Fast, Sample-Efficient Algorithms for Structured Phase Retrieval

Gauri jagatap
Electrical and Computer Engineering
Iowa State University
gauri@iastate.edu

Chinmay Hegde
Electrical and Computer Engineering
Iowa State University
chinmay@iastate.edu

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 $i = \{1, 2, \ldots, m\}$. Also known as the *phase retrieval problem*, it is a fundamental challenge in nano-, bio- and astronomical imaging systems, and speech processing. The problem is ill-posed, and therefore additional assumptions on the signal and/or the measurements are necessary.

In this paper, we first study the case where the underlying signal $\mathbf{x}^*$ is s-sparse. We develop a novel recovery algorithm that we call *Compressive Phase Retrieval with Alternating Minimization*, or *CoPRAM*. Our algorithm is simple and can 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 samples, which matches the best known existing results. It also demonstrates linear convergence in theory and practice and requires no extra tuning parameters other than the signal sparsity level s.

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, this constitutes the first end-to-end linearly convergent family of algorithms for phase retrieval where the Gaussian sample complexity has a sub-quadratic dependence on the sparsity level of the signal.

1 Introduction

1.1 Motivation

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

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

then the task is to recover $\mathbf{x}^*$ using the measurements $\mathbf{y}$ and the sampling matrix $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_m]^\top$.

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 such as diffraction imaging, X-ray crystallography, ptychography, and astronomy [1, 2, 3, 4, 5]. For such imaging systems, the optical sensors used for light acquisition can only record the intensity of the light waves but not their phase. In terms of our setup, the vector $\mathbf{x}^*$ corresponds to an image (with a resolution of n pixels) and the measurements correspond to the magnitudes of its 2D Fourier coefficients. The goal is to stably recover the image $\mathbf{x}^*$ using as few observations m as possible.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Despite the prevalence of several heuristic approaches [6, 7, 8, 9], it is generally accepted that (1) is a challenging nonlinear, ill-posed inverse problem in theory and practice. For generic $\mathbf{a}_i$ and $\mathbf{x}^*$, one can show that (1) is *NP-hard* by reduction from well-known combinatorial problems [10]. Therefore, additional assumptions on the signal $\mathbf{x}^*$ and/or the measurement vectors $\mathbf{a}_i$ are necessary.

A recent line of breakthrough results [11, 12] have provided efficient algorithms for the case where the measurement vectors arise from certain multi-variate probability distributions. The seminal paper by Netrapalli *et al.* [13] provides the first rigorous justification of classical heuristics for phase retrieval based on alternating minimization. However, all these newer results require an "overcomplete" set of observations, i.e., the number of observations m exceeds the problem dimension n ($m = \mathcal{O}(n)$ being the tightest evaluation of this bound [14]). This requirement can pose severe limitations on computation and storage, particularly when m and n are very large.

One way to mitigate the dimensionality issue is to use the fact that in practical applications, $\mathbf{x}^*$ often obeys certain *low-dimensional* structural assumptions. For example, in imaging applications $\mathbf{x}^*$ is s-*sparse* in some known basis, such as identity or wavelet. For transparency, we assume the canonical basis for sparsity throughout this paper. Similar structural assumptions form the core of sparse recovery, 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 $\mathbf{x}^*$, which is information-theoretically optimal [18].

Several approaches for solving the sparsity-constrained version of (1) have been proposed, including alternating minimization [13], methods based on convex relaxation [19, 20, 21], and iterative thresholding [22, 23]. Curiously, all of the above techniques incur a sample complexity of $\Omega(s^2 \log n)$ for stable recovery, which is quadratically worse than the information-theoretic limit [18] of $\mathcal{O}\left(s \log \frac{n}{s}\right)$[1]. Moreover, most of these algorithms have quadratic (or worse) running time [19, 22], stringent assumptions on the nonzero signal coefficients [13, 23], and require several tuning parameters [22, 23].

Finally, for specific applications, more refined structural assumptions on $\mathbf{x}^*$ 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 using block-sparsity [30, 31, 32, 33], tree sparsity [34, 30, 35, 36], clusters [37, 31, 38], and graph models [39, 38, 40]. However, these models have not been understood in the context of phase retrieval.

1.2 Our contributions

The contributions in this paper are two-fold. First, we provide a new, flexible algorithm for sparse phase retrieval that matches state of the art 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 is block-sparse, thereby *further* lowering the sample complexity of stable recovery. Our work can be viewed as a first step towards a general framework for phase retrieval of structured signals from Gaussian samples.

Sparse phase retrieval. We first study the case where the underlying signal $\mathbf{x}^*$ is s-sparse. We develop a novel recovery algorithm that we call *Compressive Phase Retrieval with Alternating Minimization*, or *CoPRAM*[2]. Our algorithm is simple and can be obtained via a natural combination of the classical alternating minimization approach for phase retrieval with the CoSaMP [41] algorithm for sparse recovery (CoSAMP also naturally extends to several sparsity models [30]). We prove that our algorithm achieves a sample complexity of $\mathcal{O}\left(s^2 \log n\right)$ with Gaussian measurement vectors $\mathbf{a}_i$ in order to achieve *linear convergence*, matching the best among all existing results. An appealing feature of our algorithm is that it requires no extra *a priori* information other than the signal sparsity level s, and no assumptions on the nonzero signal coefficients. To our knowledge, this is the first algorithm for sparse phase retrieval that simultaneously achieves all of the above properties. We use CoPRAM as the basis to formulate a block-sparse extension (Block CoPRAM).

Block-sparse phase retrieval. We consider the case where the underlying signal $\mathbf{x}^*$ arises from *structured* sparsity models, specifically *block-sparse* signals with uniform block size b (i.e., s nonzeros equally grouped into $k = s/b$ blocks). For this problem, we design a recovery algorithm that we

[1] Exceptions to this rule are [24, 25, 26, 27, 28, 29] where very carefully crafted measurements $\mathbf{a}_i$ are used.

[2] We use the terms *sparse phase retrieval* and *compressive phase retrieval* interchangeably.

Table 1: *Comparison of (Gaussian sample) sparse phase retrieval algorithms. Here,* $n, s, k = s/b$ *denote signal length, sparsity, and block-sparsity.* $\mathcal{O}_\epsilon(\cdot)$ *hides polylogarithmic dependence on* $\frac{1}{\epsilon}$.

Algorithm	Sample complexity	Running time	Assumptions	Parameters
AltMinSparse	$\mathcal{O}_\epsilon\left(s^2 \log n + s^2 \log^3 s\right)$	$\mathcal{O}_\epsilon\left(s^2 n \log n\right)$	$x^*_{\min} \approx \frac{c}{\sqrt{s}} \left\|\mathbf{x}^*\right\|_2$	none
ℓ_1-PhaseLift	$\mathcal{O}\left(s^2 \log n\right)$	$\mathcal{O}\left(\frac{n^3}{\epsilon^2}\right)$	none	none
Thresholded WF	$\mathcal{O}\left(s^2 \log n\right)$	$\mathcal{O}_\epsilon\left(n^2 \log n\right)$	none	step μ, thresholds α, β
SPARTA	$\mathcal{O}\left(s^2 \log n\right)$	$\mathcal{O}_\epsilon\left(s^2 n \log n\right)$	$x^*_{\min} \approx \frac{c}{\sqrt{s}} \left\|\mathbf{x}^*\right\|_2$	step μ, threshold γ
CoPRAM	$\mathcal{O}\left(s^2 \log n\right)$	$\mathcal{O}_\epsilon\left(s^2 n \log n\right)$	none	none
Block CoPRAM	$\mathcal{O}\left(ks \log n\right)$	$\mathcal{O}_\epsilon\left(ksn \log n\right)$	none	none

call *Block CoPRAM*. We analyze this algorithm and show that leveraging block-structure reduces the sample complexity for stable recovery to $\mathcal{O}(ks \log n)$. For sufficiently large block lengths $b = \Theta(s)$, this bound equates to $\mathcal{O}(s \log n)$. To our knowledge, this constitutes the first phase retrieval algorithm where the Gaussian sample complexity has a sub-quadratic dependence on the sparsity s of the signal.

A comparative description of the performance of our algorithms is presented in Table 1.

1.3 Techniques

Sparse phase retrieval. Our proposed algorithm, CoPRAM, is conceptually very simple. It integrates existing approaches in stable sparse recovery (specifically, the CoSaMP algorithm [41]) with the alternating minimization approach for phase retrieval proposed in [13].

A similar integration of sparse recovery with alternating minimization was also introduced in [13]; however, their approach only succeeds when the true support of the underlying signal is accurately identified during initialization, which can be unrealistic. 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 ours succeeds in the (more practical) setting of using all the available samples.

Our first challenge is to identify a good initial guess of the signal. As is the case with most non-convex techniques, CoPRAM requires an initial estimate $\mathbf{x^0}$ that is close to the true signal $\mathbf{x}^*$. The basic idea is to identify "important" co-ordinates by constructing suitable biased estimators of each signal coefficient, followed by a specific eigendecomposition. The initialization in CoPRAM is far simpler than the approaches in [22, 23]; requiring no pre-processing of the measurements and or tuning parameters other than the sparsity level s. A drawback of the theoretical results of [23] is that they impose a requirement on signal coefficients: $\min_{j \in S} |x^*_j| = C \left\|\mathbf{x}^*\right\|_2 / \sqrt{s}$. However, this assumption disobeys the power-law decay observed in real world signals. Our approach also differs from [22], where they estimate an initial support based on a parameter-dependent threshold value. Our analysis removes these requirements; we show that a coarse estimate of the support, coupled with the spectral technique in [22, 23] gives us a suitable initialization. A sample complexity of $\mathcal{O}\left(s^2 \log n\right)$ is incurred for achieving this estimate, matching the best available previous methods.

Our next challenge is to show that given a good initial guess, alternatingly estimating the phases and non-zero coefficients (using CoSaMP) gives a rapid convergence to the desired solution. To this end, we use the analysis of CoSaMP [41] and leverage a recent result by [42], to show per step decrease in the signal estimation error using the generic chaining technique of [43, 44]. In particular, we show that any "phase errors" made in the initialization, can be suitably controlled across different estimates.

Block-sparse phase retrieval. We use CoPRAM to establish its extension Block CoPRAM, which is a novel phase retrieval strategy for block sparse signals from Gaussian measurements. Again, the algorithm is based on a suitable initialization followed by an alternating minimization procedure, mirroring the steps in CoPRAM. To our knowledge, this is the first result for phase retrieval under more refined structured sparsity assumptions on the signal.

As above, the first stage consists of identifying a good initial guess of the solution. We proceed as in CoPRAM, isolating *blocks* of nonzero coordinates, by constructing a biased estimator for the "mass" of each block. We prove that a good initialization can be achieved using this procedure using only $\mathcal{O}(ks \log n)$ measurements. When the block-size is large enough ($b = \Theta(s)$), the sample complexity of the initialization is *sub-quadratic* in the sparsity level s and only a logarithmic factor above the

information-theoretic limit $\mathcal{O}(s)$ [30]. In the second stage, we demonstrate a rapid descent to the desired solution. To this end, we replace the CoSaMP sub-routine in CoPRAM with the *model-based CoSaMP* algorithm of [30], specialized to block-sparse recovery. The analysis proceeds analogously as above. To our knowledge, this constitutes the first end-to-end algorithm for phase retrieval (from Gaussian samples) that demonstrates a sub-quadratic dependence on the sparsity level of the signal.

1.4 Prior work

The phase retrieval problem has received significant attention in the past few years. Convex methodologies to solve the problem in the *lifted* framework include *PhaseLift* and its variations [11, 45, 46, 47]. Most of these approaches suffer severely in terms of computational complexity. *PhaseMax*, produces a convex relaxation of the phase retrieval problem similar to basis pursuit [48]; however it is not emperically competitive. Non-convex algorithms typically rely on finding a good initial point, followed by minimizing a quadratic (Wirtinger Flow [12, 14, 49]) or moduli ([50, 51]) measurement loss function. Arbitrary initializations have been studied in a polynomial-time trust-region setting in [52].

Some of the convex approaches in sparse phase retrieval include [19, 53], which uses a combination of trace-norm and ℓ-norm relaxation.Constrained sensing vectors have been used [25] at optimal sample complexity $\mathcal{O}\left(s \log \frac{n}{s}\right)$. Fourier measurements have been studied extensively in the convex [54] and non-convex [55] settings. More non-convex approaches for sparse phase retrieval include [13, 23, 22] which achieve Gaussian sample complexities of $\mathcal{O}\left(s^2 \log n\right)$.

Structured sparsity models such as groups, blocks, clusters, and trees can be used to model real-world signals.Applications of such models have been developed for sparse recovery [30, 33, 39, 38, 40, 56, 34, 35, 36] as well as in high-dimensional optimization and statistical learning [32, 31]. However, to the best of our knowledge, there have been no rigorous results that explore the impact of structured sparsity models for the phase retrieval problem.

2 Paper organization and notation

The remainder of the paper is organized as follows. In Sections 3 and 4, we introduce the CoPRAM and Block CoPRAM algorithms respectively, and provide a theoretical analysis of their statistical performance. In Section 5 we present numerical experiments for our algorithms.

Standard notation for matrices (capital, bold: $\mathbf{A}, \mathbf{P}$, etc.), vectors (small, bold: $\mathbf{x}, \mathbf{y}$, etc.) and scalars (α, c etc.) hold. Matrix and vector transposes are represented using $\top$ (eg. $\mathbf{x}^\top$ and $\mathbf{A}^\top$) respectively. The diagonal matrix form of a column vector $\mathbf{y} \in \mathbb{R}^m$ is represented as $\text{diag}(\mathbf{y}) \in \mathbb{R}^{m \times m}$. Operator $\text{card}(S)$ represents cardinality of S. Elements of $\mathbf{a}$ are distributed according to the zero-mean standard normal distribution $\mathcal{N}(0,1)$. The phase is denoted using $\text{sign}(\mathbf{y}) \equiv \mathbf{y}/|\mathbf{y}|$ for $\mathbf{y} \in \mathbb{R}^m$, and $\text{dist}(\mathbf{x}_1, \mathbf{x}_2) \equiv \min(\|\mathbf{x}_1 - \mathbf{x}_2\|_2, \|\mathbf{x}_1 + \mathbf{x}_2\|_2)$ for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ is used to denote "distance", upto a global phase factor (both $\mathbf{x} = \mathbf{x}^*, -\mathbf{x}^*$ satisfy $\mathbf{y} = |\mathbf{A}\mathbf{x}|$). 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 faster algorithmic implementations, $\mathbf{M}_S$ can be assumed to be a truncated matrix $\mathbf{M}_S \in \mathbb{R}^{s \times s}$, discarding all row and column elements corresponding to S^c. The element-wise inner product of two vectors $\mathbf{y}_1$ and $\mathbf{y}_2 \in \mathbb{R}^m$ is represented as $\mathbf{y}_1 \circ \mathbf{y}_2$. Unspecified large and small constants are represented by C and δ respectively. The abbreviation *w.h.p.* denotes "with high probability".

3 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 that the *sparse* recovery version of (1) is highly non-convex, and possibly has multiple local minima[22]. Therefore, as is typical in modern non-convex methods [13, 23, 57] we use an spectral technique to obtain a good initial estimate. Our technique is a modification of the initialization stages in [22, 23], but requires no tuning parameters or assumptions on signal coefficients, except for the sparsity s. Once an appropriate initial

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.
Set $\mathbf{v}_1 \leftarrow$ top singular vector of $\mathbf{M}_{\hat{S}} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_{i\hat{S}}\mathbf{a}_{i\hat{S}}^{\top} \quad \in \mathbb{R}^{s\times s}$.
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 $\mathbf{x^0}$.

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**
$\mathbf{P}^{t+1} \leftarrow \text{diag}\left(\text{sign}\left(\mathbf{A}\mathbf{x}^t\right)\right)$,
$\mathbf{x}^{t+1} \leftarrow$ CoSAMP($\frac{1}{\sqrt{m}}\mathbf{A}$,$\frac{1}{\sqrt{m}}\mathbf{P}^{t+1}\mathbf{y}$,$s$,$\mathbf{x}^t$).
end for
output $\mathbf{z} \leftarrow \mathbf{x}^{t_0}$.

estimate is chosen, we then show that a simple alternating-minimization algorithm, based on the algorithm in [13] will converge rapidly to the underlying true signal. We call our overall algorithm *Compressive Phase Retrieval with Alternating Minimization* (CoPRAM) which is divided into two stages: *Initialization* (Algorithm 1) and *Descent* (Algorithm 2).

3.1 Initialization

The high level idea of the first stage of CoPRAM is as follows; we use measurements y_i to construct a *biased* estimator, *marginal* M_{jj} corresponding to the jth signal coefficient and given by:

$$M_{jj} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 a_{ij}^2, \qquad \text{for} \qquad j \in \{1, \ldots n\}. \tag{2}$$

The marginals themselves do not directly produce signal coefficients, but the "weight" of each marginal identifies the true signal support. Then, a spectral technique based on [13, 23, 22] constructs an initial estimate $\mathbf{x^0}$. To accurately estimate support, earlier works [13, 23] assume that the magnitudes of the nonzero signal coefficients are all sufficiently large, i.e., $\Omega\left(\|\mathbf{x}^*\|_2/\sqrt{s}\right)$, which can be unrealistic, violating the power-decay law. Our analysis resolves this issue by *relaxing* the requirement of accurately identifying the support, without any tuning parameters, unlike [22]. We claim that a coarse estimate of the support is good enough, since the errors would correspond to small coefficients. Such "noise" in the signal estimate can be controlled with a sufficient number of samples. Instead, 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, if elements of $\mathbf{A}$ are distributed as per standard normal distribution $\mathcal{N}(0,1)$, a weighted correlation matrix $\mathbf{M} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_i\mathbf{a}_i^{\top}$, can be constructed, having diagonal elements M_{jj}. Then, the diagonal elements of this expectation matrix $\mathbb{E}\left[\mathbf{M}\right]$ are given by:

$$\mathbb{E}\left[M_{jj}\right] = \|\mathbf{x}^*\|^2 + 2x_j^{*2} \tag{3}$$

exhibiting a clear separation when analyzed for $j \in S$ and $j \in S^c$. We can hence claim, that signal marginals at locations on the diagonal of $\mathbf{M}$ corresponding to $j \in S$ are larger, on an average, than those for $j \in S^c$. Based on this, we evaluate the diagonal elements M_{jj} and reject $n-k$ coordinates corresponding to the smallest marginals obtain a crude approximation of signal support $\hat{S}$. Using a spectral technique, we find an initial vector in the reduced space, which is close to the true signal, if $m = \mathcal{O}\left(s^2 \log n\right)$.

Theorem 3.1. *The initial estimate $\mathbf{x^0}$, which is the output of Algorithm 1, is a small constant distance δ_0 away from the true s-sparse signal $\mathbf{x}^*$, i.e.,*

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

where $0 < \delta_0 < 1$*, as long as the number of (Gaussian) measurements satisfy,* $m \geq Cs^2 \log mn$*, with probability greater than* $1 - \frac{8}{m}$*.*

This theorem is proved via Lemmas C.1 through C.4 (Appendix C), 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 the signal coefficients obey $\min_{j \in S} |x_j^*| = C \|\mathbf{x}^*\|_2 / \sqrt{s}$, then, *w.h.p.* there exists a clear separation between the marginals M_{jj} for $j \in S$ and $j \in S^c$. Then Algorithm 1 picks up the correct support (i.e. $\hat{S} = S$); (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. The incorrect elements of $\hat{S}$ induce negligible error in estimating the intial vector. These approaches are illustrated in Figures 4 and 5 in Appendix C. The marginals $M_{jj} < \Theta$, *w.h.p.*, for $j \in S^c$ and $M_{jj} > \Theta$, $j \in S_+$, where S_+ is a big chunk of the picked support $S_+ \subseteq \hat{S}$, $S_+ = \{j \in S : x_j^{*2} \geq 15\sqrt{(\log mn)/m}\, \|\mathbf{x}^*\|_2\}$ are separated by threshold Θ (Lemmas C.1 and C.2). The identification of the support $\hat{S}$ (which provably contains a significant chunk S_+ of the true support S) is used to construct the truncated correlation matrix $\mathbf{M}_{\hat{S}}$. The top singular vector of this matrix $\mathbf{M}_{\hat{S}}$, gives us a good initial estimate $\mathbf{x^0}$.

The final step of Algorithm 1 requires a scaling of the normalized vector $\mathbf{v}_1$ by a factor ϕ, which conserves the power in the signal (Lemma F.1 in Appendix F), whp, where ϕ^2 which is defined as

$$\phi^2 = \frac{1}{m} \sum_{i=1}^{m} y_i^2. \tag{4}$$

3.2 Descent to optimal solution

After obtaining an initial estimate $\mathbf{x^0}$, we construct a method to accurately recover $\mathbf{x}^*$. For this, we adapt the alternating minimization approach from [13]. The observation model (1) can be restated as:

$$\text{sign}\left(\langle \mathbf{a}_i, \mathbf{x}^* \rangle\right) \circ y_i = \langle \mathbf{a}_i, \mathbf{x}^* \rangle \quad \text{for} \quad i = \{1, 2, \ldots m\}.$$

We introduce the *phase vector* $\mathbf{p} \in \mathbb{R}^m$ containing (unknown) signs of measurements, i.e., $p_i = \text{sign}\left(\langle \mathbf{a}_i, \mathbf{x} \rangle\right), \ \forall i$ and *phase matrix* $\mathbf{P} = \text{diag}\left(\mathbf{p}\right)$. Then our measurement model gets modified as $\mathbf{P}^*\mathbf{y} = \mathbf{A}\mathbf{x}^*$, where $\mathbf{P}^*$ is the true phase matrix. We then minimize the loss function composed of variables $\mathbf{x}$ and $\mathbf{P}$,

$$\min_{\|\mathbf{x}\|_0 \leq s, \mathbf{P} \in \mathcal{P}} \|\mathbf{A}\mathbf{x} - \mathbf{P}\mathbf{y}\|_2 . \tag{5}$$

Here $\mathcal{P}$ is a set of all diagonal matrices $\in \mathbb{R}^{m \times m}$ with diagonal entries constrained to be in $\{-1, 1\}$. Hence the problem stated above is *not convex*. Instead, we alternate between estimating $\mathbf{P}$ and $\mathbf{x}$ as follows: (1) if we fix the signal estimate $\mathbf{x}$, then the minimizer $\mathbf{P}$ is given in closed form as $\mathbf{P} = \text{diag}\left(\text{sign}\left(\mathbf{A}\mathbf{x}\right)\right)$; we call this the *phase estimation* step; (2) if we fix the phase matrix $\mathbf{P}$, the sparse vector $\mathbf{x}$ can be obtained by solving the *signal estimation* step:

$$\min_{\mathbf{x}, \|\mathbf{x}\|_0 \leq s} \|\mathbf{A}\mathbf{x} - \mathbf{P}\mathbf{y}\|_2. \tag{6}$$

We employ the CoSaMP [41] algorithm to (approximately) solve the non-convex problem (6). We do not need to explicitly obtain the minimizer for (6) 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 $\mathcal{N}\left(0, 1/\sqrt{m}\right)$. 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 (6) is given by the current signal estimate.

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

Theorem 3.2. *Given an initialization* $\mathbf{x^0}$ *satisfying Algorithm 1, if we have number of (Gaussian) measurements* $m \geq Cs \log \frac{n}{s}$*, then the iterates of Algorithm 2 satisfy:*

$$\text{dist}\left(\mathbf{x}^{t+1}, \mathbf{x}^*\right) \leq \rho_0 \text{dist}\left(\mathbf{x}^t, \mathbf{x}^*\right). \tag{7}$$

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 E.

4 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. 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 $k = \frac{s}{b}$. We use the index set $j_b = \{1, 2 \dots k\}$, to denote block-indices. 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 [30] to descend to the optimal solution.

4.1 Initialization

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:

$$M_{j_b j_b} = \sqrt{\sum_{j \in j_b} M_{jj}^2}, \tag{8}$$

to develop the initialization stage of our *Block CoPRAM* algorithm. Similar to the proof approach of CoPRAM, we 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 $\mathbf{x^0}$ (Algorithm 3, Appendix A) for the Block CoPRAM algorithm (Algorithm 4, Appendix A). 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 $\mathcal{O}(ks \log n)$. 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.

Theorem 4.1. *The initial vector $\mathbf{x^0}$, which is the output of Algorithm 3, is a small constant distance δ_b away from the true signal $\mathbf{x}^* \in \mathcal{M}_{s,b}$, i.e.,*

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

where $0 < \delta_b < 1$, as long as the number of (Gaussian) 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 D, and carries forward intuitively from the proof of the compressive phase-retrieval framework.

4.2 Descent to optimal solution

For the descent of Block CoPRAM to optimal solution, the phase-estimation step is the same as that in CoPRAM. For the signal estimation step, we attempt to solve the same minimization as in (6), except with the additional constraint that the signal $\mathbf{x}^*$ is *block sparse*,

$$\min_{\mathbf{x} \in \mathcal{M}_{s,b}} \|\mathbf{A}\mathbf{x} - \mathbf{P}\mathbf{y}\|_2, \tag{9}$$

where $\mathcal{M}_{s,b}$ describes the block sparsity model. In order to approximate the solution to (9), we use the *model-based CoSaMP* approach of [30]. 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.

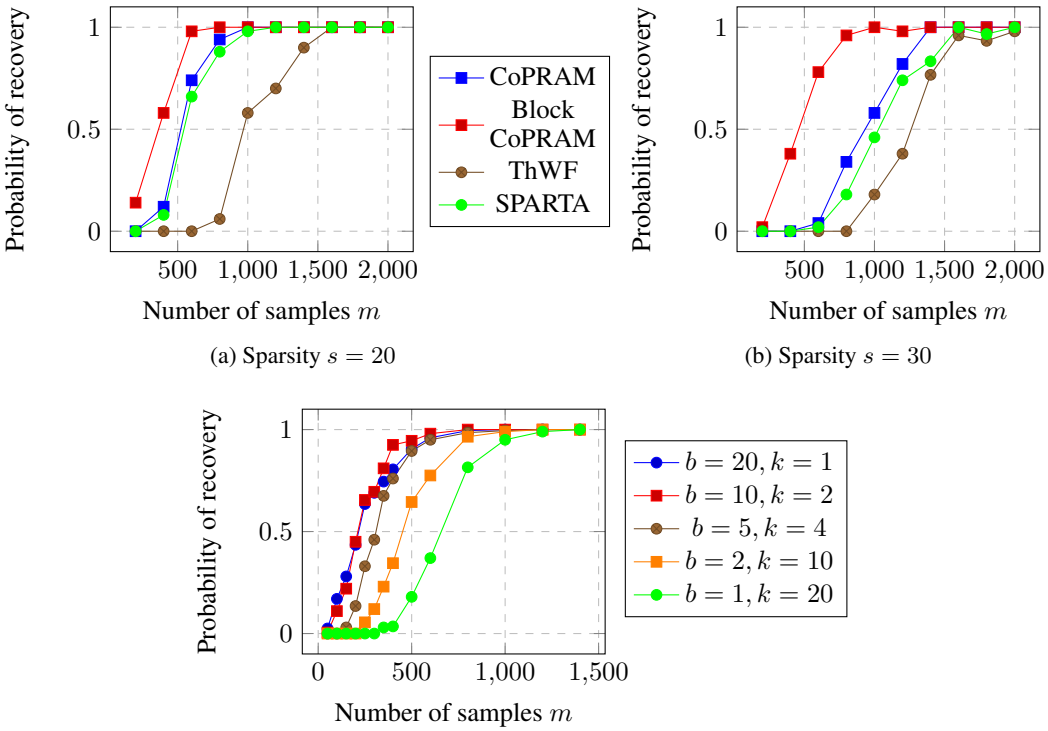
Similar to Theorem 3.2 above, we obtain the following result (the proof can be found in Appendix E):

Theorem 4.2. *Given an initialization $\mathbf{x^0}$ satisfying Algorithm 3, if we have number of (Gaussian) measurements $m \geq C\left(s + \frac{s}{b}\log\frac{n}{s}\right)$, then the iterates of Algorithm 4 satisfy:*

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

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

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 or clusters (refer Appendix A).



(a) Sparsity $s = 20$

(b) Sparsity $s = 30$

(c) Block CoPRAM, $s = 20$

Figure 1: Phase transitions for signal of length $n = 3,000$, sparsity s and block length b (a) $s = 20$, $b = 5$, (b) $s = 30$, $b = 5$, and (c) $s = 20$, $b = 20, 10, 5, 2, 1$ (Block CoPRAM only).

5 Experiments

We explore the performance of the CoPRAM and Block CoPRAM on synthetic data. All numerical experiments were conducted using MATLAB 2016a on a computer with an Intel Xeon CPU at 3.3GHz and 8GB RAM. The nonzero elements of the unit norm vector $\mathbf{x}^* \in \mathbb{R}^{3000}$ are generated from $\mathcal{N}(0, 1)$. We repeated each of the experiments (fixed n, s, b, m) in Figure 1 (a) and (b), for 50 and Figure 1 (c) for 200 independent Monte Carlo trials. For our simulations, we compared our algorithms CoPRAM and Block CoPRAM with Thresholded Wirtinger flow (Thresholded WF or ThWF) [22] and SPARTA [23]. The parameters for these algorithms were carefully chosen as per the description in their respective papers.

For the first experiment, we generated phase transition plots by evaluating the probability of empirical successful recovery, i.e. number of trials out of 50. The recovery probability for the four algorithms is displayed in Figure 1. It can be noted that increasing the sparsity of signal shifts the phase transitions to the right. However, the phase transition for Block CoPRAM has a less apparent shift (suggesting that sample complexity of m has sub-quadratic dependence on s). We see that Block CoPRAM exhibits lowest sample complexity for the phase transitions in both cases (a) and (b) of Figure 1.

For the second experiment, we study the variation of phase transition with block length, for Block CoPRAM (Figure 1(c)). 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 observe that the phase transitions improve with increase in block length. At block sparsity $\frac{s}{b} = \frac{20}{10} = 2$ (for large b, $b \to s$), we observe a saturation effect and the regime of the experiment is very close to the information theoretic limit.

Several additional phase transition diagrams can be found in Figure 2 in Appendix B. The running time of our algorithms compare favorably with Thresholded WF and SPARTA (see Table 2 in Appendix B). We also show that Block CoPRAM is more robust to noisy Gaussian measurements, in comparison to CoPRAM and SPARTA (see Figure 3 in Appendix B).

References

[1] 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.

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

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

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

[5] 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.

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

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

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

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

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

[11] 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.

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

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

[14] 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.

[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. Akcakaya, and J. Chen. Sparse phase retrieval via truncated amplitude flow. *arXiv preprint arXiv:1611.07641*, 2016.

[24] M. Iwen, A. Viswanathan, and Y. Wang. Robust sparse phase retrieval made easy. *Appl. Comput. Harmon. Anal.*, 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] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, 56(4):1982–2001, Apr. 2010.

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

[32] 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.

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

[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. *Appl. Comput. Harmon. Anal.*, 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. *Appl. Comput. Harmon. Anal.*, 42(1):37–64, 2017.

[46] E. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Appl. Comput. Harmon. Anal.*, 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] 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.

[50] 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.

[51] K. Wei. Solving systems of phaseless equations via kaczmarz methods: A proof of concept study. *Inverse Problems*, 31(12):125008, 2015.

[52] 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.

[53] X. Li and V. Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM J. Math. Anal.*, 45(5):3019–3033, 2013.

[54] 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.

[55] Y. Shechtman, A. Beck, and Y. C. Eldar. Gespar: Efficient phase retrieval of sparse signals. *IEEE Trans. Sig. Proc.*, 62(4):928–938, 2014.

[56] C. Hegde, P. Indyk, and L. Schmidt. Fast algorithms for structured sparsity. *Bulletin of the EATCS*, 1(117):197–228, Oct. 2015.

[57] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, 2010.

[58] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Stat.*, pages 1302–1338, 2000.

[59] C. Davis and W. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM J. Num. Anal.*, 7(1):1–46, 1970.

[60] V. Bentkus. An inequality for tail probabilities of martingales with differences bounded from one side. *J. Theoretical Prob.*, 16(1):161–173, 2003.

A Appendix - Block CoPRAM algorithm and extension

A.1 Block CoPRAM algorithm

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 $M_{j_b j_b} = \sqrt{\sum_{j \in j_b} M_{jj}^2} \qquad \forall j_b$, 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 $\mathbf{M}_{\hat{S}_b} = \frac{1}{m}\sum_{i=1}^{m} y_i^2 \mathbf{a}_{i\hat{S}} \mathbf{a}_{i\hat{S}}^\top \quad \in \mathbb{R}^{s \times s}$.
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 $\mathbf{x^0}$.

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**
$\mathbf{P}^{t+1} \leftarrow \text{diag}\left(\text{sign}\left(\mathbf{A}\mathbf{x}^t\right)\right)$.
$\mathbf{x}^{t+1} \leftarrow$ 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 $\mathbf{z} \leftarrow \mathbf{x}^{t_0}$.

A.2 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.

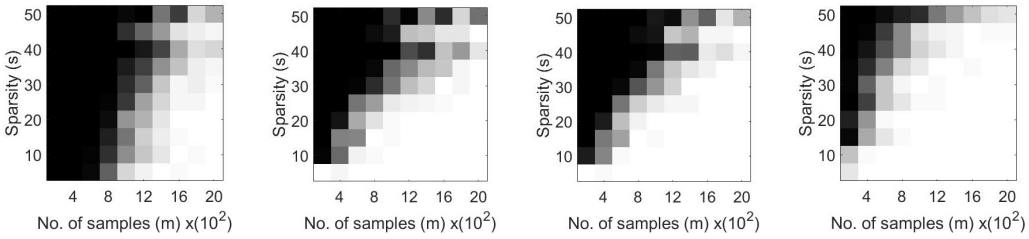
Formally, consider the *clustered sparsity* model for 1D signals in $\mathbb{R}^n$, comprising signals with s non-zeros that occur in no more than k non-overlapping blocks (clusters), each of which exhibit potentially unknown sizes and locations. The above analysis does not immediately apply to this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

B Appendix - Additional experiments

We demonstrate the benefits of our algorithms CoPRAM and Block CoPRAM through an additional set of experiments and describe our previous experiments in further detail.

For Thresholded Wirtinger flow, 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 $\text{card}(\mathcal{I}_o) = \lceil \frac{m}{6} \rceil$ as mentioned in their paper. For our set of generated signals, the AltMinSparse method mentioned in [13] 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. Figure 2 represents phase transition diagrams, where number of measurements m is spanned from $m = 200$ to $m = 2000$ in steps of 200. Similarly



(a) Thresholded WF (b) SPARTA (c) CoPRAM (d) Block CoPRAM

Figure 2: Phase transition plots for different algorithms, with signal length $n = 3000$, with uniform block length of $b = 5$.

Table 2: Mean running time of different algorithms at $s = 25$.

Algorithm	CoPRAM	Block CoPRAM	SPARTA	Thresholded WF
m at phase transition (suc freq = 1)	1,600	1,400	1,800	2,000
mean running time (s)	0.4000	0.3258	0.3080	0.5808

signal sparsity s is swept from $s = 5$ to $s = 50$ in steps of 5. The block lengths considered for all experiments in Figure 2 is $b = 5$. 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 mean running time of the algorithms for different algorithms is tabulated in Table 2. It can be noted that the running times of our algorithms CoPRAM and Block CoPRAM are at par with SPARTA and Thresholded WF.

Effect of noise: For our third experiment, we study the effect of noise on the measurements of the form $y_i = |\langle \mathbf{a}_i, \mathbf{x}^* \rangle + e_i|$, for $i \in \{1, 2 \ldots m\}$. The noise vector $\mathbf{e} \in \mathbb{R}^m$ 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 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 $\|\mathbf{z} - \mathbf{x}^*\|_2 / \|\mathbf{x}^*\|_2$ (here $\mathbf{z} = \mathbf{x}^{t_0}$) can be seen in Figure 3. Block CoPRAM outperforms CoPRAM and SPARTA in all cases considered.

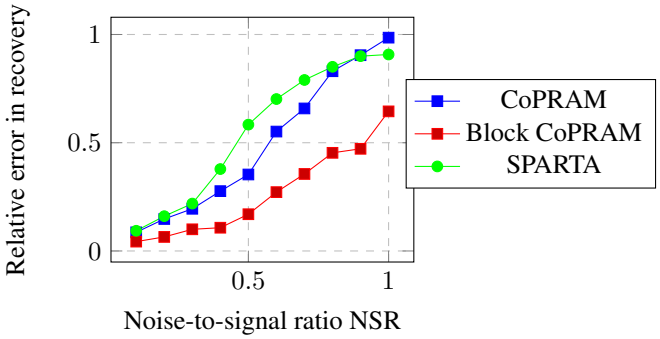


Figure 3: Variation of mean relative error in signal recovered v/s input NSR at $s = 20$ and $b = 5$,$k = 4$ for a signal of length $n = 3,000$, and number of measurements $m = 1,600$.

C Appendix - 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 C.1 - C.4 which complete the proof of Theorem 3.1.

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

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

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 $1 - \frac{1}{m}$,

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

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

$$\min_{j \in S} M_{jj} \geq \left(1 + \frac{C}{s}\right) \|\mathbf{x}^*\|_2 > \left(1 + 11\sqrt{\frac{\log mn}{m}}\right) \|\mathbf{x}^*\|_2^2 \geq \max_{j \in S^c} M_{jj}. \tag{12}$$

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 (i.e. $\Theta_1 < \Theta_2$). 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 [50], completing the proof of Theorem 3.1.

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 C.2-C.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 4 and 5.

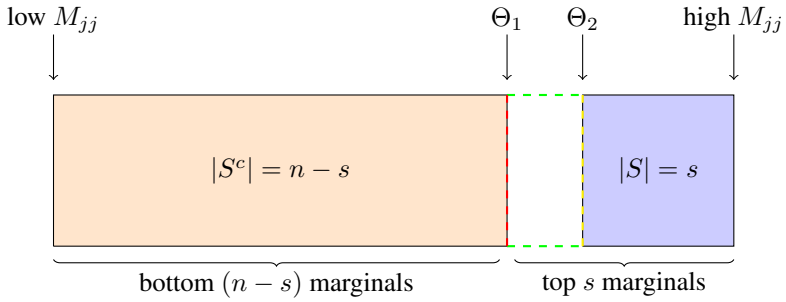


Figure 4: Partition of supports considered for analysis of proof approach 1: assumption on x_{min}^*.

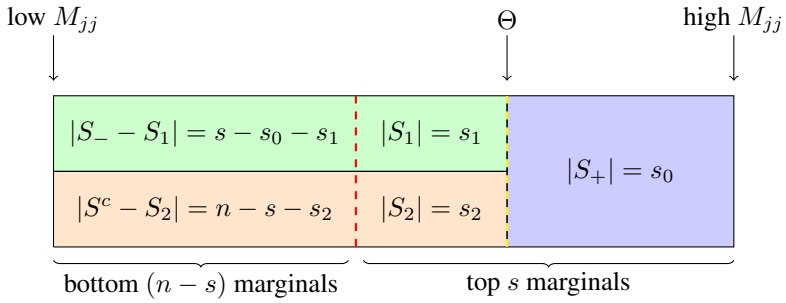


Figure 5: Partition of supports considered for analysis of proof approach 2.

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

$$M_{jj} \leq \left(1 + 11\sqrt{\frac{\log mn}{m}}\right) \|\mathbf{x}^*\|_2^2 = \Theta. \tag{13}$$

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 [58] for the χ_1^2 variables a_{ij}^2 with weights y_i^2 (here $p \equiv n - s$):

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

Further, using the Chebyshev's inequality for y_i^4:

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

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

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

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

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

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

$$\frac{1}{m}\sum_{i=1}^{m} y_i^2(a_{ij}^2 - 1) \leq 2\sqrt{3 + \sqrt{96}}\,\|\mathbf{x}^*\|_2^2 \sqrt{\frac{\log mp}{m}} + 8\,\|\mathbf{x}^*\|_2^2 \frac{(\log mp)^2}{m} \leq 8\sqrt{\frac{\log mp}{m}}\,\|\mathbf{x}^*\|_2^2. \tag{14}$$

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

$$\begin{aligned}M_{jj} &= \frac{1}{m}\sum_{i=1}^{m} y_i^2 a_{ij}^2 \leq 8\sqrt{\frac{\log mn}{m}}\,\|\mathbf{x}^*\|_2^2 + \phi^2 \\ M_{jj} &\leq \left(1 + 11\sqrt{\frac{\log mn}{m}}\right)\|\mathbf{x}^*\|_2^2 = \Theta,\end{aligned} \tag{15}$$

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

$$M_{jj} \geq \left(1 + 11\sqrt{\frac{\log mn}{m}}\right)\|\mathbf{x}^*\|_2^2 = \Theta, \tag{16}$$

where S_+ is defined as:

$$S_+ = \left\{ j \in S \mid x_j^{*2} > 15\sqrt{\frac{\log mn}{m}} \left\|\mathbf{x}^*\right\|_2^2 \right\}. \tag{17}$$

Subsequently, we can define S_- as:

$$S_- = \left\{ j \in S \mid x_j^{*2} \leq 15\sqrt{\frac{\log mn}{m}} \left\|\mathbf{x}^*\right\|_2^2 \right\}, \tag{18}$$

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:

$$\left\|\mathbf{x}^*_{S_-}\right\|_2^2 \leq 15\sqrt{\frac{s^2 \log mn}{m}} \left\|\mathbf{x}^*\right\|_2^2. \tag{19}$$

Proof. Evaluating the marginals:

$$M_{jj} - \phi^2 = \frac{1}{m}\sum_{i=1}^{m} y_i^2 \left(a_{ij}^2 - 1\right). \tag{20}$$

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 (20): $\frac{1}{m}\sum_{i=1}^{m} y_i^2 a_{ij}^2$ and $\frac{1}{m}\sum_{i=1}^{m} y_i^2$. This can be done by evaluating the expectation values:

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

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

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

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

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{m} -X_i \leq -t\right] &= \mathbb{P}\left[\sum_{i=1}^{m} y_i^2 a_{ij}^2 - m\left(\left\|\mathbf{x}^*\right\|_2^2 + 2x_j^{*2}\right) \leq -t\right] \\
&\leq \exp\left(-\frac{t^2}{192\left\|\mathbf{x}^*\right\|_2^4 m}\right) \leq \frac{1}{mk}
\end{aligned} \tag{21}$$

This requires $t = \sqrt{192}\left\|\mathbf{x}^*\right\|_2^2\sqrt{m\log mk} \approx 13.86\left\|\mathbf{x}^*\right\|_2^2\sqrt{m\log mk} \leq 13.86\left\|\mathbf{x}^*\right\|_2^2\sqrt{m\log mn}$. 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 [58], with probability greater than $1 - \frac{1}{mk}$, for each $j \in S$:

$$\frac{1}{m}\sum_{i=1}^{m} y_i^2 - \left\|\mathbf{x}^*\right\|_2^2 \leq \left(2\sqrt{\frac{\log mk}{m}} + \frac{2\log mk}{m}\right)\left\|\mathbf{x}^*\right\|_2^2 \tag{22}$$

$$\leq 3\left\|\mathbf{x}^*\right\|_2^2\sqrt{\frac{\log mk}{m}} \leq 3\left\|\mathbf{x}^*\right\|_2^2\sqrt{\frac{\log mn}{m}} \tag{23}$$

for $m > C$. Combining these two concentration bounds (21), (22), taking a union bound for all $j \in S_+$ and substituting in (20):

$$M_{jj} - \phi^2 \geq 2x_j^{*2} - 17\sqrt{\frac{\log mn}{m}}\left\|\mathbf{x}^*\right\|_2^2, \tag{24}$$

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

If the set S_+, is constructed as in (17), then evaluating the bound in (24), we get:

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

holds for all elements $j \in S_+$, with probability greater than $1 - \frac{2}{m}$, yielding the bound in (16). □

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

$$\left\|\mathbf{x}^* - \mathbf{x}^*_{\hat{S}}\right\|_2 \leq \delta_1 \|\mathbf{x}^*\|_2, \tag{25}$$

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

$$m \geq Cs^2 \log mn. \tag{26}$$

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 5 for illustration of the sets):

$$\mathbf{x}^*_{\hat{S}} = \mathbf{x}^*_{S_+} + \mathbf{x}^*_{S_1} + \mathbf{x}^*_{S_2}. \tag{27}$$

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

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

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

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

Using (19), we compute the bound,

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

which is the required condition (25). This requires sample complexity m to satisfy:

$$30\sqrt{\frac{s^2 \log mn}{m}} \leq \delta_1^2,$$
$$\implies m \geq \frac{900}{\delta_1^2} s^2 \log mn = C(\delta_1) s^2 \log mn. \tag{28}$$

□

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 C.4.

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

$$\mathrm{dist}\left(\mathbf{x}^{\mathbf{0}}, \mathbf{x}^*_{\hat{S}}\right) \equiv \min\left(\left\|\mathbf{x}^{\mathbf{0}} - \mathbf{x}^*_{\hat{S}}\right\|_2, \left\|\mathbf{x}^{\mathbf{0}} + \mathbf{x}^*_{\hat{S}}\right\|_2\right) \leq \delta_2 \|\mathbf{x}^*\|_2, \tag{29}$$

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

$$m \geq Cs \log n. \tag{30}$$

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

$$\mathbb{E}[\mathbf{M}] = \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} y_j^2 \mathbf{a}_i \mathbf{a}_i^\top\right] = \left(\mathbf{I}_{n\times n} + 2\frac{\mathbf{x}^*}{\|\mathbf{x}^*\|_2}\frac{\mathbf{x}^{*\top}}{\|\mathbf{x}^*\|_2}\right)\|\mathbf{x}^*\|_2^2,$$

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

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 $\mathbf{x^0}$ 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, $\text{card}(S_3) \leq 2s$, and,

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

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,

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

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,

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

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 [59] to bound the difference between the normalized top-singular vectors $\mathbf{x^0}$ of $\mathbf{M}_{\hat{S}}$ and $\mathbf{x}_{\hat{S}}$ of $\mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]$ as,

$$\text{dist}\left(\mathbf{x^0}, \mathbf{x}^*_{\hat{S}}\right) \leq \frac{\delta\|\mathbf{x}^*\|_2^2}{2\|\mathbf{x}^*\|_2^2} = \frac{\delta}{2}$$

$$\implies \min\left(\|\mathbf{x^0} - \mathbf{x}^*_{\hat{S}}\|_2, \left\|\mathbf{x^0} + \mathbf{x}^*_{\hat{S}}\right\|_2\right) \leq \sqrt{2\left(1 - \sqrt{1 - \frac{\delta^2}{4}}\right)} \leq \delta_2$$

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

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

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

as long as the number of measurements m follow the bound in (26). 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 (28)). This completes the proof of Theorem 3.1.

D Appendix - 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 4.1 for the initialization stage of Block CoPRAM as follows.

Theorem 4.1. *The initial vector $\mathbf{x^0}$, which is the output of Algorithm 3, is a small constant distance δ_b away from the true signal $\mathbf{x}^* \in \mathcal{M}_{s,b}$, i.e.,*

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

where $0 < \delta_b < 1$, as long as the number of (Gaussian) measurements satisfy $m \geq C\frac{s^2}{b}\log mn$ with probability greater than $1 - \frac{8}{m}$.

Proof. Evaluating the marginals $M_{j_b j_b}$, for all $j_b \in S_b^c$, from (11), with probability greater than $1 - \frac{5}{m}$, we have:

$$M_{j_b j_b} \leq \left(1 + 11\sqrt{\frac{\log mn}{m}}\right) \sqrt{b} \left\|\mathbf{x}^*\right\|_2^2 . \tag{31}$$

Evaluating the block marginals $M_{j_b j_b}$, for $j_b \in S_b$, we use a modification of (21), with probability less than $\exp\left(-\frac{mt^2}{192\|\mathbf{x}^*\|_2^4}\right) \leq \frac{1}{mn}$,

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

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

Rearranging the terms,

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

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 $1 - \frac{1}{m}$,

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

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

$$\min_{j_b\in S_b} M_{j_b j_b} \geq \left(1 + \frac{C}{\sqrt{b}k}\right)\sqrt{b}\left\|\mathbf{x}^*\right\|_2^2 . \tag{32}$$

Equating the expression for probability,

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

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 (32) and (31),

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

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 $M_{j_b j_b}$ for $j_b \in S_b^c$, we can use the result of Lemma C.1, to obtain the same bound as in (31), with probability greater than $1 - \frac{5}{m}$,

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

For evaluating block marginals $M_{j_b j_b}$ for $j_b \in S_b$ we can use equations (17) and (18), and extend this model of signal supports to block supports defined as:

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

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

Using equation (24), and LHS of (38),

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

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

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

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

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

and the inequalities:

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

This inequality gives us a bound on the number of measurements m, similar to (28),

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

with probability greater than $1 - \frac{7}{m}$. This gives us the evaluation of block-marginals for $j_b \in S_b$ 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 $1 - \frac{1}{m}$. □

E Appendix - 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 3.2 and Theorem 4.2. We prove theorem 3.2 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 $\text{dist}\left(\mathbf{x^0}, \mathbf{x}^*\right) = \|\mathbf{x^0} - \mathbf{x}^*\|_2$ at the end of initialization stage. We claim that wlog, the same results would hold, if $\text{dist}\left(\mathbf{x^0}, \mathbf{x}^*\right) = \left\|\mathbf{x^0} + \mathbf{x}^*\right\|_2$.

Theorem 3.2. *Given an initialization $\mathbf{x^0}$ satisfying Algorithm 1, if we have number of (Gaussian) measurements $m \geq Cs \log \frac{n}{s}$, then the iterates of Algorithm 2 satisfy:*

$$\text{dist}\left(\mathbf{x}^{t+1}, \mathbf{x}^*\right) \leq \rho_0 \text{dist}\left(\mathbf{x}^t, \mathbf{x}^*\right). \tag{7}$$

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

$$\begin{aligned} \mathbf{x}^{t+1,0} &\leftarrow \mathbf{x}^t && \text{initialize to best possible estimate} \\ \mathbf{r} &\leftarrow \mathbf{u} && \text{residue} \\ l &\leftarrow 0 && \text{CoSaMP internal counter} \end{aligned}$$

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

3:

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

4: **end while**

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

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:

$$\begin{aligned} \left\|\mathbf{x}^{t+1,l+1} - \mathbf{x}^*\right\|_2 &= \left\|\mathbf{x}^{t+1,l+1} - \mathbf{w} + \mathbf{w} - \mathbf{x}^*\right\|_2, \\ &\leq 2\left\|\mathbf{x}^* - \mathbf{w}\right\|_2 \end{aligned} \tag{33}$$

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

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

with high probability, where δ_s is the RIP constant. Now, analyzing the inputs to CoSaMP, in the $\mathbf{x}$-update step of Algorithm 2,

$$\begin{aligned}
\mathbf{u} &= \frac{\mathbf{P}^t\mathbf{y}}{\sqrt{m}}, \\
&= \operatorname{sign}\left(\mathbf{A}\mathbf{x}^t\right) \circ \frac{|\mathbf{A}\mathbf{x}^*|}{\sqrt{m}}, \\
&= \operatorname{sign}\left(\Phi\mathbf{x}^t\right) \circ \left\{\left(\Phi\mathbf{x}^*\right) \circ \operatorname{sign}\left(\Phi\mathbf{x}^*\right)\right\}, \\
&= \Phi\mathbf{x}^* + \left(\operatorname{sign}\left(\Phi\mathbf{x}^t\right)\operatorname{sign}\left(\Phi\mathbf{x}^*\right) - \mathbf{1}\right) \circ \Phi\mathbf{x}^*, \\
\implies \mathbf{u} - \Phi\mathbf{x}^* &= \left(\operatorname{sign}\left(\Phi\mathbf{x}^t\right)\operatorname{sign}\left(\Phi\mathbf{x}^*\right) - \mathbf{1}\right) \circ \Phi\mathbf{x}^*, \\
&= E_{ph},
\end{aligned} \tag{35}$$

where $E_{ph} \equiv \left(\operatorname{sign}\left(\Phi\mathbf{x}^t\right)\operatorname{sign}\left(\Phi\mathbf{x}^*\right) - \mathbf{1}\right) \circ \Phi\mathbf{x}^*$, is error due to failure in estimating the correct phase.

Using equation (35) and substituting into equation (34), the per-step reduction in error for each run of CoSaMP is:

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

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 $\operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^{3s}} \left\|\Phi_\Gamma\mathbf{x} - \mathbf{u}\right\|_2$, the fourth step uses (35) again, the final step uses RIP again (which holds with probability greater than $1 - e^{-\gamma_1 m}$, with γ_1 being a positive constant).

Finally, the first term in the previous inequality can be bounded using (Lemma 4.2 of CoSaMP [41], refer Lemma F.4), to yeild,

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

where ρ_3, ρ_4 are as stated in Lemma F.4. Assuming that CoSaMP is let to run a maximum of L iterations,

$$\begin{aligned}
\left\|\mathbf{x}^{t+1} - \mathbf{x}^*\right\|_2 &\leq \left(\rho_1\rho_3\right)^L \left\|\mathbf{x}^* - \mathbf{x}^t\right\|_2 + \left(\rho_1\rho_4 + \rho_2\right)\left(1 + \rho_1\rho_3 + \left(\rho_1\rho_3\right)^2 \ldots \left(\rho_1\rho_3\right)^{L-1}\right) \left\|E_{ph}\right\|_2, \\
&\leq \left(\rho_1\rho_3\right)^L \left\|\mathbf{x}^* - \mathbf{x}^t\right\|_2 + \frac{\left(\rho_1\rho_4 + \rho_2\right)}{\left(1 - \rho_1\rho_3\right)} \left\|E_{ph}\right\|_2.
\end{aligned} \tag{36}$$

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

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

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

Lemma E.1. *As long as the initial estimate is a small distance away from the true signal* $\mathbf{x}^*$*,*

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

and subsequently,

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

then the following bound holds,

$$\frac{2}{m}\sum_{i=1}^{m} \left|\mathbf{a}_i{}^\top \mathbf{x}^*\right|^2 \mathbb{1}_{\{(\mathbf{a}_i{}^\top \mathbf{x}^t)(\mathbf{a}_i{}^\top \mathbf{x}^*) \leq 0\}} \leq \frac{2}{(1-\delta_0)^2}\left(\delta + \sqrt{\frac{21}{20}}\delta_0\right)\left\|\mathbf{x}^t - \mathbf{x}^*\right\|_2^2.$$

with probability greater than $1 - e^{-\gamma_2 m}$*, where* γ_2 *is a positive constant, as long as* $m > Cs\log\frac{n}{s}$*. We can use this to bound the phase error as,*

$$\|E_{ph}\|_2 \leq \rho_5 \left\|\mathbf{x}^t - \mathbf{x}^*\right\|_2,$$

where $\rho_5 = \frac{\sqrt{2}}{(1-\delta_0)}\sqrt{\delta + \sqrt{\frac{21}{20}}\delta_0}$*,* $\delta \approx 0.001$*.*

This proof has been adapted from Lemma 7.19 of [42] and uses the generic chaining techniques of [43, 44]. Using this in addition to equation (36), we have our final per-step error reduction for a single run of CoPRAM (Algorithm 2), as:

$$\begin{aligned}\left\|\mathbf{x}^{t+1} - \mathbf{x}^*\right\|_2 &\leq \left((\rho_1\rho_3)^L + \rho_5\frac{(\rho_1\rho_4 + \rho_2)}{(1-\rho_1\rho_3)}\right)\left\|\mathbf{x}^t - \mathbf{x}^*\right\|_2, \\ &\leq \rho_0 \left\|\mathbf{x}^t - \mathbf{x}^*\right\|_2, \end{aligned} \tag{37}$$

where $\rho_0 < 1$.

Evaluating convergence parameter ρ_0**:**. To obtain per-step reduction in error, we require $\rho_0 < 1$. For sake of numerical analysis, $\delta_s, \delta_{2s}, \delta_{4s} \leq 0.0001$, then $\rho_1 \approx 1, \rho_3 \approx 0.0002$. Let $\delta_0 = 0.012$, then $\rho_5 \approx 0.16$. 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 (33) 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 4.2 for Block CoPRAM as follows.

Theorem 4.2. *Given an initialization* $\mathbf{x}^{\mathbf{0}}$ *satisfying Algorithm 3, if we have number of (Gaussian) measurements* $m \geq C\left(s + \frac{s}{b}\log\frac{n}{s}\right)$*, then the iterates of Algorithm 4 satisfy:*

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

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

The proof for this is a natural extention to the one we have proved in Theorem 3.2, and would use the results from the paper on model-based compressive sensing [30], wherever Block CoSaMP is invoked.

F Supplementary appendix

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

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

$$\left(1 - 2\sqrt{\frac{\log m}{m}}\right) \|\mathbf{x}^*\|_2^2 \leq \phi^2 \leq \left(1 + 3\sqrt{\frac{\log m}{m}}\right) \|\mathbf{x}^*\|_2^2. \tag{38}$$

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 [58] on a_{ij}^2, we can obtain the upper bound,

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

Similarly, we can obtain the lower bound,

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

The signal power ϕ^2 is then bounded as

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

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,*

$$\left\|\frac{1}{m}\sum_{i=1}^m \left|\mathbf{a}_{iS_3}^\top \mathbf{x}^*\right|^2 \mathbf{a}_{iS_3}\mathbf{a}_{iS_3}^\top - \left(\|\mathbf{x}^*\|_2^2 (\mathbf{I}_{n\times n})_{S_3} + 2\mathbf{x}^*\mathbf{x}^{*\top}\right)\right\|_2 \leq \delta \|\mathbf{x}^*\|_2^2$$

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* $X_1 \ldots X_m$ *are i.i.d. centered, bounded real-valued random variables obeying*

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

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 [60].

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,*

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

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].