ESTIMATION OF SIMULTANEOUSLY STRUCTURED COVARIANCE MATRICES FROM QUADRATIC MEASUREMENTS

Yuxin Chen[⋆] *Yuejie Chi*[†] *Andrea J. Goldsmith*[⋆]

[⋆] Department of Electrical Engineering, Stanford University
[†] Department of Electrical and Computer Engineering, The Ohio State University

ABSTRACT

This paper explores covariance estimation from energy measurements that are collected via a quadratic form of measurement vectors. A popular structural model is considered where the covariance matrices possess low-rank and sparse structures *simultaneously*. We investigate a weighted convex relaxation algorithm tailored for this joint structure, which guarantees exact and universal recovery from a small number of measurements. The algorithm is also robust against noise and imperfect structural assumptions. In particular, when the non-zero entries of the covariance matrix exhibit power-law decay, our algorithm admits *exact recovery* as soon as the number of measurements exceeds the theoretic limit. Our method is related to sparse phase retrieval: the analysis framework herein recovers and strengthens the best-known performance guarantees by extending them to approximately sparse and noisy scenarios as well as a broader class of measurement vectors, and our results are derived using much simpler analysis methods.

Index Terms— Quadratic Sampling, Convex Relaxation, Sparse, Low-Rank, RIP-ℓ_2/ℓ_1

1. INTRODUCTION

A broad class of signal processing and wireless communication tasks in stochastic environments requires reliable estimates of the spectral characteristics of random signals [1]. For instance, optimal signal transmissions are often based on the Karhunen–Loï¿œve decomposition of a random process, which requires knowledge of its second-order statistics [2]. Indeed, covariance estimation has become a core step that dictates the performance of various signal processing and communication algorithms.

In this paper, we consider the problem of recovering an unknown covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ from a small number of measurements of the form

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

where $\boldsymbol{y} := \{y_i\}_{i=1}^m$ denotes the measurements, $\boldsymbol{a}_i \in \mathbb{R}^n$ represents the measurement vector, and $\boldsymbol{\eta} := \{\eta_i\}_{i=1}^m$ denotes the noise. The measurements $\boldsymbol{a}_i^T \mathbf{\Sigma} \boldsymbol{a}_i$'s are quadratic in $\boldsymbol{a}_i$ and are referred to as *quadratic measurements*; they are also *non-negative* given the positive semidefiniteness (PSD) of $\mathbf{\Sigma}$. The measurement model (1) subsumes several sampling scenarios based on energy measurements or magnitude measurements. A partial list is provided as follows.

Spectrum Estimation of Stochastic Processes: In the high-frequency regime, an *empirical* energy measurement is often more accurate (and easier to obtain) than a phase measurement. For instance, energy measurements will become more reliable than phase measurements as communication systems shift towards extremely high carrier frequency regimes (e.g. 60GHz communications [3]), where fast phase variation degrades estimation. If we employ a measurement vector $\boldsymbol{a}_i$ and observe the average energy over N instances $\{\boldsymbol{x}_t\}_{t=1}^N$, then the measurement can be expressed as

$$y_i = \frac{1}{N} \sum\nolimits_{t=1}^{N} \left| \boldsymbol{a}_i^T \boldsymbol{x}_t \right|^2 = \boldsymbol{a}_i^T \mathbf{\Sigma}_N \boldsymbol{a}_i, \tag{2}$$

where $\mathbf{\Sigma}_N := \frac{1}{N} \sum_{t=1}^N \boldsymbol{x}_t \boldsymbol{x}_t^T$ is the sample covariance matrix.

Noncoherent Subspace Detection: Matched subspace detection [4] is widely applied in wireless communication, radar, and pattern recognition when the transmitted signal is encoded by the membership of subspaces. Our problem formulation 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 (2).

Compressive Phase Space Tomography: Phase Space Tomography [5] is an appealing method to measure the correlation matrix of a wave field in physics. However, tomography becomes challenging when the correlation matrix is large. Recently, it was proposed experimentally in [6] to recover an approximately low-rank correlation matrix [7] by only taking a few quadratic measurements in (1).

Phase Retrieval: Due to the physical constraints, one can only measure and record intensities 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, known as phase retrieval [8,9]. If we set $\mathbf{\Sigma} := \boldsymbol{x}\boldsymbol{x}^T$, then our problem formulation (1) subsumes phase retrieval as a special case.

1.1. Contributions

In all of the above applications, it is of great interest to faithfully recover the covariance matrices from only a number of quadratic measurements that is much smaller than the ambient dimension. This is made possible by exploring low-dimensional structures of the covariance matrices, such as sparsity, low rank, and so on. A general analysis framework for various structural assumptions has been recently proposed in [10]. In this paper, we consider recovery of covariance matrices that are *simultaneously* sparse and low-rank from (1). This joint structure arises in many related problems such as sparse principal component analysis [11], sparse signal recovery from random magnitude measurements [12–14], etc. See [15] and the reference therein.

We develop a convex optimization algorithm, which seeks the covariance matrix that minimizes a weighted sum of its elementwise ℓ_1 norm and trace norm satisfying the measurements (1). Based on a

Y. Chen and A. J. Goldsmith are supported by the NSF Center for Science of Information. Y. Chi is partially supported by a Google Faculty Research Award.

key and novel mixed-norm restricted isometry property (called RIP-ℓ_2/ℓ_1) introduced in [10], our analysis demonstrates that the proposed algorithm allows exact and universal covariance estimation from very few quadratic measurements in the absence of noise, provided that the sampling vectors are sub-Gaussian distributed. Once the measurement vectors are selected, a large class of covariance matrices satisfying the presumed structure can all be recovered with high probability. Our performance guarantees also extend to noisy measurements and approximate structural assumptions.

Somewhat surprisingly, when *the nonzero entries* of the structured covariance matrix experience power-law decay – a property known as compressibility in compressed sensing literature [16, 17], the proposed algorithm enables exact recovery as soon as the number of measurements exceeds the fundamental sampling limits. For the special case of rank-one matrices, our result recovers and strengthens the best-known recovery guarantees for sparse phase retrieval using convex optimization [12].

1.2. Related Work

Recovery of simultaneously structured matrices from a small number of linear measurements has recently received much attention [15]. It has been shown in [15] that convex relaxation methods that use a weighted sum of penalties with respect to each individual structure can do no better, orderwise, than a convex relaxation approach that employs only one penalty. While our algorithm confirms these results, a gain in the under-sampling ratio may be achieved when the covariance matrix is compressible, as will be made clear in Section 3.

Our covariance estimation method is inspired by recent advances in phase retrieval [13, 14, 18–24], which is equivalent to recovering rank-one covariance matrices from quadratic measurements. When the signal $\boldsymbol{x}$ is sparse with only k nonzero entries, the lifted matrix $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^T$ is simultaneously low-rank and sparse with rank one and sparsity level k^2. For this special case, our algorithm coincides with the sparse Phaselift algorithm studied in [12]. It is established in [12] that $\mathcal{O}(k^2 \log n)$ *Gaussian* measurements suffice to recover any k-sparse signal $\boldsymbol{x}$. Our paper strengthens the performance guarantee therein to hold universally for all k-sparse signals under a larger class of sub-Gaussian sensing vectors, and we demonstrate that even approximately sparse signals can be faithfully recovered with an orderwise equivalent number of measurements.

1.3. Organization

The rest of the paper is organized as follows. Section 2 describes the problem formulation and signal models. Section 3 presents the main theoretical findings, and we outline its proof in Section 4. Finally, we conclude the paper in Section 5.

2. SIGNAL MODELS

In this section, we describe the sampling model in (1), and the simultaneous structured model of the covariance matrices considered in this paper.

2.1. Measurement Model

Random sampling has been proven to be very effective in preserving information content from a small number of linear measurements, for example in compressed sensing [16, 17]. We explore the following random sampling model as presented in [10], which is able to effectively exploit the presumed low-dimensional joint structure, as will be shown in Section 3. We assume that the measurement vectors are composed of i.i.d. *sub-Gaussian* entries. In particular, $\boldsymbol{a}_i$'s, $i = 1, \ldots, m$, are i.i.d. copies of $\boldsymbol{a} = [a_1, \cdots, a_n]^T$, where each a_i is chosen i.i.d. from a distribution satisfying

$$\mathbb{E}[a_i] = 0, \quad \mathbb{E}[a_i^2] = 1, \quad \text{and} \quad \mu_4 := \mathbb{E}a_i^4 > 1. \tag{3}$$

We further assume that the noise $\boldsymbol{\eta} := [\eta_1, \cdots, \eta_m]^T$ is bounded in ℓ_1 norm *deterministically* as

$$\|\boldsymbol{\eta}\|_1 \leq \epsilon, \tag{4}$$

where ϵ is known *a priori*.

For notational simplicity, let $\boldsymbol{A}_i := \boldsymbol{a}_i\boldsymbol{a}_i^T$ represent the ith measurement matrix, and hence $y_i := \langle \boldsymbol{A}_i, \boldsymbol{\Sigma} \rangle + \eta_i$. 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 (1) as

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

2.2. Simultaneous Low-rank and Sparse Covariance Structures

We consider structured covariance matrices that are simultaneously low-rank and sparse. Specifically, suppose that the covariance matrix $\boldsymbol{\Sigma}$ can be decomposed into two components

$$\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_\Omega + \boldsymbol{\Sigma}_{\mathrm{c}}, \tag{6}$$

where $\boldsymbol{\Sigma}_\Omega$ is the main component that is simultaneously sparse and low-rank, and $\boldsymbol{\Sigma}_{\mathrm{c}}$ denotes the residual term that might arise due to imperfect structural assumptions. The rank of $\boldsymbol{\Sigma}_\Omega$ is denoted by r. More specifically, we assume that the eigen-decomposition of $\boldsymbol{\Sigma}_\Omega$ is given by

$$\boldsymbol{\Sigma}_\Omega = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \succeq 0, \tag{7}$$

where $\boldsymbol{U} := [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$ consists of all r eigenvectors of $\boldsymbol{\Sigma}_\Omega$. Without loss of generality, we assume that for all $1 \leq i \leq r$, $\mathrm{supp}\,(\boldsymbol{u}_i) \subseteq \omega$ for some index set $\omega \subseteq \{1, \cdots, n\}$, where $\mathrm{supp}\,(\boldsymbol{u}_i)$ denotes the support (set of indices containing nonzero entries) of $\boldsymbol{u}_i$, and the cardinality of ω is $|\omega| = k$. Let

$$\Omega := \omega \times \omega,$$

which is the support of $\boldsymbol{\Sigma}_\Omega$ obeying $|\Omega| = k^2$. To summarize, $\boldsymbol{\Sigma}_\Omega$ is simultaneously low-rank (with rank r) and sparse (with sparsity level k^2). The degrees of freedom underlying $\boldsymbol{\Sigma}_\Omega$ is about $\Theta\,(kr)$.

As a useful example, denote the best k-term approximation of a vector $\boldsymbol{x}$ as $\boldsymbol{x}_\Omega = \mathrm{argmin}_{\boldsymbol{z}:\|\boldsymbol{z}\|_0=k} \|\boldsymbol{x} - \boldsymbol{z}\|_{\mathrm{F}}$. Then the rank-one matrix $\boldsymbol{\Sigma} = \boldsymbol{x}\boldsymbol{x}^T$ can be decomposed into $\boldsymbol{\Sigma}_\Omega = \boldsymbol{x}_\Omega\boldsymbol{x}_\Omega^T$ and $\boldsymbol{\Sigma}_{\mathrm{c}} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_\Omega$.

3. ALGORITHM AND THEORETICAL GUARANTEES

When $\boldsymbol{\Sigma}$ is known to be simultaneously low-rank and sparse, a natural heuristic is to attempt recovery via weighted rank minimization and ℓ_0 minimization:

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} = \mathrm{argmin}_{\boldsymbol{M}} \ & \mathrm{rank}(\boldsymbol{M}) + \gamma\|\boldsymbol{M}\|_0 \\ \text{subject to} \quad & \boldsymbol{M} \succeq 0, \ \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_1 \leq \epsilon, \end{aligned} \tag{8}$$

where γ is the parameter that provides a tradeoff between the low-rank and sparse structural assumptions. However, both the rank minimization and the ℓ_0 minimization problem are in general NP-hard. Therefore, we consider an alternative convex formulation

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} = \mathrm{argmin}_{\boldsymbol{M}} \ & \mathrm{tr}(\boldsymbol{M}) + \lambda\|\boldsymbol{M}\|_1 \\ \text{subject to} \quad & \boldsymbol{M} \succeq 0, \ \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_1 \leq \epsilon. \end{aligned} \tag{9}$$

Here, λ is a regularization parameter that balances the two convex surrogates (i.e. the trace norm and the ℓ_1 norm) associated with the low-rank and sparse structural assumptions, respectively. This objective function has also been proposed in [12] for sparse phase recovery, and discussed in [15] with linear Gaussian measurements.

Our analysis ensures stable recovery of jointly low-rank and sparse covariance matrices, as stated in the following theorem.

Theorem 1. *Assume $\boldsymbol{a}_i$'s are generated i.i.d. with sub-Gaussian entries satisfying* (3). *Set λ to be any number within the interval $\left[\frac{1}{n}, \frac{1}{N_{\boldsymbol{\Sigma}}}\right]$ where*

$$N_{\boldsymbol{\Sigma}} := \max\left\{2\left\|\mathcal{P}_{\mathrm{nsd}}\left(\mathrm{sign}\left(\boldsymbol{\Sigma}_{\Omega}\right)\right)\right\|, \sqrt{\frac{k\sum_{i=1}^{r}\left\|\boldsymbol{u}_i\right\|_1^2}{2r}}\right\}. \quad (10)$$

Then with probability at least $1-\exp\left(-c_3 m\right)$, the solution $\hat{\boldsymbol{\Sigma}}$ to (9) *satisfies*

$$\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{\Omega}\right\|_{\mathrm{F}} \leq \frac{C_1}{\sqrt{r}}\left(\left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_{*}+\lambda\left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_{1}+\frac{\epsilon}{m}\right) \quad (11)$$

for all matrices $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ of the form (6), provided that

$$m > \frac{C_2}{\lambda^2} r \log n, \quad (12)$$

where C_1, C_2 and c_3 are some universal positive constants. Here, $\mathcal{P}_{\mathrm{nsd}}\left(\boldsymbol{M}\right)$ denotes the projection of $\boldsymbol{M}$ onto the negative semidefinite cone, and $\mathrm{sign}(\boldsymbol{M})$ is the sign matrix of $\boldsymbol{M}$.

Theorem 1 provides the first theoretical guarantee for stable covariance estimation under the simultaneous structure model using quadratic measurements. When the covariance matrices are simultaneously low-rank and sparse (i.e. $\boldsymbol{\Sigma}_c = \boldsymbol{0}$), and the measurements are noise-free with $\epsilon = 0$, the algorithm (9) recovers $\boldsymbol{\Sigma}$ exactly as soon as (12) is satisfied. The number m of measurements required for exact recovery depends on a quantity $N_{\boldsymbol{\Sigma}}$ as defined in (10), which consists of two terms as described below. The proof of Theorem 1 is sketched in Section 4.

3.1. Worst-Case Analysis

We first consider the general bounds. The first term $\left\|\mathcal{P}_{\mathrm{nsd}}\left(\mathrm{sign}\left(\boldsymbol{\Sigma}_{\Omega}\right)\right)\right\|$ can be bounded above by

$$\left\|\mathcal{P}_{\mathrm{nsd}}\left(\mathrm{sign}\left(\boldsymbol{\Sigma}_{\Omega}\right)\right)\right\| \leq \left\|\mathrm{sign}\left(\boldsymbol{\Sigma}_{\Omega}\right)\right\| \leq \left\|\mathrm{sign}\left(\boldsymbol{\Sigma}_{\Omega}\right)\right\|_{\mathrm{F}} = k.$$

The second term $\sqrt{k\sum_{i=1}^{r}\left\|\boldsymbol{u}_i\right\|_1^2/(2r)}$ characterizes the average compressibility of $\boldsymbol{u}_i$'s. A worst case bound can be derived as follows

$$\sqrt{\frac{k\sum_{i=1}^{r}\left\|\boldsymbol{u}_i\right\|_1^2}{2r}} \leq \sqrt{\frac{k\sum_{i=1}^{r}\left(\sqrt{k}\left\|\boldsymbol{u}_i\right\|_2\right)^2}{2r}} = k,$$

which arises when all entries of $\boldsymbol{u}_i$'s have the same magnitude. Combining the above bounds, a worst-case analysis yields an upper bound $N_{\boldsymbol{\Sigma}} \leq 2k$, which in turn implies that $m = \mathcal{O}(k^2 r \log n)$ measurements suffice for exact recovery. This indicates that there is a performance gap between the achievable performance of (9) and the theoretical limit $\mathcal{O}(kr)$. A similar observation is also made in [15] for the Gaussian linear measurement model. Furthermore, the convex algorithm (9) enables stable recovery against imperfect structural assumptions as well as against bounded noise from (11), provided that the number m of measurements is sufficiently large.

We now present a numerical example with $n = 40$. We first randomly generate ω with a support size $|\omega| = k = 6$, then generate $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\Omega} = \boldsymbol{U}\boldsymbol{U}^T$ with $\boldsymbol{U} \in \mathbb{R}^{n\times r}$. The entries of each row of $\boldsymbol{U}$ contained in the index set ω are generated with i.i.d. standard Gaussian entries with the rest of the entries set to 0. The measurement vectors are generated with i.i.d. standard Gaussian entries. Fig. 1 shows the normalized mean squared error (NMSE) $\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_{\mathrm{F}}/\|\boldsymbol{\Sigma}\|_{\mathrm{F}}$ of (9) by setting $\lambda = 1$ in the absence of noise for various ranks $r = 1, 2, 3$.

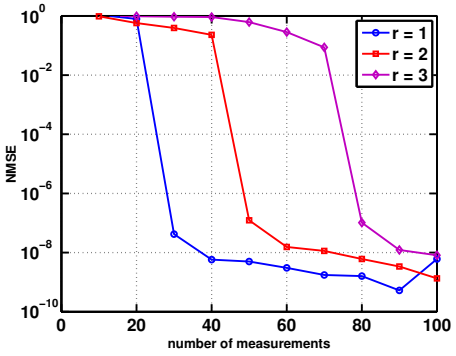

Fig. 1. The NMSE of the reconstructed covariance matrix v.s. the number of samples via (9) with $\lambda = 1$ when $n = 40$ and $k = 6$.

3.2. Near-Optimal Estimation for Compressible Models

Although there is a gap of $\mathcal{O}\left(k\log n\right)$ between the worst-case bound of Theorem 1 and the theoretic limit, in many situations we can obtain a much better bound of $N_{\boldsymbol{\Sigma}}$ that significantly improves the condition for perfect recovery.

Consider $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\Omega}$. If each row of $\mathrm{sign}\left(\boldsymbol{\Sigma}_{\Omega}\right)$ is random, then on average $\left\|\mathrm{sign}\left(\boldsymbol{\Sigma}_{\Omega}\right)\right\|$ can be as low as

$$\left\|\mathrm{sign}\left(\boldsymbol{\Sigma}_{\Omega}\right)\right\| = O\left(\sqrt{k}\,\mathrm{poly}\left(\log k\right)\right), \quad (13)$$

where $\mathrm{poly}\left(\log k\right)$ denotes any polynomial function of $\log k$. On the other hand, if the *nonzero entries* of all eigenvectors $\boldsymbol{u}_i$'s satisfy a power law decay such that the lth largest entry magnitude is bounded by $O\left(l^{-\alpha}\right)$ for some decaying exponent $\alpha \geq 1$, then one has

$$\left\|\boldsymbol{u}_i\right\|_1 = O\left(\log k\right), \quad (14)$$

which leads to $\sqrt{k\sum_{i=1}^{r}\left\|\boldsymbol{u}_i\right\|_1^2/2r} \leq \sqrt{k\log k}$.

When both (13) and (14) are satisfied, one has $N_{\boldsymbol{\Sigma}} \leq \sqrt{k}\,\mathrm{poly}\left(\log n\right)$. In this case, by setting $\lambda := 1/(\sqrt{k}\,\mathrm{poly}(\log n))$, one obtains *exact* covariance estimation as soon as m is about the order of

$$kr\,\mathrm{poly}\left(\log n\right),$$

which achieves the theoretic sampling limit except for a logarithmic factor.

3.3. Applications to Sparse Phase Retrieval

As a special case, Theorem 1 immediately leads to a theoretical guarantee for sparse signal recovery from magnitude measurements.

Specifically, suppose that the true signal $\boldsymbol{x}$ is sparse, and the goal is to recover $\boldsymbol{x}$ from a small number of phaseless measurements. The measurements we obtain can be expressed as

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

where $\boldsymbol{a}_i$ represents the random measurement vector, and the noise term $\boldsymbol{\eta}$ has bounded ℓ_1 norm $\|\boldsymbol{\eta}\|_1 \le \epsilon$. When $\boldsymbol{x}$ is exactly k-sparse, the lifted matrix $\boldsymbol{\Sigma} := \boldsymbol{x}\boldsymbol{x}^T$ has rank 1 and sparsity level k^2. For an approximately sparse signal $\boldsymbol{x}$, we let $\boldsymbol{\Sigma}_\Omega = \boldsymbol{x}_\Omega \boldsymbol{x}_\Omega^T$. Instantiating Theorem 1 to this rank-1 case, we observe that

$$\mathcal{P}_{\mathrm{nsd}}\left(\mathrm{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right) = \mathcal{P}_{\mathrm{nsd}}\left(\mathrm{sign}\left(\boldsymbol{x}_\Omega\right)\mathrm{sign}\left(\boldsymbol{x}_\Omega\right)^T\right) = \boldsymbol{0},$$

which results in $\|\mathcal{P}_{\mathrm{nsd}}\left(\mathrm{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right)\| = 0$. We then have the following corollary.

Corollary 1. *Assume the $\boldsymbol{a}_i$'s are generated i.i.d. with sub-Gaussian entries satisfying* (3)*. Set λ to be any number within the interval $\left[\frac{1}{n}, \frac{1}{\sqrt{k}}\frac{\|\boldsymbol{x}_\Omega\|_{\mathrm{F}}}{\|\boldsymbol{x}_\Omega\|_1}\right]$. Then with probability at least $1 - \exp\left(-c_3 m\right)$, the solution $\hat{\boldsymbol{X}}$ to (9) satisfies*

$$\left\|\hat{\boldsymbol{X}} - \boldsymbol{x}_\Omega \boldsymbol{x}_\Omega^T\right\|_{\mathrm{F}} \le C_1 \left\{ \left\|\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{x}_\Omega \boldsymbol{x}_\Omega^T\right\|_* + \lambda \left\|\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{x}_\Omega \boldsymbol{x}_\Omega^T\right\|_1 + \frac{\epsilon}{m} \right\} \tag{15}$$

for all signals $\boldsymbol{x} \in \mathbb{R}^n$, provided that $m > \frac{C_2 \log n}{\lambda^2}$, where C_1, C_2 and c_3 are some universal positive constants. Here, $\boldsymbol{x}_\Omega$ denotes the best k-sparse approximation of $\boldsymbol{x}$.

Corollary 1 recovers all theoretical performance guarantees of sparse phase retrieval established in [12], and improves upon them in two aspects. (i) Corollary 1 establishes the performance guarantees of the algorithm (9) when the signal is *approximately sparse* and when the measurements are *noisy*. 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. (ii) Corollary 1 established the performance guarantees for a large class of sub-Gaussian measurement vectors, while [12] only considers Gaussian measurements.

In general, by setting $\lambda = 1/k$, one can obtain *universal* recovery for all k-sparse signals from $\mathcal{O}\left(k^2 \log n\right)$ samples with exponentially high probability. Somewhat surprisingly, if the decay rate of the *nonzero entries* of a k-sparse signal $\boldsymbol{x}$ is known *a priori*, then the algorithm (9) allows near-optimal recovery. For instance, suppose that the *non-zero entries* of $\boldsymbol{x}$ satisfies the power-law decay such that the magnitude of the lth largest entry of $\boldsymbol{x}_\omega / \|\boldsymbol{x}_\omega\|_{\mathrm{F}}$ is bounded above by c_{pl}/l^α for some constants c_{pl} and exponent $\alpha > 0$. If $\alpha > 1$, then by setting $\lambda = \Theta\left(\frac{1}{\sqrt{k}\log n}\right)$, one can obtain accurate recovery from $O\left(k \log^2 n\right)$ samples under sub-Gaussian sampling, which is only a logarithmic factor from the theoretic sampling limits (which is $\Theta\left(k\right)$).

4. PROOF OUTLINE OF THEOREM 1

Due to the space limit, we only present an outline of the proof of Theorem 1. The analysis framework is established upon the novel mixed-norm restricted isometry property RIP-ℓ_2/ℓ_1 proposed in [10], for which the input and output are measured in terms of the Frobenius norm and the ℓ_1 norm, respectively. This differs subtly from the conventional restricted isometry property that measures the input and output using the same norm, and turns out to simplify the proof significantly. The RIP-ℓ_2/ℓ_1 of an operator $\mathcal{B}$ for the family of low-rank plus sparse matrices are defined as below.

Definition 1 (RIP-ℓ_2/ℓ_1 for low-rank plus sparse matrices). *For the set of matrices*

$$\mathcal{M}_{r,k} = \left\{ \boldsymbol{X}_1 + \boldsymbol{X}_2 \mid \mathit{rank}\left(\boldsymbol{X}_1\right) \le r,\ \|\boldsymbol{X}_2\|_0 \le k \right\},$$

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

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

The RIP for this class is of particular importance when the objective function is a weighted sum of $\|\cdot\|_*$ and $\|\cdot\|_1$. Note that $\mathcal{A}$ does not satisfy RIP-ℓ_2/ℓ_1. We thus introduce a set of auxiliary *zero-mean* measurement matrices

$$\boldsymbol{B}_i := \boldsymbol{A}_{2i-1} - \boldsymbol{A}_{2i}, \tag{16}$$

and let $\mathcal{B}\left(\boldsymbol{X}\right)$ represent the linear transformation that maps $\boldsymbol{X}$ to $\{\langle \boldsymbol{B}_i, \boldsymbol{X}\rangle\}_{i=1}^{\lfloor m/2 \rfloor}$. It can be shown that $\mathcal{B}$ satisfies the RIP-ℓ_2/ℓ_1 with high probability, provided that m is about the order of $r \log n/\lambda^2$. This specialized RIP-ℓ_2/ℓ_1 concept allows us to prove Theorem 1 through the following lemma.

Lemma 1. *Set λ to be any number within the interval $\left[\frac{1}{n}, \frac{1}{N_{\boldsymbol{\Sigma}}}\right]$, where $N_{\boldsymbol{\Sigma}}$ is defined in (10). If there exists a number K_1 (and $K_2 := \left\lceil \frac{K_1}{\lambda_2} \right\rceil$) such that*

$$\frac{\frac{1-\delta_{2K_1,2K_2}^{\mathrm{lb}}}{\sqrt{3}} - \frac{6\left(1+\delta_{K_1,K_2}^{\mathrm{ub}}\right)\sqrt{r}}{\sqrt{K_1}}}{\max\left\{\frac{4\left(1+\delta_{K_1,K_2}^{\mathrm{ub}}\right)\sqrt{r}}{\sqrt{K_1}}, 2\right\}} \ge \beta_3 > 0, \tag{17}$$

$$\frac{4\left(1+\delta_{K_1,K_2}^{\mathrm{ub}}\right)\sqrt{r}}{\left(1-\delta_{K_1,K_2}^{\mathrm{lb}}\right)\sqrt{K_1}} \le \frac{1}{\beta_4} \tag{18}$$

hold for some positive constants β_3 and β_4, then the solution $\hat{\boldsymbol{\Sigma}}$ to (9) satisfies

$$\left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\Omega\right\|_{\mathrm{F}} \le \frac{C}{\sqrt{r}}\left(\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_* + \lambda\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_1 + \frac{\epsilon}{m}\right) \tag{19}$$

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

The existence of β_3 and β_4 as required in (17) can be guaranteed via the establishment of RIP ℓ_2/ℓ_1 for $\mathcal{B}$.

5. CONCLUSIONS

In this paper, we investigate estimation of covariance matrices under a quadratic sampling model that yields energy measurements or magnitude measurements. We consider a popular structural model where the covariance matrix is assumed to be simultaneously sparse and low-rank, and present theoretical guarantees for a convex relaxation based estimation algorithm.

Our results indicate that jointly structured covariance matrices can be perfectly recovered from very few quadratic measurements, provided that the sampling vectors are i.i.d. drawn from sub-Gaussian distributions. Our results also demonstrate the stability and robustness of the convex program in the presence of noise and imperfect structural assumptions. Interestingly, when the nonzero entries of a covariance matrix satisfies a power-decay law, perfect recovery can be achieved as soon as the number of samples is about the order of the theoretic sampling limit. Our results subsume and improve upon the best-known performance guarantees in convex-optimization based sparse phase retrieval, with a much simpler proof.

6. REFERENCES

[1] A. Jung, G. Taubock, and F. Hlawatsch, "Compressive spectral estimation for nonstationary random processes," *IEEE Transactions on Information Theory*, vol. 59, no. 5, May 2013.

[2] R. G. Gallager, "Information theory and reliable communication," 1968.

[3] R. C. Daniels and R. W. Heath, "60 GHz wireless communications: emerging requirements and design recommendations," *IEEE Vehicular Technology Magazine*, vol. 2, no. 3, pp. 41–50, 2007.

[4] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Transactions on Signal Processing*, vol. 42, no. 8, pp. 2146–2157, 1994.

[5] M. Raymer, M. Beck, and D. McAlister, "Complex wave-field reconstruction using phase-space tomography," *Physical review letters*, vol. 72, no. 8, p. 1137, 1994.

[6] L. Tian, J. Lee, S. B. Oh, and G. Barbastathis, "Experimental compressive phase space tomography," *Optics Express*, vol. 20, no. 8, p. 8296, 2012.

[7] Y. Chi and B. L. Anderson, "Resolving spatial modes of lasers via matrix completion," *Opt. Lett.*, vol. 38, no. 19, pp. 3957–3960, Oct 2013.

[8] J. R. Fienup, "Phase retrieval algorithms: a comparison." *Applied optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[9] ——, "Reconstruction of an object from the modulus of its fourier transform," *Optics letters*, vol. 3, no. 1, pp. 27–29, 1978.

[10] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," 2013. [Online]. Available: http://arxiv.org/abs/1310.0807

[11] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of multivariate analysis*, vol. 99, no. 6, pp. 1015–1034, 2008.

[12] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM Journal on Mathematical Analysis*, 2013.

[13] A. Beck and Y. C. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," *to appear in SIAM Optimization*, 2013.

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

[15] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *arXiv preprint arXiv:1212.3753*, 2012.

[16] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289 –1306, April 2006.

[17] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[18] E. J. Candes, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, 2012.

[19] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Journal on Imaging Sciences*, vol. 6, no. 1, pp. 199–225, 2013.

[20] E. J. Candes and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Foundations of Computational Mathematics*, 2013.

[21] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," *arXiv preprint arXiv:1111.6323*, 2011.

[22] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 345–356, 2006.

[23] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, "Phase retrieval with polarization," *arXiv preprint arXiv:1210.7752*, 2012.

[24] L. Demanet and P. Hand, "Stable optimizationless recovery from phaseless linear measurements," *arXiv preprint arXiv:1208.1803*, 2012.

A. PROOF OF THEOREM 1

Before proceeding to the proof, we introduce a few notations for convenience of presentation. Suppose that the eigen-decomposition of the ground truth covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T, \tag{20}$$

where $\boldsymbol{U} := [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$ consists of all r eigenvectors. Without loss of generality, we assume that $\operatorname{supp}(\boldsymbol{u}_i) \subseteq \omega$ for some index set $\omega \subseteq \{1, \cdots, n\}$. We also denote

$$\Omega := \omega \times \omega,$$

which contains the support of all matrices $\boldsymbol{u}_i\boldsymbol{u}_i^T$ $(1 \le i \le r)$.

Let $\boldsymbol{X}_\Omega := \mathcal{P}_\Omega(\boldsymbol{X})$ and $\boldsymbol{X}_{\mathrm{c}} := \boldsymbol{X} - \mathcal{P}_\Omega(\boldsymbol{X})$. The tangent space T with respect to $\boldsymbol{X}_\Omega$ and its orthogonal complement $T^\perp$ are denoted by

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

We adopt the notations introduced in [12], namely, decomposing the entire space of symmetric matrices into the direct sum of 3 subspaces as

$$(T \cap \Omega) \oplus \left(T^\perp \cap \Omega\right) \oplus \left(\Omega^\perp\right). \tag{21}$$

In fact, one can verify that both the column and row spaces of $T^\perp \cap \Omega$ can be spanned by a set of $k - r$ orthonormal vectors that are supported on ω and orthogonal to $\boldsymbol{U}$. Interested readers are referred to [12] for a few useful properties of this decomposition. In particular, it has been pointed out by [12] that T and Ω are compatible in the sense that

$$\mathcal{P}_T\mathcal{P}_\Omega = \mathcal{P}_\Omega\mathcal{P}_T = \mathcal{P}_{T\cap\Omega}, \tag{22}$$

which immediately leads to the following identity as well

$$\mathcal{P}_{T^\perp}\mathcal{P}_\Omega = \mathcal{P}_\Omega\mathcal{P}_{T^\perp} = \mathcal{P}_{T^\perp\cap\Omega}. \tag{23}$$

Furthermore, we denote by $\mathcal{P}_{\mathrm{psd}}(\boldsymbol{M})$ and $\mathcal{P}_{\mathrm{nsd}}(\boldsymbol{M})$ the projection of $\boldsymbol{M}$ onto the positive semidefinite and negative semidefinite cones, respectively.

Suppose that $\hat{\boldsymbol{\Sigma}} := \boldsymbol{\Sigma} + \boldsymbol{H}$ is the solution to (9). Then for any $\boldsymbol{W} \in T^\perp$ and $\boldsymbol{Y} \in \Omega^\perp$ satisfying $\|\boldsymbol{W}\| \le 1$ and $\|\boldsymbol{Y}\|_\infty \le 1$, the matrix $\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{W} + \lambda\operatorname{sign}(\boldsymbol{\Sigma}_\Omega) + \lambda\boldsymbol{Y}$ forms a subgradient of the function $\|\cdot\|_* + \lambda\|\cdot\|_1$ at the point $\boldsymbol{\Sigma}_\Omega$. If we pick $\boldsymbol{W}$ and $\boldsymbol{Y}$ such that $\boldsymbol{Y} = \operatorname{sgn}(\boldsymbol{H}_{\Omega^\perp})$ and $\langle \boldsymbol{W}, \boldsymbol{H}\rangle = \|\boldsymbol{H}_{T^\perp\cap\Omega}\|_*$, then

$$\begin{aligned}
0 &\ge \|\boldsymbol{\Sigma} + \boldsymbol{H}\|_* + \lambda\|\boldsymbol{\Sigma} + \boldsymbol{H}\|_1 - \|\boldsymbol{\Sigma}\|_* - \lambda\|\boldsymbol{\Sigma}\|_1 && (24)\\
&\ge \|\boldsymbol{\Sigma}_\Omega + \boldsymbol{H}\|_* - \|\boldsymbol{\Sigma}_{\mathrm{c}}\|_* + \lambda\|\boldsymbol{\Sigma}_\Omega + \boldsymbol{H}\|_1 - \lambda\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_1 - \|\boldsymbol{\Sigma}_\Omega\|_* - \|\boldsymbol{\Sigma}_{\mathrm{c}}\|_* - \lambda\|\boldsymbol{\Sigma}_\Omega\|_1 - \lambda\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_1 && (25)\\
&\ge \left\langle \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{W}, \boldsymbol{H}\right\rangle + \langle \lambda\operatorname{sign}(\boldsymbol{\Sigma}_\Omega) + \lambda\boldsymbol{Y}, \boldsymbol{H}\rangle - 2\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_* - 2\lambda\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_1 && (26)\\
&= \left\langle \boldsymbol{U}\boldsymbol{U}^T, \boldsymbol{H}_T\right\rangle + \lambda\langle \mathcal{P}_T(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_T\rangle + \lambda\langle \mathcal{P}_{T^\perp}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T^\perp}\rangle\\
&\quad + \|\boldsymbol{H}_{T^\perp\cap\Omega}\|_* + \lambda\|\boldsymbol{H}_{\Omega^\perp}\|_1 - 2\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_* - 2\lambda\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_1, && (27)
\end{aligned}$$

where (24) follows from the optimality of $\boldsymbol{\Sigma} + \boldsymbol{H}$, (25) follows from the definitions of $\boldsymbol{\Sigma}_\Omega$ and $\boldsymbol{\Sigma}_c$ and the triangle inequality, (26) follows from the definition of sub gradient. By observing that $\boldsymbol{H}_{T^\perp\cap\Omega} \succeq 0$ and $\mathcal{P}_{T^\perp}\mathcal{P}_{\mathrm{psd}}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)) \succeq 0$, one can obtain

$$\begin{aligned}
\langle \mathcal{P}_{T^\perp}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T^\perp}\rangle &= \langle \mathcal{P}_{T^\perp}\mathcal{P}_{\mathrm{psd}}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T^\perp}\rangle + \langle \mathcal{P}_{T^\perp}\mathcal{P}_{\mathrm{nsd}}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T^\perp}\rangle\\
&\ge \langle \mathcal{P}_{T^\perp}\mathcal{P}_{\mathrm{nsd}}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T^\perp}\rangle = \langle \mathcal{P}_{T^\perp}\mathcal{P}_{\mathrm{nsd}}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T^\perp\cap\Omega}\rangle,
\end{aligned}$$

where the last equality follows from the property $\operatorname{supp}(\mathcal{P}_{\mathrm{nsd}}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega))) \subseteq \Omega$ and the identity (23).This together with (27) leads to

$$\begin{aligned}
0 &\ge \left\langle \boldsymbol{U}\boldsymbol{U}^T, \boldsymbol{H}_T\right\rangle + \lambda\langle \mathcal{P}_T(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_T\rangle + \lambda\langle \mathcal{P}_{T^\perp}\mathcal{P}_{\mathrm{nsd}}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T^\perp\cap\Omega}\rangle && (28)\\
&\quad + \|\boldsymbol{H}_{T^\perp\cap\Omega}\|_* + \lambda\|\boldsymbol{H}_{\Omega^\perp}\|_1 - 2\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_* - 2\lambda\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_1 && (29)\\
&\ge \left\langle \boldsymbol{U}\boldsymbol{U}^T + \lambda\mathcal{P}_T(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T\cap\Omega}\right\rangle + (1 - \lambda\|\mathcal{P}_{T^\perp}\mathcal{P}_{\mathrm{nsd}}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega))\|)\|\boldsymbol{H}_{T^\perp\cap\Omega}\|_* && (30)\\
&\qquad + \lambda\|\boldsymbol{H}_{\Omega^\perp}\|_1 - 2\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_* - 2\lambda\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_1 && (31)\\
&\ge \left\langle \boldsymbol{U}\boldsymbol{U}^T + \lambda\mathcal{P}_T(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T\cap\Omega}\right\rangle + \frac{1}{2}\|\boldsymbol{H}_{T^\perp\cap\Omega}\|_* + \frac{1}{2}\lambda\|\boldsymbol{H}_{\Omega^\perp}\|_1 - 2\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_* - 2\lambda\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_1, && (32)
\end{aligned}$$

where (32) follows from the compatibility property (22), which states that

$$\langle \mathcal{P}_T(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_T\rangle = \langle \mathcal{P}_{T\cap\Omega}(\operatorname{sign}(\boldsymbol{\Sigma}_\Omega)), \boldsymbol{H}_{T\cap\Omega}\rangle. \tag{33}$$

We can then bound $\mathcal{P}_T\left(\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right)$ via the parameter λ as follows

$$\begin{aligned}
\left\|\mathcal{P}_T\left(\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right)\right\|_{\mathrm{F}}^2 &\leq 2\left\|\boldsymbol{U}\boldsymbol{U}^T\left(\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right)\right\|_{\mathrm{F}}^2 \\
&= 2\sum_{i=1}^{r}\left\|\boldsymbol{u}_i^T\left(\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right)\right\|_{\mathrm{F}}^2 = 2\sum_{i=1}^{r}\left\langle \boldsymbol{u}_i\boldsymbol{u}_i^T, \operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right\rangle \\
&\leq 2\sum_{i=1}^{r}\left\|\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right\|_\infty\left\|\boldsymbol{u}_i\boldsymbol{u}_i^T\right\|_1 \qquad (34)\\
&\leq 2k\sum_{i=1}^{r}\left\|\boldsymbol{u}_i\boldsymbol{u}_i^T\right\|_1 \qquad (35)\\
&= 2kr\frac{\sum_{i:1\leq i\leq r}\left\|\boldsymbol{u}_i\right\|_1^2}{r} \leq \frac{4r}{\lambda^2}, \qquad (36)
\end{aligned}$$

where (36) follows from the assumption on λ. Also, (35) follows from Cauchy-Schwartz inequality, by observing that the ℓ_2 norm of each row / column of $\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)$ is bounded by $\sqrt{k}$. Combining (36) with (32) yields

$$\begin{aligned}
\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_* + \lambda\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1 &\leq -2\left\langle \boldsymbol{U}\boldsymbol{U}^T, \boldsymbol{H}_{T\cap\Omega}\right\rangle - 2\lambda\left\langle \mathcal{P}_T\left(\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right), \boldsymbol{H}_{T\cap\Omega}\right\rangle + 4\left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_* + 4\lambda\left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_1 \\
&\leq 2\left\|\boldsymbol{U}\boldsymbol{U}^T\right\|_{\mathrm{F}}\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\mathrm{F}} + 2\lambda\left\|\mathcal{P}_T\left(\operatorname{sign}\left(\boldsymbol{\Sigma}_\Omega\right)\right)\right\|_{\mathrm{F}}\cdot\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\mathrm{F}} + 4\left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_* + 4\lambda\left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_1 \\
&\leq 6\sqrt{r}\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\mathrm{F}} + 4\left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_* + 4\lambda\left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_1, \qquad (37)
\end{aligned}$$

where (37) results from $\left\|\boldsymbol{U}\boldsymbol{U}^*\right\|_{\mathrm{F}} = \sqrt{r}$ and (36).

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

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

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

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

This decomposition gives rise to the following bound

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

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

$$\sum_{i=2}^{M_1}\frac{1}{m}\left\|\mathcal{B}\left(\boldsymbol{H}_{T^\perp\cap\Omega}^{(i)}\right)\right\|_1 \leq \frac{\left(1+\delta_{K_1,K_2}^{\mathsf{ub}}\right)}{\sqrt{K_1}}\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_*. \qquad (38)$$

Similarly, one has

$$\sum_{i=2}^{M_2}\frac{1}{m}\left\|\mathcal{B}\left(\boldsymbol{H}_{\Omega^\perp}^{(i)}\right)\right\|_1 \leq \frac{\left(1+\delta_{K_1,K_2}^{\mathsf{ub}}\right)}{\sqrt{K_2}}\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1. \qquad (39)$$

By setting $K_2 := \lceil \frac{K_1}{\lambda^2} \rceil$, 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

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

where the last inequality follows from the fact that $\left\langle \boldsymbol{H}_{T\cap\Omega}, \boldsymbol{H}^{(1)}_{T^\perp\cap\Omega}\right\rangle = \left\langle \boldsymbol{H}_{T\cap\Omega}, \boldsymbol{H}^{(1)}_{\Omega^\perp}\right\rangle = \left\langle \boldsymbol{H}^{(1)}_{\Omega^\perp}, \boldsymbol{H}^{(1)}_{T^\perp\cap\Omega}\right\rangle = 0$. This taken collectively with (37) gives

$$
\begin{aligned}
\frac{2\epsilon}{m} \geq{}& \frac{\left(1-\delta^{\text{lb}}_{2K_1,2K_2}\right)}{\sqrt{3}}\left(\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\text{F}} + \left\|\boldsymbol{H}^{(1)}_{T^\perp\cap\Omega}\right\|_{\text{F}} + \left\|\boldsymbol{H}^{(1)}_{\Omega^\perp}\right\|_{\text{F}}\right) \\
&- \frac{\left(1+\delta^{\text{ub}}_{K_1,K_2}\right)}{\sqrt{K_1}}\left(6\sqrt{r}\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\text{F}} + 4\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_* + 4\lambda\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_1\right),
\end{aligned}
$$

which can be rearranged as

$$
\begin{aligned}
&\frac{4\left(1+\delta^{\text{ub}}_{K_1,K_2}\right)}{\sqrt{K_1}}\left(\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_* + \lambda\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_1\right) + \frac{2\epsilon}{m} \\
&\geq \left(\frac{1-\delta^{\text{lb}}_{2K_1,2K_2}}{\sqrt{3}} - \frac{6\left(1+\delta^{\text{ub}}_{K_1,K_2}\right)\sqrt{r}}{\sqrt{K_1}}\right)\left(\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\text{F}} + \left\|\boldsymbol{H}^{(1)}_{T^\perp\cap\Omega}\right\|_{\text{F}} + \left\|\boldsymbol{H}^{(1)}_{\Omega^\perp}\right\|_{\text{F}}\right).
\end{aligned}
$$

Therefore, if we know that

$$
\frac{\frac{1-\delta^{\text{lb}}_{2K_1,2K_2}}{\sqrt{3}} - \frac{6\left(1+\delta^{\text{ub}}_{K_1,K_2}\right)\sqrt{r}}{\sqrt{K_1}}}{\max\left\{\frac{4\left(1+\delta^{\text{ub}}_{K_1,K_2}\right)\sqrt{r}}{\sqrt{K_1}}, 2\right\}} \geq \beta_3 > 0
$$

for some absolute constant β_3, then

$$
\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\text{F}} + \left\|\boldsymbol{H}^{(1)}_{T^\perp\cap\Omega}\right\|_{\text{F}} + \left\|\boldsymbol{H}^{(1)}_{\Omega^\perp}\right\|_{\text{F}} \leq \frac{1}{\beta_3\sqrt{r}}\left(\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_* + \lambda\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_1 + \frac{\epsilon}{m}\right). \tag{40}
$$

On the other hand, we know from (38) and (39) that

$$
\begin{aligned}
\sum_{i=2}^{M_1}\left\|\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right\|_{\text{F}} + \sum_{i=2}^{M_2}\left\|\boldsymbol{H}^{(i)}_{\Omega^\perp}\right\|_{\text{F}} &\leq \frac{1}{1-\delta^{\text{lb}}_{K_1,K_2}}\sum_{i=2}^{M_1}\left\|\mathcal{B}\left(\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right)\right\|_1 + \sum_{i=2}^{M_2}\left\|\mathcal{B}\left(\boldsymbol{H}^{(i)}_{\Omega^\perp}\right)\right\|_1 \\
&\leq \frac{1+\delta^{\text{ub}}_{K_1,K_2}}{\left(1-\delta^{\text{lb}}_{K_1,K_2}\right)\sqrt{K_1}}\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_* + \frac{\left(1+\delta^{\text{ub}}_{K_1,K_2}\right)}{\left(1-\delta^{\text{lb}}_{K_1,K_2}\right)\sqrt{K_2}}\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1 \\
&= \frac{1+\delta^{\text{ub}}_{K_1,K_2}}{\left(1-\delta^{\text{lb}}_{K_1,K_2}\right)\sqrt{K_1}}\left(\left\|\boldsymbol{H}_{T^\perp\cap\Omega}\right\|_* + \lambda\left\|\boldsymbol{H}_{\Omega^\perp}\right\|_1\right) \\
&\leq \frac{1+\delta^{\text{ub}}_{K_1,K_2}}{\left(1-\delta^{\text{lb}}_{K_1,K_2}\right)\sqrt{K_1}}\left(6\sqrt{r}\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\text{F}} + 4\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_* + 4\lambda\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_1\right),
\end{aligned}
$$

where the last inequality arises from (37). Summing this with (40) gives

$$
\begin{aligned}
&\left(1 - \frac{6\left(1+\delta^{\text{ub}}_{K_1,K_2}\right)\sqrt{r}}{\left(1-\delta^{\text{lb}}_{K_1,K_2}\right)\sqrt{K_1}}\right)\left\|\boldsymbol{H}_{T\cap\Omega}\right\|_{\text{F}} + \sum_{i=1}^{M_1}\left\|\boldsymbol{H}^{(i)}_{T^\perp\cap\Omega}\right\|_{\text{F}} + \sum_{i=1}^{M_2}\left\|\boldsymbol{H}^{(i)}_{\Omega^\perp}\right\|_{\text{F}} \\
&\leq \left(\frac{4\left(1+\delta^{\text{ub}}_{K_1,K_2}\right)}{\left(1-\delta^{\text{lb}}_{K_1,K_2}\right)\sqrt{K_1}} + \frac{1}{\beta_3\sqrt{r}}\right)\left(\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_* + \lambda\left\|\boldsymbol{\Sigma}_{\text{c}}\right\|_1 + \frac{\epsilon}{m}\right).
\end{aligned}
$$

Therefore, if we further have

$$\frac{4\left(1+\delta^{\mathsf{ub}}_{K_1,K_2}\right)\sqrt{r}}{\left(1-\delta^{\mathsf{lb}}_{K_1,K_2}\right)\sqrt{K_1}} \leq \frac{1}{\beta_4}$$

for some constant $\beta_4 > 0$, then one has

$$\|\boldsymbol{H}\|_{\mathrm{F}} \leq c_{10}\frac{1}{\sqrt{r}}\left(\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_{*} + \lambda\|\boldsymbol{\Sigma}_{\mathrm{c}}\|_{1} + \frac{\epsilon}{m}\right)$$

for some constant c_{10} that depends only on β_3 and β_4. This completes the proof.