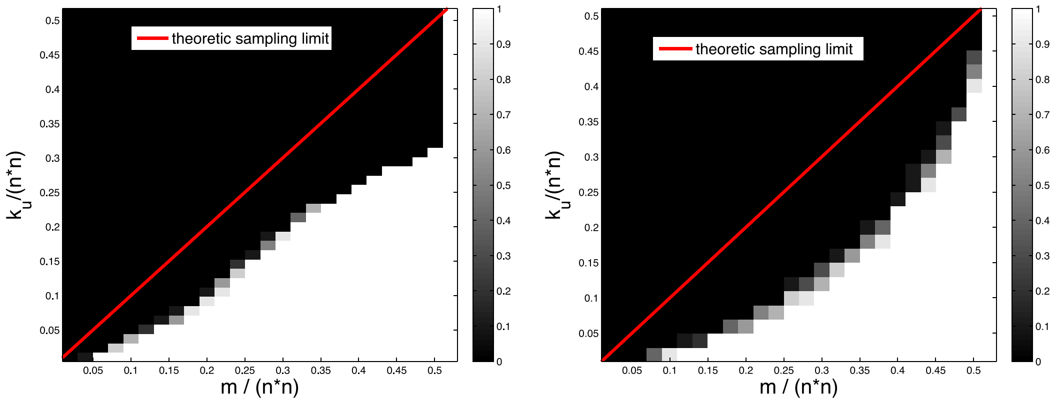


Fig. 5. Phase transition plots where frequency locations are randomly generated. The plot corresponds to the situation where $n = 50$. The empirical success rate is calculated by averaging over 50 Monte Carlo trials.

Gaussian components. The sensing vectors are generated as i.i.d. Gaussian vectors and Bernoulli vectors, and we obtain noiseless quadratic measurements $\boldsymbol{y}$. We use the off-the-shelf SDP solver SDPT3 with the modeling software CVX, and declare a matrix $\boldsymbol{\Sigma}$ to be recovered if the solution $\hat{\boldsymbol{\Sigma}}$ returned by the solver satisfies $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}}/\|\boldsymbol{\Sigma}\|_{\mathrm{F}} < 10^{-3}$. Figure 2



(a) (b)

Fig. 6. Reconstruction of sparse matrices from Gaussian quadratic measurements when $n = 50$. For ease of comparison, we let $k_{\rm u}$ denote the number of non-zero entries above or on the main diagonal, which represents the degrees of freedom for symmetric matrices. For each $(m, k_{\rm u})$ pair, we conducted Monte Carlo experiments 20 times. A PSD matrix $\boldsymbol{\Sigma}$ and m sensing vectors are selected at random. The colormap for each cell and the red line reflects the empirical probability of success and the information theoretic limit, respectively. The results are shown for (a) sparse PSD matrices, and (b) sparse symmetric matrices.

illustrates the empirical probability of successful recovery in these Monte Carlo trials, which is reflected through the color of each cell. In order to compare the optimality of the practical performance, we also plot the information theoretic limit in red lines, i.e. the fundamental lower limit on m required to recover all rank-r matrices, which is $nr - r(r-1)/2$ in our case. It turns out that the practical phase transition curve is very close to the theoretic sampling limit, which demonstrates the optimality of our algorithm.

In the second numerical example, we consider a random covariance matrix generated via the same procedure as above but with $n = 40$. We let the rank r vary as $1, 3, 5, 10$ and the number of measurements m vary from 20 to 600. For each pair of (r, m), we perform 10 independent experiments where in each run the sensing matrix is generated with i.i.d. Gaussian entries. Fig. 3 (a) shows the average Normalized Mean Squared Error (NMSE) defined as $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\rm F}^2/\|\boldsymbol{\Sigma}\|_{\rm F}^2$ with respect to m for different ranks when there is no noise. We further introduce additive bounded noise to each measurement by letting λ_i be generated from $\sigma \cdot \mathcal{U}[-1, 1]$, where $\mathcal{U}[-1, 1]$ is a uniform distribution on $[-1, 1]$, σ is the noise level. Fig. 3 (b) shows the average NMSE when $r = 5$ for different noise levels by setting $\epsilon = \sigma m$ in (7).

Interestingly, [23], [54] showed that when the covariance matrix is rank-one, if $m = O(n)$, the intersection of two convex sets, namely $\mathcal{S}_1 = \{\boldsymbol{M} : \mathcal{A}(\boldsymbol{M}) = \boldsymbol{y}\}$ and $\mathcal{S}_2 = \{\boldsymbol{M} : \boldsymbol{M} \succ 0\}$ is a singleton, with high probability. For the low-rank case, if the same conclusion holds, we can find the solution via alternating projection between two convex sets. Therefore, we experiment on the following Projection Onto Convex Sets (POCS) procedure:

$$\boldsymbol{\Sigma}_{t+1} = \mathcal{P}_{\mathcal{S}_2}\mathcal{P}_{\mathcal{S}_1}\boldsymbol{\Sigma}_t, \tag{47}$$

where $\mathcal{P}_{\mathcal{S}_2}$ denotes the projection onto the PSD cone, and

$$\mathcal{P}_{\mathcal{S}_1}\boldsymbol{\Sigma}_t := \boldsymbol{\Sigma}_t - \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}(\mathcal{A}(\boldsymbol{\Sigma_t}) - \boldsymbol{y}). \tag{48}$$

Fig. 4 (a) shows the NMSE of the reconstruction with respect to the number of iterations for $r = 3$ and different $m = 200, 250, 300, 350$. By comparing Fig. 3, we see that it requires more measurements for the POCS procedure to succeed, but the computational cost is much lower than the trace minimization. This is further validated from Fig. 4 (b), which is obtained under the same simulation setup as Fig. 3 by repeating POCS with 2000 iterations.

B. Recovery of Toeplitz Low-Rank Matrices

To justify the convex heuristic for Toeplitz low-rank matrices, we perform a series of numerical experiments for matrices of dimension $n = 50$. By Caratheodory's theorem, each PSD Toeplitz matrix can be uniquely decomposed into a linear combination of line spectrum [55]. Thus, we generate the PSD Toeplitz matrix by randomly generating the frequencies and amplitudes of each line spectra. In the real-valued case, the underlying spectral spikes occur in conjugate pairs (i.e. $(f_1, -f_1)$, $(f_2, -f_2)$, $\cdots$). We independently generate $r/2$ frequency pairs within the unit disk uniformly at random, and the amplitudes are generated as the absolute values of i.i.d. Gaussian variables. Figure 5 illustrates the phase transition diagram for varying choices of (m, r). Each trial is declared successful if the estimate $\hat{\boldsymbol{\Sigma}}$ satisfies $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\rm F}/\|\boldsymbol{\Sigma}\|_{\rm F} < 10^{-3}$. The empirical success rate is calculated by averaging over 50 Monte Carlo trials, and is reflected by the color of each cell. While there are in total r degrees of freedom, our algorithm exhibits an approximately linear phase transition curve, which confirms our theoretical prediction in the absence of noise.

C. Recovery of Sparse Matrices

We perform a series of Monte Carlo trials for various parameters for matrices of dimensions 50×50. We first generate PSD sparse covariance matrices in the following way. For each sparsity value k, we generate a $\sqrt{k} \times \sqrt{k}$ matrix via $\boldsymbol{\Sigma}_k = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a $\sqrt{k} \times \sqrt{k}$ matrix with independent Gaussian components. We then randomly select $\sqrt{k}$ rows and columns of $\boldsymbol{\Sigma}$ and embed $\boldsymbol{\Sigma}_k$ into the corresponding $\sqrt{k} \times \sqrt{k}$ submatrix; all other entries of $\boldsymbol{\Sigma}$ are set to 0. In addition,

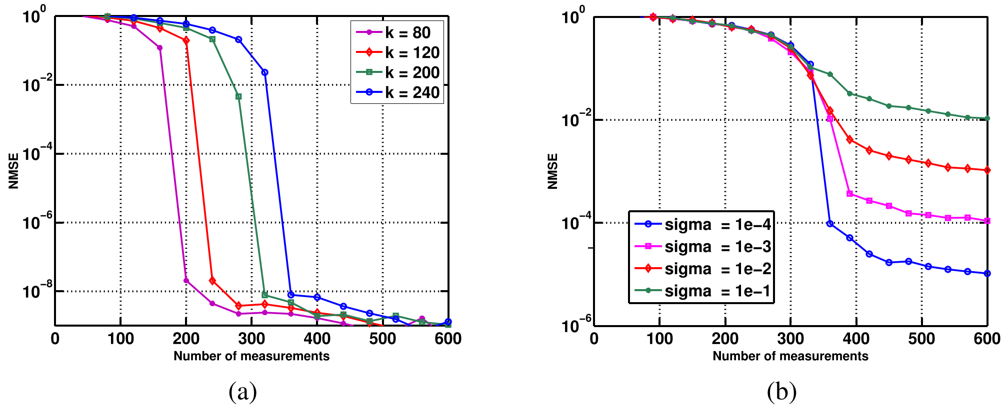


Fig. 7. The NMSE of the reconstructed sparse matrix via ℓ_1 minimization vs. the number of measurements when $n = 40$: (a) for different sparsity level when no noise is present; (b) for different noise levels when $k = 240$.

we also conduct numerical simulations for general symmetric sparse matrices, where the non-zero entries are drawn from an i.i.d. Gaussian distribution and the support is randomly chosen. For each (m, k) pair in each scenario, we repeated the experiments 20 times, and solve it using CVX. Again, a matrix $\mathbf{\Sigma}$ is claimed to be recovered if the solution $\hat{\mathbf{\Sigma}}$ returned by the solver satisfies $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathrm{F}}/\|\mathbf{\Sigma}\|_{\mathrm{F}} < 10^{-3}$. Figure 6 illustrates the empirical success probability in these Monte Carlo experiments. For ease of comparison, we also plot the degrees of freedom in red lines, which is $\frac{\sqrt{k}\left(\sqrt{k}+1\right)}{2}$ in our case. It turns out that the practical phase transition curve is close to the theoretic sampling limit, which demonstrates the optimality of our algorithm.

Another numerical example concerns recovery of a random symmetric sparse matrix (not necessarily PSD). We randomly generated a symmetric sparse matrix of sparsity level k with $n = 40$, and sketched it with i.i.d. Gaussian vectors. For each pair of (r, m), we perform 10 independent runs where in each run the sensing matrix is generated with i.i.d. standard Gaussian entries. Fig. 7 (a) shows the average NMSE with respect to m for different sparsity levels when there is no noise. We further introduce additive bounded noise to each measurement by letting λ_i be generated from $\sigma \cdot \mathcal{U}[-1, 1]$, and run 10 trials for each pair of (σ, m). Fig. 7 (b) shows the average NMSE when $k = 240$ for different noise levels by setting $\epsilon = \sigma m$ in (14).

VI. Conclusions and Future Work

We have investigated a general covariance estimation problem under a quadratic (rank-one) sampling model. This sampling model acts as an effective signal processing method for real-time data with limited processing power and memory at the data acquisition devices, and subsumes many sampling strategies where we can only obtain magnitude or energy samples. Three of the most popular covariance structures, i.e. sparsity, low rank, and jointly Toeplitz and low-rank structure, have been explored as well as sparse phase retrieval.

Our results indicate that covariance matrices under the above structural assumptions can be perfectly recovered from a small set of quadratic measurements and minimal storage, as long as the sensing vectors are i.i.d. drawn from sub-Gaussian distributions. The recovery can be achieved via efficient convex programming as soon as the memory complexity exceeds the fundamental sampling theoretic limit. We also observe universal recovery phenomena, in the sense that once the sensing vectors are chosen, all covariance matrices possessing the presumed structure can be recovered. Our results highlight the stability and robustness of the convex program in the presence of noise and imperfect structural assumptions. The performance guarantees for low-rank, sparse and jointly rank-one and sparse models are established via a novel notion of a mixed-norm restricted isometry property (RIP-ℓ_2/ℓ_1), which significantly simplifies the proof. Our contribution also includes a systematic approach to analyze Toeplitz low-rank structure, which relies on RIP-ℓ_2/ℓ_2 under near-isotropic and bounded operators.

Several future directions of interest are as follows.

- Another covariance structure of interest is an approximately *sparse inverse covariance matrix* rather than a sparse covariance matrix. In particular, when the signals are jointly Gaussian, the inverse covariance matrix encodes the conditional independence, which is often sparse. It remains to be seen whether the measurement scheme in (1) can be used to recover a sparse inverse covariance matrix.
- It will be interesting to explore whether more general types of sampling models satisfy RIP-ℓ_2/ℓ_1. For instance, when the sensing vectors do not have i.i.d. entries, more delicate mathematical tricks are necessary to establish RIP-ℓ_2/ℓ_1.
- In the case where RIP-ℓ_2/ℓ_1 is difficult to evaluate (e.g. the case with random Fourier sensing vectors), it would be interesting to develop an RIP-less theory in a similar flavor for linear measurement models [52].

Appendix A
Proof of Proposition 1

To prove Proposition 1, we will first derive an upper bound and a lower bound on $\mathbb{E}[|\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle|]$, and then apply the Bernstein-type inequality [56, Proposition 5.16] to establish the large deviation bound.

In order to derive an upper bound on $\mathbb{E}[|\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle|]$, the key step is to apply the Hanson-Wright inequality [57], [58], which characterizes the concentration of measure for quadratic forms in sub-Gaussian random variables. We adopt the version in [58] and repeat it below for completeness.

Lemma 7 (Hanson-Wright Inequality): Let $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ *be a random vector with independent components* X_i *which satisfy* $\mathbb{E}[X_i] = 0$ *and* $\|X_i\|_{\psi_2} \leq K$. *Let* $\boldsymbol{A}$ *be an* $n \times n$ *matrix. Then for any* $t > 0$,

$$\mathbb{P}\left\{\left|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}\right]\right| > t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4\|\boldsymbol{A}\|_{\mathrm{F}}^2}, \frac{t}{K^2\|\boldsymbol{A}\|}\right)\right]. \tag{49}$$

Remark 6: Here, $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm

$$\|X\|_{\psi_2} := \min_{p\geq 1} p^{-1/2}\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}.$$

Similarly, the sub-exponential norm $\|\cdot\|_{\psi_1}$ is defined as

$$\|X\|_{\psi_1} := \min_{p\geq 1} p^{-1}\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}.$$

See [56, Secs. 5.2.3 and 5.2.4] for an introduction.

Observe that $\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle$ can be written as a symmetric quadratic form in $2n$ i.i.d. sub-Gaussian random variables

$$\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle = \begin{bmatrix}\boldsymbol{a}_{2i-1}^\top & \boldsymbol{a}_{2i}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{X} & \\ & -\boldsymbol{X}\end{bmatrix}\begin{bmatrix}\boldsymbol{a}_{2i-1}\\ \boldsymbol{a}_{2i}\end{bmatrix}.$$

The Hanson-Wright inequality (49) then asserts that: there exists an absolute constant $c > 0$ such that for any matrix $\boldsymbol{X}$, $|\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle| \leq t$ with probability at least

$$1 - 2\exp\left[-c\min\left(\frac{t^2}{4K^4\|\boldsymbol{X}\|_{\mathrm{F}}^2}, \frac{t}{K^2\|\boldsymbol{X}\|}\right)\right].$$

This indicates that $\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle$ is a sub-exponential random variable [56, Sec. 5.2.4] satisfying

$$\mathbb{E}\left[|\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle|\right] \leq c_1\|\boldsymbol{X}\|_{\mathrm{F}} \tag{50}$$

for some positive constant c_1.

On the other hand, to derive a lower bound on $\mathbb{E}[|\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle|]$, we notice that for a random variable ξ, repeatedly applying the Cauchy-Schwartz inequality yields

$$\left(\mathbb{E}\left[\xi^2\right]\right)^2 \leq \mathbb{E}\left[|\xi|\right]\mathbb{E}\left[|\xi|^3\right] \leq \mathbb{E}\left[|\xi|\right]\sqrt{\mathbb{E}\left[\xi^2\right]\mathbb{E}\left[\xi^4\right]},$$

which further leads to

$$\mathbb{E}\left[|\xi|\right] \geq \sqrt{\frac{\left(\mathbb{E}\left[\xi^2\right]\right)^3}{\mathbb{E}\left[\xi^4\right]}}. \tag{51}$$

Let $\xi := \langle \boldsymbol{B}_i, \boldsymbol{X}\rangle$, of which the second moment can be expressed as

$$\mathbb{E}\left[\xi^2\right] = \mathbb{E}\left[|\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle|^2\right] = \left\langle \boldsymbol{X}, \mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i(\boldsymbol{X})\right]\right\rangle.$$

Simple algebraic manipulation yields

$$\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i(\boldsymbol{X})\right] = 4\boldsymbol{X} + 2(\mu_4 - 3)\,\mathrm{diag}(\boldsymbol{X}),$$

and hence

$$\mathbb{E}\left[\xi^2\right] = 4\|\boldsymbol{X}\|_{\mathrm{F}}^2 + 2(\mu_4 - 3)\sum_{i=1}^n |\boldsymbol{X}_{ii}|^2 \geq \min\{4, 2(\mu_4 - 1)\}\|\boldsymbol{X}\|_{\mathrm{F}}^2 = c_2\|\boldsymbol{X}\|_{\mathrm{F}}^2, \tag{52}$$

where $c_2 := \min\{4, 2(\mu_4 - 1)\}$. Furthermore, since $\xi := \langle \boldsymbol{B}_i, \boldsymbol{X}\rangle$ has been shown to be sub-exponential with sub-exponential norm $\Theta(\|\boldsymbol{X}\|_{\mathrm{F}})$, one can derive [56]

$$\mathbb{E}\left[\xi^4\right] = \left(4\|\xi\|_{\psi_1}\right)^4 \leq c_3\|\boldsymbol{X}\|_{\mathrm{F}}^4 \tag{53}$$

for some constant $c_7 > 0$. This taken collectively with (51) and (52) gives rise to

$$\mathbb{E}\left[|\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle|\right] \geq \sqrt{\frac{c_2^3\|\boldsymbol{X}\|_{\mathrm{F}}^6}{c_3\|\boldsymbol{X}\|_{\mathrm{F}}^4}} = c_4\|\boldsymbol{X}\|_{\mathrm{F}}$$

for some constant $c_4 > 0$.

Now, we are ready to characterize the concentration of $\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle$, which is a simple consequence of the following sub-exponential variant of Bernstein inequality.

Lemma 8 ([56, Proposition 5.16]): Let $X_1, \ldots, X_m$ *be independent sub-exponential random variables with* $\mathbb{E}[X_i] = 0$ *and* $K = \max_i\|X_i\|_{\psi_1}$. *Then for every* $t > 0$, *we have*

$$\mathbb{P}\left\{\frac{1}{m}\left|\sum_{i=1}^m X_i\right| \geq t\right\} \leq 2\exp\left[-cm\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right] \tag{54}$$

where c *is an absolute constant.*

Recall that it has been shown in (50) that the sub-exponential norm of $X_i := |\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle| - \mathbb{E}[|\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle|]$ satisfies $\|X_i\|_{\psi_1} \leq c'\|\boldsymbol{X}\|_{\mathrm{F}}$ for some universal constant c'. Therefore, Lemma 8 implies that for any $\epsilon > 0$, one has

$$\left|\frac{1}{m}\|\mathcal{B}(\boldsymbol{X})\|_1 - \frac{1}{m}\mathbb{E}\left[\|\mathcal{B}(\boldsymbol{X})\|_1\right]\right| \leq \epsilon\|\boldsymbol{X}\|_{\mathrm{F}}$$

with probability exceeding $1 - 2\exp(-cm\epsilon^2)$ for some absolute constant $c > 0$. This yields

$$\frac{1}{m}\|\mathcal{B}(\boldsymbol{X})\|_1 \leq \frac{1}{m}\mathbb{E}\left[\|\mathcal{B}(\boldsymbol{X})\|_1\right] + \epsilon\|\boldsymbol{X}\|_{\mathrm{F}} \leq (c_1 + \epsilon)\|\boldsymbol{X}\|_{\mathrm{F}}$$

and

$$\frac{1}{m}\|\mathcal{B}(\boldsymbol{X})\|_1 \geq \frac{1}{m}\mathbb{E}\left[\|\mathcal{B}(\boldsymbol{X})\|_1\right] - \epsilon\|\boldsymbol{X}\|_{\mathrm{F}} \geq (c_4 - \epsilon)\|\boldsymbol{X}\|_{\mathrm{F}}$$

with probability at least $1 - 2\exp(-cm\epsilon^2)$, where the constants c, c_1 and c_4 depend only on the sub-Gaussian norm of a_i. Renaming the universal constants establishes Proposition 1.

Appendix B
Proof of Theorem 5

The proof of Theorem 5 follows the entropy method introduced in [25] for compressed sensing and [27] for Pauli measurements. Note, however, that in our case, the measurement measurements do not form a basis, and are not even bounded. Our theorem extend the results in [27] (which focuses on Pauli basis) to general near-isotropic measurements.

Specifically, the RIP-ℓ_2/ℓ_2 constant can be bounded by

$$\begin{aligned}\delta_r &= \sup_{\|\boldsymbol{X}\|_{\mathrm{F}}\leq 1,\mathrm{rank}(\boldsymbol{X})\leq r}\left|\frac{1}{m}\sum_{i=1}^{m}|\langle\boldsymbol{B}_i,\boldsymbol{X}\rangle|^2-\|\boldsymbol{X}\|_{\mathrm{F}}^2\right|\\ &= \sup_{T\in\mathcal{M}_r^{\mathrm{t}},\boldsymbol{X}\in T,\|\boldsymbol{X}\|_{\mathrm{F}}\leq 1}\left|\left\langle\boldsymbol{X},\left(\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\mathcal{I}\right)\boldsymbol{X}\right\rangle\right| \quad (55)\\ &\leq \sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left(\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\mathcal{I}\right)\mathcal{P}_T\right\|\\ &\leq \sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\mathcal{B}_i^*\mathcal{B}_i\right)\right\}\mathcal{P}_T\right\|+\frac{c_5}{n}, \quad (56)\end{aligned}$$

where

$$\mathcal{M}_r^{\mathrm{t}}:=\left\{\text{tangent space w.r.t. }\boldsymbol{M}\mid\forall\boldsymbol{M}:\ \mathrm{rank}\left(\boldsymbol{M}\right)\leq r\right\}. \quad (57)$$

and hence (55) arises since the supremum is taken over all tangent space T associated with rank-r matrices. The last inequality (56) follows from the near-isotropic assumption of $\mathcal{B}_i$ (i.e. (30)).

The first step is to prove that $\mathbb{E}[\delta_r]\leq\epsilon$ for some small constant $\epsilon>0$. For sufficiently large n, it suffices to prove that

$$E:=\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]\right)\right\}\mathcal{P}_T\right\|\right]\leq\delta. \quad (58)$$

This can be established by a Gaussian process approach as follows.

Observe that $\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]\right)$ is a zero-mean operator, which can be reduced to symmetric operators via the symmetrization argument (see [53]). Specifically, let $\tilde{\mathcal{B}}_i$ be an independent copy of $\mathcal{B}_i$. Conditioning on $\mathcal{B}_i$ we have

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\frac{1}{m}\sum_{i=1}^{m}\tilde{\mathcal{B}}_i^*\tilde{\mathcal{B}}_i\,\middle|\,\mathcal{B}_i\ (1\leq i\leq m)\right]\\ &\quad=\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right].\end{aligned}$$

Since the function $f(\mathcal{X}):=\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\|\mathcal{P}_T\mathcal{X}\mathcal{P}_T\|$ is convex in $\mathcal{X}$, applying Jensen's inequality yields

$$\begin{aligned}&\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]\right)\right\}\mathcal{P}_T\right\|\\ &=\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathbb{E}\left[\mathcal{P}_T\left(\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\frac{1}{m}\sum_{i=1}^{m}\tilde{\mathcal{B}}_i^*\tilde{\mathcal{B}}_i\right)\mathcal{P}_T\,\middle|\,\mathcal{B}_i\right]\right\|\\ &\leq\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left(\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\frac{1}{m}\sum_{i=1}^{m}\tilde{\mathcal{B}}_i^*\tilde{\mathcal{B}}_i\right)\mathcal{P}_T\right\|\,\middle|\,\mathcal{B}_i\right].\end{aligned}$$

Undoing conditioning over $\mathcal{B}_i$ we get

$$\begin{aligned}&\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]\right)\right\}\mathcal{P}_T\right\|\right]\\ &\leq\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left(\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\frac{1}{m}\sum_{i=1}^{m}\tilde{\mathcal{B}}_i^*\tilde{\mathcal{B}}_i\right)\mathcal{P}_T\right\|\right]\\ &\leq 2\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|\right], \quad (59)\end{aligned}$$

where ϵ_i's are i.i.d. symmetric Bernoulli random variables. Moreover, if we generate a set of i.i.d. random variables $g_i\sim\mathcal{N}(0,1)$, then the conditional expectation obeys

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}|g_i|\,\epsilon_i\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\,\middle|\,\epsilon_i,\mathcal{B}_i\right]\\ &\quad=\sqrt{\frac{2}{\pi}}\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T.\end{aligned}$$

Similarly, by convexity of $f(\mathcal{X}):=\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\|\mathcal{P}_T\mathcal{X}\mathcal{P}_T\|$, one can obtain

$$\begin{aligned}&\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|\right]\\ &=\sqrt{\frac{\pi}{2}}\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}|g_i|\,\epsilon_i\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\,\middle|\,\epsilon_i,\mathcal{B}_i\right]\right\|\right]\\ &\leq\sqrt{\frac{\pi}{2}}\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\frac{1}{m}\sum_{i=1}^{m}g_i\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|\right]. \quad (60)\end{aligned}$$

Putting (59) and (60) together we obtain

$$\begin{aligned}&\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]\right)\right\}\mathcal{P}_T\right\|\right]\\ &\leq\sqrt{2\pi}\,\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\frac{1}{m}\sum_{i=1}^{m}g_i\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|\right]\\ &=\sqrt{2\pi}\,\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}},\boldsymbol{X}\in T,\|\boldsymbol{X}\|_{\mathrm{F}}=1}\left|\frac{1}{m}\sum_{i=1}^{m}g_i\left|\mathcal{B}_i\left(\boldsymbol{X}\right)\right|^2\right|\right]. \quad (61)\end{aligned}$$

It then boils down to characterizing the supremum of a Gaussian process.

We now prove the following lemma.

Lemma 9: Suppose that $g_i\sim\mathcal{N}(0,1)$ are i.i.d. random variables, and that $K\leq n^2$. Conditional on $\mathcal{B}_i$'s, we have

$$\begin{aligned}&\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left(\sum_{i=1}^{m}g_i\mathcal{B}_i^*\mathcal{B}_i\right)\mathcal{P}_T\right\|\,\middle|\,\mathcal{B}_i\ (1\leq i\leq m)\right]\\ &\quad\leq C_{14}\sqrt{r}K\log^3 n\sup_{T:T\in\mathcal{M}_r^{\mathrm{t}}}\sqrt{\left\|\sum_{i=1}^{m}\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|}. \quad (62)\end{aligned}$$

Proof: See Appendix I. □

Combining Lemma 9 with (61) and undoing the conditioning on $\mathcal{B}_i$'s yield

$$\begin{aligned}
&\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\mathcal{B}_i^*\mathcal{B}_i\right)\right\}\mathcal{P}_T\right\|\right]\\
&\leq \frac{C_{15}\sqrt{r}K\log^3 n}{m}\cdot\mathbb{E}\left[\sqrt{\sup_{T:T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\sum_{i=1}^{m}\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|}\right]\\
&\leq \frac{C_{15}\sqrt{r}K\log^3 n}{\sqrt{m}}\sqrt{\mathbb{E}\left[\sup_{T:T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\frac{1}{m}\sum_{i=1}^{m}\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|\right]}
\end{aligned}$$

for some universal constant $C_{15}>0$, where the last inequality follows from Jensen's inequality. Recall the definition of E in (58), then the above inequality implies

$$E\leq C_{15}\left(\frac{\sqrt{r}K\log^3 n}{\sqrt{m}}\right)\sqrt{E+1},$$

or more concretely,

$$\mathbb{E}\left[\delta_r\right]\leq E\leq 2C_{15}\frac{\sqrt{r}K\log^3 n}{\sqrt{m}}<1 \tag{63}$$

as soon as $m>\left(2C_{15}\sqrt{r}K\log^3 n\right)^2$.

Now that we have established that $\mathbb{E}\left[\delta_r\right]$ can be a small constant if $m>4C_{15}^2rK^2\log^6 n$, it remains to show that δ_r sharply concentrates around $\mathbb{E}\sigma_r$. To this end, consider the Banach space Υ of operators $\mathcal{H}:\mathbb{R}^{n\times n}\mapsto\mathbb{R}^{n\times n}$ equipped with the norm

$$\|\mathcal{H}\|_{\Upsilon}:=\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\|\mathcal{P}_T\mathcal{H}\mathcal{P}_T\|.$$

Let ε_i's be i.i.d. symmetric Bernoulli variables, then the symmetrization trick (see [53]) yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]\right\|_{\Upsilon}\right]\\
&\leq\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\mathcal{B}_i^*\mathcal{B}_i\right\|_{\Upsilon}\right]\\
&\leq 2\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]\right\|_{\Upsilon},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\left\{\left\|\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]\right\|_{\Upsilon}\right.\\
&\qquad\left.>2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\mathcal{B}_i^*\mathcal{B}_i-\mathbb{E}\left[\mathcal{B}_i^*\mathcal{B}_i\right]\right\|_{\Upsilon}\right]+u\right\}\\
&\leq\mathbb{P}\left\{\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i^*\mathcal{B}_i-\tilde{\mathcal{B}}_i^*\tilde{\mathcal{B}}_i\right)\right\|_{\Upsilon}>u\right\}\\
&\leq 2\mathbb{P}\left\{\left\|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\mathcal{B}_i^*\mathcal{B}_i\right\|_{\Upsilon}>\frac{u}{2}\right\},
\end{aligned}$$

where $\tilde{\mathcal{B}}_i$ is an independent copy of $\mathcal{B}_i$. Note that $\varepsilon_i\mathcal{B}_i^*\mathcal{B}_i$'s are i.i.d. zero-mean random operators.

In addition, for any $1\leq i\leq m$, we know that

$$\begin{aligned}
\left\|\epsilon_i\mathcal{B}_i^*\mathcal{B}_i\right\|_{\Upsilon}&=\max_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|\\
&=\max_{T\in\mathcal{M}_r^{\mathrm{t}},\|\boldsymbol{X}\|_{\mathrm{F}}=1}\left|\langle\boldsymbol{B}_i,\mathcal{P}_T(\boldsymbol{X})\rangle\right|^2\\
&\leq\max_{T\in\mathcal{M}_r^{\mathrm{t}},\|\boldsymbol{X}\|_{\mathrm{F}}=1}\|\boldsymbol{B}_i\|^2\|\mathcal{P}_T(\boldsymbol{X})\|_*^2\leq K^2r.
\end{aligned}$$

[25, Th. 3.10] asserts that there is a universal constant $C_{12}>0$ such that

$$\begin{aligned}
&\mathbb{P}\left\{\left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\mathcal{B}_i^*\mathcal{B}_i\right\|_{\Upsilon}>8q\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\mathcal{B}_i^*\mathcal{B}_i\right\|_{\Upsilon}+\frac{2K^2r}{m}l+t\right\}\\
&\leq\left(\frac{C_{12}}{q}\right)^l+2\exp\left(-\frac{t^2}{256q\left(\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\mathcal{B}_i^*\mathcal{B}_i\right\|_{\Upsilon}\right]\right)^2}\right).
\end{aligned}$$

If we take $q=2C_{12}$, $l=C_{13}\log n$ and $t=C_{14}\sqrt{\log n}\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\mathcal{B}_i^*\mathcal{B}_i\right\|_{\Upsilon}\right]$, then for sufficiently large C_{13} and C_{14}, there exists an absolute constant $C_{20}>0$ such that if $m>C_{20}rK^2\log^7 n$, then for any small positive constant δ we have

$$\begin{aligned}
\left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\mathcal{B}_i^*\mathcal{B}_i\right\|_{\Upsilon}&<C_{15}\sqrt{\log n}\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i\mathcal{B}_i^*\mathcal{B}_i\right\|_{\Upsilon}\right]\\
&<\delta
\end{aligned}$$

with probability exceeding $1-n^{-2}$.

Now that we have ensured a small RIP-ℓ_2/ℓ_2 constant, repeating the argument as in [44] and [49] implies

$$\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_{\mathrm{F}}\leq C_2\frac{\epsilon_2}{\sqrt{m}} \tag{64}$$

for all $\boldsymbol{\Sigma}$ of rank at most r. This concludes the proof.

APPENDIX C
PROOF OF LEMMA 1

We first introduce a few mathematical notations before proceeding to the proof. Let the singular value decomposition of a rank-r matrix $\boldsymbol{\Sigma}$ be $\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, then the tangent space T at the point $\boldsymbol{\Sigma}$ is defined as $T:=\left\{\boldsymbol{U}\boldsymbol{M}_1+\boldsymbol{M}_2\boldsymbol{V}^\top\mid\boldsymbol{M}_1\in\mathbb{R}^{r\times n},\boldsymbol{M}_2\in\mathbb{R}^{n\times r}\right\}$. We denote by $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$ the orthogonal projection onto T and its orthogonal complement, respectively. For notational simplicity, we denote $\boldsymbol{H}_T:=\mathcal{P}_T(\boldsymbol{H})$ and $\boldsymbol{H}_{T^\perp}:=\boldsymbol{H}-\mathcal{P}_T(\boldsymbol{H})$ for any matrices $\boldsymbol{H}\in\mathbb{R}^{n\times n}$. The proof is inspired by the techniques introduced for operators satisfying RIP-ℓ_2/ℓ_2 [44], [49].

Write $\boldsymbol{\Sigma}:=\boldsymbol{\Sigma}_r+\boldsymbol{\Sigma}_{\mathrm{c}}$, where $\boldsymbol{\Sigma}_r$ represents the best rank-r approximation of $\boldsymbol{\Sigma}$. Denote by T the tangent space with respect to $\boldsymbol{\Sigma}_r$. Suppose that the solution to (7) is given by $\hat{\boldsymbol{\Sigma}}=\boldsymbol{\Sigma}+\boldsymbol{H}$ for some matrix $\boldsymbol{H}$. The optimality of $\hat{\boldsymbol{\Sigma}}$ yields

$$\begin{aligned}
0&\geq\|\boldsymbol{\Sigma}+\boldsymbol{H}\|_*-\|\boldsymbol{\Sigma}\|_*\\
&\geq\|\boldsymbol{\Sigma}_r+\boldsymbol{H}\|_*-\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_*-\|\boldsymbol{\Sigma}\|_*\\
&\geq\left\|\boldsymbol{\Sigma}_r+\boldsymbol{H}_{T^\perp}\right\|_*-\|\boldsymbol{H}_T\|_*-\|\boldsymbol{\Sigma}_r\|_*-2\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_*\\
&=\|\boldsymbol{\Sigma}_r\|_*+\left\|\boldsymbol{H}_{T^\perp}\right\|_*-\|\boldsymbol{H}_T\|_*-\|\boldsymbol{\Sigma}_r\|_*-2\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_*,
\end{aligned}$$

which leads to

$$\left\|\boldsymbol{H}_{T^\perp}\right\|_*\leq\|\boldsymbol{H}_T\|_*+2\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_*. \tag{65}$$

We then divide $\boldsymbol{H}_{T^\perp}$ into $M = \left\lceil \frac{n-r}{K_1} \right\rceil$ orthogonal matrices $\boldsymbol{H}_1$, $\boldsymbol{H}_2$, $\cdots$, $\boldsymbol{H}_M$ satisfying the following: (i) the largest singular value of $\boldsymbol{H}_{i+1}$ does not exceed the smallest non-zero singular value of $\boldsymbol{H}_i$, and (ii)

$$\left\| \boldsymbol{H}_{T^\perp} \right\|_* = \sum_{i=1}^{M} \| \boldsymbol{H}_i \|_* \tag{66}$$

and $\operatorname{rank}(\boldsymbol{H}_i) = K_1$ for $1 \le i \le M-1$. Along with the bound (65), this yields that

$$\begin{aligned}\sum_{i\ge 2} \| \boldsymbol{H}_i \|_{\mathrm{F}} &\le \frac{1}{\sqrt{K_1}} \sum_{i\ge 2} \| \boldsymbol{H}_{i-1} \|_* \le \frac{1}{\sqrt{K_1}} \left\| \boldsymbol{H}_{T^\perp} \right\|_* \\ &\le \frac{1}{\sqrt{K_1}} \left(\| \boldsymbol{H}_T \|_* + 2 \| \boldsymbol{\Sigma}_{\mathrm{c}} \|_* \right). \end{aligned} \tag{67}$$

Since the feasibility constraint requires $\| \mathcal{A}(\boldsymbol{\Sigma}) - \boldsymbol{y} \|_1 \le \epsilon_1$, we have $\| \mathcal{A}(\boldsymbol{H}) \|_1 \le \| \mathcal{A}(\boldsymbol{\Sigma}) - \boldsymbol{y} \|_1 + \left\| \mathcal{A}\left(\hat{\boldsymbol{\Sigma}} \right) - \boldsymbol{y} \right\|_1 \le 2\epsilon_1$, and then following from the definition $\boldsymbol{B}_i = \boldsymbol{A}_{2i-1} - \boldsymbol{A}_{2i}$ that

$$\frac{1}{m} \| \mathcal{B}(\boldsymbol{H}) \|_1 \le \frac{1}{m} \| \mathcal{A}(\boldsymbol{H}) \|_1 \le \frac{2\epsilon_1}{m},$$

yielding

$$\begin{aligned}\frac{2\epsilon_1}{m} &\ge \frac{1}{m} \| \mathcal{B}(\boldsymbol{H}) \|_1 \\ &\ge \frac{1}{m} \| \mathcal{B}(\boldsymbol{H}_T + \boldsymbol{H}_1) \|_1 - \sum_{i\ge 2} \frac{1}{m} \| \mathcal{B}(\boldsymbol{H}_i) \|_1 \\ &\ge \left(1 - \delta^{\mathrm{lb}}_{2r+K_1}\right) \| \boldsymbol{H}_T + \boldsymbol{H}_1 \|_{\mathrm{F}} - \left(1 + \delta^{\mathrm{ub}}_{K_1}\right) \sum_{i\ge 2} \| \boldsymbol{H}_i \|_{\mathrm{F}} \\ &\ge \frac{\left(1 - \delta^{\mathrm{lb}}_{2r+K_1}\right)}{\sqrt{2}} \left(\| \boldsymbol{H}_T \|_{\mathrm{F}} + \| \boldsymbol{H}_1 \|_{\mathrm{F}} \right) \\ &\quad - \frac{\left(1 + \delta^{\mathrm{ub}}_{K_1}\right)}{\sqrt{K_1}} \left(\| \boldsymbol{H}_T \|_* + 2 \| \boldsymbol{\Sigma}_{\mathrm{c}} \|_* \right). \end{aligned}$$

By reorganizing the terms and using the fact that $\| \boldsymbol{H}_T \|_* \le \sqrt{2r} \| \boldsymbol{H}_T \|_{\mathrm{F}}$, one can derive

$$\begin{aligned}&\left[\frac{\left(1 - \delta^{\mathrm{lb}}_{2r+K_1}\right)}{\sqrt{2}} - \frac{\left(1 + \delta^{\mathrm{ub}}_{K_1}\right)\sqrt{2r}}{\sqrt{K_1}} \right] \| \boldsymbol{H}_T \|_{\mathrm{F}} \\ &\quad + \frac{\left(1 - \delta^{\mathrm{lb}}_{2r+K_1}\right)}{\sqrt{2}} \| \boldsymbol{H}_1 \|_{\mathrm{F}} \le \frac{2\left(1 + \delta^{\mathrm{ub}}_{K_1}\right)}{\sqrt{K_1}} \| \boldsymbol{\Sigma}_{\mathrm{c}} \|_* + \frac{2\epsilon_1}{m}. \end{aligned} \tag{68}$$

The bound (68) allows us to see that if $\frac{1-\delta^{\mathrm{lb}}_{2r+K_1}}{\sqrt{2}} - \left(1 + \delta^{\mathrm{ub}}_{K_1}\right)\sqrt{\frac{2r}{K_1}} \ge \beta_1 > 0$ for some absolute constant β_1, then one has

$$\| \boldsymbol{H}_T \|_{\mathrm{F}} + \| \boldsymbol{H}_1 \|_{\mathrm{F}} \le \frac{2}{\beta_1} \left(\frac{\left(1 + \delta^{\mathrm{ub}}_{K_1}\right)}{\sqrt{K_1}} \| \boldsymbol{\Sigma}_{\mathrm{c}} \|_* + \frac{\epsilon_1}{m} \right). \tag{69}$$

On the other hand, (67) allows us to bound

$$\begin{aligned}\sum_{i\ge 2} \| \boldsymbol{H}_i \|_{\mathrm{F}} &\le \frac{1}{\sqrt{K_1}} \left(\| \boldsymbol{H}_T \|_* + 2 \| \boldsymbol{\Sigma}_{\mathrm{c}} \|_* \right) \\ &\le \sqrt{\frac{2r}{K_1}} \| \boldsymbol{H}_T \|_{\mathrm{F}} + \frac{2}{\sqrt{K_1}} \| \boldsymbol{\Sigma}_{\mathrm{c}} \|_*. \end{aligned} \tag{70}$$

Putting the above computation together establishes

$$\begin{aligned}\| \boldsymbol{H} \|_{\mathrm{F}} &\le \| \boldsymbol{H}_T \|_{\mathrm{F}} + \| \boldsymbol{H}_1 \|_{\mathrm{F}} + \sum_{i\ge 2} \| \boldsymbol{H}_i \|_{\mathrm{F}} \\ &\le \left(1 + \sqrt{\frac{2r}{K_1}} \right) \left(\| \boldsymbol{H}_T \|_{\mathrm{F}} + \| \boldsymbol{H}_1 \|_{\mathrm{F}} \right) + \frac{2}{\sqrt{K_1}} \| \boldsymbol{\Sigma}_{\mathrm{c}} \|_* \\ &\le \left(\frac{C_1}{\beta_1} + C_3 \right) \frac{\| \boldsymbol{\Sigma}_{\mathrm{c}} \|_*}{\sqrt{K_1}} + \frac{C_2}{\beta_1} \cdot \frac{\epsilon_1}{m} \end{aligned}$$

for some positive universal constants C_1, C_2 and C_3.

APPENDIX D
PROOF OF LEMMA 2

For an index set Ω, let $\mathcal{P}_\Omega$ as the orthogonal projection onto the index set Ω. We denote $\boldsymbol{H}_\Omega$ as the matrix supported on $\boldsymbol{H}_\Omega = \mathcal{P}_\Omega(\boldsymbol{H})$ and $\boldsymbol{H}_{\Omega^\perp}$ as the projection onto the complement support set $\Omega^\perp$. Write $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \boldsymbol{H}$, and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\Omega_0} + \boldsymbol{\Sigma}_{\Omega_0^c}$, where Ω_0 denotes the support of the k largest entries of $\boldsymbol{\Sigma}$. The feasibility constraint yields

$$\frac{1}{m} \| \mathcal{B}(\boldsymbol{H}) \|_1 \le \frac{2}{m} \left\| \mathcal{A}(\hat{\boldsymbol{\Sigma}}) - \mathcal{A}(\boldsymbol{\Sigma}) \right\|_1 \le \frac{2\epsilon_1}{m}.$$

The triangle inequality of ℓ_1 norm gives

$$\| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \|_1 \le \| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{\Omega_0} \|_1 + \| \boldsymbol{\Sigma}_{\Omega_0^{\mathrm{c}}} \|_1$$

Decompose $\boldsymbol{H}_{\Omega_0^{\mathrm{c}}}$ into a collection of M_2 matrices $\boldsymbol{H}_{\Omega_1}$, $\boldsymbol{H}_{\Omega_2}$, $\ldots$, $\boldsymbol{H}_{\Omega_{M_2}}$, where $\left\| \boldsymbol{H}_{\Omega_i} \right\|_0 = K_2$ for all $1 \le i < M_2$, $\boldsymbol{H}_{\Omega_1}$ consists of the K_2 largest entries of $\boldsymbol{H}_{\Omega_0^c}$, $\boldsymbol{H}_{\Omega_2}$ consists of the K_2 largest entries of $\boldsymbol{H}_{(\Omega_0 \cup \Omega_1)^c}$, and so on. A similar argument as in [46] implies

$$\sum_{i\ge 2} \| \boldsymbol{H}_{\Omega_i} \|_{\mathrm{F}} \le \frac{1}{\sqrt{K_2}} \sum_{i\ge 1} \| \boldsymbol{H}_{\Omega_i} \|_1 = \frac{1}{\sqrt{K_2}} \| \boldsymbol{H}_{\Omega_0^{\mathrm{c}}} \|_1. \tag{71}$$

The optimality of $\hat{\boldsymbol{\Sigma}}$ yields

$$\begin{aligned}\| \boldsymbol{\Sigma} \|_1 &\ge \| \boldsymbol{\Sigma} + \boldsymbol{H} \|_1 = \| \boldsymbol{\Sigma}_{\Omega_0} + \boldsymbol{H} \|_1 - \| \boldsymbol{\Sigma}_{\Omega_0^{\mathrm{c}}} \|_1 \\ &\ge \| \boldsymbol{\Sigma}_{\Omega_0} \|_1 + \| \boldsymbol{H}_{\Omega_0^{\mathrm{c}}} \|_1 - \| \boldsymbol{H}_{\Omega_0} \|_1 - \| \boldsymbol{\Sigma}_{\Omega^c} \|_1, \end{aligned}$$

which gives

$$\| \boldsymbol{H}_{\Omega_0^{\mathrm{c}}} \|_1 \le \| \boldsymbol{H}_{\Omega_0} \|_1 + 2 \| \boldsymbol{\Sigma}_{\Omega^c} \|_1.$$

Combining the above bound and (71) leads to

$$\begin{aligned}\sum_{i\ge 2} \| \boldsymbol{H}_{\Omega_i} \|_{\mathrm{F}} &\le \frac{1}{\sqrt{K_2}} \left(\| \boldsymbol{H}_{\Omega_0} \|_1 + 2 \| \boldsymbol{\Sigma}_{\Omega^c} \|_1 \right) \\ &\le \frac{1}{\sqrt{K_2}} \left(\sqrt{k} \| \boldsymbol{H}_{\Omega_0} \|_{\mathrm{F}} + 2 \| \boldsymbol{\Sigma}_{\Omega^c} \|_1 \right). \end{aligned} \tag{72}$$

It then follows that

$$\begin{aligned}\frac{2\epsilon_1}{m} &\ge \frac{1}{m} \| \mathcal{B}(\boldsymbol{H}) \|_1 \\ &\ge \frac{1}{m} \left\| \mathcal{B}\left(\boldsymbol{H}_{\Omega_0} + \boldsymbol{H}_{\Omega_1} \right) \right\|_1 - \frac{1}{m} \sum_{i\ge 2} \left\| \mathcal{B}\left(\boldsymbol{H}_{\Omega_i} \right) \right\|_1 \\ &\ge \left(1 - \gamma^{\mathrm{lb}}_{k+K_2}\right) \left\| \boldsymbol{H}_{\Omega_0} + \boldsymbol{H}_{\Omega_1} \right\|_{\mathrm{F}} - \left(1 + \gamma^{\mathrm{ub}}_{K_2}\right) \sum_{i\ge 2} \left\| \boldsymbol{H}_{\Omega_i} \right\|_{\mathrm{F}} \\ &\ge \frac{\left(1 - \gamma^{\mathrm{lb}}_{k+K_2}\right)}{\sqrt{2}} \left(\left\| \boldsymbol{H}_{\Omega_0} \right\|_{\mathrm{F}} + \left\| \boldsymbol{H}_{\Omega_1} \right\|_{\mathrm{F}} \right) \\ &\quad - \frac{\left(1 + \gamma^{\mathrm{ub}}_{K_2}\right)}{\sqrt{K_2}} \left(\sqrt{k} \left\| \boldsymbol{H}_{\Omega_0} \right\|_{\mathrm{F}} + 2 \| \boldsymbol{\Sigma}_{\Omega^c} \|_1 \right). \end{aligned}$$

Reorganizing the above equation yields

$$\left[\frac{(1-\gamma_{k+K_2}^{\mathrm{lb}})}{\sqrt{2}}-\frac{\left(1+\gamma_{K_2}^{\mathrm{ub}}\right)\sqrt{k}}{\sqrt{K_2}}\right]\|\boldsymbol{H}_{\Omega_0}\|_{\mathrm{F}} + \frac{(1-\gamma_{k+K_2}^{\mathrm{lb}})}{\sqrt{2}}\|\boldsymbol{H}_{\Omega_1}\|_{\mathrm{F}} \leq \frac{2\left(1+\gamma_{K_2}^{\mathrm{ub}}\right)}{\sqrt{K_2}}\|\boldsymbol{\Sigma}_{\Omega^c}\|_1+\frac{2\epsilon_1}{m}.$$

Recalling Assumption (26), one has

$$\|\boldsymbol{H}_{\Omega_0}\|_{\mathrm{F}}+\|\boldsymbol{H}_{\Omega_1}\|_{\mathrm{F}} \leq \frac{2}{\beta_2}\left[\frac{\left(1+\gamma_{K_2}^{\mathrm{ub}}\right)}{\sqrt{K_2}}\|\boldsymbol{\Sigma}_{\Omega^c}\|_1+\frac{\epsilon_1}{m}\right].$$

This along with (72) gives

$$\|\boldsymbol{H}\|_{\mathrm{F}} \leq \left(\frac{C_1}{\beta_2}+C_3\right)\frac{\|\boldsymbol{\Sigma}_{\Omega^c}\|_1}{\sqrt{K_2}}+\frac{C_2}{\beta_2}\frac{\epsilon_1}{m}$$

for some universal constants C_1, C_2 and C_3, as claimed.

APPENDIX E
PROOF OF LEMMA 3

Before proceeding to the proof, we introduce a few notations for convenience of presentation. Let $\boldsymbol{X} := \boldsymbol{x}\boldsymbol{x}^\top$, $\boldsymbol{X}_\Omega := \boldsymbol{x}_\Omega\boldsymbol{x}_\Omega^\top$ and $\boldsymbol{X}_{\mathrm{c}} := \boldsymbol{X}-\boldsymbol{X}_\Omega$, where $\boldsymbol{x}_\Omega$ denotes the best k-term approximation of $\boldsymbol{x}$. We set $\boldsymbol{u} := \frac{1}{\|\boldsymbol{x}_\Omega\|_2}\boldsymbol{x}_\Omega$, and hence the tangent space T with respect to $\boldsymbol{X}_\Omega$ and its orthogonal complement $T^\perp$ are characterized by

$$T := \left\{\boldsymbol{u}\boldsymbol{z}^\top+\boldsymbol{z}\boldsymbol{u}^\top \mid \boldsymbol{z}\in\mathbb{R}^n\right\},$$
$$T^\perp := \left\{\left(\boldsymbol{I}-\boldsymbol{u}\boldsymbol{u}^\top\right)\boldsymbol{M}\left(\boldsymbol{I}-\boldsymbol{u}\boldsymbol{u}^\top\right) \mid \boldsymbol{M}\in\mathbb{R}^{n\times n}\right\}.$$

We adopt the notations introduced in [24] as follows: let Ω denote the support of $\boldsymbol{X}_\Omega$, and decompose the entire matrix space into the direct sum of 3 subspaces as

$$(T\cap\Omega)\oplus\left(T^\perp\cap\Omega\right)\oplus\left(\Omega^\perp\right). \tag{73}$$

In fact, one can verify that

$$T\cap\Omega=\left\{\boldsymbol{u}\boldsymbol{z}^\top+\boldsymbol{z}\boldsymbol{u}^\top \mid \boldsymbol{z}_{\Omega^c}=\boldsymbol{0}\right\},$$

and that both the column and row spaces of $T^\perp\cap\Omega$ can be spanned by a set of $k-1$ orthonormal vectors that are supported on Ω and orthogonal to $\boldsymbol{u}$. As pointed out by [24], T and Ω are compatible in the sense that

$$\mathcal{P}_T\mathcal{P}_\Omega=\mathcal{P}_\Omega\mathcal{P}_T=\mathcal{P}_{T\cap\Omega}. \tag{74}$$

In the following, we will use $\delta_{r,l}^{\mathrm{lb}}$ and $\delta_{r,l}^{\mathrm{ub}}$ to represent $\delta_{k,r,l}^{\mathrm{lb}}$ and $\delta_{k,r,l}^{\mathrm{ub}}$ for brevity, whenever the value of k is clear from the context.

Suppose that $\hat{\boldsymbol{X}}=\boldsymbol{x}\boldsymbol{x}^\top+\boldsymbol{H}$ is the solution to (16). Then for any $\boldsymbol{W}\in T^\perp$ and $\boldsymbol{Y}\in\Omega^\perp$ satisfying $\|\boldsymbol{W}\|\leq 1$ and $\|\boldsymbol{Y}\|_\infty\leq 1$, the matrix $\boldsymbol{u}\boldsymbol{u}^\top+\boldsymbol{W}+\lambda\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top+\lambda\boldsymbol{Y}$ forms a subgradient of the function $\|\cdot\|_*+\lambda\|\cdot\|_1$ at point $\boldsymbol{X}_\Omega$. If we pick $\boldsymbol{W}$ and $\boldsymbol{Y}$ such that $\boldsymbol{Y}=\mathrm{sgn}\left(\boldsymbol{H}_{\Omega^\perp}\right)$ and $\langle\boldsymbol{W},\boldsymbol{H}\rangle=\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_*$, then

$$0 \geq \|\boldsymbol{X}+\boldsymbol{H}\|_*+\lambda\|\boldsymbol{X}+\boldsymbol{H}\|_1-\|\boldsymbol{X}\|_*-\lambda\|\boldsymbol{X}\|_1 \tag{75}$$
$$\begin{aligned}\geq{}& \|\boldsymbol{X}_\Omega+\boldsymbol{H}\|_*-\|\boldsymbol{X}_{\mathrm{c}}\|_*+\lambda\|\boldsymbol{X}_\Omega+\boldsymbol{H}\|_1-\lambda\|\boldsymbol{X}_{\mathrm{c}}\|_1\\ &-\|\boldsymbol{X}_\Omega\|_*-\|\boldsymbol{X}_{\mathrm{c}}\|_*-\lambda\|\boldsymbol{X}_\Omega\|_1-\lambda\|\boldsymbol{X}_{\mathrm{c}}\|_1\end{aligned} \tag{76}$$
$$\begin{aligned}\geq{}& \left\langle\boldsymbol{u}\boldsymbol{u}^\top+\boldsymbol{W},\boldsymbol{H}\right\rangle+\left\langle\lambda\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top+\lambda\boldsymbol{Y},\boldsymbol{H}\right\rangle\\ &-2\|\boldsymbol{X}_{\mathrm{c}}\|_*-2\lambda\|\boldsymbol{X}_{\mathrm{c}}\|_1\end{aligned} \tag{77}$$
$$\begin{aligned}={}& \left\langle\boldsymbol{u}\boldsymbol{u}^\top,\boldsymbol{H}_T\right\rangle+\lambda\left\langle\mathcal{P}_T\left(\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right),\boldsymbol{H}_T\right\rangle\\ &+\lambda\left\langle\mathcal{P}_{T^\perp}\left(\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right),\boldsymbol{H}_{T^\perp}\right\rangle\\ &+\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_*+\lambda\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1-2\|\boldsymbol{X}_{\mathrm{c}}\|_*-2\lambda\|\boldsymbol{X}_{\mathrm{c}}\|_1\\ \geq{}& \left\langle\boldsymbol{u}\boldsymbol{u}^\top+\lambda\mathcal{P}_T\left(\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right),\boldsymbol{H}_{T\cap\Omega}\right\rangle+\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_*\\ &+\lambda\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1-2\|\boldsymbol{X}_{\mathrm{c}}\|_*-2\lambda\|\boldsymbol{X}_{\mathrm{c}}\|_1,\end{aligned} \tag{78}$$

where (75) follows from the optimality of $\hat{\boldsymbol{X}}$, (76) follows from the definitions of $\boldsymbol{X}_\Omega$ and $\boldsymbol{X}_{\mathrm{c}}$ and the triangle inequality, (77) follows from the definition of subgradient. Finally, (78) follows from the following two facts:

(i) $\boldsymbol{H}_{T^\perp}\succeq 0$, a consequence of the feasibility constraint of (16). This further gives

$$\begin{aligned}&\left\langle\mathcal{P}_{T^\perp}\left(\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right),\boldsymbol{H}_{T^\perp}\right\rangle\\ &=\left\langle\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top,\boldsymbol{H}_{T^\perp}\right\rangle\\ &=\mathrm{sign}(\boldsymbol{u})^\top\boldsymbol{H}_{T^\perp}\mathrm{sign}(\boldsymbol{u})\geq 0.\end{aligned}$$

(ii) It follows from (74) and the fact $\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\in\Omega$ that

$$\begin{aligned}&\left\langle\mathcal{P}_T\left(\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right),\boldsymbol{H}_T\right\rangle\\ &=\left\langle\mathcal{P}_{T\cap\Omega}\left(\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right),\boldsymbol{H}_{T\cap\Omega}\right\rangle.\end{aligned} \tag{79}$$

Since any matrix in T has rank at most 2, one can bound

$$\left\|\mathcal{P}_T\left(\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right)\right\|_*^2\leq 2\left\|\mathcal{P}_T\left(\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right)\right\|_{\mathrm{F}}^2$$
$$\leq 4\left\|\boldsymbol{u}\boldsymbol{u}^\top\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right\|_{\mathrm{F}}^2 \tag{80}$$
$$\begin{aligned}&=4\left|\langle\boldsymbol{u},\mathrm{sign}(\boldsymbol{u})\rangle\right|^2\left\|\boldsymbol{u}\cdot\mathrm{sign}(\boldsymbol{u})^\top\right\|_{\mathrm{F}}^2\\ &=4\left|\langle\boldsymbol{u},\mathrm{sign}(\boldsymbol{u})\rangle\right|^2\|\mathrm{sign}(\boldsymbol{u})\|_{\mathrm{F}}^2\\ &\leq 4k\|\boldsymbol{u}\|_1^2\|\mathrm{sign}(\boldsymbol{u})\|_\infty^2\leq 4k\|\boldsymbol{u}\|_1^2\leq\frac{4}{\lambda^2},\end{aligned} \tag{81}$$

where (80) follows from the definition of $\mathcal{P}_T$ that

$$\begin{aligned}&\left\|\mathcal{P}_T\left(\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right)\right\|_{\mathrm{F}}^2\\ &=\left\|\boldsymbol{u}\boldsymbol{u}^\top\mathrm{sign}(\boldsymbol{u})\mathrm{sign}(\boldsymbol{u})^\top+(\boldsymbol{I}-\boldsymbol{u}\boldsymbol{u}^\top)\mathrm{sign}(\boldsymbol{u})\mathrm{sign}(\boldsymbol{u})^\top\boldsymbol{u}\boldsymbol{u}^\top\right\|_{\mathrm{F}}^2\\ &=\left\|\boldsymbol{u}\boldsymbol{u}^\top\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right\|_{\mathrm{F}}^2\\ &\quad+\left\|\left(\boldsymbol{I}-\boldsymbol{u}\boldsymbol{u}^\top\right)\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\boldsymbol{u}\boldsymbol{u}^\top\right\|_{\mathrm{F}}^2\\ &\leq\left\|\boldsymbol{u}\boldsymbol{u}^\top\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right\|_{\mathrm{F}}^2+\left\|\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\boldsymbol{u}\boldsymbol{u}^\top\right\|_{\mathrm{F}}^2\\ &=2\left\|\boldsymbol{u}\boldsymbol{u}^\top\mathrm{sign}(\boldsymbol{u})\,\mathrm{sign}(\boldsymbol{u})^\top\right\|_{\mathrm{F}}^2,\end{aligned}$$

and (81) arises from the assumption on λ. Combining (81) with (78) yields

$$\begin{aligned}
&\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_* + \lambda\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1 \\
&\leq -\left\langle \boldsymbol{u}\boldsymbol{u}^\top, \boldsymbol{H}_{T\cap\Omega}\right\rangle - \lambda\left\langle \mathcal{P}_T\left(\operatorname{sign}(\boldsymbol{u})\operatorname{sign}(\boldsymbol{u})^\top\right), \boldsymbol{H}_{T\cap\Omega}\right\rangle \\
&\quad + 2\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_* + 2\lambda\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_1 \\
&\leq \left|\boldsymbol{u}^\top\boldsymbol{H}_{T\cap\Omega}\boldsymbol{u}\right| + \lambda\left\|\mathcal{P}_T\left(\operatorname{sign}(\boldsymbol{u})\operatorname{sign}(\boldsymbol{u})^\top\right)\right\|_* \cdot \left\|\boldsymbol{H}_{T\cap\Omega}\right\| \\
&\quad + 2\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_* + 2\lambda\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_1 \\
&\leq 3\left\|\boldsymbol{H}_{T\cap\Omega}\right\| + 2\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_* + 2\lambda\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_1, \qquad (82)
\end{aligned}$$

where (82) results from $\|\boldsymbol{u}\|_2 = 1$ and (81).

Divide $\boldsymbol{H}_{T^\perp\cap\Omega}$ into $M_1 := \left\lceil \frac{k-2}{K_1} \right\rceil$ orthogonal matrices $\boldsymbol{H}^{(1)}_{T^\perp\cap\Omega}, \boldsymbol{H}^{(2)}_{T^\perp\cap\Omega}, \cdots, \boldsymbol{H}^{(M_1)}_{T^\perp\cap\Omega} \in T^\perp\cap\Omega$ satisfying the following properties: (i) the largest singular value of $\boldsymbol{H}^{(i+1)}_{T^\perp\cap\Omega}$ does not exceed the smallest non-zero singular value of $\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}$, and (ii)

$$\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_* = \sum_{i=1}^{M}\left\|\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right\|_*,$$
$$\operatorname{rank}\left(\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right) = K_1 \quad (1 \leq i \leq M_1 - 1).$$

In the meantime, divide $\boldsymbol{H}_{\Omega^\perp}$ into $M_2 = \left\lceil \frac{n^2-k^2}{K_2} \right\rceil$ orthogonal matrices $\boldsymbol{H}^{(1)}_{\Omega^\perp}, \boldsymbol{H}^{(2)}_{\Omega^\perp}, \cdots, \boldsymbol{H}^{(M_2)}_{\Omega^\perp} \in \Omega^\perp$ of *non-overlapping* support such that (i) the largest entry magnitude of $\boldsymbol{H}^{(i+1)}_{\Omega^\perp}$ does not exceed the magnitude of the smallest non-zero entry of $\boldsymbol{H}^{(i)}_{\Omega^\perp}$, and (ii)

$$\left\|\boldsymbol{H}^{(i)}_{\Omega^\perp}\right\|_0 = K_2 \ (1 \leq i \leq M_2 - 1).$$

This decomposition gives rise to the following bound

$$\sum_{i=2}^{M_1}\left\|\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right\|_{\mathrm{F}} \leq \sum_{i=2}^{M_1}\frac{1}{\sqrt{K_1}}\left\|\boldsymbol{H}^{(i-1)}_{T^\perp\cap\Omega}\right\|_* \leq \frac{1}{\sqrt{K_1}}\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_*,$$

which combined with the RIP-ℓ_2/ℓ_1 property of $\mathcal{B}$ yields

$$\begin{aligned}
\sum_{i=2}^{M_1}\frac{1}{m}\left\|\mathcal{B}\left(\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right)\right\|_1 &\leq \sum_{i=2}^{M_1}\frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)}{m}\left\|\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right\|_{\mathrm{F}} \\
&\leq \frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)}{\sqrt{K_1}}\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_*, \qquad (83)
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\sum_{i=2}^{M_2}\frac{1}{m}\left\|\mathcal{B}\left(\boldsymbol{H}^{(i)}_{\Omega^\perp}\right)\right\|_1 &\leq \sum_{i=2}^{M_1}\frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)}{m}\left\|\boldsymbol{H}^{(i)}_{\Omega^\perp}\right\|_{\mathrm{F}} \\
&\leq \frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)}{\sqrt{K_2}}\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1. \qquad (84)
\end{aligned}$$

The above argument relies on our construction scheme that $\operatorname{rank}(\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}) \leq K_1$, $\operatorname{supp}\left(\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right) \subseteq \Omega$, and $\left\|\boldsymbol{H}^{(i)}_{\Omega^\perp}\right\|_0 \leq K_2$, and hence all of $\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}$ and $\boldsymbol{H}^{(i)}_{\Omega^\perp}$ ($i \geq 1$) belong to $\mathcal{M}_{k,K_1,K_2}$.

Set $K_2 := \left\lceil \frac{K_1}{\lambda^2} \right\rceil$, and hence $\sqrt{\frac{K_1}{K_2}} \leq \lambda$. Recalling $\boldsymbol{H} = \boldsymbol{H}_{T\cap\Omega} + \boldsymbol{H}_{T^\perp\cap\Omega} + \boldsymbol{H}_{\Omega^\perp}$, one can proceed as follows

$$\begin{aligned}
\frac{2\epsilon_1}{m} &\geq \frac{1}{m}\left\|\mathcal{B}(\boldsymbol{H})\right\|_1 \\
&\geq \frac{1}{m}\left\|\mathcal{B}\left(\boldsymbol{H}_{T\cap\Omega} + \boldsymbol{H}^{(1)}_{T^\perp\cap\Omega} + \boldsymbol{H}^{(1)}_{\Omega^\perp}\right)\right\|_1 \\
&\quad - \sum_{i=2}^{M_1}\frac{1}{m}\left\|\mathcal{B}\left(\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right)\right\|_1 - \sum_{i=2}^{M_2}\frac{1}{m}\left\|\mathcal{B}\left(\boldsymbol{H}^{(i)}_{\Omega^\perp}\right)\right\|_1 \\
&\geq \left(1-\delta^{\mathrm{lb}}_{2K_1,2K_2}\right)\left\|\boldsymbol{H}_{T\cap\Omega} + \boldsymbol{H}^{(1)}_{T^\perp\cap\Omega} + \boldsymbol{H}^{(1)}_{\Omega^\perp}\right\|_{\mathrm{F}} \\
&\quad - \frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_*}{\sqrt{K_1}} - \frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1}{\sqrt{K_2}} \\
&\geq \left(1-\delta^{\mathrm{lb}}_{2K_1,2K_2}\right)\left\|\boldsymbol{H}_{T\cap\Omega} + \boldsymbol{H}^{(1)}_{T^\perp\cap\Omega} + \boldsymbol{H}^{(1)}_{\Omega^\perp}\right\|_{\mathrm{F}} \\
&\quad - \frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)}{\sqrt{K_1}}\left(\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_* + \lambda\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1\right) \\
&\geq \frac{\left(1-\delta^{\mathrm{lb}}_{2K_1,2K_2}\right)}{\sqrt{3}}\left(\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\mathrm{F}} + \left\|\boldsymbol{H}^{(1)}_{T^\perp\cap\Omega}\right\|_{\mathrm{F}} + \left\|\boldsymbol{H}^{(1)}_{\Omega^\perp}\right\|_{\mathrm{F}}\right) \\
&\quad - \frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)}{\sqrt{K_1}}\left(\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_* + \lambda\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1\right).
\end{aligned}$$

This taken collectively with (82) gives

$$\begin{aligned}
&\frac{2\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)}{\sqrt{K_1}}\left(\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_* + \lambda\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_1\right) + \frac{2\epsilon}{m} \\
&\qquad \geq \left(\frac{1-\delta^{\mathrm{lb}}_{2K_1,2K_2}}{\sqrt{3}} - \frac{3\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)}{\sqrt{K_1}}\right) \cdot \\
&\qquad\quad \left(\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\mathrm{F}} + \left\|\boldsymbol{H}^{(1)}_{T^\perp\cap\Omega}\right\|_{\mathrm{F}} + \left\|\boldsymbol{H}^{(1)}_{\Omega^\perp}\right\|_{\mathrm{F}}\right).
\end{aligned}$$

Therefore, if we know that

$$\frac{\frac{1-\delta^{\mathrm{lb}}_{2K_1,2K_2}}{\sqrt{3}} - \frac{3\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)}{\sqrt{K_1}}}{2\max\left\{\frac{1+\delta^{\mathrm{ub}}_{K_1,K_2}}{\sqrt{K_1}}, 1\right\}} \geq \beta_3 > 0$$

for some absolute constant β_3, then

$$\begin{aligned}
&\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\mathrm{F}} + \left\|\boldsymbol{H}^{(1)}_{T^\perp\cap\Omega}\right\|_{\mathrm{F}} + \left\|\boldsymbol{H}^{(1)}_{\Omega^\perp}\right\|_{\mathrm{F}} \\
&\qquad \leq \frac{1}{\beta_3}\left(\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_* + \lambda\left\|\boldsymbol{X}_{\mathrm{c}}\right\|_1 + \frac{\epsilon_1}{m}\right). \qquad (85)
\end{aligned}$$

On the other hand, we know from (83) and (84) that

$$\begin{aligned}
&\sum_{i=2}^{M_1}\left\|\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right\|_{\mathrm{F}} + \sum_{i=2}^{M_2}\left\|\boldsymbol{H}^{(i)}_{\Omega^\perp}\right\|_{\mathrm{F}} \\
&\leq \frac{1}{1-\delta^{\mathrm{lb}}_{K_1,K_2}}\sum_{i=2}^{M_1}\left\|\mathcal{B}\left(\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right)\right\|_1 + \sum_{i=2}^{M_2}\left\|\mathcal{B}\left(\boldsymbol{H}^{(i)}_{\Omega^\perp}\right)\right\|_1 \\
&\leq \frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_*}{\left(1-\delta^{\mathrm{lb}}_{K_1,K_2}\right)\sqrt{K_1}} + \frac{\left(1+\delta^{\mathrm{ub}}_{K_1,K_2}\right)\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1}{\left(1-\delta^{\mathrm{lb}}_{K_1,K_2}\right)\sqrt{K_2}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1+\delta^{\mathrm{ub}}_{K_1,K_2}}{\left(1-\delta^{\mathrm{lb}}_{K_1,K_2}\right)\sqrt{K_1}} \left(\left\| \boldsymbol{H}_{T^\perp\cap\Omega} \right\|_* + \lambda \left\| \boldsymbol{H}_{\Omega^\perp} \right\|_1\right) \\
&\leq \frac{1+\delta^{\mathrm{ub}}_{K_1,K_2}}{\left(1-\delta^{\mathrm{lb}}_{K_1,K_2}\right)\sqrt{K_1}} \left(3\left\| \boldsymbol{H}_{T\cap\Omega} \right\| + 2\left\| \boldsymbol{X}_{\mathrm{c}} \right\|_* + 2\lambda\left\| \boldsymbol{X}_{\mathrm{c}} \right\|_1\right),
\end{aligned} \tag{86}
$$

where (86) is a consequence of (83) and (84), and the last inequality arises from (82). This together with (85) completes the proof.

APPENDIX F
PROOF OF LEMMA 4

Simple calculation yields that

$$\mathbb{E}\left[\boldsymbol{A}_i \left\langle \boldsymbol{A}_i, \boldsymbol{X} \right\rangle\right] = 2\boldsymbol{X} + \left(1 + \frac{\mu_4 - 3}{n}\right) \operatorname{tr}\left(\boldsymbol{X}\right)\cdot \boldsymbol{I}. \tag{87}$$

When $\mu_4 = 3$, one can see that

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{B}_i \left\langle \boldsymbol{B}_i, \boldsymbol{X} \right\rangle\right] &= \frac{1}{4}\mathbb{E}\left[\left(\boldsymbol{A}_{2i-1} - \boldsymbol{A}_{2i}\right)\left\langle \boldsymbol{A}_{2i-1} - \boldsymbol{A}_{2i}, \boldsymbol{X}\right\rangle\right] \\
&= \boldsymbol{X}.
\end{aligned} \tag{88}
$$

When $\mu_4 \neq 3$, consider the linear combination

$$\boldsymbol{B} = a\boldsymbol{A}_1 + b\boldsymbol{A}_2 + c\boldsymbol{A}_3,$$

where we aim to find the coefficients a, b and c that makes $\boldsymbol{B}$ isotropic. If we further require

$$\mathbb{E}\left[\boldsymbol{B}\right] = a + b + c = \frac{\epsilon}{\sqrt{n}}, \tag{89}$$

then one can compute

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{B}\left\langle \boldsymbol{B}, \boldsymbol{X}\right\rangle\right] &= 2\left(a^2+b^2+c^2\right)\boldsymbol{X} \\
&\quad + \left[\left(1+\frac{\mu_4-3}{n}\right)\left(a^2+b^2+c^2\right)\right. \\
&\qquad \left. + 2\left(ab+bc+ac\right)\right]\operatorname{tr}\left(\boldsymbol{X}\right)\cdot\boldsymbol{I}.
\end{aligned}
$$

Our goal is thus to determine a, b and c that satisfy

$$\left(1+\frac{\mu_4-3}{n}\right)\left(a^2+b^2+c^2\right) + 2\left(ab+bc+ac\right) = 0,$$

which combined with (89) gives

$$\frac{\mu_4-3}{n}\left(a^2+b^2+c^2\right) + \frac{\epsilon^2}{n} = 0. \tag{90}$$

If we set $a = 1$, then (90) reduces to

$$
\begin{aligned}
&\frac{\mu_4-3}{n}\left(1+b^2+\left(\frac{\epsilon}{\sqrt{n}}-1-b\right)^2\right) + \frac{\epsilon^2}{n} = 0 \\
&\quad\Rightarrow b^2 + b\left(1-\frac{\epsilon}{\sqrt{n}}\right) + \frac{1}{2}\left(1-\frac{\epsilon}{\sqrt{n}}\right)^2 \\
&\qquad + \frac{1}{2} + \frac{\epsilon^2}{2\left(\mu_4-3\right)} = 0.
\end{aligned}
$$

Solving this quadratic equation yields

$$b = \frac{-\left(1-\frac{\epsilon}{\sqrt{n}}\right)+\sqrt{\Delta}}{2}; \quad c = \frac{-\left(1-\frac{\epsilon}{\sqrt{n}}\right)-\sqrt{\Delta}}{2}, \tag{91}$$

where

$$
\begin{aligned}
\Delta &:= \left(1-\frac{\epsilon}{\sqrt{n}}\right)^2 - 4\left(\frac{1}{2}\left(1-\frac{\epsilon}{\sqrt{n}}\right)^2 + \frac{1}{2} + \frac{\epsilon^2}{2\left(\mu_4-3\right)}\right) \\
&= -\left(1-\frac{\epsilon}{n}\right)^2 - 2 - \frac{2\epsilon^2}{\mu_4-3}.
\end{aligned}
$$

Note that $\Delta > 0$ when $\epsilon^2 > 1.5\cdot\left|3-\mu_4\right|$. Also, b and c satisfy

$$1 + b^2 + c^2 = \frac{\epsilon^2}{3-\mu_4}. \tag{92}$$

By choosing $\alpha = \sqrt{\frac{3-\mu_4}{2\epsilon^2}}$, $\beta = b\alpha$, and $\gamma = c\alpha$, we derive the form of $\boldsymbol{B}_i$ as introduced in (39), which satisfies

$$\mathbb{E}\left[\boldsymbol{B}_i\left\langle \boldsymbol{B}_i, \boldsymbol{X}\right\rangle\right] = \boldsymbol{X}.$$

Finally, we remark that for *any* norm $\left\|\cdot\right\|_{\mathrm{n}}$. This can be easily bounded as follows

$$
\begin{aligned}
\left\|\boldsymbol{B}_i\right\|_{\mathrm{n}} &\leq \sqrt{\frac{\left|3-\mu_4\right|}{2\epsilon^2}}\left(1+|b|+|c|\right)\max_{i:1\leq i\leq m}\left\|\boldsymbol{A}_i\right\|_{\mathrm{n}} \\
&\leq \sqrt{3}\sqrt{\frac{\left|3-\mu_4\right|}{2\epsilon^2}\left(1+b^2+c^2\right)}\max_{i:1\leq i\leq m}\left\|\boldsymbol{A}_i\right\|_{\mathrm{n}} && (93)\\
&= \sqrt{3}\max_{i:1\leq i\leq m}\left\|\boldsymbol{A}_i\right\|_{\mathrm{n}}. && (94)
\end{aligned}
$$

This concludes the proof.

APPENDIX G
PROOF OF LEMMA 5

Let $\boldsymbol{M}$ represent the symmetric Toeplitz matrix as follows

$$\boldsymbol{M} = \left[\boldsymbol{M}_{|i-l|}\right]_{1\leq i,l\leq n} := \mathcal{T}\left(\boldsymbol{z}\boldsymbol{z}^\top\right),$$

and since each descending diagonal of a Toeplitz matrix is constant, the entry $\boldsymbol{M}_k$ is given by the average of the corresponding diagonal, i.e.

$$\boldsymbol{M}_k := \frac{1}{n-k}\sum_{l=k+1}^{n} z_l z_{l-k}, \quad 0 \leq k < n.$$

Apparently, one has $\mathbb{E}\left[\boldsymbol{M}_0\right] = 1$ and $\mathbb{E}\left[\boldsymbol{M}_k\right] = 0$ $(1 \leq k < n)$.

The harmonic structure of the Toeplitz matrix $\boldsymbol{M}$ motivates us to embed it into a circulant matrix $\boldsymbol{C}_{\boldsymbol{M}}$. Specifically, a $(2n-1)\times(2n-1)$ circulant matrix

$$\boldsymbol{C}_{\boldsymbol{M}} := \begin{bmatrix} c_0 & c_1 & \cdots & c_{2n-2} \\ c_{2n-2} & c_0 & c_1 & c_2 \\ \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & \cdots & c_0 \end{bmatrix}$$

is constructed such that

$$c_i := \begin{cases} \boldsymbol{M}_i, & \text{if } 0 \leq i < n; \\ \boldsymbol{M}_{2n-i-1}, & \text{if } n \leq i \leq 2n-2. \end{cases}$$

Since $\boldsymbol{M}$ is a submatrix of $\boldsymbol{C}_{\boldsymbol{M}}$, it suffices to bound the spectral norm of $\boldsymbol{C}_{\boldsymbol{M}}$. Define $\omega_i := \exp\left(\frac{2\pi j}{2n-1}\cdot i\right)$, then the corresponding eigenvalues of $\boldsymbol{C_M}$ are given by

$$\begin{aligned}\lambda_i &:= \sum_l c_l \omega_i^l = \boldsymbol{M}_0 + \sum_{l=1}^{n-1} \boldsymbol{M}_l \omega_i^l + \sum_{l=n}^{2n-2} \boldsymbol{M}_{2n-l-1} \omega_i^l \\ &= \boldsymbol{M}_0 + 2\sum_{l=1}^{n-1} \boldsymbol{M}_l \cos\left(\frac{2\pi i l}{2n-1}\right),\end{aligned}$$

for $i = 0, 1, \cdots, 2n-2$, which satisfies $\mathbb{E}\lambda_i = \mathbb{E}\boldsymbol{M}_0 = 1$. This leads to an upper bound as follows

$$\|\boldsymbol{M}\| \le \|\boldsymbol{C_M}\| \le \max_{0\le i\le 2n-2} |\lambda_i|. \tag{95}$$

Note that λ_i is a quadratic form in $\{z_1, z_2, \cdots, z_n\}$. Define the symmetric coefficient matrix $\boldsymbol{G}^{(i)}$ such that for any $1 \le \alpha, \beta \le n$,

$$\boldsymbol{G}_{\alpha,\beta}^{(i)} = \frac{1}{n-|l|}\cos\left(\frac{2\pi i\,|l|}{2n-1}\right), \quad \text{if } \alpha - \beta = l,$$

which satisfies

$$\begin{aligned}\lambda_i &= \mathbb{E}[\boldsymbol{M}_0] + \sum_{1\le\alpha,\beta\le n} \boldsymbol{G}_{\alpha,\beta}^{(i)}\left(z_\alpha z_\beta - \mathbb{E}\left[z_\alpha z_\beta\right]\right) \\ &= 1 + \sum_{1\le\alpha,\beta\le n} \boldsymbol{G}_{\alpha,\beta}^{(i)}\left(z_\alpha z_\beta - \mathbb{E}\left[z_\alpha z_\beta\right]\right).\end{aligned}$$

When $\boldsymbol{z}$ are drawn from a sub-Gaussian measure, Lemma 7 asserts that there exists an absolute constant $c_{10} > 0$ such that

$$\mathbb{P}\{|\lambda_i - 1| \ge t\} \le \exp\left(-c_{10}\min\left\{\frac{t}{\|\boldsymbol{G}^{(i)}\|}, \frac{t^2}{\|\boldsymbol{G}^{(i)}\|_{\mathrm{F}}^2}\right\}\right) \tag{96}$$

holds for any $t > 0$.

It remains to compute $\|\boldsymbol{G}^{(i)}\|_{\mathrm{F}}$ and $\|\boldsymbol{G}^{(i)}\|$. Since $\boldsymbol{G}^{(i)}$ is a symmetric Toeplitz matrix, we have

$$\|\boldsymbol{G}^{(i)}\|_{\mathrm{F}}^2 = \sum_{\alpha,\beta=1}^{n} \left|\boldsymbol{G}_{\alpha,\beta}\right|^2 \le 2\sum_{l=0}^{n-1}\frac{1}{n-l} \le 2\log n. \tag{97}$$

It then follows that

$$\|\boldsymbol{G}^{(i)}\| \le \|\boldsymbol{G}^{(i)}\|_{\mathrm{F}} \le \sqrt{2\log n}. \tag{98}$$

Substituting these two bounds into (96) immediately yields that there exists a constant $c_{12} > 0$ such that

$$\lambda_i \le c_{12}\log^{\frac{3}{2}} n, \quad 1 \le i \le 2n-2 \tag{99}$$

holds with probability exceeding $1 - \frac{1}{n^{10}}$. This taken collectively with (95) concludes the proof.

Appendix H
Proof of Lemma 6

For technical convenience, we introduce another collection of events

$$\forall 1 \le i \le m: \quad F_i := \{\|\boldsymbol{B}_i\|_{\mathrm{F}} \le 20n\log n\}.$$

Since the restriction of $\mathcal{B}_i$ to Toeplitz matrices is isotropic and $\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T} \succeq 0$, we have $\mathcal{T} = \mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\right] \succeq \mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_E\right] \succeq \mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{E\cap F_i}\right]$, which yields

$$\left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_E\right] - \mathcal{T}\right\| \le \left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{E\cap F_i}\right] - \mathcal{T}\right\|. \tag{100}$$

Thus, it is sufficient to evaluate $\left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathbf{1}_{E\cap F_i}\right] - \mathcal{T}\right\|$. To this end, we adopt an argument of similar spirit as [52, Appendix B]. Write

$$\begin{aligned}\mathcal{T} &= \mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\right] \\ &= \mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{E\cap F_i}\right] + \mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{E^c\cup F_i^c}\right],\end{aligned}$$

and, consequently,

$$\begin{aligned}&\left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{E\cap F_i}\right] - \mathcal{T}\right\| \\ &\quad= \left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{E^c\cup F_i^c}\right]\right\| \\ &\quad\le \left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{F_i\cap E^c}\right]\right\| + \left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{F_i^c}\right]\right\|, \end{aligned}\tag{101}$$

which allows us to bound $\left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{F_i\cap E^c}\right]\right\|$ and $\left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{F_i^c}\right]\right\|$ separately.

First, it follows from the identity $\left\|\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\right\| = \|\mathcal{T}(\boldsymbol{B}_i)\|_{\mathrm{F}}^2$ and the definition of the event F_i that

$$\left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{F_i\cap E^c}\right]\right\| \le (20n\log n)^2\,\mathbb{P}\left(E^c\right) < \frac{1}{n^2}. \tag{102}$$

Second, applying the tail inequality on the quadratic form (see [59, Proposition 1.1]) yields

$$\mathbb{P}\left(\|\boldsymbol{A}_i\|_{\mathrm{F}} \ge c_{20}\left(n + 2\sqrt{nt} + 2t\right)\right) \le e^{-t}. \tag{103}$$

Thus, for any $t > (20n\log n)^2$, one has

$$\mathbb{P}\left(\|\boldsymbol{A}_i\|_{\mathrm{F}} \ge \sqrt{\frac{t}{3}}\right) \le e^{-c_{21}\sqrt{t}} \tag{104}$$

for some absolute constant $c_{21} > 0$. Recall that $\|\boldsymbol{B}_i\|_{\mathrm{F}} \le \sqrt{3}\max\left\{\|\boldsymbol{A}_{3i-2}\|_{\mathrm{F}}, \|\boldsymbol{A}_{3i-1}\|_{\mathrm{F}}, \|\boldsymbol{A}_{3i}\|_{\mathrm{F}}\right\}$, which indicates

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{B}_i\|_{\mathrm{F}}^2 \ge t\right) &\le \mathbb{P}\left(\|\boldsymbol{A}_{3i-1}\|_{\mathrm{F}}^2 \ge \frac{t}{3}\right) \\ &\quad+\mathbb{P}\left(\|\boldsymbol{A}_{3i-2}\|_{\mathrm{F}}^2 \ge \frac{t}{3}\right) \\ &\quad+\mathbb{P}\left(\|\boldsymbol{A}_{3i}\|_{\mathrm{F}}^2 \ge \frac{t}{3}\right) \\ &\le 3\mathbb{P}\left(\|\boldsymbol{A}_i\|_{\mathrm{F}} \ge \sqrt{\frac{t}{3}}\right) \\ &\le 3e^{-c_{21}\sqrt{t}} := g(t).\end{aligned}$$

A similar approach as introduced in [52, Appendix B] gives

$$\begin{aligned}&\left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{F_i^c}\right]\right\| \le \mathbb{E}\left[\|\boldsymbol{B}_i\|_{\mathrm{F}}^2\mathbf{1}_{F_i^c}\right] \\ &\quad\le (20n\log n)^2\, g\left((20n\log n)^2\right) + \int_{(20n\log n)^2}^{\infty} g(t)\,\mathrm{d}t \\ &\quad< (20n\log n)^2\, g\left((20n\log n)^2\right) + \int_{(20n\log n)^2}^{\infty} \frac{1}{t^5}\mathrm{d}t \\ &\quad< \frac{c_{15}}{n^2}\end{aligned}\tag{105}$$

for some absolute constant $c_{15} > 0$. This taken collectively with (100), (101) and (102) yields

$$\left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_E\right] - \mathcal{T}\right\| \le \left\|\mathbb{E}\left[\mathcal{T}\mathcal{B}_i^*\mathcal{B}_i\mathcal{T}\mathbf{1}_{E\cap F_i}\right] - \mathcal{T}\right\| \le \frac{\tilde{c}_{15}}{n^2}$$

for some absolute constant $\tilde{c}_{15} > 0$.

APPENDIX I
PROOF OF LEMMA 9

Dudley's inequality [60, Th. 11.17] allows us to bound the supremum of the Gaussian process as follows

$$\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}},X\in T,\|X\|_{\mathrm{F}}=1}\left|\sum_{i=1}^m g_i\left|\mathcal{B}_i\left(X\right)\right|^2\right|\,\middle|\,\mathcal{B}_i\ (1\leq i\leq m)\right]$$
$$\leq 24\int_0^\infty \log^{\frac{1}{2}} N\left(\mathcal{D}_{2r}^2, d\left(\cdot,\cdot\right),u\right)\mathrm{d}u, \tag{106}$$

where $\mathcal{D}_r^2 := \{X \mid \|X\|_{\mathrm{F}}=1, \mathrm{rank}\,(X)\leq 2r\}$. Here, $N\left(\mathcal{Z}, d\left(\cdot,\cdot\right),u\right)$ denotes the smallest number of balls of radius u centered in points of $\mathcal{Z}$ needed to cover the set $\mathcal{Z}$, under the pseudo metric $d\left(\cdot,\cdot\right)$ defined as follows

$$d\left(X,Y\right) := \sqrt{\sum_{i=1}^m\left(\left|\mathcal{B}_i\left(X\right)\right|^2-\left|\mathcal{B}_i\left(Y\right)\right|^2\right)^2}.$$

For any (X,Y) that satisfy $\|X\|_{\mathrm{F}}=\|Y\|_{\mathrm{F}}=1$, $\mathrm{rank}\,(X)\leq r$ and $\mathrm{rank}\,(Y)\leq r$, the pseudo metric satisfies

$$\begin{aligned}
d\left(X,Y\right) &\leq \sqrt{\left(\max_{i:1\leq i\leq m}\left|\mathcal{B}_i\left(X-Y\right)\right|^2\right)\sum_{i=1}^m\left|\mathcal{B}_i\left(X+Y\right)\right|^2}\\
&\leq \sqrt{2}\sqrt{\sum_{i=1}^m\left|\mathcal{B}_i\left(X\right)\right|^2+\left|\mathcal{B}_i\left(Y\right)\right|^2}\max_{i:1\leq i\leq m}\left|\mathcal{B}_i\left(X-Y\right)\right|\\
&\leq \left\{\sqrt{\left\langle X,\left(\sum_{i=1}^m\mathcal{B}_i^*\mathcal{B}_i\right)(X)\right\rangle}+\sqrt{\left\langle Y,\left(\sum_{i=1}^m\mathcal{B}_i^*\mathcal{B}_i\right)(Y)\right\rangle}\right\}\\
&\quad\cdot\sqrt{2}\max_{i:1\leq i\leq m}\left|\mathcal{B}_i\left(X-Y\right)\right|\\
&\leq 2\sqrt{2}\sup_{T:T\in\mathcal{M}_r^{\mathrm{t}}}\sqrt{\left\|\sum_{i=1}^m\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|}\max_{i:1\leq i\leq m}\left|\langle B_i, X-Y\rangle\right|,
\end{aligned}$$

where the last inequality relies on the observation that $\|X\|_{\mathrm{F}}=\|Y\|_{\mathrm{F}}=1$.

If we introduce the quantity

$$R := \sup_{T:T\in\mathcal{M}_r^{\mathrm{t}}}\sqrt{\left\|\sum_{i=1}^m\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|} \tag{107}$$

and define another pseudo metric $\|\cdot\|_{\mathcal{B}}$ as

$$\|X\|_{\mathcal{B}} := \max_{i:1\leq i\leq m}\left|\langle B_i, X\rangle\right|, \tag{108}$$

then $d\left(X,Y\right)\leq 2\sqrt{2}R\left\|X-Y\right\|_{\mathcal{B}}$, which allows us to bound

$$\begin{aligned}
&\int_0^\infty\log^{\frac{1}{2}} N\left(\mathcal{D}_{2r}^2, d\left(\cdot,\cdot\right),u\right)\mathrm{d}u\\
&\quad\leq\int_0^\infty\log^{\frac{1}{2}} N\left(\mathcal{D}_{2r}^2, 2\sqrt{2}R\left\|\cdot\right\|_{\mathcal{B}},u\right)\mathrm{d}u\\
&\quad=\int_0^\infty\log^{\frac{1}{2}} N\left(\frac{1}{\sqrt{2r}}\mathcal{D}_{2r}^2, \left\|\cdot\right\|_{\mathcal{B}},\frac{u}{4R\sqrt{r}}\right)\mathrm{d}u\\
&\quad\leq\int_0^\infty\log^{\frac{1}{2}} N\left(\mathcal{D}_{2r}^1, \left\|\cdot\right\|_{\mathcal{B}},\frac{u}{4R\sqrt{r}}\right)\mathrm{d}u\\
&\quad\leq\ 4R\sqrt{r}\int_0^\infty\log^{\frac{1}{2}} N\left(\mathcal{D}^1, \left\|\cdot\right\|_{\mathcal{B}},u\right)\mathrm{d}u.
\end{aligned} \tag{109}$$

Here, $\mathcal{D}_r^1$ and $\mathcal{D}^1$ stand for

$$\begin{aligned}
\mathcal{D}_r^1 &:= \left\{X \mid \|X\|_*\leq 1, \mathrm{rank}\,(X)\leq r\right\},\\
\mathcal{D}^1 &:= \left\{X \mid \|X\|_*\leq 1\right\},
\end{aligned}$$

and we have exploited the containment

$$\frac{1}{\sqrt{2r}}\mathcal{D}_{2r}^2\subseteq\mathcal{D}_{2r}^1\subseteq\mathcal{D}^1.$$

Hence it suffices to bound

$$E_2 := 4R\sqrt{r}\int_0^\infty\log^{\frac{1}{2}} N\left(\mathcal{D}^1, \left\|\cdot\right\|_{\mathcal{B}},u\right)\mathrm{d}u.$$

It remains to bound the covering number (or metric entropy) of the nuclear-norm ball $\mathcal{D}^1$. Repeating the well-known procedure as in [61, p. 1113] yields

$$\begin{aligned}
\int_0^\infty\sqrt{\log N\left(\mathcal{D}^1, \left\|\cdot\right\|_{\mathcal{B}},u\right)}\mathrm{d}u &\leq C_{10}K\left(\log n\right)^{5/2}\sqrt{\log m}\\
&\leq C_{11}K\log^3 n
\end{aligned}$$

for some constants $C_{10}, C_{11}>0$. This taken collectively with (106) and (109) gives that conditioning on $\mathcal{B}_i$'s, one has

$$\mathbb{E}\left[\sup_{T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\mathcal{P}_T\left(\sum_{i=1}^m g_i\mathcal{B}_i^*\mathcal{B}_i\right)\mathcal{P}_T\right\|\,\middle|\,\mathcal{B}_i\ (1\leq i\leq m)\right]$$
$$\leq C_{14}\sqrt{r}K\log^3 n\sqrt{\sup_{T:T\in\mathcal{M}_r^{\mathrm{t}}}\left\|\sum_{i=1}^m\mathcal{P}_T\mathcal{B}_i^*\mathcal{B}_i\mathcal{P}_T\right\|}. \tag{110}$$

for some absolute constant $C_{14}>0$.

ACKNOWLEDGMENTS

The authors would like to thank Emmanuel Candes for stimulating discussions. We would also like to thank Yihong Wu for his helpful comments and suggestions on Theorem 4, Yudong Chen for fruitful discussions about statistical consistency, and Xiaodong Li for helpful discussion on sparse phase retrieval.

REFERENCES

[1] Y. Chen, Y. Chi, and A. J. Goldsmith, "Estimation of simultaneously structured covariance matrices from quadratic measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 7669–7673.

[2] Y. Chen, Y. Chi, and A. J. Goldsmith, "Robust and universal covariance estimation from quadratic measurements via convex programming," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, Jun./Jul. 2014, pp. 2017–2021.

[3] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[4] N. E. Karoui, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *Ann. Statist.*, vol. 36, no. 6, pp. 2717–2756, 2008.

[5] S. Muthukrishnan, *Data Streams: Algorithms and Applications*. Norwell, MA, USA: NOW Pub., 2005.

[6] R. C. Daniels and R. W. Heath, "60 GHz wireless communications: Emerging requirements and design recommendations," *IEEE Veh. Technol. Mag.*, vol. 2, no. 3, pp. 41–50, Sep. 2007.

[7] A. Jung, G. Taubock, and F. Hlawatsch, "Compressive spectral estimation for nonstationary random processes," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 3117–3138, May 2013.

[8] G. Leus and Z. Tian, "Recovering second-order statistics from compressive measurements," in *Proc. 4th IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Dec. 2011, pp. 337–340.

[9] D. D. Ariananda and G. Leus, "Compressive wideband power spectrum estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4775–4789, Sep. 2012.

[10] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[11] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 520–544, Jan. 2010.

[12] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 2146–2157, Aug. 1994.

[13] M. G. Raymer, M. Beck, and D. McAlister, "Complex wave-field reconstruction using phase-space tomography," *Phys. Rev. Lett.*, vol. 72, no. 8, p. 1137, Feb. 1994.

[14] L. Tian, J. Lee, S. B. Oh, and G. Barbastathis, "Experimental compressive phase space tomography," *Opt. Exp.*, vol. 20, no. 8, pp. 8296–8308, 2012.

[15] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Commun. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, Aug. 2013.

[16] Y. Shechtman, Y. C. Eldar, A. Szameit, and M. Segev, "Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing," *Opt. Exp.*, vol. 19, no. 16, pp. 14807–14822, 2011.

[17] I. Waldspurger, A. d'Aspremont, and S. Mallat. (2012). "Phase recovery, MaxCut and complex semidefinite programming." [Online]. Available: http://arxiv.org/abs/1206.0102

[18] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[19] Y. Shechtman, A. Beck, and Y. C. Eldar. (2013). "GESPAR: Efficient phase retrieval of sparse signals." [Online]. Available: http://arxiv.org/abs/1301.1018

[20] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 2796–2804.

[21] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1043–1055, Feb. 2015.

[22] Y. Chen, X. Yi, and C. Caramanis, "A convex formulation for mixed regression with two components: Minimax optimal rates," in *Proc. Conf. Learn. Theory (COLT)*, 2014, pp. 1–45.

[23] E. J. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Found. Comput. Math.*, vol. 14, no. 5, pp. 1017–1026, Oct. 2013.

[24] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM J. Math. Anal.*, vol. 45, no. 5, pp. 3019–3033, 2013.

[25] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, vol. 61, no. 8, pp. 1025–1045, Aug. 2008.

[26] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, Mar. 2011.

[27] Y.-K. Liu, "Universal low-rank matrix recovery from Pauli measurements," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1638–1646.

[28] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[29] Y. Chen and Y. Chi, "Robust spectral compressed sensing via structured matrix completion," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6576–6601, Oct. 2014.

[30] I. M. Johnstone, "High dimensional statistical inference and random matrices," in *Proc. Int. Congr. Math.*, Madrid, Spain, Aug. 2006, pp. 307–333.

[31] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[32] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[33] Z. Tian, Y. Tafesse, and B. M. Sadler, "Cyclic feature detection with sub-Nyquist sampling for wideband spectrum sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 1, pp. 58–69, Feb. 2012.

[34] D. Romero and G. Leus, "Wideband spectrum sensing from compressed measurements using spectral prior information," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6232–6246, Dec. 2013.

[35] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. Nowak. (2013). "Sketching sparse matrices." [Online]. Available: http://arxiv.org/abs/1303.6544

[36] A. Beck and Y. C. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," *SIAM J. Optim.*, vol. 23, no. 3, pp. 1480–1509, 2013.

[37] K. Jaganathan, S. Oymak, and B. Hassibi, "Recovery of sparse 1-D signals from the magnitudes of their Fourier transform," in *Proc. IEEE ISIT*, Jul. 2012, pp. 1473–1477.

[38] E. Oja, *Subspace Methods of Pattern Recognition*, vol. 4. U.K.: Research Studies Press, 1983.

[39] I. Mitliagkas, C. Caramanis, and P. Jain, "Memory limited, streaming PCA," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 2886–2894.

[40] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. Allerton*, Sep./Oct. 2010, pp. 704–711.

[41] Y. Chi, Y. C. Eldar, and R. Calderbank, "PETRELS: Parallel subspace estimation and tracking by recursive least squares from partial observations," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5947–5959, Dec. 2013.

[42] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," in *Computer Vision—ECCV*. Berlin, Germany: Springer-Verlag, 2002, pp. 707–720.

[43] T. T. Cai and A. Zhang. (2014). "ROP: Matrix recovery via rank-one projections." [Online]. Available: http://arxiv.org/abs/1310.5791

[44] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[45] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry. (2011). "Compressive phase retrieval from squared output measurements via semidefinite programming." [Online]. Available: http://arxiv.org/abs/1111.6323

[46] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comp. Rendus Math.*, vol. 346, nos. 9–10, pp. 589–592, May 2008.

[47] G. W. Stewart and J.-G. Sun, "Matrix perturbation theory," 1990.

[48] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, "Efficient and robust compressed sensing using optimized expander graphs," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4299–4308, Sep. 2009.

[49] E. J. Candes and Y. Plan, "Tight Oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2342–2359, Apr. 2011.

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

[51] M. Talagrand, "Majorizing measures: The generic chaining," *Ann. Probab.*, vol. 24, no. 3, pp. 1049–1103, Jul. 1996.

[52] E. J. Candes and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7235–7254, Nov. 2011.

[53] T. Tao, *Topics in Random Matrix Theory* (Graduate Studies in Mathematics). Providence, RI, USA: AMS, 2012.

[54] L. Demanet and P. Hand. (2012). "Stable optimizationless recovery from phaseless linear measurements." [Online]. Available: http://arxiv.org/abs/1208.1803

[55] U. Grenander and G. Szegő, *Toeplitz Forms and Their Applications*. Oakland, CA, USA: Univ. California Press, 1958.

[56] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Compress. Sens., Theory Appl.*, pp. 210–268, 2012.

[57] D. L. Hanson and F. T. Wright, "A bound on tail probabilities for quadratic forms in independent random variables," *Ann. Math. Statist.*, vol. 42, no. 3, pp. 1079–1083, 1971.

[58] M. Rudelson and R. Vershynin. (Jun. 2013). "Hanson–Wright inequality and sub-Gaussian concentration." [Online]. Available: http://arxiv.org/abs/1306.2872

[59] D. Hsu, S. M. Kakade, and T. Zhang, "A tail inequality for quadratic forms of sub-Gaussian random vectors," *Electron. Commun. Probab.*, vol. 17, no. 52, pp. 1–6, Nov. 2012.

[60] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, vol. 23. New York, NY, USA: Springer-Verlag, 1991.

[61] G. Aubrun, "On almost randomizing channels with a short Kraus decomposition," *Commun. Math. Phys.*, vol. 288, no. 3, pp. 1103–1116, Jun. 2009.

Yuxin Chen (S'09) received the B.S. in Microelectronics with High Distinction from Tsinghua University in 2008, the M.S. in Electrical and Computer Engineering from the University of Texas at Austin in 2010, the M.S. in Statistics from Stanford University in 2013, and the Ph.D. in Electrical Engineering from Stanford University in 2015. He is currently a postdoctoral scholar in the Department of Statistics at Stanford University. His research interests include high-dimensional structured estimation, information theory, compressed sensing, and network science.

Yuejie Chi (S'09–M'12) received the Ph.D. degree in Electrical Engineering from Princeton University in 2012, and the B.E. (Hon.) degree in Electrical Engineering from Tsinghua University, Beijing, China, in 2007. Since September 2012, she has been an assistant professor with the department of Electrical and Computer Engineering and the department of Biomedical Informatics at the Ohio State University.

She is the recipient of the IEEE Signal Processing Society Young Author Best Paper Award in 2013 and the Best Paper Award at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2012. She received AFOSR Young Investigator Program Award in 2015, the Ralph E. Powe Junior Faculty Enhancement Award from Oak Ridge Associated Universities in 2014, a Google Faculty Research Award in 2013, the Roberto Padovani scholarship from Qualcomm Inc. in 2010, and an Engineering Fellowship from Princeton University in 2007. She has held visiting positions at Colorado State University, Stanford University and Duke University, and interned at Qualcomm Inc. and Mitsubishi Electric Research Lab. Her research interests include statistical signal processing, information theory, machine learning and their applications in high-dimensional data analysis, active sensing and bioinformatics.

Andrea J. Goldsmith (S'90–M'93–SM'99–F'05) is the Stephen Harris professor of Electrical Engineering at Stanford University. She was previously on the faculty of Electrical Engineering at Caltech. Her research interests are in information theory and communication theory, and their application to wireless communications and related fields. Dr. Goldsmith co-founded and served as CTO for two wireless companies: Accelera, Inc., which develops software-defined wireless network technology for cloud-based management of WiFi access points, and Quantenna Communications, Inc., which develops high-performance WiFi chipsets. She has also held industry positions at Maxim Technologies, Memorylink Corporation, and AT&T Bell Laboratories. She is a Fellow of the IEEE and of Stanford, and has received several awards for her work, including the IEEE ComSoc Edwin H. Armstrong Achievement Award as well as Technical Achievement Awards in Communications Theory and in Wireless Communications, the National Academy of Engineering Gilbreth Lecture Award, the IEEE ComSoc and Information Theory Society Joint Paper Award, the IEEE ComSoc Best Tutorial Paper Award, the Alfred P. Sloan Fellowship, and the Silicon Valley/San Jose Business Journal's Women of Influence Award. She is author of the book "*Wireless Communications*" and co-author of the books "*MIMO Wireless Communications*" and "*Principles of Cognitive Radio*," all published by Cambridge University Press, as well as an inventor on 28 patents. She received the B.S., M.S. and Ph.D. degrees in Electrical Engineering from U.C. Berkeley.

Dr. Goldsmith has served as editor for the IEEE TRANSACTIONS ON INFORMATION THEORY, the *Journal on Foundations and Trends in Communications and Information Theory and in Networks*, the IEEE TRANSACTIONS ON COMMUNICATIONS, and the *IEEE Wireless Communications Magazine* as well as on the Steering Committee for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. She participates actively in committees and conference organization for the IEEE Information Theory and Communications Societies and has served on the Board of Governors for both societies. She has also been a Distinguished Lecturer for both societies, served as President of the IEEE Information Theory Society in 2009, founded and chaired the student committee of the IEEE Information Theory society, and chaired the Emerging Technology Committee of the IEEE Communications Society. At Stanford she received the inaugural University Postdoc Mentoring Award, served as Chair of Stanfords Faculty Senate in 2009, and currently serves on its Faculty Senate, Budget Group, and Task Force on Women and Leadership.