Sample-Efficient Algorithms for Recovering Structured Signals from Magnitude-Only Measurements

Gauri Jagatap and Chinmay Hegde

***Abstract*—We consider the problem of recovering a signal $\mathbf{x}^* \in \mathbb{R}^n$, from magnitude-only measurements, $y_i = |\langle \mathbf{a}_i, \mathbf{x}^* \rangle|$ for $i = \{1, 2, \ldots, m\}$. This is a stylized version of the classical *phase retrieval problem*, and is a fundamental challenge in nano- and bio-imaging systems, astronomical imaging, and speech processing. It is well known that the above problem is ill-posed, and therefore some additional assumptions on the signal and/or the measurements are necessary.**

In this paper, we consider the case where the underlying signal $\mathbf{x}^*$ is s-sparse. For this case, we develop a novel recovery algorithm that we call *Compressive Phase Retrieval with Alternating Minimization*, or *CoPRAM*. Our algorithm is simple and be obtained via a natural combination of the classical alternating minimization approach for phase retrieval with the CoSaMP algorithm for sparse recovery. Despite its simplicity, we prove that our algorithm achieves a sample complexity of $\mathcal{O}\left(s^2 \log n\right)$ with Gaussian measurements $\mathbf{a}_i$, which matches the best known existing results; moreover, it also demonstrates linear convergence in theory and practice. An appealing feature of our algorithm is that it requires no extra tuning parameters other than the signal sparsity level s. Moreover, we show that our algorithm is robust to noise.

The quadratic dependence of sample complexity on the sparsity level is sub-optimal, and we demonstrate how to alleviate this via *additional* assumptions beyond sparsity. First, we study the (practically) relevant case where the sorted coefficients of the underlying sparse signal exhibit a power law decay. In this scenario, we show that the CoPRAM algorithm achieves a sample complexity of $\mathcal{O}\left(s \log n\right)$, which is close to the information-theoretic limit.

We then consider the case where the underlying signal $\mathbf{x}^*$ arises from *structured* sparsity models. We specifically examine the case of *block-sparse* signals with uniform block size of b and block sparsity $k = s/b$. For this problem, we design a recovery algorithm that we call *Block CoPRAM* that further reduces the sample complexity to $\mathcal{O}\left(ks \log n\right)$. For sufficiently large block lengths of $b = \Theta(s)$, this bound equates to $\mathcal{O}\left(s \log n\right)$.

To our knowledge, our approach constitutes the first family of *linearly convergent* algorithms for signal recovery from magnitude-only Gaussian measurements that exhibit a sub-quadratic dependence on the signal sparsity level.

***Index Terms*—Phase retrieval, sparsity, non-convex optimization, alternating minimization, structured sparsity, block-sparsity.**

The authors are with the Electrical and Computer Engineering Department at Iowa State University, Ames, IA 50010. Email: {gauri,chinmay}@iastate.edu. The authors thank Piotr Indyk, Thanh Nguyen, and the anonymous reviewers of this manuscript for their helpful feedback. A conference version of this manuscript appeared in the Annual Conference on Neural Information Processing Systems (NIPS) in December 2017 [1]; this version has an expanded set of theoretical results as well as numerical experiments. This work was supported in part by the National Science Foundation under the grants CCF-1566281, CCF-1750920, and CCF-1815101.

I. Introduction

A. Motivation

In this paper, we consider the problem of recovering a high-dimensional vector $\mathbf{x}^* \in \mathbb{R}^n$ from (possibly noisy) *magnitude-only* linear measurements (or samples). That is, for $\mathbf{a}_i \in \mathbb{R}^n$, if

$$y_i \;=\; \left|\langle \mathbf{a_i}, \mathbf{x}^* \rangle\right|, \quad \text{for } i = 1, \ldots, m, \tag{1}$$

then the task is to recover $\mathbf{x}^*$ using the samples $\mathbf{y}$ and the matrix $\mathbf{A} = \left[\mathbf{a_1}\ \mathbf{a_2}\ \ldots\ \mathbf{a_m}\right]^\top$.

Problems of this kind arise in numerous scenarios in machine learning, imaging, and statistics. For example, the classical problem of *phase retrieval* is encountered in imaging systems including diffraction imaging, X-ray crystallography, ptychography, and astronomical imaging [2], [3], [4], [5], [6]. For such systems, the physics of light acquisition are such that the optical sensors can only record the intensity of the light waves but not its phase. In terms of our setup, the vector $\mathbf{x}^*$ would correspond to an image (possessing a resolution of n pixels) and the measurements correspond to the magnitudes of its 2D Fourier transform coefficients. The goal is to stably recover the image $\mathbf{x}^*$ from the measurements, ideally with as few observations (i.e., as small m as possible).

Despite the prevalence of several heuristic approaches [7], [8], [9], [10] to solve (1), it is generally accepted that (1) is a very challenging nonlinear, ill-posed inverse problem both in theory and practice. Indeed, for generic $\mathbf{a}_i$ and $\mathbf{x}^*$, one can show that (1) is *NP-hard* by reduction from certain well-known combinatorial problems [11]. Therefore, additional assumptions on the vector $\mathbf{x}^*$ and/or the measurements $\mathbf{a}_i$ are necessary.

A recent line of breakthrough results [12], [13], [14] have provided provably efficient algorithmic procedures for the special case where the measurement vectors are randomly drawn from certain multi-variate probability distributions (such as i.i.d. Gaussian distributions). By convention, we will continue to term such methods as "phase retrieval" algorithms. However, all these newer results require an "overcomplete" set of observations, i.e., the number of observations m exceeds the problem dimension n, sometimes by a significant amount. This requirement can pose severe limitations on computation, storage, and processing the measurements, particularly in the high-dimensional regime when m and n are very large. Our focus in this paper is to address the following:

***Challenge**: Can we solve the phase retrieval problem using very few samples (in particular, significantly fewer samples than the problem dimension)?*

A possible solution to the above challenge is to leverage the fact that in many practical applications, $\mathbf{x}^*$ often obeys certain *low-dimensional* structural assumptions. A common structural assumption used in imaging applications is that $\mathbf{x}^*$ is *s-sparse* in some known basis, such as the identity or the wavelet basis. For transparency, we assume that the sparsity basis is the canonical basis throughout this paper, unless otherwise specified. Similar structural assumptions form the core of sparse recovery, compressive sensing, and streaming algorithms [15], [16], [17], and it has been established that only $\mathcal{O}\left(s \log \frac{n}{s}\right)$ samples are necessary for stable recovery of $\mathbf{x}^*$; moreover, the dependence of the number of samples on n and s is information-theoretically optimal [18].

Solving the sparsity-constrained version of (1) (sometimes referred to as *sparse phase retrieval* in the literature) is therefore a natural next step, and numerous approaches have been proposed in this regard. These include a variant of alternating minimization [14], methods based on convex relaxation [19], [20], [21], and iterative thresholding-based techniques [22], [23]. However, all existing methods suffer from one (or more) of the following drawbacks:

1) Somewhat curiously, all of the above algorithms incur a sample complexity of $\Omega(s^2 \log n)$ for stable recovery, which is *quadratically worse* than the information-theoretic limit of $\mathcal{O}\left(s \log \frac{n}{s}\right)$[1].
2) Most algorithms suffer from a running time that is quadratic (or worse) in the dimension of the signal [19], [22].
3) Many algorithms require stringent assumptions on the minimum (absolute) value of the nonzero signal coefficients [14], [23].
4) Typically, these algorithms require tuning of several parameters for their proper functioning [19], [22], [23].

Finally, in specific applications, more refined structural assumptions on $\mathbf{x}^*$ *beyond sparsity* are applicable. For example, point sources in astronomical images often produce *clusters* of nonzero pixels in a given image, while wavelet coefficients of natural images often can be organized as connected *sub-trees*. Algorithms that leverage such *structured sparsity* assumptions have been shown to achieve considerably improved sample-complexity in statistical learning and sparse recovery problems [30], [31], [32]. Indeed, a plethora of algorithms for modeling several types of structured sparsity constraints, including block-sparsity [31], [33], tree sparsity [34], [31], [35], [36], clusters [37], [33], [38], and graph-based models [39], [38], [40]. However, a systematic approach that leverage structured sparsity models in the context of phase retrieval does not seem to have been studied in the literature.

[1]Exceptions to this rule are the approaches of [24], [25], [26], [27], [28], [29], which indeed achieve near-optimal sample complexity and/or running time; however, these schemes are applicable only for very carefully designed measurements a_i.

B. Our contributions

In this paper, we establish a flexible algorithmic framework that systematically leverages (structured) sparsity-based signal models for the phase retrieval problem. We rigorously show that our approach matches the best available state-of-the-art sparse phase retrieval[2] methods both from a statistical as well as computational viewpoint. Next, we show that it is possible to extend this algorithm to the case where the signal obeys certain types of block-sparsity structures, thereby *further* lowering the sample complexity of stable signal recovery.

1) We first consider the standard case where the underlying signal $\mathbf{x}^*$ is s-sparse (i.e. it has underlying model $\mathcal{M}_s$, where $\mathcal{M}_s$ consists of all s-sparse vectors of dimension n). For this case, we develop a novel recovery algorithm that we call *Compressive Phase Retrieval with Alternating Minimization*, or *CoPRAM*. Our algorithm is simple and be obtained via a natural combination of the classical alternating minimization approach for phase retrieval with the CoSaMP algorithm for sparse recovery, together with a smart initialization. Despite its simplicity, we prove that our algorithm achieves a sample complexity of $\mathcal{O}\left(s^2 \log n\right)$ with Gaussian sampling vectors $\mathbf{a}_i$ in order to achieve linear convergence, which matches the best among all available existing results. An appealing feature of our algorithm is that it requires no extra *a priori* information or tuning parameters other than the signal sparsity level s, and that it requires no assumptions whatsoever on the nonzero signal coefficients. Finally, we show that our algorithm is stable with respect to noise in the measurements. To our knowledge, this is the first algorithm for sparse phase retrieval that simultaneously achieves all of the above properties[3].
2) Next, following the setup of [20], we consider the case where the signal coefficients exhibit a *power-law decay*. Specifically, without loss of generality, suppose that the indices of $\mathbf{x}^*$ are such that $|x_1^*| \geq |x_2^*| \geq \ldots |x_s^*| > 0$ and $x_j^{*2} \leq \frac{C(\alpha)}{j^\alpha}$. Then, we can prove that our CoPRAM algorithm exhibits a sample complexity of $m > \mathcal{O}\left(s \log n\right)$, which is very close to the information theoretic limit.
3) Finally, we consider the case where the underlying signal $\mathbf{x}^*$ belongs to an *a priori* specified structured sparsity model. We specifically examine the case of *block-sparse* signals with uniform block size b (i.e., the s non-zeros can be equally grouped into $k = s/b$ non-overlapping blocks). We can equivalently say that the signal $\mathbf{x}^*$ has underlying model $\mathcal{M}_s^b$. For this problem, we design a recovery algorithm that we call *Block CoPRAM*. We analyze this algorithm and show that leveraging block-structure further reduces the sample complexity of stable recovery to $\mathcal{O}\left(ks \log n\right)$. For sufficiently large block lengths of $b = \omega(s)$, or block sparsity $k \approx 1$, this bound

[2]Note that we use the terms *sparse phase retrieval* and *compressive phase retrieval* interchangeably throughout the course of this paper.

[3]During peer review, it was pointed out to us that a careful combination of the techniques from [22] and [23] with additional analysis will also lead to similar results.

equates to $\mathcal{O}(s \log n)$ which, again, is very close to the information theoretic limit. We also demonstrate that the more challenging case of *overlapping* blocks can also be solved using our technique, with a constant factor increase in sample complexity.

To our knowledge, this constitutes the first *linearly convergent* series of algorithms for phase retrieval where the (Gaussian) sample complexity has a sub-quadratic dependence on the sparsity level of the signal. A comparative description of the performance of our algorithms is presented in Table I.

C. Techniques

Sparse phase retrieval. Our proposed CoPRAM algorithm is conceptually very simple. It integrates existing approaches in stable sparse recovery (specifically, the CoSaMP algorithm [41]) for sparse signal estimation with the alternating minimization approach for phase retrieval proposed in [14][4].

A similar integration of sparse recovery with alternating minimization was also introduced in the work of [14]; however, their approach only succeeds when the true support of the underlying signal is accurately identified during initialization. This can be fairly challenging to achieve in realistic situations when the signal coefficients are of differing magnitudes. Instead, CoPRAM permits the support of the estimate to evolve across iterations, and therefore can iteratively "correct" for any errors made during the initialization. Moreover, their analysis requires using fresh samples for every new update of the estimate, while our analysis succeeds in the (more practically useful) setting of using all the available samples at our disposal.

Our first challenge is to identify a good initial guess of the signal. As is the case with most non-convex algorithmic techniques, CoPRAM requires an initial estimate $\mathbf{x^0}$ that is relatively close to the true vector $\mathbf{x}^*$. To this end, we use a variant of the spectral initialization procedure previously proposed in [23]. The basic idea is to identify "important" co-ordinates by constructing suitable biased estimators of each signal coefficient, followed by a specific eigendecomposition. However, the initialization in CoPRAM is far simpler than the approach in [23]; we perform no pre-processing of the measurements and our method requires no tuning parameters other than the sparsity level s. We also provide a novel analysis of this modified initialization procedure. A drawback of the theoretical results of [23] is that they impose a minimum requirement on every non-zero entry of the true vector $\mathbf{x}^*$: $x^*_{min} \equiv \min_{j \in S} |x^*_j| = C \|\mathbf{x}^*\|_2 / \sqrt{s}$. However, this assumption is equivalent to supposing that all nonzero coefficients are approximately the same magnitude, which can be unrealistic; in the case of real-world signal and image data, the (sorted) coefficients usually obey a power-law decay, which violates these constraints. Our analysis removes this requirement; the high level idea in our approach is to show that a coarse estimate of the support will suffice, since any errors in support identification necessarily have to coincide with small coefficients. Our approach also differs from the method adopted in [22], which selects indices corresponding to large coefficients based on a parameter-dependent threshold value. The support estimation step of our algorithm, coupled with the spectral decomposition method in [22] gives us a suitable initialization. We prove that the sample complexity for achieving this initial estimate of $\mathbf{x}^*$ is $\mathcal{O}\left(s^2 \log n\right)$, matching that of the best available previous methods.

Our next challenge is to show that starting from a good initial guess, an alternating procedure that switches between estimating the phases and estimating the sparse signal (using CoSaMP) converges rapidly to the desired solution. To this end, we unpack the analysis of the CoSaMP algorithm provided in [41]. In particular, we show that any "phase errors" made in the initialization step can be suitably controlled across different estimates. As a key step in our analysis, we leverage a recent result by [42] that shows sufficient decrease in the signal estimation error using the generic chaining technique of [43], [44]. Here too, our algorithm requires no tuning parameters other than the sparsity level s.

Block-sparse phase retrieval. We can then use CoPRAM to establish its extension Block CoPRAM, which is a novel phase retrieval strategy for block sparse signals, which have been sampled using generic Gaussian measurements. Again, the algorithm is based on a suitable initialization followed by an alternating minimization procedure, and the algorithmic steps exactly mirror those of CoPRAM. To our knowledge, this is the first results for phase retrieval under more refined structured sparsity assumptions on the signal.

As above, the first challenge is to identify a good initial guess of the solution in the first stage. We proceed as in CoPRAM, but instead of identifying important co-ordinates, we instead isolate *blocks* of nonzero coordinates. The high level idea is to construct a different, specially chosen biased estimator for the "mass" of each block. We prove that a good initialization can be achieved using this procedure using only $\mathcal{O}(ks \log n)$ generic measurements. When the block-size is large enough, the sample complexity of the initialization can be *sub-quadratic* in the sparsity s. Specifically, for $b = \Theta(s)$ the sample complexity is only a logarithmic factor away from the information-theoretic limit $\mathcal{O}(s)$.

The second challenge is to demonstrate rapid descent to the desired solution in the second stage. To this end, we replace the CoSaMP sub-routine in CoPRAM with the *model-based CoSaMP* algorithm of [31], specialized to block-sparse recovery. The analysis proceeds analogously as above. To our knowledge, this constitutes the first end-to-end linearly convergent algorithm for phase retrieval (with generic Gaussian measurements) that demonstrates a sub-quadratic dependence on the sparsity level of the signal.

D. Paper organization

The remainder of the paper is organized as follows. In Section II we provide a brief overview of prior work. In Section III, we present preliminaries and notation used for our analysis. In Sections IV and V, we introduce the CoPRAM and Block-CoPRAM algorithms respectively, and provide a

[4]It is worthwhile to note that the high level idea of alternately estimating the phase and the signal is classical, dating back to the work of Gerchberg and Saxton [7].

TABLE I: *Comparison of our proposed methods with existing approaches for sparse phase retrieval using Gaussian measurements. Here, n denotes signal length, s denotes sparsity, and $k = s/b$ denotes block-sparsity. $\mathcal{O}_\epsilon(\cdot)$ hides polylogarithmic dependence on $\frac{1}{\epsilon}$.*

Algorithm	Sample complexity	Running time	Assumptions	Parameters
AltMinSparse [14]	$\mathcal{O}_\epsilon\left(s^2 \log n + s^2 \log^3 s\right)$	$\mathcal{O}_\epsilon\left(s^2 n \log n\right)$	$x^*_{\min} \approx \frac{c}{\sqrt{s}} \|\mathbf{x}^*\|_2$	none
ℓ_1-PhaseLift [19]	$\mathcal{O}\left(s^2 \log n\right)$	$\mathcal{O}\left(\frac{n^3}{\epsilon^2}\right)$	none	none
Thresholded WF [22]	$\mathcal{O}\left(s^2 \log n\right)$	$\mathcal{O}_\epsilon\left(n^2 \log n\right)$	none	stepsize μ, thresholds α, β
SPARTA [23]	$\mathcal{O}\left(s^2 \log n\right)$	$\mathcal{O}_\epsilon\left(s^2 n \log n\right)$	$x^*_{\min} \approx \frac{c}{\sqrt{s}} \|\mathbf{x}^*\|_2$	stepsize μ, threshold γ
CoPRAM (this paper) CoPRAM (this paper)	$\mathcal{O}\left(s^2 \log n\right)$ $\mathcal{O}\left(s \log n\right)$	$\mathcal{O}_\epsilon\left(s^2 n \log n\right)$ $\mathcal{O}_\epsilon\left(sn \log n\right)$	none power-law decay	none none
Block CoPRAM (this paper)	$\mathcal{O}\left(ks \log n\right)$	$\mathcal{O}_\epsilon\left(ksn \log n\right)$	none	none

theoretical analysis of their statistical as well as computational performance. In Section VI we provide a series of numerical experiments demonstrating the performance of our algorithms, and in Section VII we provide concluding remarks.

II. Related work

A. Prior work

The phase retrieval problem has received significant attention in the past few years. Attempts to solve this problem have mainly fallen into one of two broad solution approaches: convex and non-convex.

Convex approaches involve linearizing the problem by lifting the signal $\mathbf{x}^*$ into a higher-dimensional space and solving a constrained optimization problem. Popular methodologies to solve the problem in the lifted framework include the seminal *PhaseLift* approach and its variations [12], [45], [46]; along similar lines is the PhaseCut approach [47] which proposes a phase retrieval approach based on an SDP relaxation of the MaxCut problem. However, most lifting based approaches suffer severely in terms of computational complexity. A recent convex approach that does not use the lifting procedure is *PhaseMax*, which produces a novel relaxation of the phase retrieval problem similar to basis pursuit [48]. While theoretically sound, the empirical performance of PhaseMax is not competitive with other lifting-based approaches.

On the other hand, nonconvex algorithms typically consist of two stages: finding a good initial point, followed by minimizing a loss function via a procedure similar to gradient-descent. The loss function being minimized can be either a function of a quadratic form involving the unknown signal, or a function involving the modulus of the inner products with the signal with the measurement vectors. Approaches based on Wirtinger Flow (WF) [13], [49], [22], [50] popularly use the quadratic form, while approaches based on Amplitude Flow (AF) [51], [23] as well as stochastic algorithms based on the Kaczmarz method [52] use the modulus form. In [53] Sun *et al.* describe a polynomial-time trust-region algorithm that uses arbitrary initializations to find the global optimum.

Recent works have adapted the phase retrieval framework for the case when the underlying signal is sparse. Some of the convex approaches include [19], [54], which uses a combination of trace-norm and ℓ-norm relaxation, to solve an SDP problem. Constrained sensing vectors have been used by Bahmani and Romberg [25] to effectively de-couple the problem into a phase retrieval stage followed by a sparse signal recovery stage, at optimal sample complexity of $\mathcal{O}\left(s \log \frac{n}{s}\right)$. Fourier measurements have been studied extensively in [55] in the convex setting. Similarly, non-convex approaches for sparse phase retrieval include [14], [23], [22] which achieve sample complexities of $\mathcal{O}\left(s^2 \log n\right)$. Fourier measurements have been evaluated in the non-convex setting in [56], where they use a local search method to solve the sparse phase retrieval problem. Separate from this line of work, Schniter and Rangan in [57] have proposed an approximate message passing algorithm to experimentally exhibit a sample complexity of $\mathcal{O}\left(s \log \frac{n}{s}\right)$.

Going beyond sparsity, a natural approach is to *refine* the assumptions on the nonzero signal coefficients so as to better model various types of real-world phenomena. Structured sparsity models have been proposed to leverage combinatorial interactions such as groups, blocks, clusters, trees, and various other refinements that can be used to model the signal of interest. Applications of structured sparsity models have been developed for sparse recovery [31], [32], [39], [38], [40], [58], [34], [31], [35], [36] as well as in high-dimensional optimization and statistical learning [30], [33]. However, to the best of our knowledge, there has been no rigorous results that explore the impact of structured sparsity models for the phase retrieval problem.

B. Subsequent work

Since the initial appearance of this paper on Arxiv and its subsequent conference publication [1], numerous related works have emerged. Of notable interest has been Waldsperger's new result for phase retrieval which shows that standard alternating minimization provably converges *without* any special initialization, albeit with much higher sample complexity [59]. At the moment, we do not know how to design initialization-free algorithms in the case of *sparse* phase retrieval, and leave that as potential future work. Other

recent works include re-weighted amplitude flow [60] and convolutional phase retrieval [61].

From a theoretical standpoint, the motivation behind choosing to analyze Gaussian measurements is well justified by the numerous papers in literature [20], [49], [22]. Typically Gaussian and sub-Gaussian measurements have proven to be easier to analyze in terms of establishing theoretical guarantees, and random Fourier-based variants [49] exist in literature that are able to emulate the performance of these results and are more realistic. However, from an application standpoint, the framework proposed in this paper has spurred follow-up work in an optical imaging application called Fourier ptychography with promising improvements in sample complexity by utilizing the underlying structure of signals; see [62], [63] for details.

Additionally, we were also able to extend our analysis for phase retrieval of structured sparse signals for a general class of sparsity models, with a specific extension to rooted tree sparse signals [64].

III. Preliminaries

We introduce some notation that will be used throughout the paper. We use bold capital-case letters ($\mathbf{A}, \mathbf{P}$, etc.) to denote matrices, bold small-case letters ($\mathbf{x}, \mathbf{y}$, etc.) to denote vectors, and non-bold letters (α, c etc.) for scalars. We use $\mathbf{x}^\top$ and $\mathbf{A}^\top$ to denote the transpose of the vector $\mathbf{x}$ and the matrix $\mathbf{A}$ respectively. The diagonal matrix form of a column vector $\mathbf{y} \in \mathbb{R}^m$ is represented as $\text{diag}(\mathbf{y})$, which is the matrix in $\mathbb{R}^{m\times m}$ with its diagonal elements as $\mathbf{y}$ and all off-diagonal entries are zero. The cardinality of set S is expressed using the operator $\text{card}(S)$.

In this paper we use the standard Gaussian (or normal) distribution over $\mathbb{R}^n$ (i.e., the elements of $\mathbf{a}$ are distributed according to the distribution $\mathcal{N}(0,1)$). The vector ℓ_2 norm is defined as $\|\mathbf{x}\|_2$, for a vector $\mathbf{x}$. However, if the argument is a matrix $\mathbf{M}$, then, $\|\mathbf{M}\|_2$ denotes the *spectral* norm of the matrix. We define $\text{sign}(x) \equiv \frac{x}{|x|}$ for every $x \in \mathbb{R}, x \neq 0$, with the convention that $\text{sign}(0) = 0$. The distance between two vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$, can be expressed in terms of operator $\text{dist}(\mathbf{x}_1, \mathbf{x}_2) \equiv \min(\|\mathbf{x}_1 - \mathbf{x}_2\|_2, \|\mathbf{x}_1 + \mathbf{x}_2\|_2)$[5]. The projection of vector $\mathbf{x} \in \mathbb{R}^n$ onto a set of coordinates S is represented as $\mathbf{x}_S \in \mathbb{R}^n$, $x_{Sj} = x_j$ for $j \in S$, and 0 elsewhere. Projection of matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ onto S is $\mathbf{M}_S \in \mathbb{R}^{n\times n}$, $M_{Sij} = M_{ij}$ for $i, j \in S$, and 0 elsewhere. For the sake of improved computational complexity, for algorithmic implementations, $\mathbf{x}_S$ can be assumed to be a truncated vector $\mathbf{x} \in \mathbb{R}^s$, discarding all elements in S^c. The element-wise product (also called Hadamard product) of two vectors $\mathbf{x}_1$ and $\mathbf{x}_2 \in \mathbb{R}^n$ is represented as $\mathbf{x}_1 \circ \mathbf{x}_2$. C has been used to denote unspecified constants that are *large enough*. Similarly δ has been used for small constants. The abbreviations *wlog* and *whp* denote "without loss of generality" and "with high probability" respectively.

[5]For complex signals, the distance between two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ can be expressed as $\text{dist}(\mathbf{x}_1, \mathbf{x}_2) \equiv \min_{\varphi\in[0,2\pi)} \|\mathbf{x}_1 - e^{i\varphi}\cdot\mathbf{x}_2\|_2$ where $e^{i\varphi}$ is a global phase error. In this paper, since we specifically study real-valued measurements, φ takes discrete values, $\varphi \in \{0, \pi\}$.

Algorithm 1 CoPRAM: Initialization.

input $\mathbf{A}, \mathbf{y}, s$

Compute signal power: $\phi^2 = \frac{1}{m}\sum_{i=1}^{m} y_i^2$.

Compute signal marginals: $M_{jj} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 a_{ij}^2 \quad \forall j$.

Set $\hat{S} \leftarrow j$'s corresponding to top-s M_{jj}'s.

$\mathbf{v}_1 =$ top singular vector of $\mathbf{M}_{\hat{S}} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_{i\hat{S}}\mathbf{a}_{i\hat{S}}^\top \in \mathbb{R}^{s\times s}$.

$\mathbf{x^0} = \phi\mathbf{v}$, where $\mathbf{v} \leftarrow \mathbf{v}_1$ for $\hat{S}$ and $\mathbf{0} \in \mathbb{R}^{n-s}$ for $\hat{S}^c$.

output $\mathbf{x^0}$.

Algorithm 2 CoPRAM: Descent.

input $\mathbf{A}, \mathbf{y}, \mathbf{x^0}, s, t_0$

Initialize $\mathbf{x^0}$ according to Algorithm 1.

for $t = 0, \cdots, t_0 - 1$ **do**

$\mathbf{P}^{t+1} \leftarrow \text{diag}\left(\text{sign}\left(\mathbf{A}\mathbf{x^t}\right)\right)$,

$\mathbf{x^{t+1}} = \text{CoSaMP}(\frac{1}{\sqrt{m}}\mathbf{A}, \frac{1}{\sqrt{m}}\mathbf{P}^{t+1}\mathbf{y}, s, \mathbf{x}^t)$.

end for

output $\mathbf{z} \leftarrow \mathbf{x}^{t_0}$.

IV. Compressive phase retrieval

In this section, we propose a new algorithm for solving the sparse phase retrieval problem and analyze its performance. Later, we will show how to extend this algorithm to the case of more refined structural assumptions about the underlying sparse signal.

We first provide a brief outline of our proposed algorithm. It is clear from the discussion in the introduction that the recovery problem (1) is highly non-convex, and multiple locally optimal solutions might exist. Therefore, as is typical in modern non-convex methods [14], [23], [65] we use an spectral technique to obtain a good initial estimate. This technique itself is a modification of the initialization stage of Algorithm 1 of [23]; however, as we discuss below, our method requires no special tuning parameters except for knowledge of the underlying sparsity s. Moreover, in contrast with [23] our theoretical analysis requires no extra assumptions on the signal coefficients.

Once an appropriate initial estimate is chosen, we then show that a simple alternating-minimization algorithm will converge rapidly to the underlying true signal. Our proposed algorithm is new, and builds upon the original alternating-minimization algorithm proposed in [14]. In a departure from existing sparse phase retrieval methods [23], [22], our method is *parameter free* except for knowledge of the sparsity level s.

We call our overall algorithm *Compressive Phase Retrieval with Alternating Minimization* (CoPRAM). As described above, the algorithm is divided into two stages: an *Initialization* stage and a *Descent* stage. The stages are presented in pseudocode form as Algorithms 1 and 2.

A. Initialization

The first stage of CoPRAM uses a similar approach as those provided in previous sparse phase retrieval methods. The high

level idea is to use the measurements y_i to construct a *biased* estimator of the (squared) absolute values of the true signal coefficients. For the j^{th} signal coefficient, the *marginal* M_{jj} is given by:

$$M_{jj} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 a_{ij}^2,$$

and the set of all M_{jj}'s can be calculated in $\mathcal{O}(mn)$ time. The marginals themselves do not directly produce the signal coefficients, but the "weight" of each marginal (with sufficiently many samples) can enable identification of the coordinates of the true signal support. Once the support is accurately identified, a spectral technique (e.g., the methods of [14], [23], [22]) can be used to construct a good initial estimate $\mathbf{x^0}$.

However, accurate support identification can be tricky in general, particularly in the presence of very small signal coefficients. Indeed, to avoid this issue, earlier works [14], [23] assume that the magnitudes of the nonzero signal coefficients are all sufficiently large, i.e., $\Omega\left(\|\mathbf{x}^*\|_2/\sqrt{s}\right)$. As discussed earlier, this assumption can be unrealistic, violating the power-decay law.

Our analysis resolves this issue by *relaxing* the requirement of accurately identifying the support. The basic intuition is that even a coarse estimate of the support suffices to achieve a good estimate, since the errors are all going to correspond to small coefficients anyway. Such "noise" in the signal estimate can be controlled with a sufficient number of samples. A similar argument has been made in the analysis of the initialization stage of [22]; however, their estimate is a strict subset of the true support and their method requires tuning of real-valued parameters that can be hard to estimate in practice. Instead, we use a simpler estimation procedure. Indeed, we show that a simple pruning step that rejects the smallest $(n-k)$ coordinates, followed by the spectral procedure of [23], gives us the initialization that we need.

Concretely, we leverage the following fact: if elements of $\mathbf{A}$ are distributed as per standard normal distribution $\mathcal{N}(0,1)$, a weighted correlation matrix $\mathbf{M}$ can be constructed with diagonal elements M_{jj},

$$\mathbf{M} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_i \mathbf{a}_i{}^\top,$$
$$M_{jj} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 a_{ij}^2. \tag{2}$$

Then the expectation of this matrix $\mathbf{M}$ is,

$$\begin{aligned}\mathbb{E}[\mathbf{M}] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_i \mathbf{a}_i{}^\top\right], \\ &= \left(\mathbf{I}_{n\times n} + 2\frac{\mathbf{x}^*}{\|\mathbf{x}^*\|_2}\cdot\frac{\mathbf{x}^{*\top}}{\|\mathbf{x}^*\|_2}\right)\|\mathbf{x}^*\|_2^2,\end{aligned} \tag{3}$$

where $\mathbf{M}, \mathbb{E}[\mathbf{M}] \in \mathbb{R}^{n\times n}$. The diagonal elements of this expectation matrix $\mathbb{E}[\mathbf{M}]$ are given by:

$$\mathbb{E}[M_{jj}] = \begin{cases} \|\mathbf{x}^*\|^2 + 2x_j^{*2} & \text{for} \quad j \in S, \\ \|\mathbf{x}^*\|^2 & \text{for} \quad j \in S^c. \end{cases} \tag{4}$$

Intuitively, the signal marginals at locations on the diagonal of $\mathbf{M}$ corresponding to $j \in S$ are larger, on an average, than those corresponding to the zero-locations ($j \in S^c$). Using this as the baseline, we evaluate the diagonal elements of the matrix $\mathbf{M}$ (which we refer to as *marginals*) and establish a threshold value Θ which separates the indices j corresponding to these marginals into sets S and S^c.

We formalize the above argument. Our first result shows that given a sufficient number of measurements of the form (5), this method produces an estimate that is close enough to the true underlying signal.

Theorem IV.1. *The output of Algorithm 1, $\mathbf{x^0} \in \mathcal{M}_s$ is a small constant distance $0 < \delta_0 < 1$ away from the true signal $\mathbf{x}^* \in \mathcal{M}_s$, i.e.,*

$$\text{dist}\left(\mathbf{x^0}, \mathbf{x}^*\right) \leq \delta_0 \|\mathbf{x}^*\|_2,$$

as long as the number of (Gaussian) measurements 'm' satisfies the following bound,

$$m \geq Cs^2 \log mn, \tag{5}$$

with probability greater than $1 - \frac{8}{m}$.

This theorem is proved via Lemmas A.1 through A.4, and the argument proceeds as follows. We evaluate the marginals of the signal M_{jj}, in broadly two cases: $j \in S$ and $j \in S^c$. The key idea is to establish one of the following:

1) If there is a restriction on the minimum element of the true $\mathbf{x}^*$ (i.e., if it is bounded away from zero by a specific amount), then there exists a clear separation between the marginals M_{jj} for $j \in S$ and $j \in S^c$, with high probability. Then one would, whp, pick up the correct support in Algorithm 1 (i.e. $\hat{S} = S$). The top-singular vector of the truncated matrix $\mathbf{M}_S$ gives a good initial estimate $\mathbf{x^0}$.
2) If there is no such restriction, even then the support picked up in Algorithm 1, $\hat{S}$, contains a bulk of the correct support S. Some fraction of the elements picked up in $\hat{S}$ are incorrect, but we prove that they induce negligible error in estimating the initial vector $\mathbf{x^0}$.

These approaches are illustrated in Figures 7 and 8 in Appendix A. The marginals M_{jj} for $j \in S^c$ are upper bounded as stated in Lemma A.1. Similarly, the marginals M_{jj} for $j \in S$ are lower bounded as stated in Lemma A.2. The identification of the support $\hat{S}$ (which provably contains a significant chunk of the true support S) serves as a basis to construct the truncated correlation matrix $\mathbf{M}_{\hat{S}}$. The top singular vector of this matrix $\mathbf{M}_{\hat{S}}$, $\mathbf{x^0}$ gives us a good initial estimate of the true signal $\mathbf{x}^*$.

The final step of Algorithm 1 requires a scaling by a factor ϕ. This ensures that the power of the initial estimate $\mathbf{x^0}$ is close to the power of the true signal $\mathbf{x}^*$ (this is required because the calculation of the top-singular vector gives us a normalized vector $\mathbf{v}$). Provided sufficiently many samples, the signal power $\|\mathbf{x}^*\|_2^2$ is well approximated by the average power in the measurements ϕ^2 which is defined as

$$\phi^2 = \frac{1}{m}\sum_{i=1}^{m} y_i^2. \tag{6}$$

Using Lemma F.1 in Appendix F we can show that

$$1-\delta \leq \frac{\phi^2}{\|\mathbf{x}^*\|_2^2} \leq 1+\delta,$$

for small constant $0 < \delta \ll 1$ with probability greater than $1-\frac{1}{m}$.

B. Descent to optimal solution

Once we obtain a good enough initial estimate $\mathbf{x^0}$ such that $\operatorname{dist}\left(\mathbf{x^0}, \mathbf{x}^*\right) \leq \delta_0 \|\mathbf{x}^*\|_2$, whp, we now construct a method to accurately recovery the true $\mathbf{x}^* \in \mathcal{M}_s$. To achieve this, we adapt the alternating minimization approach from [14].

We introduce some notation. The observation model in (1) can be restated as follows:

$$\operatorname{sign}\left(\langle \mathbf{a}_i, \mathbf{x}^* \rangle\right) \circ y_i = \langle \mathbf{a}_i, \mathbf{x}^* \rangle,$$

for all $i = \{1, 2, \dots m\}$. We denote the *phase vector* $\mathbf{p} \in \mathbb{R}^m$ as a vector that contains the (unknown) signs of the measurements, i.e., $p_i = \operatorname{sign}\left(\langle \mathbf{a}_i, \mathbf{x} \rangle\right)$ for all $i = \{1, 2, \dots m\}$. We can also define a diagonal *phase matrix* $\mathbf{P} \in \mathbb{R}^{m \times m}$, such that $\mathbf{P} = \operatorname{diag}(\mathbf{p})$. Then our measurement model gets modified as:

$$\mathbf{P}^*\mathbf{y} = \mathbf{A}\mathbf{x}^*,$$

where $\mathbf{P}^*$ denotes the true phase matrix. Consider minimizing the loss function composed of two variables $\mathbf{x}$ and $\mathbf{P}$,

$$\min_{\|\mathbf{x}\|_0 \leq s, \mathbf{P} \in \mathcal{P}} \|\mathbf{A}\mathbf{x} - \mathbf{P}\mathbf{y}\|_2. \tag{7}$$

Note that the problem above is *not convex*, because $\mathbf{P}$ is restricted to be a diagonal matrix $\in \mathcal{P}$, where $\mathcal{P}$ is a set of all diagonal matrices with diagonal entries constrained to be in $\{-1, 1\}$. Instead, we alternate between estimating $\mathbf{P}$ and $\mathbf{x}$. We perform two estimation steps:

1) If we fix the signal estimate $\mathbf{x}$, then the minimizer $\mathbf{P} \in \mathcal{P}$ is given in closed form as:

$$\mathbf{P} = \operatorname{diag}\left(\operatorname{sign}\left(\mathbf{A}\mathbf{x}\right)\right). \tag{8}$$

We call this the *phase estimation* step.

2) If we fix the phase matrix $\mathbf{P} \in \mathcal{P}$, the signal vector $\mathbf{x}$ can be obtained by solving a sparse recovery problem,

$$\min_{\mathbf{x}, \|\mathbf{x}\|_0 \leq s} \|\mathbf{A}\mathbf{x} - \mathbf{P}\mathbf{y}\|_2, \tag{9}$$

if $m < n$ (and each entry of $\mathbf{A}$, a_{ij} is sampled from independent Gaussian $\mathcal{N}(0,1)$, such that $\frac{\mathbf{A}}{\sqrt{m}}$ satisfies the *restricted isometry property*). We call this the *signal estimation* step.

We employ the CoSaMP [41] algorithm to (approximately) solve (9). Note that since (9) itself is a non-convex problem and exact minimization can be hard. However, we do not need to explicitly obtain the minimizer but only show a sufficient descent criterion, which we achieve by performing a careful analysis of the CoSaMP algorithm. For analysis reasons, we require that the entries of the input sensing matrix are distributed according to $\mathcal{N}\left(0, \mathbf{I}/\sqrt{m}\right)$. This can be achieved by scaling down the inputs to CoSaMP: $\mathbf{A}^t, \mathbf{P}^{t+1}\mathbf{y}$ by a factor of $\sqrt{m}$ (see $\mathbf{x}$-update step of Algorithm 2). Another distinction is that we use a "warm start" CoSaMP routine for each iteration where the initial guess of the solution to (9) is given by the current signal estimate.

We now analyze our proposed descent scheme. We obtain the following theoretical result:

Theorem IV.2. *Given an initialization $\mathbf{x^0} \in \mathcal{M}_s$ satisfying $\operatorname{dist}\left(\mathbf{x}^0, \mathbf{x}^*\right) \leq \delta_0 \|\mathbf{x}^*\|_2$, for $0 < \delta_0 < 1$, if we have number of (Gaussian) measurements $m > Cs \log \frac{n}{s}$, then the iterates of Algorithm 2 satisfy:*

$$\operatorname{dist}\left(\mathbf{x}^{t+1}, \mathbf{x}^*\right) \leq \rho_0 \operatorname{dist}\left(\mathbf{x}^t, \mathbf{x}^*\right), \tag{10}$$

where $0 < \rho_0 < 1$ is a constant, with probability greater than $1 - e^{-\gamma m}$, for positive constant γ.

The proof of this theorem can be found in Appendix C.

Combining both stages, the number of measurements are required to obey the following lower bound:

$$m_0 > \max\left(C_1 s^2 \log mn, C_2 s \log \frac{n}{s}\right) \equiv Cs^2 \log mn, \tag{11}$$

for the overall CoPRAM algorithm to succeed.

C. Robustness to noise

The above analysis assumes that the measurements are pristine (noiseless). We can also demonstrate that the CoPRAM algorithm are sufficiently robust in the presence of noise. This is established in the following theorem.

Theorem IV.3. *Given Gaussian measurements $a_{ij} \in \mathcal{N}(0,1)$, CoPRAM can recover the model sparse signal $\mathbf{x}^{t_0} \in \mathcal{M}_s$ from noisy measurements $\mathbf{y}$ of the form*

$$\mathbf{y} = |\mathbf{A}\mathbf{x}^*| + \epsilon,$$

where $\epsilon \in \mathbb{R}^m$ is a scaled sub-exponential. This retains the previously derived expression for sample complexity as in Theorem IV.1 up to a constant factor. The algorithm converges according to the iteration invariant:

$$\left\|\mathbf{x}^{t_o} - \mathbf{x}^*\right\|_2 \leq c_1 \|\mathbf{x}^*\|_2 + c_2 \|\epsilon\|_2$$

where t_o is the number of outer iterations of CoPRAM and Block CoPRAM, $c_1 < 1$ and $c_2 = 200$.

The proof for this theorem can be found in Appendix D. An identical analysis holds for the block-sparse case which we elaborate in more detail below.

D. Sparse signals exhibiting power law decay

The quadratic dependence of the sample complexity on the signal sparsity level s, as derived in Theorem IV.1, is typical (and also shared by the other works [49], [23]) but somewhat problematic. In particular, if the signal sparsity exceeds the square-root of the dimension n, the result becomes moot and one may as well as standard phase retrieval techniques!

In this section, we demonstrate a method to break through this quadratic barrier, albeit under somewhat more stringent assumptions on the signal. Specifically, we analyze the scenario hypothesized in [20] in which the signal $\mathbf{x}^*$ follows a power-law decay. That is, suppose that the signal coefficients

Algorithm 3 Block CoPRAM: Initialization.

input $\mathbf{A}, \mathbf{y}, b, k$.
Compute signal power $\phi^2 = \frac{1}{m}\sum_{i=1}^{m} y_i^2$.
Compute block marginals $M_{j_b j_b} = \sqrt{\sum_{j\in j_b} M_{jj}^2} \quad \forall j_b$, where M_{jj} is as in (2).
Select $\hat{S}_b \leftarrow j_b$'s corresponding to top-k $M_{j_b j_b}$'s, $\hat{S}$ is signal support corresponding to blocks $\hat{S}_b$.
Compute $\mathbf{v}_1 \leftarrow$ top singular vector of $\mathbf{M}_{\hat{S}_b} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_{i\hat{S}} \mathbf{a}_{i\hat{S}}^\top \in \mathbb{R}^{s\times s}$.
Compute $\mathbf{x^0} \leftarrow \phi\mathbf{v}$ where $\mathbf{v} \leftarrow \mathbf{v}_1$ for $\hat{S}$, and $\mathbf{0} \in \mathbb{R}^{n-s}$ for $\hat{S}^c$.
output $\mathbf{x^0}$.

have been suitably re-indexed such that x_j^{*2}, for $j \in 1, 2, \ldots, s$ can be arranged in non-increasing order:

$$x_j^{*2} \leq \frac{C(\alpha)}{j^\alpha} \tag{12}$$

where and $\alpha > 1$. Due to the isotropic nature of the Gaussian measurement scheme, such a re-indexing can be assumed without loss of generality.

We can show that this extra power-law decay assumption results in *far fewer* samples, $\mathcal{O}(s \log n)$ for the CoPRAM initialization step to achieve a sufficiently good initial guess. Combined with Theorem IV.2, we obtain the following result:

Theorem IV.4. *Given Gaussian measurements $a_{ij} \in \mathcal{N}(0,1)$, then CoPRAM can recover the s-sparse signal $\mathbf{x}^{t_0} \in \mathcal{M}_s$, with $\|\mathbf{x}^{t_0} - \mathbf{x}^*\|_2 \leq \delta \|\mathbf{x}^*\|_2$, where t_0 is the number of outer iterations of CoPRAM, from $m > Cs\log n$ measurements, as long as the coefficients of the signal follow a power-law decay as described in (12).*

The proof for this theorem can be found in Appendix E.

V. Block-sparse phase retrieval

The analysis of the proofs mentioned so far, as well as experimental results suggest that we can reduce sample complexity for successful sparse phase retrieval by exploiting further structural information about the signal. We assume that the true signal $\mathbf{x}^*$, is block sparse with uniform block length b and effective block sparsity $k = \frac{s}{b}$. We introduce the concept of *block marginals*, a block-analogue to signal marginals, which can be analyzed to crudely estimate the block support of the signal in consideration. We use this formulation, along with the alternating minimization approach that uses model-based CoSaMP [31] to descend to the optimal solution. In the next subsections, we discuss the initialization and descent of the Block CoPRAM algorithm. The pseudo-code for Block CoPRAM is stated in Algorithms 3 and 4.

A. Initialization

Block-sparse signals $\mathbf{x}^*$, can be said to be following a sparsity model $\mathcal{M}_s^b$, where $\mathcal{M}_s^b$ describes the set of all block-sparse signals with s non-zeros being grouped into uniform pre-determined blocks of size b, such that block-sparsity

Algorithm 4 Block CoPRAM: Descent.

input $\mathbf{A}, \mathbf{y}, \mathbf{x^0}, b, k, t_0$
Initialize $\mathbf{x^0}$ according to Algorithm 3.
for $t = 0, \cdots, t_0 - 1$ **do**
$\mathbf{P}^{t+1} \leftarrow \text{diag}\left(\text{sign}\left(\mathbf{A}\mathbf{x^t}\right)\right)$.
$\mathbf{x}^{t+1} \leftarrow \text{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left\|\mathbf{A}\mathbf{x} - \mathbf{P}^{t+1}\mathbf{y}\right\|_2 =$ BlockCoSaMP($\frac{1}{\sqrt{m}}\mathbf{A}, \frac{1}{\sqrt{m}}\mathbf{P}^{t+1}\mathbf{y}, b, k, \mathbf{x}^t$).
end for
output $\mathbf{z} \leftarrow \mathbf{x}^{t_0}$.

$k = \frac{s}{b}$. The effective sparsity of the signal is still s, however the non-zero elements are constrained to appear in blocks. We use the index set $j_b = \{1, 2 \ldots k\}$, to denote block-indices.

Analogous to the concept of marginals defined above, we introduce *block marginals* $M_{j_b j_b}$, where M_{jj} is defined as in (2). For block index j_b, we define:

$$M_{j_b j_b} = \sqrt{\sum_{j\in j_b} M_{jj}^2}, \tag{13}$$

to develop the initialization stage of our *Block CoPRAM* algorithm. Similar to the proof approach of CoPRAM, we show that there exists a threshold that separates the block marginals $\mathbf{M}_{j_b j_b}$, for $j_b \in S_b$ and $\mathbf{M}_{j_b j_b}$, for $j_b \in S_b^c$ respectively. Here, S_b represents the "block support", i.e., the set of active block-indices. We can then evaluate the block marginals, and use the top-k such marginals to obtain a crude approximation $\hat{S}_b$ of the true block support S_b. This support can be used to construct the truncated correlation matrix $\mathbf{M}_{\hat{S}_b}$. The top singular vector of this matrix $\mathbf{M}_{\hat{S}_b}$ gives a good initial estimate $\mathbf{x^0}$ for the Block CoPRAM algorithm (Algorithm 4). Through the evaluation of block marginals, we proceed to prove that the sample complexity required for a good initial estimate (and subsequently, successful signal recovery of block sparse signals) is given by $\mathcal{O}\left((s^2/b)\log n\right) = \mathcal{O}(ks\log n)$. This essentially reduces the sample complexity of signal recovery by a factor equal to the block-length b over the sample complexity required for standard sparse phase retrieval.

Formally, we obtain the following result:

Theorem V.1. *The initial vector $\mathbf{x^0} \in \mathcal{M}_s^b$, which is the output from Algorithm 3, is a small constant distance away from the true signal $\mathbf{x}^* \in \mathcal{M}_s^b$, i.e.*

$$\text{dist}\left(\mathbf{x^0}, \mathbf{x}^*\right) \leq \delta_b \|\mathbf{x}^*\|_2,$$

for $0 < \delta_b < 1$, as long as the number of measurements satisfy

$$m \geq C\frac{s^2}{b}\log mn,$$

with probability greater than $1 - \frac{8}{m}$.

The proof can be found in Appendix B, and carries forward intuitively from the proof of the sparse phase-retrieval framework.

B. Descent to optimal solution

For the descent of Block CoPRAM to optimal solution, the phase-estimation step is the same as that in CoPRAM (8).

For the signal estimation step, we attempt to solve the same minimization as in (9), except with the additional constraint that the signal $\mathbf{x}^*$ is *block sparse*,

$$\min_{\mathbf{x}\in\mathcal{M}_s^b} \|\mathbf{A}\mathbf{x}-\mathbf{P}\mathbf{y}\|_2, \tag{14}$$

where $\mathcal{M}_s^b$ describes the block sparsity model. In order to approximate the solution to (14), we use the *model-based CoSaMP* approach of [31]. This is a straightforward specialization of the CoSaMP algorithm and has been shown to achieve improved sample complexity over existing approaches for standard sparse recovery.

Similar to Theorem IV.2 above, we obtain the following result:

Theorem V.2. *Given an initialization $\mathbf{x^0}\in\mathcal{M}_s^b$, satisfying $\text{dist}\left(\mathbf{x}^t,\mathbf{x}^*\right)\leq\delta_b\|\mathbf{x}^*\|_2$, where $0<\delta_b<1$, if we have number of measurements $m\geq C\left(s+\frac{s}{b}\log\frac{n}{s}\right)$, then the iterates of Algorithm 4 satisfy:*

$$\text{dist}\left(\mathbf{x}^{t+1},\mathbf{x}^*\right)\leq\rho_b\text{dist}\left(\mathbf{x}^t,\mathbf{x}^*\right). \tag{15}$$

where $0<\rho_b<1$ is a constant, with probability greater than $1-e^{-\gamma m}$, for positive constant γ.

The proof of this theorem can be found in Appendix C.

C. Extension to blocks of non-uniform sizes

The analysis so far has been made for uniform blocks of size b. However the same algorithm can be extended to the case of sparse signals with *non-uniform* blocks. Such a model is particularly useful for time-series signals where the nonzeros occur in "bursts" of variable lengths and start times.

Formally, consider the *clustered sparsity* model for 1D signals in $\mathbb{R}^n$, comprising signals with s non-zeros that occur in no more than k_c non-overlapping blocks (clusters), each of which exhibit potentially unknown sizes and locations. The above analysis does not immediately apply to this case; however, by the analysis approach of [37], we can show that any such clustered-sparse signal with parameters (s,k_c) can be *simulated* using a *uniform* block-sparse signal with parameters $(s,3k_c)$.

This can be demonstrated as follows. Assuming that the non-zeros coefficients exist in k_c non-overlapping clusters, in the best case, all k_c clusters are uniformly sized. In the worst case, we would have (k_c-1) clusters with $2-$ coefficients lying on either side of the boundary of two uniform block supports (therefore corresponding to two active blocks). The remaining $s-2(k_c-1)$ coefficients constitute the last cluster. In this case, the number of active blocks can be bounded as:

$$k\leq 2\cdot(k_c-1)+\frac{s-2(k_c-1)}{s/k_c}\leq 3k$$

where k is effective block sparsity, and uniform blocks of size $b=s/k_c$ are considered.

Therefore, the only price to be paid is a tripling of the block sparsity parameter k. Provided we are willing to tolerate this increase, we can use exactly the same Block CoPRAM algorithm (including both the initialization as well as the descent stages) as described above, with only a constant factor increase in the sample complexity.

We note that this argument is only applicable to block-sparse 1D signals (such as time-domain signals); extending this argument to general clustered-sparse images and higher-dimensional data is much more involved, and we leave this to future work.

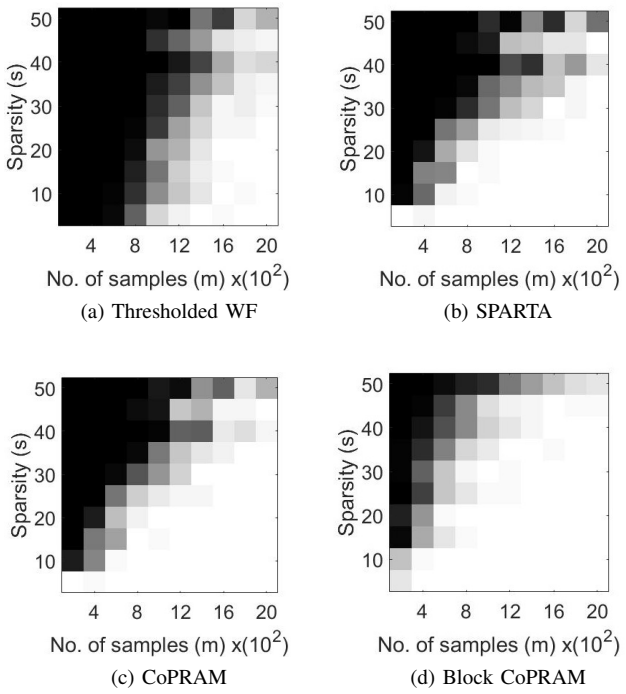
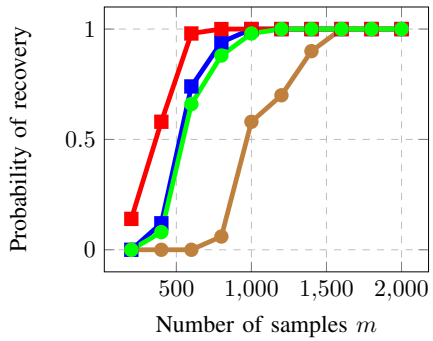


Fig. 1: *Phase transition plots for different algorithms, with signal length $n=3000$, having uniform block length of $b=5$.*

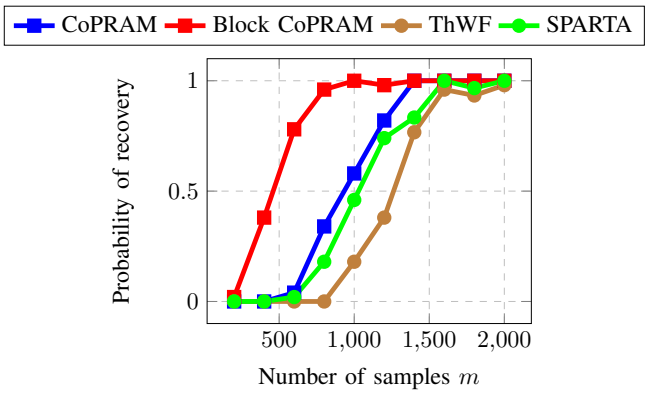
VI. Numerical experiments

In this section, we present the results of a range of simulations supporting our algorithms and demonstrate their benefits over the state-of-the-art in sparse phase retrieval. All numerical experiments were conducted using MATLAB 2016a on a desktop computer with an Intel Xeon CPU at 3.3GHz and 8GB RAM.

Our experiments explores the performance of the CoPRAM and Block CoPRAM algorithms on synthetic data. The non-zero elements of the test signal $\mathbf{x}^*\in\mathbb{R}^n$, with $n=3,000$ are generated using zero-mean Gaussian distribution $\mathcal{N}(0,1)$ and normalized, such that $\|\mathbf{x}^*\|=1$. The elements of sensing matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$, a_{ij} are also generated using the zero-mean Gaussian distribution $\mathcal{N}(0,1)$. The sparsity levels s are chosen in steps of 5 with a maximum value of $s=50$ such that $n=3000\gtrsim 50^2$ (for values of $s>\sqrt{n}$, the effect of sparsity is minimal and standard non-sparsity based phase retrieval algorithms perform equally well). A block length of $b=5$ is considered for all generated signals in experiments in Figures 1, 2, and 4. The number of measurements m is swept from $m=200$ to $m=2,000$ in steps of 200. We repeated each of the experiments (fixed n,s,m) in Figures 1, 2, and 4



(a) $s = 20$



(b) $s = 30$

Fig. 2: *Phase transition graph for (a)* $s = 20$ *and (b)* $s = 30$*, for a signal of length* $n = 3,000$ *and block length* $b = 5$.

for 50 independent Monte Carlo trials, and the experiments in Figure 3 for 200 independent Monte Carlo trials.

For our simulations, we compared our algorithms CoPRAM and Block CoPRAM with two other sparse phase retrieval algorithms: Thresholded Wirtinger Flow [22] and SPARTA [23]. For our set of generated signals, the AltMinSparse method mentioned in [14] does not recover the signal in most cases (if the initialization stage fails to pick the correct support, the subsequent AltMinPhase procedure can never give a good solution). We therefore do not include this algorithm for comparisons.

For Thresholded WF, we set parameters which were optimized based on a number of trial cases and were kept constant throughout all experiments, with values $\alpha = 1.5$, $\mu = 0.23$ and $\beta = 0.3$. Similarly, for SPARTA, we set the parameters to be $\gamma = 0.7$, $\mu = 1$ and $\text{card}(\mathcal{I}_l) = \lceil \frac{m}{6} \rceil$ as mentioned in their paper. For the first experiment, we generated phase transition plots by evaluating the probability of successful recovery, i.e. number of trials out of 50, that gave a relative error in retrieval $\frac{\|\mathbf{x}^{t_0} - \mathbf{x}^*\|_2}{\|\mathbf{x}^*\|_2} < 0.05$. We let each of the algorithms to run for a total of $t_0 = 30$ iterations. The recovery probability for varying values of s and m has been illustrated in Figure 1 through phase transition diagrams. It can be noted that CoPRAM (1(c)) and SPARTA (1(b)) perform comparably, while Block CoPRAM (1(d)) performs the best among all four algorithms, in terms of sample complexity.

The phase transition graphs for $s = 20$ and $s = 30$ for the four algorithms is displayed in Figure 2.

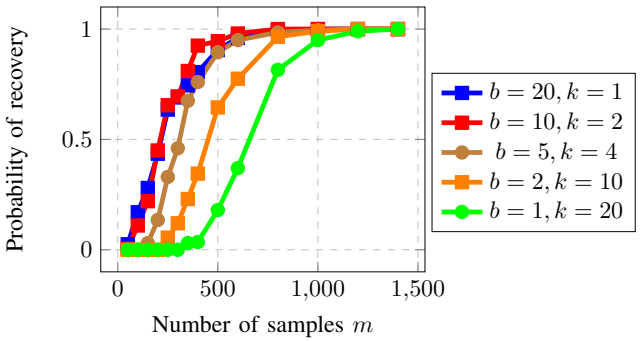


Fig. 3: *Variation of phase transition for Block CoPRAM at* $s = 20$ *and* $b = 20, 10, 5, 2, 1$ *for a signal of length* $n = 3,000$.

It can be noted that increasing the sparsity of signal shifts the phase transitions to the right (sample complexity of m has a quadratic dependence on s for CoPRAM, SPARTA and Thresholded WF). However the phase transition for Block CoPRAM has a less apparent shift, as compared to other algorithms (sample complexity of m has sub-quadratic dependence on s). It can be noted that as sparsity s increases, the *gap* between the phase transition of Block CoPRAM and other algorithms in consideration, increases. As demonstrated in Figure 1, the Block CoPRAM approach exhibits lowest sample complexity for the phase transitions in both cases (a) and (b) of Figure 2.

The mean running time of the algorithms for different algorithms is tabulated in Table II. It can be noted that the running times of our algorithms CoPRAM and Block CoPRAM are at par with SPARTA and Thresholded WF.

TABLE II: *Mean run time of different algorithms at* $s = 25$.

Algorithm	CoPRAM	Block CoPRAM	SPARTA	ThWF
m at phase trans	1,600	1,400	1,800	2,000
mean run time (s)	0.4000	0.3258	0.3080	0.5808

Leveraging block-sparsity. For the second experiment, we study the variation of phase transition with block-length, for Block CoPRAM (refer Figure 3). For this experiment we fixed a signal of length $n = 3,000$, sparsities $s = 20, k = 1$ for a block length of $b = 20$. We observed that the phase transitions improve with increase in block length (used to estimate the signal in the algorithm) up to a point. At block sparsities $\frac{s}{b} = \frac{20}{10} = 2$ and $\frac{s}{b} = \frac{20}{20} = 1$, there is little difference in the phase transitions, as the regime of the experiment is very close to the information theoretic bound of $s \log \frac{n}{s}$.

Effect of noise. For our third experiment, we study the effect of noise on the measurements of the form:

$$y_i = |\langle \mathbf{a}_i, \mathbf{x}^* \rangle| + \epsilon_i,$$

for $i \in \{1, 2 \dots m\}$, to verify the claims in Theorem IV.3. The noise vector $\epsilon \in \mathbb{R}^m$ is sampled from a zero-mean Gaussian distribution $\mathcal{N}(0, \sigma^2)$, where σ^2 is determined using the input noise-signal-ratio (NSR). We compared CoPRAM, Block CoPRAM and SPARTA to analyze robustness to noisy measurements for amplitude only measurements (ThWF is

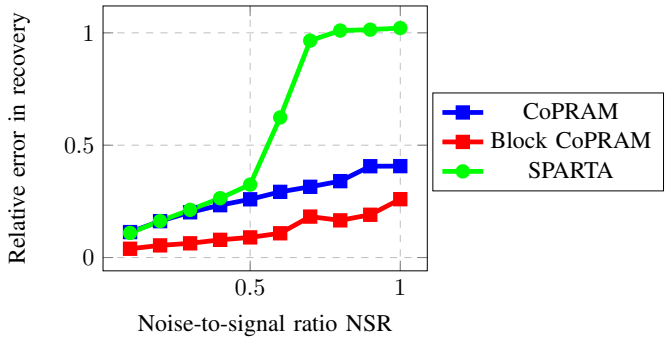
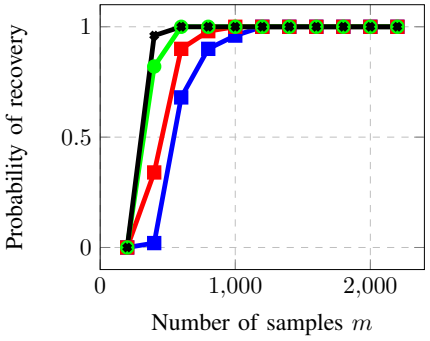
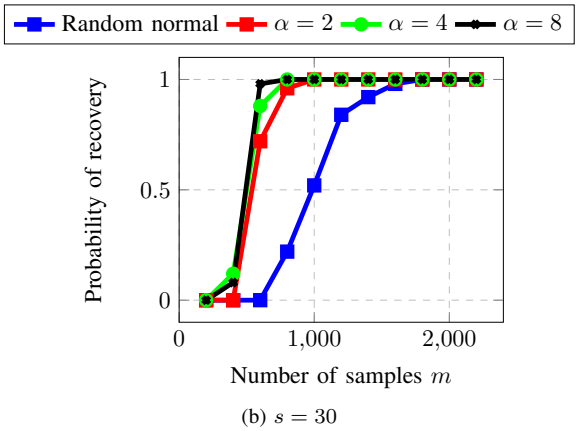


Fig. 4: *Variation of mean relative error in signal recovered v/s input NSR at* $s = 20$ *and* $b = 5, k = 4$ *for a signal of length* $n = 3,000$*, and number of measurements* $m = 1,600$.

excluded because they use quadratic measurements). We vary the input NSR $= \sigma^2 / \|\mathbf{x}^*\|_2^2$, from 0.1 to 1 in steps of 0.1. We fix signal parameters $n = 3,000, s = 20, b = 5, k = 4$ and number of measurements to $m = 1,600$. This experiment was run for 50 independent Monte Carlo trials. The variation of mean relative error $\|\mathbf{x}^{t_0} - \mathbf{x}^*\|_2 / \|\mathbf{x}^*\|_2$ can be seen in Figure 4. We observe that Block CoPRAM exhibits greater robustness to noise compared to CoPRAM and SPARTA in all cases considered.



(a) $s = 20$

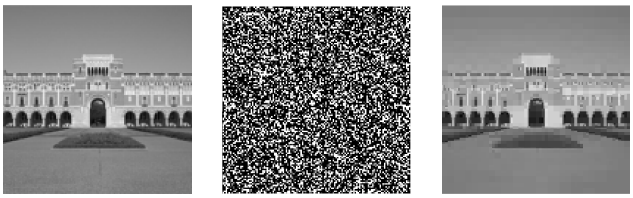


(b) $s = 30$

Fig. 5: *Variation of phase transition for CoPRAM at (a)* $s = 20$ *and (b)* $s = 30$ *at decay rates* $\alpha = 2, 4, 8$*, for a signal of length* $n = 3,000$*, compared to standard* s*-sparse signal with coefficients picked random normally.*

Power-law decay. For our fourth experiment, we verify the claims in Theorem IV.4. We set the signal length to $n = 3,000$. We analyze the effect of power law decay on signals with sparsities $s = 20$ and $s = 30$ for different rates of decay $\alpha = 2, 4, 8$ and compare this to the case with no power law decay (coefficients picked random normally). We observe an improvement in sample complexity with respect to the "no powerlaw decay" case, as seen in Figure 5. The improvement is more prominent as we increase the sparsity from $s = 20$ to $s = 30$.

Experiments on real images. For our final experiment, we evaluated the performance of our algorithm on a real image, with induced sparsity in the wavelet basis (db1). We used a 128×128 image of Lovett Hall, and used the thresholded wavelet transform (using Haar wavelet) of this image as the sparse signal with $s = 0.09n$. This image was reconstructed using $m = 16,384$ samples, using CoPRAM and the standard AltMinPhase algorithm described in [14]. We used signal-to-noise ratio (SNR) to quantify quality of reconstruction. In Figure 6, we demonstrate how enforcing a sparsity constraint enables recovery of the same image using fewer samples and better reconstruction quality.



(a) Original (b) AltMinPhase SNR=-0.71dB (c) CoPRAM SNR=82.86dB

Fig. 6: *Reconstruction of the original Lovett (a) image using (b) AltMinPhase for* $m = 16,384$*, (c) CoPRAM for* $m = 16,384$ *measurements, where* $f = m/n$.

VII. Discussion

In this paper, we have introduced a set of new algorithmic approaches for sparse as well as structured sparse phase retrieval. Our algorithms are conceptually simple and indeed are reminiscent of classical heuristics for phase retrieval; however, our analysis also shows an asymptotic reduction in sample complexity when additional structures on top of standard sparsity are leveraged within the reconstruction process.

In this paper, we chose to study real-valued Gaussian measurements, however this study can plausibly be extended to complex-valued Gaussian measurements as well. In this case, the initialization stage, would remain as is, as it essentially bounds the norm of the *sine* of *angle* between $\mathbf{x^0}$ and $\mathbf{x}^*$ (refer proof of Lemma (A.4)). However, there exist additional challenges in analyzing the descent stage of our algorithm, explicitly in bounding the phase error term, which we intend to address in future work.

Similarly, several open questions remain, including expanding our analysis to more sophisticated sparsity models, such as clusters, trees, groups, and graphical models [58]; this has in part been examined in our follow-up work [64]. Moreover,

our analysis only applies to the case of Gaussian samples, and extending our results to more realistic measurement schemes such as Fourier samples and coded diffraction patterns [46] will be an interesting direction of future study; indeed, our preliminary experimental results in [62], [63] have shown the empirical benefits of our algorithms for such challenging measurement setups.

APPENDIX A
COPRAM INITIALIZATION

In this section we state the proofs related to the *initialization* in Algorithm 1, for compressive phase retrieval. This includes the proofs of Lemmas A.1 - A.4 which complete the proof of Theorem IV.1.

The outline of the proof is sketched out as follows. Using Lemma A.1, we can find an upper bound on marginals M_{jj} for $j \in S$. Consequently,

$$\max_{j \in S^c} M_{jj} \leq \left(1 + 11\sqrt{\frac{\log mn}{m}}\right) \|\mathbf{x}^*\|_2^2 = \Theta_1 \tag{16}$$

with probability greater than $1 - \frac{5}{m}$. Marginals M_{jj} for $j \in S$ can be evaluated in two ways:

1) Assuming a bound on the minimum element of $\mathbf{x}^*$: $x_{min}^{*2} \equiv \min_{j \in S} x_j^{*2} = \frac{C}{s} \|\mathbf{x}^*\|_2^2$. The proof then carries forward from the work in [23], where they arrive at the lower bound on the minimum marginal for $j \in S$, with probability greater than $1 - \frac{1}{m}$,

$$\begin{aligned}\min_{j \in S} M_{jj} &\geq \|\mathbf{x}^*\|_2^2 + x_{min}^{*2} \\ &= \left(1 + \frac{C}{s}\right) \|\mathbf{x}^*\|_2^2 = \Theta_2,\end{aligned}$$

given that $m \geq C_0 s^2 \log(mn)$. This proof is similar to that mentioned in Lemma A.2. Piecing these two together,

$$\begin{aligned}\min_{j \in S} M_{jj} &\geq \left(1 + \frac{C}{s}\right) \|\mathbf{x}^*\|_2 \\ &> \left(1 + 11\sqrt{\frac{\log mn}{m}}\right) \|\mathbf{x}^*\|_2^2 \\ &\geq \max_{j \in S^c} M_{jj}.\end{aligned}$$

which implies that the support picked up using the top s-marginals M_{jj} is the true support with probability greater than $1 - \frac{6}{m}$, given $m \geq C_0 s^2 \log(mn)$, as long as there is a clear separation between Θ_1 and Θ_2. They proceed to show that with a high probability, $\|\mathbf{x}^0 - \mathbf{x}^*\|_2 \leq \delta_0 \|\mathbf{x}^*\|_2$, using Proposition 1 of [51], completing the proof of Theorem IV.1.

2) If there is no such assumption on the minimum entry x_{min}^{*2}, we proceed with a longer proof, as stated below using Lemmas A.2-A.4. The idea is to show that $\mathbf{x}^* \approx \mathbf{x}_{\hat{S}}^*$ and subsequently $\mathbf{x}_{\hat{S}}^* \approx \mathbf{x}^0$, effectively implying that $\mathbf{x}^0 \approx \mathbf{x}^*$.

This idea and the partition of support sets used in the proof have been illustrated in Figures 7 and 8.

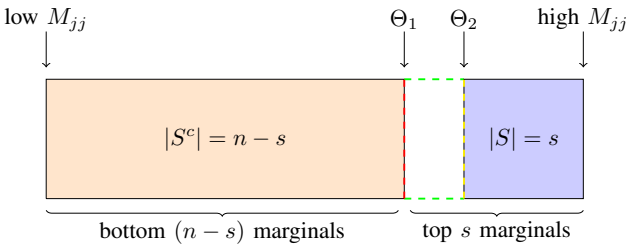


Fig. 7: Partition of supports considered for analysis of proof approach 1: assumption on x_{min}^*.

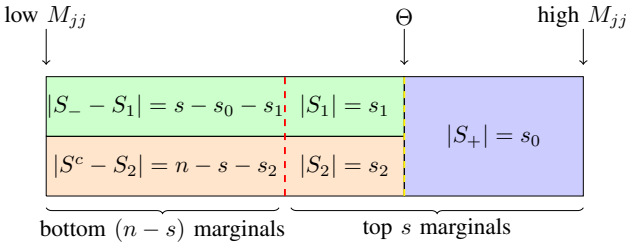


Fig. 8: Partition of supports considered for analysis of proof approach 2.

Lemma A.1. *For all $j \in S^c$, with probability greater than $1 - \frac{5}{m}$, the corresponding marginals are upper-bounded as*

$$M_{jj} \leq \left(1 + 11\sqrt{\frac{\log mn}{m}}\right) \|\mathbf{x}^*\|_2^2 = \Theta. \tag{17}$$

Proof. Evaluating the marginals:

$$M_{jj} - \phi^2 = \frac{1}{m} \sum_{i=1}^{m} y_i^2 \left(a_{ij}^2 - 1\right),$$

where y_i is independent of a_{ij} for all $j \in S^c$. Evaluating the tail bound in terms of a series of tail bounds for independent random variables y_i and a_{ij}, one can use Lemma 4.1 of [66] for the χ_1^2 variables a_{ij}^2 with weights y_i^2 (here $p \equiv n - s$):

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{m} y_i^2 \left(a_{ij}^2 - 1\right) > 2\sqrt{t_1}\left(\sum_{i=1}^{m} y_i^4\right)^{\frac{1}{2}} + 2\left(\max_i y_i^2\right) t_1\right] \\ \leq \exp(-t_1) = \frac{1}{mp}.\end{aligned}$$

where t_1 is chosen *large enough*, such that $\exp(-t_1) = 1/mp$.

Further, using the Chebyshev's inequality for y_i^2:

$$\mathbb{P}\left[\sum_{i=1}^{m} \frac{y_i^4}{\|\mathbf{x}^*\|_2^4} > 3m + \sqrt{96mt_2}\right] \leq \frac{1}{t_2^2} = \frac{1}{mp}.$$

Using the Gaussian tail bound for y_i^2 followed by union bound:

$$\mathbb{P}\left[\max_i \frac{y_i^2}{\|\mathbf{x}^*\|_2^2} > t_2\right] \leq 2m \exp\left(\frac{-t_2}{2}\right) = \frac{2}{mp^2} \leq \frac{2}{mp}.$$

where t_2 is chosen to be large enough, such that $\exp(-t_2/2) = 1/(mp)^2$.

With probability at most $\frac{4}{mp}$, for each $j \in S^c$, using a union bound on these three tail bounds,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m} y_i^2(a_{ij}^2-1) &> 2\sqrt{3+\sqrt{96p}}\,\|\mathbf{x}^*\|_2^2\sqrt{\frac{\log mp}{m}} \\ &\quad + 8\,\|\mathbf{x}^*\|_2^2\,\frac{(\log mp)^2}{m}, \\ &> 2\sqrt{3+\sqrt{96}}\,\|\mathbf{x}^*\|_2^2\sqrt{\frac{\log mp}{m}} \\ &\quad + 8\,\|\mathbf{x}^*\|_2^2\,\frac{(\log mp)^2}{m}.\end{aligned}$$

Using a union bound for all $j \in S^c$ (p such), with probability at least $1-\frac{4}{m}$,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m} y_i^2(a_{ij}^2-1) &\le 2\sqrt{3+\sqrt{96}}\,\|\mathbf{x}^*\|_2^2\sqrt{\frac{\log mp}{m}} \\ &\quad + 8\,\|\mathbf{x}^*\|_2^2\,\frac{(\log mp)^2}{m}, \\ &\le 8\sqrt{\frac{\log mp}{m}}\,\|\mathbf{x}^*\|_2^2.\end{aligned}$$

Using Lemma F.1, for $m > C$, and using the fact that $p \le n$:

$$\begin{aligned}M_{jj} &= \frac{1}{m}\sum_{i=1}^{m} y_i^2 a_{ij}^2 \le 8\sqrt{\frac{\log mn}{m}}\,\|\mathbf{x}^*\|_2^2 + \phi^2, \\ M_{jj} &\le \left(1+11\sqrt{\frac{\log mn}{m}}\right)\|\mathbf{x}^*\|_2^2 = \Theta, \end{aligned} \tag{18}$$

which establishes the upper bound on marginals associated with the zero-locations $j \in S^c$, with probability greater than $1-\frac{5}{m}$. □

Lemma A.2. *For $j \in S_+ \subset S$, with probability greater than $1-\frac{2}{m}$, the corresponding marginals are lower-bounded as*

$$M_{jj} \ge \left(1+11\sqrt{\frac{\log mn}{m}}\right)\|\mathbf{x}^*\|_2^2 = \Theta, \tag{19}$$

where S_+ is defined as

$$S_+ = \left\{ j \in S \mid x_j^{*2} > 15\sqrt{\frac{\log mn}{m}}\,\|\mathbf{x}^*\|_2^2 \right\}. \tag{20}$$

Subsequently, we can define S_- as

$$S_- = \left\{ j \in S \mid x_j^{*2} \le 15\sqrt{\frac{\log mn}{m}}\,\|\mathbf{x}^*\|_2^2 \right\}, \tag{21}$$

with S_+ and S_- forming a partition of S and the corresponding energy in the elements $x_j, j \in S_-$ is lower-bounded as

$$\left\|\mathbf{x}^*_{S_-}\right\|_2^2 \le 15\sqrt{\frac{s^2\log mn}{m}}\,\|\mathbf{x}^*\|_2^2. \tag{22}$$

Proof. Evaluating the marginals:

$$M_{jj} - \phi^2 = \frac{1}{m}\sum_{i=1}^{m} y_i^2\left(a_{ij}^2-1\right). \tag{23}$$

For $j \in S$, y_i and a_{ij} are dependent random variables. The marginal M_{jj} can be evaluated through a concentration bounds on the two terms that compose the RHS of (23): $\frac{1}{m}\sum_{i=1}^{m} y_i^2 a_{ij}^2$ and $\frac{1}{m}\sum_{i=1}^{m} y_i^2$. This can be done by evaluating the expectation values:

$$\begin{aligned}\mathbb{E}\left[y_i^2\right] &= \|\mathbf{x}^*\|_2^2, \\ \mathbb{E}\left[y_i^2 a_{ij}^2\right] &= \|\mathbf{x}^*\|_2^2 + 2x_j^{*2}, \\ \mathbb{E}\left[y_i^4 a_{ij}^4\right] &= 105x_j^{*4} + 90x_j^{*2}\left(\|\mathbf{x}^*\|_2^2 - x_j^{*2}\right) \\ &\quad + 9\left(\|\mathbf{x}^*\|_2^2 - x_j^{*2}\right)^2.\end{aligned}$$

Constructing variable $X_i = \|\mathbf{x}^*\|_2^2 + 2x_j^{*2} - y_i^2 a_{ij}^2$ which is upper bounded, with zero mean and bounded variance, we can use Lemma F.3 to establish a concentration bound with parameters:

$$\begin{aligned}X_i &\le \|\mathbf{x}^*\|_2^2 + 2x_j^{*2} \le 3\,\|\mathbf{x}^*\|_2^2, \\ \mathbb{E}\left[X_i\right] &= 0, \\ \mathbb{E}\left[X_i^2\right] &= 20x_j^{*4} + 68\,\|\mathbf{x}^*\|_2^2\,x_j^{*2} + 8\,\|\mathbf{x}^*\|_2^4 \le 96\,\|\mathbf{x}^*\|_2^4.\end{aligned}$$

Using Lemma F.3, for each $j \in S$,

$$\begin{aligned}&\mathbb{P}\left[\sum_{i=1}^{m} -X_i \le -t\right] \\ &\quad = \mathbb{P}\left[\sum_{i=1}^{m} y_i^2 a_{ij}^2 - m\left(\|\mathbf{x}^*\|_2^2 + 2x_j^{*2}\right) \le -t\right], \\ &\quad \le \exp\left(-\frac{t^2}{192\,\|\mathbf{x}^*\|_2^4\,m}\right) \le \frac{1}{ms}.\end{aligned} \tag{24}$$

Here, t is chosen to be large enough, such that $t = \sqrt{192}\,\|\mathbf{x}^*\|_2^2\sqrt{m\log ms} \approx 13.86\,\|\mathbf{x}^*\|_2^2\sqrt{m\log ms}$ $\le 13.86\,\|\mathbf{x}^*\|_2^2\sqrt{m\log mn}$ and s is the sparsity level. This establishes the bound on the first term $\frac{1}{m}\sum_{i=1}^{m} y_i^2 a_{ij}^2$. Similarly, we can establish a bound on the second term $\frac{1}{m}\sum_{i=1}^{m} y_i^2$, which requires Lemma 4.1 of [66], with probability greater than $1-\frac{1}{ms}$, for each $j \in S$:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m} y_i^2 - \|\mathbf{x}^*\|_2^2 &\le \left(2\sqrt{\frac{\log ms}{m}} + \frac{2\log ms}{m}\right)\|\mathbf{x}^*\|_2^2, \\ &\le 3\,\|\mathbf{x}^*\|_2^2\sqrt{\frac{\log ms}{m}}, \\ &\le 3\,\|\mathbf{x}^*\|_2^2\sqrt{\frac{\log mn}{m}}.\end{aligned} \tag{25}$$

for $m > C$. Combining these two concentration bounds (24), (25), taking a union bound for all $j \in S_+$ and substituting in (23):

$$M_{jj} - \phi^2 \ge 2x_j^{*2} - 17\sqrt{\frac{\log mn}{m}}\,\|\mathbf{x}^*\|_2^2, \tag{26}$$

which holds with probability at least $1-\frac{2}{m}$.

If the set S_+, is constructed as in (20), then evaluating the bound in (26), we get:

$$\begin{aligned} M_{jj} - \phi^2 &\geq 2x_j^{*2} - 17\sqrt{\frac{\log mn}{m}}\left\|\mathbf{x}^*\right\|_2^2, \\ M_{jj} &\geq \left(1 + 2x_j^{*2} - 19\sqrt{\frac{\log mn}{m}}\right)\left\|\mathbf{x}^*\right\|_2^2, \\ &\geq \left(1 + 11\sqrt{\frac{\log mn}{m}}\right)\left\|\mathbf{x}^*\right\|_2^2, \end{aligned}$$

holds for all elements $j \in S_+$, with probability greater than $1 - \frac{2}{m}$, yielding the bound in (19). □

Lemma A.3. *If $\hat{S}$ is chosen as in Algorithm 1, with probability greater than $1 - \frac{2}{m}$,*

$$\left\|\mathbf{x}^* - \mathbf{x}^*_{\hat{S}}\right\|_2 \leq \delta_1 \left\|\mathbf{x}^*\right\|_2, \tag{27}$$

as long as the number of measurements m follow the following bound

$$m \geq Cs^2 \log mn. \tag{28}$$

Proof. If $\hat{S}$ is chosen such that it corresponds to the top-s marginals M_{jj}, then it will pick up S_+ corresponding to large marginals $M_{jj} > \Theta$, $S_1 = S_- \cap \hat{S}$ and $S_2 = S^c \cap \hat{S}$ corresponding to small marginals $M_{jj} < \Theta$ (S_+, S_1, S_2 form a partition of $\widehat{S}$ and $\mathrm{card}(\hat{S}) = s$, refer Figure 8 for illustration of the sets):

$$\mathbf{x}^*_{\hat{S}} = \mathbf{x}^*_{S_+} + \mathbf{x}^*_{S_1} + \mathbf{x}^*_{S_2}. \tag{29}$$

By definition $\mathbf{x}_{S^c} = \mathbf{0}$ and therefore $\mathbf{x}_{S_2} = \mathbf{0}$. If we can prove that $\mathbf{x}^* \approx \mathbf{x}^*_{\hat{S}}$ and $\mathbf{x}^*_{\hat{S}} \approx \mathbf{x}^{\mathbf{0}}$, then we can claim that $\mathbf{x}^{\mathbf{0}} \approx \mathbf{x}^*$. First, we prove that $\left\|\mathbf{x}^* - \mathbf{x}^*_{\hat{S}}\right\|_2 \leq \delta_1 \left\|\mathbf{x}^*\right\|_2$:

$$\begin{aligned} \left\|\mathbf{x}^* - \mathbf{x}^*_{\hat{S}}\right\|_2^2 &= \left\|\mathbf{x}^* - \mathbf{x}^*_{S_+} - \mathbf{x}^*_{S_1}\right\|_2^2, \\ &\leq \left\|\mathbf{x}^* - \mathbf{x}^*_{S_+}\right\|_2^2 + \left\|\mathbf{x}^*_{S_1}\right\|_2^2, \\ &\leq \left\|\mathbf{x}^* - \mathbf{x}^*_{S_+}\right\|_2^2 + \left\|\mathbf{x}^*_{S_-}\right\|_2^2. \end{aligned}$$

By construction, S_- and S_+ form a partion of S:

$$\begin{aligned} \mathbf{x}^* &= \mathbf{x}^*_{S_-} + \mathbf{x}^*_{S_+}, \\ \implies \left\|\mathbf{x}^* - \mathbf{x}^*_{\hat{S}}\right\|_2^2 &\leq 2\left\|\mathbf{x}^*_{S_-}\right\|_2^2. \end{aligned}$$

Using (22), we compute the bound,

$$\begin{aligned} \left\|\mathbf{x}^* - \mathbf{x}^*_{\hat{S}}\right\|_2^2 &\leq 30\sqrt{\frac{s^2 \log mn}{m}}\left\|\mathbf{x}^*\right\|_2^2, \\ &\leq \delta_1^2 \left\|\mathbf{x}^*\right\|_2^2. \end{aligned} \tag{30}$$

since there are *at most* s elements in S_-, which is the required condition (27). This requires sample complexity m to satisfy:

$$\begin{aligned} 30\sqrt{\frac{s^2 \log mn}{m}} &\leq \delta_1^2, \\ \implies m \geq \frac{900}{\delta_1^2} s^2 \log mn &= C(\delta_1) s^2 \log mn. \end{aligned} \tag{31}$$

□

We have proved that $\mathbf{x}^* \approx \mathbf{x}^*_{\hat{S}}$. Now we need to prove that $\mathbf{x}^*_{\hat{S}} \approx \mathbf{x}^{\mathbf{0}}$, which we do using Lemma A.4.

Lemma A.4. *With probability greater than $1 - \frac{8}{m}$*

$$\begin{aligned} \mathrm{dist}\left(\mathbf{x}^{\mathbf{0}}, \mathbf{x}^*_{\hat{S}}\right) &\equiv \min\left(\left\|\mathbf{x}^{\mathbf{0}} - \mathbf{x}^*_{\hat{S}}\right\|_2, \left\|\mathbf{x}^{\mathbf{0}} + \mathbf{x}^*_{\hat{S}}\right\|_2\right), \\ &\leq \delta_2 \left\|\mathbf{x}^*\right\|_2, \end{aligned}$$

as long as the number of measurements m follow the following bound

$$m \geq Cs \log n.$$

Proof. The top singular vector of $\mathbb{E}[\mathbf{M}]$ is equal to true $\mathbf{x}^*$, from (3):

$$\begin{aligned} \mathbb{E}[\mathbf{M}] &= \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} y_j^2 \mathbf{a}_i \mathbf{a}_i{}^\top\right], \\ &= \left(\mathbf{I}_{n\times n} + 2\frac{\mathbf{x}^*}{\left\|\mathbf{x}^*\right\|_2}\frac{\mathbf{x}^{*\top}}{\left\|\mathbf{x}^*\right\|_2}\right)\left\|\mathbf{x}^*\right\|_2^2, \\ \text{similarly, } \mathbb{E}[\mathbf{M}_S] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_{iS} \mathbf{a}_{iS}{}^\top\right], \\ &= \left((\mathbf{I}_{n\times n})_S + 2\frac{\mathbf{x}^*}{\left\|\mathbf{x}^*\right\|_2}\frac{\mathbf{x}^{*\top}}{\left\|\mathbf{x}^*\right\|_2}\right)\left\|\mathbf{x}^*\right\|_2^2, \\ &= \mathbb{E}[\mathbf{M}]. \end{aligned}$$

We then define $\mathbf{M}_{\hat{S}} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_{i\hat{S}} \mathbf{a}_{i\hat{S}}^\top$ and $\mathbf{x}^{\mathbf{0}}$ is the top singular vector of $\mathbf{M}_{\hat{S}}$.

Defining $S_3 \equiv (S \cup S_2) \subset (S \cup \hat{S})$, where $S_2 = \hat{S} \cap S^c$, then, $\mathrm{card}(S_3) \leq 2s$, and,

$$\begin{aligned} \mathbb{E}[\mathbf{M}_{S_3}] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_{iS_3} \mathbf{a}_{iS_3}^\top\right], \\ &= \left((\mathbf{I}_{n\times n})_{S_3} + 2\frac{\mathbf{x}^*}{\left\|\mathbf{x}^*\right\|_2}\frac{\mathbf{x}^{*\top}}{\left\|\mathbf{x}^*\right\|_2}\right)\left\|\mathbf{x}^*\right\|_2^2. \end{aligned}$$

At this stage, we can invoke the proof idea from [22], as stated in Lemma F.2 from Appendix F, to give the following bound,

$$\left\|\mathbf{M}_{S_3} - \mathbb{E}[\mathbf{M}_{S_3}]\right\|_2 \leq \delta \left\|\mathbf{x}^*\right\|_2^2,$$

with probability at least $1 - \frac{1}{m}$, as long as $m \geq Cs \log n$. Now we can use the fact that $\hat{S} \subset S_3$, so that,

$$\left\|\mathbf{M}_{\hat{S}} - \mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]\right\|_2 \leq \left\|\mathbf{M}_{S_3} - \mathbb{E}[\mathbf{M}_{S_3}]\right\|_2 \leq \delta \left\|\mathbf{x}^*\right\|_2^2.$$

Since $\mathbf{M}_{\hat{S}}$ can be seen as a perturbation of $\mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]$, where the top two singular values of $\mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]$ are spaced $2\left\|\mathbf{x}^*_{\hat{S}}\right\|_2^2$ apart, we can use the Sin-Theta theorem [67] to bound the difference between the normalized top-singular vectors $\mathbf{x}^{\mathbf{0}}$ of $\mathbf{M}_{\hat{S}}$ and $\mathbf{x}_{\hat{S}}$ of $\mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]$ as,

$$\begin{aligned} \left\|\sin\angle(\mathbf{x}^{\mathbf{0}}, \mathbf{x}^*_{\hat{S}})\right\|_2 &\leq \frac{\delta \left\|\mathbf{x}^*\right\|_2^2}{2\left\|\mathbf{x}^*\right\|_2^2} = \frac{\delta}{2}. \\ \Rightarrow \min\left(\left\|\mathbf{x}^{\mathbf{0}} - \mathbf{x}^*_{\hat{S}}\right\|_2, \left\|\mathbf{x}^{\mathbf{0}} + \mathbf{x}^*_{\hat{S}}\right\|_2\right) &\leq \sqrt{2 - \sqrt{4 - \delta^2}}, \\ &\leq \delta_2. \end{aligned}$$

Hence, with probability greater than $1-\frac{8}{m}$, Lemma A.4 holds. □

Combining Lemmas A.3 and A.4, we have the final result:

$$\begin{aligned}\text{dist}\left(\mathbf{x^0},\mathbf{x}^*\right) &= \min\left(\left\|\mathbf{x^0}-\mathbf{x}^*\right\|_2,\left\|\mathbf{x^0}+\mathbf{x}^*\right\|_2\right),\\ &\leq \delta_0\left\|\mathbf{x}^*\right\|_2,\end{aligned}$$

as long as the number of measurements m follow the bound in (28). Hence the initial vector $\mathbf{x^0}$ is upto a constant factor away from the true vector $\mathbf{x}^*$. The constant $\delta_0 \leq \delta_1+\delta_2$ can be decreased by increasing the number of samples (see equation (31)). This completes the proof of Theorem IV.1.

Appendix B
Block CoPRAM initialization

In this section we state the proofs related to the *initialization* for Block CoPRAM in Algorithm 3, for block sparse signals.

We prove theorem V.1 for the initialization stage of Block CoPRAM as follows.

Theorem V.1. *The initial vector $\mathbf{x^0} \in \mathcal{M}_s^b$, which is the output from Algorithm 3, is a small constant distance away from the true signal $\mathbf{x}^* \in \mathcal{M}_s^b$, i.e.*

$$\text{dist}\left(\mathbf{x^0},\mathbf{x}^*\right) \leq \delta_b\left\|\mathbf{x}^*\right\|_2,$$

for $0<\delta_b<1$, as long as the number of measurements satisfy

$$m \geq C\frac{s^2}{b}\log mn,$$

with probability greater than $1-\frac{8}{m}$.

Proof. Evaluating the marginals $M_{j_bj_b}$, for all $j_b \in S_b^c$, from (16), with probability greater than $1-\frac{5}{m}$, we have:

$$M_{j_bj_b} \leq \left(1+11\sqrt{\frac{\log mn}{m}}\right)\sqrt{b}\left\|\mathbf{x}^*\right\|_2^2. \tag{32}$$

Evaluating the block marginals $M_{j_bj_b}$, for $j_b \in S_b$, we use a modification of (24), with probability less than $\exp\left(-\frac{mt^2}{192\|\mathbf{x}^*\|_2^4}\right) \leq \frac{1}{mn}$,

$$\frac{1}{m}\sum_{i=1}^{m} -X_i \leq -t$$

$$\frac{1}{m}\sum_{i=1}^{m} y_i^2a_{ij}^2 - \left(\left\|\mathbf{x}^*\right\|_2^2+2x_j^{*2}\right) \leq -t$$

Rearranging the terms,

$$\begin{aligned}\sum_{j\in j_b} M_{jj}^2 &\leq \sum_{j\in j_b}\left[\left(\left\|\mathbf{x}^*\right\|_2^2-t\right)+2x_j^{*2}\right]^2,\\ &\leq b\left(\left\|\mathbf{x}^*\right\|_2^2-t\right)^2+4\left\|\mathbf{x}_{j_b}^*\right\|_2^4\\ &\qquad +4\sqrt{b}\left\|\mathbf{x}_{j_b}^*\right\|_2^2\left(\left\|\mathbf{x}^*\right\|_2^2-t\right),\\ \implies M_{j_bj_b} &\leq \sqrt{b}\left(\left\|\mathbf{x}^*\right\|_2^2-t\right)+2\left\|\mathbf{x}_{j_b}^*\right\|_2^2,\end{aligned}$$

where the final expression holds with probability less than $\frac{b}{mn}$. Here, we have used he shorthand $\left\|\mathbf{x}_{j_b}^*\right\|_2^2 \equiv \sum_{j\in j_b} x_j^{*2}$.

Finally, taking a minimum over all such block marginals $j_b \in S_b$, with probability greater than $1-\frac{1}{m}$,

$$\begin{aligned}M_{j_bj_b} &\geq \sqrt{b}\left(\left\|\mathbf{x}^*\right\|_2^2-t\right)+2\left\|\mathbf{x}_{j_b}^*\right\|_2^2,\\ &\geq \sqrt{b}\left\|\mathbf{x}^*\right\|_2^2+\left\|\mathbf{x}_{b_{min}}^*\right\|_2^2,\end{aligned}$$

if $\sqrt{b}t = \left\|\mathbf{x}_{b_{min}}^*\right\|_2^2 \equiv \min_{j_b\in S_b}\left\|\mathbf{x}_{j_b}^*\right\|_2^2$. Assuming that $\left\|\mathbf{x}_{b_{min}^*}\right\|_2^2 = \frac{C}{k}\left\|\mathbf{x}^*\right\|_2^2$, the following holds

$$\min_{j_b\in S_b} M_{j_bj_b} \geq \left(1+\frac{C}{\sqrt{bk}}\right)\sqrt{b}\left\|\mathbf{x}^*\right\|_2^2. \tag{33}$$

Equating the expression for probability,

$$\begin{aligned}m &\geq 192\frac{\left\|\mathbf{x}^*\right\|_2^4}{t^2}\log mn,\\ &\geq Cbk^2\log mn = C\frac{s^2}{b}\log mn,\end{aligned}$$

which puts a bound on the block marginals for $j_b \in S_b$.

Hence, as long as $m \geq C\frac{s^2}{b}\log n$, there is a clear separation in the marginals, using (33) and (32),

$$\begin{aligned}\min_{j_b\in S_b} M_{j_bj_b} &\geq \left(1+\frac{C}{\sqrt{bk}}\right)\sqrt{b}\left\|\mathbf{x}^*\right\|_2^2,\\ &> \left(1+11\sqrt{\frac{\log mn}{m}}\right)\sqrt{b}\left\|\mathbf{x}^*\right\|_2^2,\\ &\geq \max_{j_b\in S_b^c} M_{j_bj_b},\end{aligned}$$

where C is large enough. Given that there is a clear separation in the marginals, the block support $\hat{S}_b$ as picked up as in Algorithm 3, is exactly the true block support S_b.

It is then straightforward to show that the top singular vector of the truncated covariance matrix $\mathbf{M}_{\hat{S}_b}$ is actually close to the true block sparse vector $\mathbf{x}^*$, which holds with probability greater than $1-\frac{1}{m}$.

Thus far, the proof requires an assumption on $\left\|\mathbf{x}_{b_{min}}^*\right\|_2$. We do away with this assumption as follows:

For evaluating block marginals $M_{j_bj_b}$ for $j_b \in S_b^c$, we can use the result of Lemma A.1, to obtain the same bound as in (32), with probability greater than $1-\frac{5}{m}$,

$$M_{j_bj_b} \leq \left(1+11\sqrt{\frac{\log mn}{m}}\right)\sqrt{b}\left\|\mathbf{x}^*\right\|_2^2.$$

For evaluating block marginals $M_{j_bj_b}$ for $j_b \in S_b$ we can use equations (20) and (21), and extend this model of signal supports to block supports defined as:

$$S_{b-} = \left\{j_b \in S_b \mid \left\|\mathbf{x}_{j_b}^*\right\|_2^2 \equiv \sum_{j\in j_b} x_j^{*2} \leq 15\sqrt{\frac{b\log mn}{m}}\left\|\mathbf{x}^*\right\|_2^2\right\},$$

$$S_{b+} = \left\{j_b \in S_b \mid \left\|\mathbf{x}_{j_b}^*\right\|_2^2 \equiv \sum_{j\in j_b} x_j^{*2} > 15\sqrt{\frac{b\log mn}{m}}\left\|\mathbf{x}^*\right\|_2^2\right\}.$$

Using equation (26), and LHS of (45),

$$M_{jj} \geq 2x_j^{*2} - 17\sqrt{\frac{\log mn}{m}}\|\mathbf{x}^*\|_2^2 + \phi^2,$$
$$\geq 2x_j^{*2} + \left(1 - 19\sqrt{\frac{\log mn}{m}}\right)\|\mathbf{x}^*\|_2^2.$$

Constructing block marginals as $M_{j_b j_b} \equiv \sqrt{\sum_{j\in j_b} M_{jj}^2}$,

$$M_{j_b j_b} \geq \sqrt{b}\left(1 - 19\sqrt{\frac{\log mn}{m}}\right)\|\mathbf{x}^*\|_2^2 + 2\left\|\mathbf{x}^*_{j_b}\right\|_2^2,$$
$$\implies M_{j_b j_b} \geq \left(1 + 11\sqrt{\frac{b\log mn}{m}}\right)\|\mathbf{x}^*\|_2^2.$$

We can then extend the proof of Lemma A.3 to give the partitions,

$$\mathbf{x}^*_{\hat{S}_b} = \mathbf{x}^*_{S_{b+}} + \mathbf{x}^*_{S_1} + \mathbf{x}^*_{S_2},$$
$$\mathbf{x}^* = \mathbf{x}^*_{S_{b-}} + \mathbf{x}^*_{S_{b+}}.$$

and the inequalities:

$$\left\|\mathbf{x}^* - \mathbf{x}^*_{\hat{S}_b}\right\|_2^2 \leq 2\left\|\mathbf{x}^*_{S_{b-}}\right\|_2^2,$$
$$= 2\sum_{j_b\in S_{b-}}\|\mathbf{x}_{j_b}\|_2^2,$$
$$\leq 15k\sqrt{\frac{b\log mn}{m}}\|\mathbf{x}^*\|_2^2 \leq \delta\|\mathbf{x}^*\|_2^2.$$

This inequality gives us a bound on the number of measurements m, similar to (31),

$$m \geq \frac{15^2}{\delta^2}k^2 b\log mn = C(\delta)\frac{s^2}{b}\log mn,$$

with probability greater than $1 - \frac{7}{m}$. This gives us the evaluation of block-marginals for $j_b \in \hat{S}_b$ and S_b^c, respectively. It is then straightforward to show that the top singular vector of the truncated covariance matrix $\mathbf{M}_{\hat{S}_b}$, given $\hat{S}_b$ is actually close to the true block sparse vector $\mathbf{x}^*$ with probability greater than $1 - \frac{1}{m}$. □

APPENDIX C
COPRAM AND BLOCK COPRAM DESCENT

In this section we state the proofs related to the *descent to optimal solution* in Algorithm 2 (CoPRAM), for sparse signals and Algorithm 4 (Block CoPRAM), for block sparse signals. This includes the proof of Theorem IV.2 and Theorem V.2. We prove theorem IV.2 to show descent of the CoPRAM algorithm, as follows.

Note: For evaluation of the distance measure $\text{dist}(\cdot,\cdot)$, we only consider $\text{dist}(\mathbf{x}^t, \mathbf{x}^*) = \|\mathbf{x}^t - \mathbf{x}^*\|_2$, assuming that $\text{dist}(\mathbf{x}^0, \mathbf{x}^*) = \|\mathbf{x}^0 - \mathbf{x}^*\|_2$ at the end of initialization stage. We claim that wlog, the same results would hold, if $\text{dist}(\mathbf{x}^0, \mathbf{x}^*) = \left\|\mathbf{x}^0 + \mathbf{x}^*\right\|_2$.

Theorem IV.2. *Given an initialization $\mathbf{x}^0 \in \mathcal{M}_s$ satisfying $\text{dist}(\mathbf{x}^0, \mathbf{x}^*) \leq \delta_0\|\mathbf{x}^*\|_2$, for $0 < \delta_0 < 1$, if we have number of (Gaussian) measurements $m > Cs\log\frac{n}{s}$, then the iterates of Algorithm 2 satisfy:*

$$\text{dist}(\mathbf{x}^{t+1}, \mathbf{x}^*) \leq \rho_0 \text{dist}(\mathbf{x}^t, \mathbf{x}^*), \tag{10}$$

where $0 < \rho_0 < 1$ is a constant, with probability greater than $1 - e^{-\gamma m}$, for positive constant γ.

Algorithm 5 CoSaMP

input $\Phi = \frac{\mathbf{A}}{\sqrt{m}}, \mathbf{u} = \frac{\mathbf{P}^t\mathbf{y}}{\sqrt{m}}, s, \mathbf{x}^t$.

1: Initialize
$\mathbf{x}^{t+1,0} \leftarrow \mathbf{x}^t$ initialize to best possible estimate
$\mathbf{r} \leftarrow \mathbf{u}$ residue
$l \leftarrow 0$ CoSaMP internal counter

2: **while** halting condition not true, **do**

3:
$l \leftarrow l + 1$
$\mathbf{v} \leftarrow \Phi^T\mathbf{r}$ signal proxy
$\Omega \leftarrow \text{supp}(\mathbf{v}_{2s})$
$\Gamma \leftarrow \Omega \cup \text{supp}(\mathbf{x}^{t+1,l-1})$
$\mathbf{w} \leftarrow \Phi_\Gamma^\dagger\mathbf{u}$ corresponding to Γ, 0 elsewhere
$\mathbf{x}^{t+1,l} \leftarrow$ Truncate to top s values of $\mathbf{w}$, call this support Γ_s
$\mathbf{r} \leftarrow \mathbf{u} - \Phi\mathbf{x}^{t+1,l}$

4: **end while**

5: $\mathbf{x}^{t+1,L} \leftarrow \Phi^\dagger_{\Gamma_s} u$.

output $\mathbf{x}^{t+1} \leftarrow \mathbf{x}^{t+1,L}$

To show the descent of our alternating minimization algorithm using CoSaMP, we need to analyze the reduction in error, per step of CoSaMP, (refer Algorithm 5) first:

$$\left\|\mathbf{x}^{t+1,l+1} - \mathbf{x}^*\right\|_2 = \left\|\mathbf{x}^{t+1,l+1} - \mathbf{w} + \mathbf{w} - \mathbf{x}^*\right\|_2,$$
$$\leq 2\|\mathbf{x}^* - \mathbf{w}\|_2 \tag{34}$$

where $\mathbf{w}$ corresponds to the ℓ'th run of CoSaMP for the $(t+1)^{th}$ update of $\mathbf{x}$. Using RIP of $\Phi = \frac{\mathbf{A}}{\sqrt{m}}$,

$$\left\|\mathbf{x}^{t+1,l+1} - \mathbf{x}^*\right\|_2 \leq \frac{2}{\sqrt{1-\delta_{2s}}}\|\Phi\mathbf{x}^* - \Phi\mathbf{w}\|_2, \tag{35}$$

with high probability, where δ_{2s} is the RIP constant. Now, analyzing the inputs to CoSaMP, in step 4 of Algorithm 2,

$$\mathbf{u} = \frac{\mathbf{P}^t\mathbf{y}}{\sqrt{m}},$$
$$= \text{sign}(\mathbf{A}\mathbf{x}^t) \circ \frac{|\mathbf{A}\mathbf{x}^*|}{\sqrt{m}},$$
$$= \text{sign}(\Phi\mathbf{x}^t) \circ \{(\Phi\mathbf{x}^*) \circ \text{sign}(\Phi\mathbf{x}^*)\},$$
$$= \Phi\mathbf{x}^* + (\text{sign}(\Phi\mathbf{x}^t) \pm \text{sign}(\Phi\mathbf{x}^*)) \circ \Phi\mathbf{x}^*,$$
$$\implies \mathbf{u} - \Phi\mathbf{x}^* = \pm(\text{sign}(\Phi\mathbf{x}^t) - \text{sign}(\Phi\mathbf{x}^*)) \circ \Phi\mathbf{x}^*, \tag{36}$$
$$= E_{ph},$$

where $E_{ph} \equiv (\text{sign}(\Phi\mathbf{x}^t) \pm \text{sign}(\Phi\mathbf{x}^*)) \circ \Phi\mathbf{x}^*$, is error due to failure in estimating the correct phase.

Using equation (36) and substituting into equation (35), the per-step reduction in error for each run of CoSaMP is:

$$\begin{aligned}
&\left\|\mathbf{x}^{t+1,l+1}-\mathbf{x}^{*}\right\|_2 \\
&\leq \frac{2}{\sqrt{1-\delta_{2s}}}\left\|\mathbf{u}-E_{ph}-\Phi\mathbf{w}\right\|_2, \\
&\leq \frac{2}{\sqrt{1-\delta_{2s}}}\left\|\mathbf{u}-\Phi\mathbf{w}\right\|_2+\frac{2}{\sqrt{1-\delta_{2s}}}\left\|E_{ph}\right\|_2, \\
&\leq \frac{2}{\sqrt{1-\delta_{2s}}}\left\|\mathbf{u}-\Phi_{\Gamma}\mathbf{w}_{\Gamma}\right\|_2+\frac{2}{\sqrt{1-\delta_{2s}}}\left\|E_{ph}\right\|_2, \\
&\leq \frac{2}{\sqrt{1-\delta_{2s}}}\left\|\mathbf{u}-\Phi_{\Gamma}\mathbf{x}^{*}_{\Gamma}\right\|_2+\frac{2}{\sqrt{1-\delta_{2s}}}\left\|E_{ph}\right\|_2, \\
&\leq \frac{2}{\sqrt{1-\delta_{2s}}}\left\|\Phi\mathbf{x}^{*}+E_{ph}-\Phi_{\Gamma}\mathbf{x}^{*}_{\Gamma}\right\|_2+\frac{2}{\sqrt{1-\delta_{2s}}}\left\|E_{ph}\right\|_2, \\
&\leq \frac{2}{\sqrt{1-\delta_{2s}}}\left\|\Phi\mathbf{x}^{*}-\Phi_{\Gamma}\mathbf{x}^{*}_{\Gamma}\right\|_2+\frac{4}{\sqrt{1-\delta_{2s}}}\left\|E_{ph}\right\|_2, \\
&\leq \frac{2}{\sqrt{1-\delta_{2s}}}\left\|\Phi_{\Gamma^c}\mathbf{x}^{*}_{\Gamma^c}\right\|_2+\frac{4}{\sqrt{1-\delta_{2s}}}\left\|E_{ph}\right\|_2, \\
&\leq 2\sqrt{\frac{1+\delta_{2s}}{1-\delta_{2s}}}\left\|\left(\mathbf{x}^{*}-\mathbf{x}^{t+1,l}\right)_{\Gamma^c}\right\|_2+\frac{4}{\sqrt{1-\delta_{2s}}}\left\|E_{ph}\right\|_2, \\
&:= \rho_1\left\|\left(\mathbf{x}^{*}-\mathbf{x}^{t+1,l}\right)_{\Gamma^c}\right\|_2+\rho_2\left\|E_{ph}\right\|_2,
\end{aligned}$$

where the first step is from using triangle inequality, the second step is from using the fact that $\mathbf{w}$ is exactly $3s$-sparse with support Γ. The third step is using the fact that truncation of $\mathbf{w}$ in $\Gamma, \in \mathbb{R}^{3s}$, is the minimizer of the LS problem $\text{argmin}_{\mathbf{x}\in\mathbb{R}^{3s}}\left\|\Phi_{\Gamma}\mathbf{x}-\mathbf{u}\right\|_2$, the fourth step uses (36) again. This is followed by a triangle inequality, and another use of RIP (which holds with probability greater than $1-e^{-\gamma_1 m}$, with γ_1 being a positive constant). Finally, in the last step, the first term is obtained by bounding $\left\|(\mathbf{x}^{*}-\mathbf{x}^{t+1,l})_{\Gamma^c}\right\|_2$ using (Lemma 4.2 of CoSaMP [41], refer Lemma F.4), to yeild,

$$\begin{aligned}
\left\|\mathbf{x}^{t+1,l+1}-\mathbf{x}^{*}\right\|_2 \leq \rho_1\rho_3\left\|\mathbf{x}^{*}-\mathbf{x}^{t+1,l}\right\|_2 \\
+\left(\rho_1\rho_4+\rho_2\right)\left\|E_{ph}\right\|_2,
\end{aligned}$$

where $\rho_1 := 2\sqrt{\frac{1+\delta_{2s}}{1-\delta_{2s}}}, \rho_2 := \frac{4}{\sqrt{1-\delta_{2s}}}$ and ρ_3, ρ_4 are as stated in Lemma F.4. Assuming that CoSaMP is let to run a maximum of L iterations,

$$\begin{aligned}
&\left\|\mathbf{x}^{t+1}-\mathbf{x}^{*}\right\|_2 \\
&\leq \left(\rho_1\rho_3\right)^{L}\left\|\mathbf{x}^{*}-\mathbf{x}^{t}\right\|_2 \\
&\qquad +\left(\rho_1\rho_4+\rho_2\right)\left(1+\rho_1\rho_3\ldots\left(\rho_1\rho_3\right)^{L-1}\right)\left\|E_{ph}\right\|_2, \\
&\leq \left(\rho_1\rho_3\right)^{L}\left\|\mathbf{x}^{*}-\mathbf{x}^{t}\right\|_2+\frac{\left(\rho_1\rho_4+\rho_2\right)}{\left(1-\rho_1\rho_3\right)}\left\|E_{ph}\right\|_2. \qquad (37)
\end{aligned}$$

The second part of this proof requires a bound on the phase error term $\left\|E_{ph}\right\|_2$:

$$E_{ph} = \pm\left(\text{sign}\left(\Phi\mathbf{x}^{t}\right)-\text{sign}\left(\Phi\mathbf{x}^{*}\right)\right)\circ\Phi\mathbf{x}^{*}.$$

We proceed to finish this proof by invoking Lemma C.1.

Lemma C.1. *As long as the initial estimate is a small distance away from the true signal* $\mathbf{x}^{*}\in\mathcal{M}_s$, $\text{dist}\left(\mathbf{x}^{\mathbf{0}},\mathbf{x}^{*}\right)\leq\delta_0\left\|\mathbf{x}^{*}\right\|_2$, *and subsequently,* $\text{dist}\left(\mathbf{x}^{t},\mathbf{x}^{*}\right)\leq\delta_0\left\|\mathbf{x}^{*}\right\|_2$, *where* $\mathbf{x}^{t}$ *is the* t^{th} *update of Algorithm 2, then the following bound holds,*

$$\frac{2}{m}\sum_{i=1}^{m}\left(\mathbf{a}_i{}^{T}\mathbf{x}^{*}\right)^2\cdot\mathbf{1}_{\{(\mathbf{a}_i{}^{T}\mathbf{x}^{t})(\mathbf{a}_i{}^{T}\mathbf{x}^{*})\leq 0\}}\leq\rho_5^2\left\|\mathbf{x}^{t}-\mathbf{x}^{*}\right\|_2^2,$$

with probability greater than $1-e^{-\gamma_2 m}$, *where* γ_2 *is a positive constant, as long as* $m > C\left(s+\log\left(\text{card}(\mathbb{M}_{4s})\right)\right)$ *and* $\rho_5^2 = 0.0128$.

The complete proof of Lemma C.1 can be found in Appendix F.

Using this in addition to equation (37), we have our final per-step error reduction for a single run of CoPRAM (Algorithm 2), as:

$$\begin{aligned}
\left\|\mathbf{x}^{t+1}-\mathbf{x}^{*}\right\|_2 &\leq \left(\left(\rho_1\rho_3\right)^{L}+\rho_5\frac{\left(\rho_1\rho_4+\rho_2\right)}{\left(1-\rho_1\rho_3\right)}\right)\left\|\mathbf{x}^{t}-\mathbf{x}^{*}\right\|_2, \\
&\leq \rho_0\left\|\mathbf{x}^{t}-\mathbf{x}^{*}\right\|_2, \qquad (38)
\end{aligned}$$

where $\rho_0 < 1$.

A. Evaluating convergence parameter ρ_0

To obtain per-step reduction in error, we require $\rho_0 < 1$. For sake of numerical analysis, $\delta_{2s}, \delta_{4s} \leq 0.0001$, then $\rho_1 \approx 2, \rho_3 \approx 0.0002$. Let $\delta_0 = 0.01$, then $\rho_5 \approx 0.12$. Similarly, $\rho_2 \approx 4$ and $\rho_4 \approx 2$. Suppose CoSaMP is allowed to run for $L = 5$ iterations then, $\rho_0 \approx 0.96 < 1$.

The inequalities used for CoSaMP, particularly (34) can be made tighter, which would give less tight restrictions on the factor δ_0, that controls how close the intial estimate is to the true signal $\mathbf{x}^{*}$.

We now restate theorem V.2 for Block CoPRAM as follows.

Theorem V.2. *Given an initialization* $\mathbf{x}^{\mathbf{0}}\in\mathcal{M}_s^b$, *satisfying* $\text{dist}\left(\mathbf{x}^{t},\mathbf{x}^{*}\right)\leq\delta_b\left\|\mathbf{x}^{*}\right\|_2$, *where* $0<\delta_b<1$, *if we have number of measurements* $m\geq C\left(s+\frac{s}{b}\log\frac{n}{s}\right)$, *then the iterates of Algorithm 4 satisfy:*

$$\text{dist}\left(\mathbf{x}^{t+1},\mathbf{x}^{*}\right)\leq\rho_b\text{dist}\left(\mathbf{x}^{t},\mathbf{x}^{*}\right). \qquad (15)$$

where $0<\rho_b<1$ *is a constant, with probability greater than* $1-e^{-\gamma m}$, *for positive constant* γ.

Proof. The proof for this theorem is a natural extension to the one we have proved in Theorem IV.2, adapted for *block* sparse signals. For the first part, the sequence of steps are the same as (35) - (37) in the proof of Theorem IV.2. The per-iteration error for the t^{th} iteration of Block CoPRAM, with L iterations of Model-based (block) CoSaMP, can be derived as:

$$\left\|\mathbf{x}^{t+1}-\mathbf{x}^{*}\right\|_2\leq\left(\rho_1\rho_3\right)^{L}\left\|\mathbf{x}^{*}-\mathbf{x}^{t}\right\|_2+\frac{\left(\rho_1\rho_4+\rho_2\right)}{\left(1-\rho_1\rho_3\right)}\left\|E_{ph}\right\|_2, \qquad (39)$$

where $\rho_1 = 2\sqrt{\frac{1+\delta_{M_{2s}}}{1-\delta_{M_{2s}}}}$ and $\rho_2 = \frac{4}{\sqrt{1-\delta_{M_{2s}}}}$ and $\delta_{M_{2s}}$ is the model-RIP [31] constant with parameter $2s$. Similarly, $\rho_3 = \frac{\delta_{M_{2s}}+\delta_{M_{4s}}}{1-\delta_{M_{2s}}}$ and $\rho_4 = \frac{2\sqrt{1+\delta_{M_{2s}}}}{1-\delta_{M_{2s}}}$ are obtained from a model-based extension of Lemma F.4, via Corollary F.5.

Finally, E_{ph} is the error in estimating phase in the t^{th} run of Block CoPRAM. The second part of this proof requires a bound on the phase error term $\|E_{ph}\|_2$:

$$\|E_{ph}\|_2^2 = \frac{2}{m}\sum_{i=1}^{m}\left(\mathbf{a}_i{}^T\mathbf{x}^*\right)^2 \cdot \mathbf{1}_{\{\text{sign}(\mathbf{a}_i\mathbf{x}^t)\,\text{sign}(\mathbf{a}_i\mathbf{x}^*)=-1\}}.$$

such that $\mathbf{x}^*, \mathbf{x}^t \in \mathcal{M}_s^b$. We obtain this via Corollary (C.2) of Lemma C.1:

Corollary C.2. *As long as the initial estimate is a small distance away from the true signal $\mathbf{x}^* \in \mathcal{M}_s^b$, $\text{dist}\left(\mathbf{x}^0, \mathbf{x}^*\right) \leq \delta_b \|\mathbf{x}^*\|_2$, and subsequently, $\text{dist}\left(\mathbf{x}^t, \mathbf{x}^*\right) \leq \delta_b \|\mathbf{x}^*\|_2$, where $\mathbf{x}^t$ is the t^{th} update of Algorithm 4, then the following bound holds,*

$$\frac{2}{m}\sum_{i=1}^{m}\left(\mathbf{a}_i{}^T\mathbf{x}^*\right)^2 \cdot \mathbf{1}_{\{(\mathbf{a}_i{}^T\mathbf{x}^t)(\mathbf{a}_i{}^T\mathbf{x}^*)\leq 0\}} \leq \rho_5^2 \left\|\mathbf{x}^t - \mathbf{x}^*\right\|_2^2,$$

with probability greater than $1-e^{-\gamma_2 m}$, where γ_2 is a positive constant, as long as $m > C\left(s + \frac{s}{b}\log\frac{n}{s}\right)$ and $\rho_5^2 = 0.0128$.

Proof. For the case of block sparse signals, we use the approach from section IV.D of [31]. In this case, the cardinality of $\mathcal{M}_s$ is:

$$\text{card}(\mathbb{M}_s) = \binom{\frac{n}{b}}{\frac{s}{b}} \leq \left(\frac{e\cdot(n/b)}{(s/b)}\right)^{\frac{s}{b}}$$
$$\implies \log\left(\text{card}(\mathbb{M}_{4s})\right) \leq \frac{4s}{b}\log\frac{n}{4s}$$

Hence the sample complexity is $m > C\left(s + \frac{s}{b}\log\frac{n}{s}\right)$. □

□

APPENDIX D
NOISE ROBUSTNESS

In this section, we show that both CoPRAM and Block CoPRAM are robust to noise, and establish the proof of Theorem IV.3.

Theorem IV.3. *Given Gaussian measurements $a_{ij} \in \mathcal{N}(0,1)$, CoPRAM can recover the model sparse signal $\mathbf{x}^{t_0} \in \mathcal{M}_s$ from noisy measurements $\mathbf{y}$ of the form*

$$\mathbf{y} = |\mathbf{A}\mathbf{x}^*| + \epsilon,$$

where $\epsilon \in \mathbb{R}^m$ is a scaled sub-exponential. This retains the previously derived expression for sample complexity as in Theorem IV.1 up to a constant factor. The algorithm converges according to the iteration invariant:

$$\left\|\mathbf{x}^{t_o} - \mathbf{x}^*\right\|_2 \leq c_1\|\mathbf{x}^*\|_2 + c_2\|\epsilon\|_2$$

where t_o is the number of outer iterations of CoPRAM and Block CoPRAM, $c_1 < 1$ and $c_2 = 200$.

We assume the modified version of (1) acquisition model:

$$\tilde{\mathbf{y}} = |\mathbf{A}\mathbf{x}^*| + \epsilon = \mathbf{y} + \epsilon \tag{40}$$

where ϵ is distributed according to a scaled sub-exponential random variable. If the variance of the noise is much smaller in comparison to the magnitude of the measurements, then the following approximation holds:

$$\tilde{y_i}^2 = y_i^2 + \eta_i \tag{41}$$

where η_i are sub-exponentail random variables (special case is Gaussian with distribution $\mathcal{N}(\mathbf{0}, \sigma^2)$).

A. CoPRAM Initialization

The marginals used for initialization will get modified as:

$$\tilde{M}_{jj} = M_{jj} + \frac{1}{m}\sum_{i=1}^{m}\eta_i a_{ij}^2 \tag{42}$$

Similarly signal power gets modified as:

$$\tilde{\phi}^2 = \phi^2 + \frac{1}{m}\sum_{i=1}^{m}\eta_i \tag{43}$$

Then much of the analysis for the initialization, follows from [22]. Key points in the proof stated in Appendix A get modified as follows:

Modified Lemma (A.1):

$$\tilde{M}_{jj} - \tilde{\phi}^2 = \left(M_{jj} - \phi^2\right) + \left(\frac{1}{m}\sum_{i=1}^{m}\eta_i(a_{ij}^2 - 1)\right)$$

where the second term is bounded as Equation (6.4) in [22] as

$$\max_{1\leq j\leq n}\left|\frac{1}{m}\sum_{i=1}^{m}\eta_i(a_{ij}^2-1)\right| \leq C\sigma\sqrt{\frac{\log mp}{m}}$$

with probability greater than $1 - \frac{4}{m}$ and the first term is bounded using (18).

Since we have assumed that the noise variance is much lesser than the signal power, $C\sigma \leq \alpha\|\mathbf{x}^*\|_2$. Hence, the new marginal threshold for $j \in S^c$ is:

$$\tilde{M}_{jj} - \tilde{\phi}^2 \leq (8+\alpha)\sqrt{\frac{\log mp}{m}}\|\mathbf{x}^*\|_2^2$$
$$\Rightarrow \tilde{M}_{jj} \leq \left(1 + (11+\alpha)\sqrt{\frac{\log mn}{n}}\right)\|\mathbf{x}^*\|_2^2 = \tilde{\Theta}$$

with probability greater than $1 - \frac{9}{m}$.

Modified Lemma (A.2):

Similarly, $\tilde{M}_{jj} - \tilde{\phi}^2$ is lower bounded for $j \in S_+$, using (26) along with Equation (6.4) in [22]:

$$\tilde{M}_{jj} \geq \tilde{\Theta}$$

where we redefine S_+ and S_- as

$$S_+ = \left\{j \in S \mid x_j^{*2} > \left(15 + \frac{\alpha}{2}\right)\sqrt{\frac{\log mn}{m}}\|\mathbf{x}^*\|_2^2\right\}.$$

Subsequently, we can define S_- as

$$S_- = \left\{j \in S \mid x_j^{*2} \leq \left(15 + \frac{\alpha}{2}\right)\sqrt{\frac{\log mn}{m}}\|\mathbf{x}^*\|_2^2\right\},$$

Lemmas (A.3) and (A.4) hold using these modifications, up to a constant factor. Theorem IV.1 holds with probability greater than $1 - \frac{12}{m}$.

B. CoPRAM Descent

In the Descent stage, apart from the *phase estimation* error and *signal estimation* error, we also have an additional measurement error term. This modification reflects in Equation (36) of Appendix C as:

$$\begin{aligned}\mathbf{u} - \Phi\mathbf{x}^* &= \pm\left(\text{sign}\left(\Phi\mathbf{x}^t\right) - \text{sign}\left(\Phi\mathbf{x}^*\right)\right) \circ \Phi\mathbf{x}^* \pm \text{sign}\left(\Phi\mathbf{x}^*\right) \circ \epsilon, \\ &= E_{ph} + E_m,\end{aligned}$$

where $E_m \equiv \pm\,\text{sign}\left(\Phi\mathbf{x}^*\right) \circ \epsilon$ is the measurement error. This error propagates to modify Equation (37) as:

$$\begin{aligned}&\left\|\mathbf{x}^{t+1} - \mathbf{x}^*\right\|_2 \\ &\leq (\rho_1\rho_3)^L \left\|\mathbf{x}^* - \mathbf{x}^t\right\|_2 + \frac{(\rho_1\rho_4 + \rho_2)}{(1-\rho_1\rho_3)} \left\|E_{ph} + E_m\right\|_2, \\ &\leq (\rho_1\rho_3)^L \left\|\mathbf{x}^* - \mathbf{x}^t\right\|_2 \\ &\qquad + \frac{(\rho_1\rho_4 + \rho_2)}{(1-\rho_1\rho_3)} \left\|E_{ph}\right\|_2 + \frac{(\rho_1\rho_4 + \rho_2)}{(1-\rho_1\rho_3)} \left\|\epsilon\right\|_2.\end{aligned}$$

Finally, the main convergence result from Equation (38) gets modified as:

$$\left\|\mathbf{x}^{t+1} - \mathbf{x}^*\right\|_2 \leq \rho_0 \left\|\mathbf{x}^t - \mathbf{x}^*\right\|_2 + \rho_m \left\|\epsilon\right\|_2,$$

where $\rho_m = \frac{(\rho_1\rho_4+\rho_2)}{(1-\rho_1\rho_3)} \approx 8$, as follows from the discussion in section (C-A) of Appendix C. After a total of t_o iterations of CoPRAM, we get the convergence result:

$$\begin{aligned}\left\|\mathbf{x}^{t_o} - \mathbf{x}^*\right\|_2 &\leq \rho_0^{t_o} \left\|\mathbf{x^0} - \mathbf{x}^*\right\|_2 + \frac{\rho_m}{1-\rho_0} \left\|\epsilon\right\|_2, \\ &\leq \rho_0^{t_o}\delta_0 \left\|\mathbf{x}^*\right\|_2 + \frac{\rho_m}{1-\rho_0} \left\|\epsilon\right\|_2 \approx 200 \left\|\epsilon\right\|_2,\end{aligned}$$

where the last approximation holds if one utilizes the analysis in section (C-A) of Appendix C, with $\rho_0 \approx 0.96$ and $\rho_m \approx 8$. However the coefficient of $\left\|\epsilon\right\|_2$ can be further lowered via tighter analysis of the bounds in Appendix C. Thus the quality of reconstruction only depends on level of input noise $\left\|\epsilon\right\|_2$, which is bounded with high probability.

Moreover, the number of iterations t_o depends on the accuracy of reconstruction expected. For example, if one intends to recover $\mathbf{x}_0^t$, such that $\left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2 / \left\|\mathbf{x}^*\right\|_2 \leq \epsilon_0$. Then the number of iteration varies logarithmically with $1/\epsilon_0$:

$$\begin{aligned}\epsilon_0 &= \rho_0^{t_o}\delta_0 + 200\frac{\left\|\epsilon\right\|_2}{\left\|\mathbf{x}^*\right\|_2} \\ \rho_0^{t_o} &= \frac{\epsilon_0 - 200\frac{\left\|\epsilon\right\|_2}{\left\|\mathbf{x}^*\right\|_2}}{\delta_0} \\ t_o &= \frac{\log(\delta_0) + \log(1/(\epsilon_0 - 200\frac{\left\|\epsilon\right\|_2}{\left\|\mathbf{x}^*\right\|_2}))}{\log(1/\rho_0)}.\end{aligned}$$

In the case of noiseless measurements, $t_o = \frac{\log\left(\frac{\delta_0}{\epsilon_0}\right)}{\log(1/\rho_0)}$.

Appendix E
Power Law Decay

In this section we state the proof of the sample complexity required for sparse signals following power-law decay, as described in Theorem IV.4.

Theorem IV.4. *Given Gaussian measurements $a_{ij} \in \mathcal{N}(0,1)$, then CoPRAM can recover the s-sparse signal $\mathbf{x}^{t_0} \in \mathcal{M}_s$, with $\left\|\mathbf{x}^{t_0} - \mathbf{x}^*\right\|_2 \leq \delta \left\|\mathbf{x}^*\right\|_2$, where t_0 is the number of outer iterations of CoPRAM, from $m > Cs\log n$ measurements, as long as the coefficients of the signal follow a power-law decay as described in (12).*

Note that a power-law decaying signal can be normalized as follows:

$$\begin{aligned}x_j^{*2} &\leq \frac{C(\alpha,s)}{j^\alpha}, \\ \left\|\mathbf{x}^*\right\|_2^2 &\leq C(\alpha,s)\sum_{j=1}^{s}\frac{1}{j^\alpha} := C(\alpha,s)\zeta(s,\alpha), \\ \left\|\mathbf{x}^*\right\|_2^2 + e &= C(\alpha,s)\zeta(s,\alpha), \\ \implies C(\alpha,s) &= \frac{\left\|\mathbf{x}^*\right\|_2^2 + e}{\zeta(s,\alpha)},\end{aligned}$$

where e is an indicator of the tightness of power-law inequality (we assume that $e < \left\|\mathbf{x}^*\right\|_2^2$) and $\zeta(s,\alpha) := \sum_{j=1}^{s}\frac{1}{j^\alpha}$. For $\alpha = 0$, $C(0,s) = \frac{\left\|\mathbf{x}^*\right\|_2 + e}{s}$. Numerical evaluation suggests that $\zeta(s,\alpha)$ is monotonically decreasing in α and monotonically increasing in s.

Additionally, we can find the index at which all power-law decaying signal elements fall bellow threshold Θ_0:

$$j' = \left\lfloor \left(\frac{C(\alpha,s)}{\Theta_0}\right)^{\frac{1}{\alpha}} \right\rfloor. \tag{44}$$

The task is to find a tighter bound for $\left\|\mathbf{x}^*{}_{S_-}\right\|_2$, as compared to (30). This bound can be established using our additional assumption of power law decay. We analyze this bound using the threshold criteria $\Theta_0 = 15\sqrt{\frac{\log mn}{m}}\left\|\mathbf{x}^*\right\|_2^2$, as in (20) and $\alpha > 1$:

$$\begin{aligned}&\left\|\mathbf{x}^*{}_{S_-}\right\|_2^2 \\ &\leq j'\Theta_0 + C\sum_{j=j'+1}^{s}\frac{1}{j^\alpha}, \\ &\leq j'\Theta_0 + C\int_{j'}^{s} j^{-\alpha}dj, \\ &\leq j'\Theta_0 + \frac{C}{\alpha-1}\left(\frac{1}{j'^{\alpha-1}} - \frac{1}{s^{\alpha-1}}\right), \\ &\leq j'\Theta_0 + \frac{C}{\alpha-1}\frac{1}{j'^{\alpha-1}} = \frac{\alpha}{\alpha-1}j'\Theta_0, \\ &= \frac{\alpha}{\alpha-1}\left(\frac{\left\|\mathbf{x}^*\right\|_2^2 + e}{\zeta(s,\alpha)}\frac{1}{15\left\|\mathbf{x}^*\right\|_2^2}\sqrt{\frac{m}{\log mn}}\right)^{1/\alpha}\Theta_0, \\ &= \frac{\alpha}{\alpha-1}\zeta(s,\alpha)^{-\frac{1}{\alpha}}\Theta_0^{1-\frac{1}{\alpha}} \\ &\quad + \frac{1}{\alpha-1}\left(\frac{e}{\left\|\mathbf{x}^*\right\|_2^2}\frac{1}{\zeta(s,\alpha)}\right)^{\frac{1}{\alpha}}\Theta_0^{1-\frac{1}{\alpha}} + (\text{higher order terms}) \\ &\lessapprox \frac{\alpha}{\alpha-1}\zeta(s,\alpha)^{-\frac{1}{\alpha}}\Theta_0^{1-\frac{1}{\alpha}} \quad (OR) < \frac{2^{\frac{1}{\alpha}}\alpha}{\alpha-1}\zeta(s,\alpha)^{-\frac{1}{\alpha}}\Theta_0^{1-\frac{1}{\alpha}} \\ &\leq \frac{\alpha}{\alpha-1}\Theta_0 \qquad\qquad\qquad\qquad (OR) \leq \frac{2^{\frac{1}{\alpha}}\alpha}{\alpha-1}\Theta_0\end{aligned}$$

where in the first step, we upper bound the summation with the integral of a curve lying above it; in the second step, we evaluate the integral; in the third step, we drop the $\left(\frac{1}{s^{\alpha-1}}\right)$ term; this is followed by substituting the value of j' via (44); subsequently, we do a binomial expansion. The inexactness in the inequality is originates from the terms in the expansion that depend on e, and is an exact inequality for $e = 0$ (tighter). We can also trivially bound e as $e < \|\mathbf{x}^*\|_2^2$, as per the setting of our problem (looser). In the penultimate step, we use the fact that $\zeta(s,\alpha)^{-1/\alpha} \leq 1$.

Finally, we equate this bound to $\delta_1^2 \|\mathbf{x}^*\|_2^2$ to obtain the following sample complexity:

$$\implies m \geq C(\delta_1, \alpha) \log mn.$$

In this regime, the sample complexity for the overall algorithm is dominated by the sample complexity for the descent stage ($\mathcal{O}(s \log n)$ = max($C \log mn$,$Cs \log n/s$)).

Appendix F
Supplementary Lemmas

In this section we state some of the lemmas with or without proofs, used in Appendices A and C.

Lemma F.1. *With probability of at least* $1 - \frac{1}{m}$,

$$\left(1 - 2\sqrt{\frac{\log m}{m}}\right) \|\mathbf{x}^*\|_2^2 \leq \phi^2 \leq \left(1 + 3\sqrt{\frac{\log m}{m}}\right) \|\mathbf{x}^*\|_2^2. \tag{45}$$

Proof. Rotational invariance property of Gaussian distributions imply that $\mathbf{y}_i^2 \equiv (\sum_{j=1}^n a_{ij} x_j^*)^2$ has the same distrubution as $a_{ij}^2 \|\mathbf{x}^*\|_2^2$. Using Lemma 4.1 of [66] on a_{ij}^2, we can obtain the upper bound,

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m a_{ij}^2 - 1 \geq 2\frac{\sqrt{m \log m}}{m} + 2\frac{\log m}{m}\right] \leq \exp(-\log m) = \frac{1}{m}.$$

Similarly, we can obtain the lower bound,

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m a_{ij}^2 - 1 \leq -2\frac{\sqrt{m \log m}}{m}\right] \leq \exp(-\log m) = \frac{1}{m}.$$

The signal power ϕ^2 is then bounded from below as

$$\left(1 - 2\sqrt{\frac{\log m}{m}}\right) \|\mathbf{x}^*\|_2^2 \leq \phi^2,$$

and similarly, it is bounded from above as,

$$\begin{aligned}\phi^2 &\leq \left(1 + 2\sqrt{\frac{\log m}{m}} + 2\frac{\log m}{m}\right) \|\mathbf{x}^*\|_2^2, \\ &< \left(1 + 3\sqrt{\frac{\log m}{m}}\right) \|\mathbf{x}^*\|_2^2,\end{aligned}$$

with probability at least $1 - \frac{1}{m}$, for $m > C$, large enough. If $m \approx 1000$, then the bounds are,

$$(1-\delta)\|\mathbf{x}^*\|_2^2 \leq \phi^2 \leq (1+\delta)\|\mathbf{x}^*\|_2^2,$$

where $\delta = 0.0207$. □

Lemma F.2. *With probability at least* $1 - \frac{1}{m}$, *the following holds,*

$$\left\|\frac{1}{m}\sum_{i=1}^m \left|\mathbf{a}_{i S_3}^\top \mathbf{x}^*\right|^2 \mathbf{a}_{i S_3}\mathbf{a}_{i S_3}^\top - \left(\|\mathbf{x}^*\|_2^2 \mathbf{I}_{S_3} + 2\mathbf{x}^*\mathbf{x}^{*\top}\right)\right\|_2 \leq \delta \|\mathbf{x}^*\|_2^2$$

where $\mathrm{card}(S_3) \leq 2s$, *provided* $m > C(\delta)(2s)\log(2s)$.

This proof has been adapted from Lemma A.6 of [22].

Lemma F.3. *Suppose* $X_1 \dots X_m$ *are i.i.d. centered, bounded real-valued random variables obeying*

$$\begin{aligned}X_i &\leq b, \\ \mathbb{E}[X_i] &= 0, \\ \mathbb{E}\left[X_i^2\right] &= v^2, \\ \sigma^2 &= \max\left\{b^2, v^2\right\},\end{aligned}$$

with cumulative distribution function of the standard normal distribution being denoted as

$$\begin{aligned}\Phi(x) &= \int_{-\infty}^x \phi(t)dt, \\ \phi(t) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right),\end{aligned}$$

then

$$\mathbb{P}\left[\sum_{i=1}^m X_i \geq t\right] \leq \min\left\{\exp\left(\frac{-t^2}{2\sigma^2}\right), 25\left(1 - \Phi\left(\frac{t}{\sigma}\right)\right)\right\}.$$

This establishes the tail probability of martingale with differences bounded from one side [68].

Lemma F.4. *The* $2s$*-sparse residual error* $\left\|(\mathbf{x}^* - \mathbf{x}^{t+1,l})_{\Gamma^c}\right\|_2$ *can be upper bounded as,*

$$\begin{aligned}\left\|(\mathbf{x}^* - \mathbf{x}^{t+1,l})_{\Gamma^c}\right\|_2 &\leq \left\|(\mathbf{x}^* - \mathbf{x}^{t+1,l})_{\Omega^c}\right\|_2 \\ &\leq \rho_3 \left\|(\mathbf{x}^* - \mathbf{x}^{t+1,l})\right\|_2 + \rho_4 \|E_{ph}\|_2\end{aligned}$$

where $\rho_3 = \frac{\delta_{2s}+\delta_{4s}}{1-\delta_{2s}}$ *and* $\rho_4 = \frac{2\sqrt{1+\delta_{2s}}}{1-\delta_{2s}}$.

This lemma has been adapted from Lemmas 4.2 and 4.3 of [41].

Corollary F.5. *The* $2s$*-sparse residual error* $\left\|(\mathbf{x}^* - \mathbf{x}^{t+1,l})_{\Gamma^c}\right\|_2$, *for* $\mathbf{x}^*, \mathbf{x}^t, \mathbf{x}^{t+1} \in \mathcal{M}_s$ *can be upper bounded as,*

$$\begin{aligned}\left\|(\mathbf{x}^* - \mathbf{x}^{t+1,l})_{\Gamma^c}\right\|_2 &\leq \left\|(\mathbf{x}^* - \mathbf{x}^{t+1,l})_{\Omega^c}\right\|_2 \\ &\leq \rho_3 \left\|(\mathbf{x}^* - \mathbf{x}^{t+1,l})\right\|_2 + \rho_4 \|E_{ph}\|_2\end{aligned}$$

where $\rho_3 = \frac{\delta_{M_{2s}}+\delta_{M_{4s}}}{1-\delta_{M_{2s}}}$ *and* $\rho_4 = \frac{2\sqrt{1+\delta_{M_{2s}}}}{1-\delta_{M_{2s}}}$ *and* $\delta_{M_{2s}}, \delta_{M_{4s}}$ *are model-RIP constants.*

Lemma C.1. *As long as the initial estimate is a small distance away from the true signal* $\mathbf{x}^* \in \mathcal{M}_s$, $\mathrm{dist}(\mathbf{x}^0, \mathbf{x}^*) \leq \delta_0 \|\mathbf{x}^*\|_2$, *and subsequently,* $\mathrm{dist}(\mathbf{x}^t, \mathbf{x}^*) \leq \delta_0 \|\mathbf{x}^*\|_2$, *where*

$\mathbf{x}^t$ is the t^{th} update of Algorithm 2, then the following bound holds,

$$\frac{2}{m}\sum_{i=1}^{m}\left(\mathbf{a}_i{}^T\mathbf{x}^*\right)^2\cdot\mathbf{1}_{\{(\mathbf{a}_i{}^T\mathbf{x}^t)(\mathbf{a}_i{}^T\mathbf{x}^*)\leq 0\}}\leq\rho_5^2\left\|\mathbf{x}^t-\mathbf{x}^*\right\|_2^2,$$

with probability greater than $1-e^{-\gamma_2 m}$, where γ_2 is a positive constant, as long as $m>C\left(s+\log\left(\mathrm{card}(\mathbb{M}_{4s})\right)\right)$ and $\rho_5^2=0.0128$.

Proof. This proof has been adapted from Lemma 3 of [50] and uses the generic chaining techniques of [43], [44]. A longer version can also be found in the full version of [64].

We are required to bound the following term:

$$\begin{aligned}\|E_{ph}\|_2^2&\leq\frac{2}{m}\sum_{i=1}^{m}\left(\mathbf{a}_i{}^\top\mathbf{x}^*\right)^2\cdot\mathbf{1}_{\{|\mathbf{a}_i{}^\top\mathbf{x}^*|<|\mathbf{a}_i{}^\top\mathbf{h}|\}}\\&\leq\frac{2}{m}\sum_{i=1}^{m}\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\cdot\mathbf{1}_{\{|\mathbf{a}_i{}^\top\mathbf{x}^*|<|\mathbf{a}_i{}^\top\mathbf{h}|\}},\\&\leq\frac{2}{m}\sum_{i=1}^{m}\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\right)\qquad(46)\\&\leq\frac{2}{m}\sum_{i=1}^{m}\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\cdot\mathbf{1}_{\{(1-\delta)|\mathbf{a}_i{}^\top\mathbf{x}^*|<|\mathbf{a}_i{}^\top\mathbf{h}|\}},\qquad(47)\\&:=\frac{2\|\mathbf{h}\|_2^2}{m}\sum_{i=1}^{m}\gamma_i\end{aligned}$$

where we have a fixed $\mathbf{h}$ defined as $\mathbf{h}=\mathbf{x}^t\pm\mathbf{x}^*$ ($\pm$ corresponds to sign of minimum ℓ_2 norm) and satisfying $\|\mathbf{h}\|_2\leq\delta_0\|\mathbf{x}^*\|_2$, δ is a small constant, and the pre-final steps in equations (46) and (47) can be obtained via auxiliary random Lipschitz approximations $\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\right)$, as in Eq. 52 of Section C.1 (refer Proof of Lemma 3) of [50].

Here we invoke Lemma 3 of [50], which we modify to suit our problem formulation. Firstly, we relax the constraint for the initial separation δ_0. Secondly, we calculate the expectation of random variable $\gamma_i:=\frac{(\mathbf{a}_i{}^\top\mathbf{h})^2}{\|\mathbf{h}\|_2^2}\mathbf{1}_{\{(1-\delta)|\mathbf{a}_i{}^\top\mathbf{x}^*|<|\mathbf{a}_i{}^\top\mathbf{h}|\}}$, by setting $\delta_0=0.0035$ and $\delta=0.01$. We therefore evaluate the integral expansion of $\mathbb{E}\left[\gamma_i\right]$, (Section C.1, proof of Lemma 3 of [50]) and this expression can be bounded as:

$$\begin{gathered}\mathbb{E}\left[\gamma_i\right]\leq 0.0063\quad\text{for}\quad\delta_0<0.0035\quad\text{and}\quad\delta=0.01,\\\implies\mathbb{E}\left[\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\right)\right]\leq 0.0063\|\mathbf{h}\|_2^2,\\(\text{for}\quad\delta_0<0.0035\quad\text{and}\quad\delta=0.01).\end{gathered}$$

Using Bernstein type inequality [69] on sub-exponential variable $\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\right)$,

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\right)}{\|\mathbf{h}\|_2^2}>(0.0063+\epsilon)\right]<\exp\left(-cm\epsilon^2\right).$$

At this point, we leverage the sparsity of the problem and consider a union bound over all $2s$-sparse $\mathbf{h}$'s (such that $\mathbf{x}^t$ and $\mathbf{x^0}$ are sparse and contained in $\mathcal{M}_s$) lying in an ϵ'- net $\mathcal{N}_{\epsilon'}$ sphere of radius $\delta_0\|\mathbf{x}^*\|_2$ and $\epsilon'=\epsilon\delta_0\|\mathbf{x}^*\|_2$. The ϵ'-net has cardinality $\mathrm{card}(\mathcal{N}_{\epsilon'})\leq\mathrm{card}(\mathbb{M}_{2s})\left(1+\frac{2}{\epsilon}\right)^{2s}$. For example, $\mathrm{card}(\mathcal{N}_{\epsilon'})\leq\binom{n}{2s}\left(1+\frac{2}{\epsilon}\right)^{2s}$ for general $2s$-sparse signals $\left(\mathrm{card}(\mathbb{M}_{2s})=\binom{n}{2s}\right)$.

Now the union bound over all such $\mathbf{h}_0\in\mathcal{N}_{\epsilon'}$, such that $\|\mathbf{h}-\mathbf{h}_0\|_2\leq\epsilon\|\mathbf{h}\|_2$ is:

$$\begin{aligned}&\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}_0\right)^2\right)}{\|\mathbf{h}_0\|_2^2}\leq(0.0063+\epsilon)\right]\qquad(48)\\&>1-\mathrm{card}(\mathbb{M}_{2s})\left(1+\frac{2}{\epsilon}\right)^{2s}\exp\left(-cm\epsilon^2\right),\qquad(49)\\&\forall\mathbf{h}_0\in\mathcal{N}_{\epsilon'}.\end{aligned}$$

Now, we bound the RHS of Eq.(46) as follows:

$$\begin{aligned}&\frac{2}{m}\sum_{i=1}^{m}\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\right)-\frac{2}{m}\sum_{i=1}^{m}\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}_0\right)^2\right)\\&\leq\frac{2}{m}\left|\sum_{i=1}^{m}\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\right)-\frac{2}{m}\sum_{i=1}^{m}\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}_0\right)^2\right)\right|\\&\leq\frac{2}{m}\sum_{i=1}^{m}\left|\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\right)-\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}_0\right)^2\right)\right|\\&\leq\frac{2}{m}\cdot\frac{1}{\delta}\sum_{i=1}^{m}\left|\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2-\left(\mathbf{a}_i{}^\top\mathbf{h}_0\right)^2\right|\qquad(50)\\&\leq 2\cdot\frac{c}{\delta}\left\|\mathbf{h}\mathbf{h}^\top-\mathbf{h}_0\mathbf{h}_0^\top\right\|_F\qquad(51)\\&\leq 2\cdot\frac{3c}{\delta}\|\mathbf{h}-\mathbf{h}_0\|_2\cdot\|\mathbf{h}\|_2\leq\frac{6c\epsilon}{\delta}\|\mathbf{h}\|_2^2\qquad(52)\end{aligned}$$

where (50) is due to the χ_i's being Lipschitz functions with constant $\frac{1}{\delta}$ and (51) and (52) are through Lemma F.6 and Lemma 2 of [49] respectively, with probability $1-c\,\mathrm{card}(\mathbb{M}_{4s})\exp(-Cm)$.

Lemma F.6. *For all symmetric rank-2 matrices $\mathbf{H}\in\mathbb{R}^{4s\times 4s}$, if $m>Cs$, then with probability $1-c\exp\left(-Cm\right)$,*

$$\frac{1}{m}\sum_{i=1}^{m}\left|\mathbf{a}_{i_\Omega}\mathbf{H}\mathbf{a}_{i_\Omega}^\top\right|\leq c\|\mathbf{H}\|_F,\qquad(53)$$

where Ω is a $4s$-dimensional support vector and $\mathbf{a}_{i_\Omega}\in\mathbb{R}^{4s}$ is a sub-vector of $\mathbf{a}_i$ (adapted from Lemma 1 of [49]).

Consequently, taking a union bound over all $4s$-dimensional subspaces in lying in n-dimension, the bound in (53) holds with probability at least $1-c\,\mathrm{card}(\mathbb{M}_{4s})\exp(-Cm)$, where $\mathbf{H}:=\left(\mathbf{h}_\Omega\mathbf{h}_\Omega^\top-\mathbf{h}_{0_\Omega}\mathbf{h}_{0_\Omega}^\top\right)$ and $\mathbf{h}_{0_\Omega}\in\mathbb{R}^{4s}$ and $\mathbf{h}_\Omega\in\mathbb{R}^{4s}$ are sub-vectors of $\mathbf{h}$ and $\mathbf{h}_0$, such that $\Omega:=supp(\mathbf{h})\cup supp(\mathbf{h}_0)$.

Effectively, we evaluate the sample complexity, by considering the probability with which the final expression in Equation 52 holds,

$$\begin{gathered}\mathrm{card}(\mathbb{M}_{4s})\exp\left(-cm\epsilon^2\right)<\delta,\\\implies m>C\left(s+\log\left(\mathrm{card}(\mathbb{M}_{4s})\right)\right).\end{gathered}$$

Specifically, for sparse signals, $\mathrm{card}(\mathbb{M}_{4s})=\binom{n}{4s}\leq\left(\frac{e\cdot n}{4s}\right)^{4s}$,

$$m>C\left(s+4s\log\frac{n}{4s}\right)>C's\log\frac{n}{s}.$$

Using the result at the end of (52), and combining with (48) we have,

$$\frac{2}{m}\sum_{i=1}^{m}\chi_i\left(\left(\mathbf{a}_i{}^\top\mathbf{h}\right)^2\right) \le 2\left(0.0063+\epsilon+\frac{3c\epsilon}{\delta}\right)\|\mathbf{h}\|_2^2 < 0.0128\,\|\mathbf{h}\|_2^2\,.$$

since ϵ can be chosen to be as small as required, hence concluding the proof of Lemma C.1. □

REFERENCES

[1] G. Jagatap and C. Hegde. Fast, sample-efficient algorithms for structured phase retrieval. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 4922–4932, 2017.

[2] Y. Shechtman, Y. Eldar, O. Cohen, H. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Sig. Proc. Mag.*, 32(3):87–109, 2015.

[3] R. Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

[4] A. Maiden and J. Rodenburg. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy*, 109(10):1256–1262, 2009.

[5] R. Harrison. Phase problem in crystallography. *JOSA a*, 10(5):1046–1055, 1993.

[6] J. Miao, T. Ishikawa, Q Shen, and T. Earnest. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annu. Rev. Phys. Chem.*, 59:387–410, 2008.

[7] R. Gerchberg and W. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35(237), 1972.

[8] J. Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[9] S. Marchesini. Phase retrieval and saddle-point optimization. *JOSA A*, 24(10):3289–3296, 2007.

[10] K. Nugent, A. Peele, H. Chapman, and A. Mancuso. Unique phase recovery for nonperiodic objects. *Physical review letters*, 91(20):203902, 2003.

[11] M. Fickus, D. Mixon, A. Nelson, and Y. Wang. Phase retrieval from very few measurements. *Linear Alg. Appl.*, 449:475–499, 2014.

[12] E. Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.*, 66(8):1241–1274, 2013.

[13] E. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.

[14] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 2796–2804, 2013.

[15] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[16] D. Needell, J. Tropp, and R. Vershynin. Greedy signal recovery review. In *Proc. Asilomar Conf. Sig. Sys. Comput.*, pages 1048–1050. IEEE, 2008.

[17] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[18] K. Do Ba, P. Indyk, E. Price, and D. Woodruff. Lower bounds for sparse recovery. In *Proc. ACM Symp. Discrete Alg. (SODA)*, pages 1190–1197, 2010.

[19] H. Ohlsson, A. Yang, R. Dong, and S. Sastry. Cprl–an extension of compressive sensing to the phase retrieval problem. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 1367–1375, 2012.

[20] Y. Chen, Y. Chi, and A. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Trans. Inform. Theory*, 61(7):4034–4059, 2015.

[21] K. Jaganathan, S. Oymak, and B. Hassibi. Sparse phase retrieval: Convex algorithms and limitations. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pages 1022–1026. IEEE, 2013.

[22] T. Cai, X. Li, and Z. Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *Ann. Stat.*, 44(5):2221–2251, 2016.

[23] G. Wang, L. Zhang, G. Giannakis, M. Akçakaya, and J. Chen. Sparse phase retrieval via truncated amplitude flow. *IEEE Trans. Sig. Proc.*, PP(99):1–1, 2017.

[24] M. Iwen, A. Viswanathan, and Y. Wang. Robust sparse phase retrieval made easy. *Ap. Comp. Har. An.*, 42(1):135–142, 2017.

[25] S. Bahmani and J. Romberg. Efficient compressive phase retrieval with constrained sensing vectors. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 523–531, 2015.

[26] H. Qiao and P. Pal. Sparse phase retrieval using partial nested Fourier samplers. In *Proc. IEEE Global Conf. Signal and Image Processing (GlobalSIP)*, pages 522–526. IEEE, 2015.

[27] S. Cai, M. Bakshi, S. Jaggi, and M. Chen. Super: Sparse signals with unknown phases efficiently recovered. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pages 2007–2011. IEEE, 2014.

[28] D. Yin, R. Pedarsani, X. Li, and K. Ramchandran. Compressed sensing using sparse-graph codes for the continuous-alphabet setting. In *Proc. Allerton Conf. on Comm., Contr., and Comp.*, pages 758–765. IEEE, 2016.

[29] R. Pedarsani, D. Yin, K. Lee, and K. Ramchandran. Phasecode: Fast and efficient compressive phase retrieval based on sparse-graph codes. *IEEE Trans. Inform. Theory*, 2017.

[30] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc. Stat. Meth.*, 68(1):49–67, 2006.

[31] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Info. Theory*, 56(4):1982–2001, Apr. 2010.

[32] Y. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans. Sig. Proc.*, 58(6):3042–3054, 2010.

[33] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *J. Machine Learning Research*, 12(Nov):3371–3412, 2011.

[34] M. Duarte, C. Hegde, V. Cevher, and R. Baraniuk. Recovery of compressible signals from unions of subspaces. In *Proc. IEEE Conf. Inform. Science and Systems (CISS)*, March 2009.

[35] C. Hegde, P. Indyk, and L. Schmidt. A fast approximation algorithm for tree-sparse recovery. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, June 2014.

[36] C. Hegde, P. Indyk, and L. Schmidt. Nearly linear-time model-based compressive sensing. In *Proc. Intl. Colloquium on Automata, Languages, and Programming (ICALP)*, July 2014.

[37] V. Cevher, P. Indyk, C. Hegde, and R. Baraniuk. Recovery of clustered sparse signals from compressive measurements. In *Proc. Sampling Theory and Appl. (SampTA)*, May 2009.

[38] C. Hegde, P. Indyk, and L. Schmidt. A nearly linear-time framework for graph-structured sparsity. In *Proc. Int. Conf. Machine Learning (ICML)*, July 2015.

[39] V. Cevher, M. Duarte, C. Hegde, and R. Baraniuk. Sparse signal recovery using Markov Random Fields. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, Dec. 2008.

[40] C. Hegde, P. Indyk, and L. Schmidt. Approximation-tolerant model-based compressive sensing. In *Proc. ACM Symp. Discrete Alg. (SODA)*, Jan. 2014.

[41] D. Needell and J. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Ap. Comp. Har. An.*, 26(3):301–321, 2009.

[42] M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *arXiv preprint arXiv:1702.06175*, 2017.

[43] M. Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2006.

[44] S. Dirksen. Tail bounds via generic chaining. *Electronic J. Probability*, 20, 2015.

[45] D. Gross, F. Krahmer, and R. Kueng. Improved recovery guarantees for phase retrieval from coded diffraction patterns. *Ap. Comp. Har. An.*, 42(1):37–64, 2017.

[46] E. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Ap. Comp. Har. An.*, 39(2):277–299, 2015.

[47] I. Waldspurger, A.d'Aspremont, and S. Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[48] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis pursuit. *arXiv preprint arXiv:1610.07531*, 2016.

[49] Y. Chen and E. Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 739–747, 2015.

[50] H. Zhang and Y. Liang. Reshaped wirtinger flow for solving quadratic system of equations. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 2622–2630, 2016.
[51] G. Wang and G. Giannakis. Solving random systems of quadratic equations via truncated generalized gradient flow. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 568–576, 2016.
[52] K. Wei. Solving systems of phaseless equations via kaczmarz methods: A proof of concept study. *Inverse Problems*, 31(12):125008, 2015.
[53] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pages 2379–2383. IEEE, 2016.
[54] X. Li and V. Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM J. Math. Anal.*, 45(5):3019–3033, 2013.
[55] K. Jaganathan, S. Oymak, and B. Hassibi. Recovery of sparse 1-d signals from the magnitudes of their fourier transform. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pages 1473–1477. IEEE, 2012.
[56] Y. Shechtman, A. Beck, and Y. C. Eldar. Gespar: Efficient phase retrieval of sparse signals. *IEEE Trans. Sig. Proc.*, 62(4):928–938, 2014.
[57] P. Schniter and S. Rangan. Compressive phase retrieval via generalized approximate message passing. *IEEE Trans. Sig. Proc.*, 63(4):1043–1055, 2015.
[58] C. Hegde, P. Indyk, and L. Schmidt. Fast algorithms for structured sparsity. *Bul. of the EATCS*, 1(117):197–228, Oct. 2015.
[59] Irène Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. *arXiv preprint arXiv:1609.03088*, 2016.
[60] G. Wang, G. Giannakis, Y. Saad, and J. Chen. Solving (almost) all systems of random quadratic equations. pages 1865–1875, 2017.
[61] Q. Qu, Y. Zhang, Y. Eldar, and J. Wright. Convolutional phase retrieval. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 6082–6092, 2017.
[62] G. Jagatap, Z. Chen, C. Hegde, and N. Vaswani. Sub-diffraction imaging using fourier ptychography and structured sparsity. *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc. (ICASSP)*, 2018.
[63] Z. Chen, G. Jagatap, S. Nayer, C. Hegde, and N. Vaswani. Low rank fourier ptychography. *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc. (ICASSP)*, 2018.
[64] G. Jagatap and C. Hegde. Towards sample-optimal methods for solving random quadratic equations with structure. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*. IEEE, 2018.
[65] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, 2010.
[66] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Stat.*, pages 1302–1338, 2000.
[67] C. Davis and W. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM J. Num. Anal.*, 7(1):1–46, 1970.
[68] V. Bentkus. An inequality for tail probabilities of martingales with differences bounded from one side. *J. Theoretical Prob.*, 16(1):161–173, 2003.
[69] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Gauri Jagatap is a PhD student in the Department of Electrical and Computer Engineering at Iowa State University, where she works with Dr. Chinmay Hegde. Her area of research comprises of developing fast and efficient algorithms for machine learning and signal processing applications. Prior to ISU, she completed dual degrees, with a Bachelor's in Electrical and Electronics Engineering and Master's in Physics from BITS Pilani University (India), in 2015.

Chinmay Hegde (S,'07, M, '12, SM '18) is an assistant professor, and Black and Veatch Faculty Fellow, in the Department of Electrical and Computer Engineering at Iowa State University. His research focuses on developing fast and robust algorithms for machine learning and statistical signal processing, with applications to imaging problems. Prior to this, he was a Shell-MIT Postdoctoral Fellow in CSAIL at the Massachusetts Institute of Technology. He is the recipient of several awards, including best paper awards at SPARS and ICML, a best poster award at MMLS, the Budd Award for Best Engineering Thesis at Rice University in 2013, the NSF CRII Award in 2016, the Boast-Nillson Award for Educational Impact in 2018, and the NSF CAREER Award in 2018.