EFFICIENT AND GUARANTEED ALGORITHMS FOR SPARSE INVERSE PROBLEMS

BY

KIRYUNG LEE

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical and Computer Engineering in the Graduate College of the University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Yoram Bresler, Chair
Professor Marius Junge
Associate Professor Minh Do
Associate Professor Olgica Milenkovic
Adjunct Professor Sean Meyn
ABSTRACT

Compressed sensing is a new data acquisition paradigm enabling universal, simple, and reduced-cost acquisition, by exploiting a sparse signal model. Most notably, recovery of the signal by computationally efficient algorithms is guaranteed for certain randomized acquisition systems. However, there is a discrepancy between the theoretical guarantees and practical applications. In applications, including Fourier imaging in various modalities, the measurements are acquired by inner products with vectors selected randomly (sampled) from a frame. Currently available guarantees are derived using a so-called restricted isometry property (RIP), which has only been shown to hold under ideal assumptions. For example, the sampling from the frame needs to be independent and identically distributed with the uniform distribution, and the frame must be tight. In practice though, one or more of the ideal assumptions is typically violated and none of the existing guarantees applies. Motivated by this discrepancy, we propose two related changes in the existing framework: (i) a generalized RIP called the restricted biorthogonality property (RBOP); and (ii) correspondingly modified versions of existing greedy pursuit algorithms, which we call oblique pursuits. Oblique pursuits are guaranteed using the RBOP without requiring ideal assumptions; hence, the guarantees apply to practical acquisition schemes. Numerical results show that oblique pursuits also perform competitively with, or sometimes better than their conventional counterparts.

We also propose robust and efficient algorithms for the joint sparse recovery problem in compressed sensing, which simultaneously recover the supports of jointly sparse signals from their multiple measurement vectors obtained through a common sensing matrix. In a favorable situation, the unknown matrix, which consists of the jointly sparse signals, has linearly independent nonzero rows. In this case, the Multiple Signal Classification (MUSIC) algorithm, originally proposed by Schmidt for the direction of arrival estimation
problem in sensor array processing and later proposed and analyzed for joint sparse recovery by Feng and Bresler, provides a guarantee with the minimum number of measurements. We focus instead on the unfavorable but practically significant case of rank defect or ill-conditioning. This situation arises with a limited number of measurement vectors, or with highly correlated signal components. In this case, MUSIC fails and, in practice, none of the existing methods can consistently approach the fundamental limit. We propose subspace-augmented MUSIC (SA-MUSIC), which improves on MUSIC so that the support is reliably recovered under such unfavorable conditions. Combined with a subspace-based greedy algorithm, Orthogonal Subspace Matching Pursuit, which is also proposed and analyzed in Chapter 3, SA-MUSIC provides a computationally efficient algorithm with a performance guarantee. The performance guarantees are given in terms of a version of the restricted isometry property. In particular, we also present a non-asymptotic perturbation analysis of the signal subspace estimation step, which has been missing in the previous studies of MUSIC.

Finally, we address compressed sensing of a low-rank matrix posing the inverse problem as an approximation problem with a specified target rank of the solution. A simple search over the target rank then provides the minimum rank solution satisfying a prescribed data approximation bound. We propose an atomic decomposition providing an analogy between parsimonious representations of a sparse vector and a low-rank matrix and extending efficient greedy algorithms from the vector to the matrix case. In particular, we propose an efficient and guaranteed algorithm named ADMiRA that extends Needell and Tropp’s compressive sampling matching pursuit (CoSaMP) algorithm from the sparse vector to the low-rank matrix case. The performance guarantee is given in terms of the rank-restricted isometry property and bounds both the number of iterations and the error in the approximate solution for the general case of noisy measurements and approximately low-rank solution. With a sparse measurement operator as in the matrix completion problem, the computation in ADMiRA is linear in the number of measurements. Numerical experiments for the matrix completion problem show that, although the rank-restricted isometry property is not satisfied in this case, ADMiRA is a competitive algorithm for matrix completion.
ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Yoram Bresler. He kindly guided my study with patience and encouragement, and taught me his insight, philosophy, and enthusiasm. Most important, he showed me how to enjoy doing study and research. I have never seen him losing his smile even in his busiest moments. He has been and will be the greatest role model in my life.

My thanks also go to my co-advisor Marius Junge, who taught me advanced math of his expertise. It was a great help to my research and I enjoyed his courses and private lectures a lot. I also thank other professors on my committee. Minh Do and Sean Meyn were great teachers and gave me helpful advice. I thank Olgica Milenkovic for the fruitful discussions that we had about my research. For the same reason, I thank my colleague Wei Dai. I thank Angelia Nedich for her great course on convex optimization.

The years that I spent in my study were enjoyable since I had great colleagues. Thanks to Ashvin George, Nitin Aggarwal, Behzad Sharif, Saiprasad Ravishankar, and Luke Pfister for their kind support and the happy memories here in Urbana-Champaign that we share.

My former advisors, Taejeong Kim and Dong Sik Kim, gave me good advice for my PhD study and I appreciate it. I appreciate Jong Chul Ye for his recommendation of me to Yoram, opening up the opportunity to start the study in this thesis.

Last, but not the least, I deeply appreciate the support of my beloved family. My mother devoted her life to make me who I am today. Thank you mom. My acknowledgments cannot be complete without mentioning my wife. Thank you Soomin for your patience and love. My thesis would not have been completed without your encouragement and support.
TABLE OF CONTENTS

LIST OF TABLES ..................................................... vii

LIST OF FIGURES ................................................... viii

CHAPTER 1 INTRODUCTION ......................................... 1
  1.1 Compressed Sensing ........................................... 2
  1.2 Joint Sparse Recovery ......................................... 14
  1.3 Low Rank Matrix Recovery .................................... 18

CHAPTER 2 OBLIQUE PURSUITS FOR COMPRESSED SENSING 24
  2.1 Notation ....................................................... 24
  2.2 Oblique Pursuit Algorithms .................................. 25
  2.3 Restricted Biorthogonality Property ......................... 40
  2.4 Numerical Results ........................................... 50

CHAPTER 3 SUBSPACE METHODS FOR JOINT SPARSE RECOVERY .. 57
  3.1 Notation ....................................................... 57
  3.2 Problem Statement ............................................ 58
  3.3 MUSIC Revisited .............................................. 59
  3.4 Subspace-Augmented MUSIC .................................. 66
  3.5 Weak Restricted Isometry Property ......................... 73
  3.6 Performance Guarantees ...................................... 78
  3.7 Analysis of Signal Subspace Estimation ..................... 87
  3.8 Numerical Experiments ....................................... 92
  3.9 Discussion .................................................... 96

CHAPTER 4 ADMIRA: ATOMIC DECOMPOSITION FOR MINIMUM RANK APPROXIMATION ........................................... 101
  4.1 Vector vs. Matrix ............................................. 101
  4.2 Algorithm ..................................................... 105
  4.3 Main Results: Performance Guarantee ....................... 106
  4.4 Properties of the Rank-Restricted Isometry ................ 109
  4.5 Proof of Theorem 4.3.1 ...................................... 110
  4.6 Implementation and Scalability .............................. 114
  4.7 Numerical Experiment ....................................... 117
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The RBOP condition required for linear convergence in Theorem 2.2.9.</td>
<td>38</td>
</tr>
<tr>
<td>2.2</td>
<td>Quality (PSNR in dB) of images reconstructed from noisy variable density Fourier samples with measurement SNR = 30 (dB). Results averaged over 100 random sampling patterns.</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Completion of random matrices by ADMiRA: $n = m$, $r = 2$, $p = 10[n^{1.2-r}\log_{10} n]$.</td>
<td>118</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparison of ADMiRA and SVT: no noise, $n = m = 1000$.</td>
<td>119</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

2.1 Phase transition of support recovery by various greedy pursuit algorithms (the horizontal and vertical axes denote the ratio $m/n$ of number of measurements to number of unknowns and ratio $s/m$ of sparsity level to number of measurements, respectively): signal $x^*$ is exactly $s$-sparse with nonzero entries that are $\pm 1$ with random sign. $n = 1024$, SNR = 30dB, $\kappa = 2$ .............................................. 52

2.2 Error images and PSNR for recovery by various algorithms from noisy measurements (the maximum intensity of the input image is normalized as 1): SNR = 30dB, downsample by 3. ................................................................. 55

3.1 Comparison of MUSIC for full row rank case versus SA-MUSIC with known partial support (SA-MUSIC+Oracle) in the rank-defective case: tradeoff between parameter $\delta$ (for the weak-1 RIP) and $\eta$ (for the subspace estimate perturbation). The region below the curve provides a guarantee. 82

3.2 Comparison of SA-MUSIC and OSMP when a partial support of size $s - r$ is given: tradeoff between parameter $\delta$ (for the weak-1 RIP) and $\eta$ (for the subspace estimate perturbation). The region below the curve provides a guarantee. 84

3.3 Required $\delta$ (for the weak-1 RIP) for the guarantee of SA-MUSIC+OSMP in Theorem 3.6.11 in the noiseless case ($\eta = 0$). The dot at the top right of the plot corresponds to $\alpha_{s+1}^\text{weak}(A; J_0) > 0$, or $\delta_{s+1}^\text{weak}(A; J_0) < 1$, required in the full row rank case, $r = s$ (in this case SA-MUSIC+OSMP reduces to MUSIC). ......................................................... 85

3.4 Tradeoff between parameters $\delta$ (for the weak-1 RIP) and $\eta$ (for the subspace estimate perturbation) for the guarantee of SA-MUSIC+OSMP in Theorem 3.6.11 in the noisy case. Values $(\eta, \delta)$ in the region below the curve provide a guarantee. 86

3.5 Constant factor $C_\delta$ in (3.6.18) for the guarantee of SA-MUSIC+OSMP when $A$ is a partial Fourier matrix. Values $(\eta, C_\delta)$ in the region above the curve provide a guarantee. ........................................ 88
3.6 Test on rank defect, \( n = 128, s = 8, N = 256 \), Left column noiseless data; right column SNR = 30 dB: (a),(b) \( \text{rank}(X_0^h) = 4 \). (c),(d) \( \text{rank}(X_0^h) = 6 \). (e),(f) \( \text{rank}(X_0^h) = 8 \) (full row rank).

3.7 Comparison of runtime \( N = 256, \text{SNR} = 30 \) dB.

3.8 Test on varying \( N, n = 128, s = 8, r = 6 \), Left column noiseless; right column SNR = 30 dB: (a),(b) \( N = 16 \). (c),(d) \( N = 64 \).

3.9 Test on large condition number, \( n = 128, s = 8, \text{rank}(X_0^h) = s \) (full row rank), \( N = 256, \text{SNR} = 30 \) dB. (a) \( \kappa = 10 \). (b) \( \kappa = 50 \).

3.10 Required weak-1 RIC for the guarantees of M-BP (average case analysis with the multichannel model with error probability \( \epsilon = 10^{-3} \)), and SA-MUSIC (worst case analysis) for the noiseless case (a) \( n = 128, s = 8 \). (b) \( n = 1024, s = 64 \).

4.1 Completion of random matrices by ADMiRA: \( n = m = 500, r = 2 \).

4.2 Phase transition of matrix completion: \( n = m = 100 \).
Linear inverse problems are ubiquitous in mathematics, statistics, engineering, etc. The linearity of the system is either inherent to the problem, or is used as an approximation to simplify solutions and their analysis. Solutions to a finite-dimensional linear system are well characterized by conventional linear algebra, in terms of uniqueness and stability with respect to perturbations.

Solving an underdetermined system, which has fewer independent equations than unknowns, is often of interest. For example, in acquiring data about a signal or image, the cost is proportional to the number of measurements, or equivalently, to the number of equations describing the relationship between the unknown signal and the measurements. It is therefore desired to recover the signal from as few measurements as possible. When the measurements are fewer than the dimension of the space in which the signals of interest live, the recovery problem is underdetermined. In another example of matrix completion, only a few entries of the unknown matrix are revealed; hence, the resulting linear system is also underdetermined.

In conventional linear algebra, the solution to an underdetermined system is neither unique nor stable. Fortunately, the situation is fundamentally different when sufficient prior information is available about the solution. In particular, solutions sought by applications are often “sparse” in a certain sense. For example:

- The solution is sparse in the standard basis, i.e., it has few nonzero elements.
- The solution is sparse in a certain transform domain or over a dictionary.
- Solutions to a common system with different observations are jointly sparse, i.e., share the same pattern of non-zero locations.
• The solution is a matrix of low rank, i.e., has a sparse spectrum over left/right transforms determined by its own singular vectors.

Indeed, the study of a sparse solution to linear systems dates back to 1970s. Instead of finding the sparsest solution by exhaustive search, which is usually computationally infeasible, and sometimes not even an option (for example, impossible for the matrix completion problem, because the space of solutions can not be enumerated), efficient algorithms have been proposed to find a sparse solution. For example, $\ell_1$-norm-based convex optimization formulations admit efficient algorithms. Forward greedy algorithms is another group of popular method with lower computational cost. Notable contributions in recent years are to the performance guarantee of computationally efficient algorithms. The analyses were developed using the tools of non-commutative probability theory and convex analysis.

In this thesis, we further explore the following three sparse linear inverse problems:

1. compressed sensing in practical scenarios, in which existing recovery guarantees fail;
2. joint sparse recovery;
3. low-rank solution to matrix-valued inverse problem.

For the three problems, we propose computationally efficient algorithms with performance guarantees. In the remainder of this chapter, we provide an introduction and a summary of our contribution for each problem.

1.1 Compressed Sensing

1.1.1 Compressed Sensing

Many natural and man-made signals admit sparse representations [1]. Compressed sensing is a new paradigm of data acquisition that takes advantage of this property to reduce the amount of data that needs to be acquired to recover the signal of interest. Unlike the conventional paradigm, in which large quantities of data are acquired, often followed by compression, compressed
sensing acquires minimally redundant data directly in a universal way that does not depend on the data [2–4].

The model for the acquisition is formally stated as the following linear system: Let \( f \in \mathbb{K}^d \) (where \( \mathbb{K} = \mathbb{R} \) or \( \mathbb{K} = \mathbb{C} \)) be the unknown signal. The measurement vector \( y \in \mathbb{K}^m \) obtained by sensing matrix \( A \in \mathbb{K}^{m \times d} \) is

\[
y = Af + w
\]

where \( w \in \mathbb{K}^m \) denotes additive noise. In the conventional paradigm, arbitrary signal \( f \in \mathbb{K}^d \) is stably reconstructed when the rows of \( A \) constitute a frame for \( \mathbb{K}^d \), which requires redundant measurements \((m \geq d)\). In contrast, compressed sensing aims to reconstruct signals that are (approximately) \( s \)-sparse over a dictionary \( D \in \mathbb{K}^{d \times n} \) (cf. [3, 5]) from compressive measurements \((m < d)\). Let \( x \in \mathbb{K}^n \) be the coefficient vector of \( f \) over \( D \) such that \( f \approx Dx \) with \( x \) being \( s \)-sparse.\(^1\) Then, the composition \( \Psi = AD \) can be viewed as a sensing matrix for \( x \) that produces the measurement vector \( y \). Once an estimate \( \hat{x} \) of \( x \) is computed, \( D\hat{x} \) provides an estimate of the unknown signal \( f \). Hence, we may focus on the recovery of sparse \( x \).

In an ideal case with exact sparse signal model and noise-free measurements, if any \( 2s \) columns of \( \Psi \) are linearly independent, the unknown \( s \)-sparse \( x \) is recovered as the unique solution to the linear system \( \Psi x = y \) [6–8]. In typical examples of compressed sensing (e.g., \( \Psi \) is a matrix with independently and identically distributed (i.i.d.) Gaussian entries), this is often achieved with \( m = 2s \). However, this algebraic guarantee only shows the uniqueness of the solution. Furthermore, it is only valid in the absence of measurement noise and no error in the sparse signal model.

In practice, both computational cost of signal recovery, and its robustness against noise and model error are of interest. For certain matrices \( \Psi \), the unknown \( x \) is stably recovered using efficient algorithms from compressive measurements. The required number of measurements for a stable recovery is quantified through a property of \( \Psi \) called the \textit{restricted isometry property} (RIP) [9].

\textbf{Definition 1.1.1.} The \( s \)-restricted isometry constant \( \delta_s(\Psi) \) of \( \Psi \) is defined

\(^1\)When \( w \) is assumed arbitrary, the model error term \( A(f - Dx) \) can be absorbed into \( w \). Alternatively, \( x \) can be assumed approximately sparse. We consider the former case in this paper.
as the smallest $\delta$ that satisfies
\[
(1 - \delta)\|x\|_2^2 \leq \|\Psi x\|_2^2 \leq (1 + \delta)\|x\|_2^2, \quad \forall s\text{-sparse } x.
\] (1.1.1)

Matrix $\Psi$ satisfies the RIP of order $s$ if $\delta_s(\Psi) < c$ for some constant $c \in (0, 1)$. Intuitively, smaller $\delta_s(\Psi)$ implies that $\Psi^*\Psi x$ is closer to $x$ for all $s$-sparse $x$. Although, in general, the recovery of $s$-sparse $x$ from compressive measurements is NP hard even in the noiseless case, the recovery can be accomplished efficiently (in polynomial time) and with guaranteed accuracy, when $\Psi$ satisfies the RIP with certain parameters (order and threshold). Such results are among the major achievements of compressed sensing theory. For example, when $\delta_{2s}(\Psi) < \sqrt{2} - 1$, the solution to an $\ell_1$-norm-based convex optimization formulation provides a good approximation of the unknown $s$-sparse $x$ [10]. The approximation error in this result is guaranteed to be small, and vanishes in the noiseless case. A computationally efficient alternative is provided by iterative greedy algorithms [11–14], which exploit the RIP of $\Psi$ to compute an approximation of $x$. These iterative greedy algorithms provide similar approximation guarantees when $\delta_{ks}(\Psi) < c$, where $k \in \{2, 3, 4\}$ and $c \in (0, 1)$ are constants specified by the algorithms. Different applications of the RIP require different values for the parameters $k$ and $c$. Henceforth, we assume that $k$ and $c$ are arbitrarily fixed constants as above.

The question of feasibility of compressed sensing then reduces to determining whether, and with how many measurements, $\Psi$ satisfies the RIP.\(^2\) Certain random matrices $\Psi \in \mathbb{K}^{m \times n}$ satisfy $\delta_s(\Psi) < \delta$ with high probability when the number of measurements $m$ satisfies $m = O(\delta^{-2}s \ln^\alpha n)$ for some small integer $\alpha$ [17–20]. This result, when combined with the aforementioned RIP-based guarantees of the recovery algorithms, enables “compressive sensing” ($m < d$). For example, if $\Psi$ satisfies the strong concentration property, that is, $\|\Psi x\|_2^2$ is highly concentrated around its expectation for all $x$, then $\delta_s(\Psi) < \delta$ holds with $m = O(\delta^{-2}s \ln(n/s))$ [17]. In words, a number $m$ of measurements that is proportional to the number $s$ of nonzeros, and only logarithmic in the number $n$ of unknowns, suffices for stable and computationally efficient recovery. This celebrated result of compressed sensing has been extended to

\(^2\)There also exist analyses not in terms of the RIP (e.g., [15], [16]). However, these analyses only apply to certain ideal random matrices such as an i.i.d. Gaussian matrix, which although reasonable in models for regression problems in statistics, is rarely used in practical acquisition systems.
the case where $A$ satisfies the strong concentration property with $\delta_s(A) < \delta$ and $D$ satisfies the RIP, stating that $\delta_s(AD) < \delta(D) + \delta + \delta \cdot \delta_s(D)$ holds with $m = O(\delta^{-2} s \ln(n/s))$ [21]. Now, the RIP of $D$ is often relatively easy to satisfy. Recall that the role of $D$ is to provide a sparse representation of $f$. Although redundant $D$ (with $n \geq d$) performs better in this respect, it is often the case that $f$ is sparse over a $D$ that is a basis (e.g., a piecewise smooth signal $f$ over a wavelet basis $D$). In this case, $\delta_s(D)$ is easily bounded using the condition number of $D$. Furthermore, if $D$ is an orthonormal basis, then $\delta_s(D) = 0$ for any $s \leq n$. As for the strong concentration property of $A$, it is satisfied by an i.i.d. Gaussian or Bernoulli matrix [17]. This has been extended recently to any matrix satisfying the RIP with certain parameters, when postmultiplied by a random diagonal matrix of $\pm 1$ [22]. When implementing such a sensing system is technically feasible, it would provide a sensing matrix $A$ that admits efficient computation [23].

However, although the aforementioned random matrix models are interesting in theory, they are rarely used in practice. In most practical signal acquisition systems, the linear functionals used for acquiring the measurements (rows of $A$) are determined by the physics of the specific modality and by design constraints of the sensor. In compressed sensing applied to these systems [2,24], the sensing matrix $A$ does not follow the aforementioned random matrix models; instead its rows are i.i.d. samples from the uniform distribution on a set that constitutes a frame in $\mathbb{K}^d$.

To describe the sensing matrix more precisely, we recall the definition of a frame [25]. We denote by $L_2(\Omega, \mu)$ the Hilbert space of functions defined on a compact set $\Omega$ that are square integrable with respect to a probability measure $\nu$ on $\Omega$, and by $\ell^d_2$ the $d$-dimensional Euclidean space.

**Definition 1.1.2.** Let $\mu$ denote the uniform probability measure on a compact set $\Omega$. Let $(\phi_\omega)_{\omega\in\Omega}$ be a set of vectors in $\mathbb{K}^d$. Let $\Phi : L_2(\Omega, \mu) \to \ell^d_2$ be the
synthesis operator associated with \((\phi_\omega)_{\omega \in \Omega}\) defined as

\[
\Phi h = \int_{\Omega} \phi_\omega h(\omega) d\mu(\omega), \quad \forall h \in L_2(\Omega, \mu),
\]

with its adjoint \(\Phi^* : \ell^d_2 \to L_2(\Omega, \mu)\), which is the corresponding analysis operator given by

\[
(\Phi^* f)(\omega) = \langle \phi_\omega, f \rangle, \quad \forall \omega \in \Omega, \forall f \in \ell^d_2.
\]

Then, \((\phi_\omega)_{\omega \in \Omega}\) is a frame, if the frame operator \(\Phi \Phi^* \) satisfies \(\alpha \leq \lambda_{\min}(\Phi \Phi^*) \leq \lambda_{\max}(\Phi \Phi^*) \leq \beta\) for some positive real numbers \(\alpha\) and \(\beta\). In particular, if the frame operator \(\Phi \Phi^* \) is a scaled identity, then \((\phi_\omega)_{\omega \in \Omega}\) is a tight frame.

Let \(\nu\) be a probability measure on \(\Omega\). Let \(\bar{z}\) denote the complex conjugate of \(z \in \mathbb{C}\) and \([m]\) denote the set \(\{1, \ldots, m\}\). The sensing matrix \(A \in \mathbb{K}^{m \times d}\) is constructed from a frame \((\phi_\omega)_{\omega \in \Omega}\) as

\[
A_{k,\ell} = \frac{1}{\sqrt{m}} (\phi_{\omega_k})_\ell, \quad \forall k \in [m], \ell \in [d]
\]

for random indices \((\omega_k)_{k=1}^m\) in \(\Omega\) chosen i.i.d. with respect to \(\nu\). We call this type of matrix a random frame matrix. It is the model for a sensing matrix of primary interest in this paper, and we will assume henceforth that \(A\) is defined by (1.1.4).

Random frame matrices arise in numerous applications of compressed sensing. We list a few below. For simplicity, they are described for the 1D case.

**Example 1.1.3.** An important example of a random frame matrix is a random partial discrete Fourier transform (DFT) matrix. Let

\[
\phi_\omega \triangleq [1, e^{-j2\pi \omega}, \ldots, e^{-j2\pi(d-1)\omega}]^T
\]

be defined for \(\omega \in \Omega \triangleq \{1/d, \ldots, (d-1)/d, 1\}\). In this setup, \(\nu : \Omega \to [0,1]\) is a cumulative density function on \(\Omega\) and \(\frac{d\nu}{d\mu}(\omega)\) denotes the probability that \(\omega\) will be chosen, multiplied by \(d\). Then, an \(m \times d\) random partial DFT matrix is constructed from \((\phi_\omega)_{\omega \in \Omega}\) using \(\nu\) by (1.1.4). The frame \((\phi_\omega)_{\omega \in \Omega}\) in this example is a tight frame, and \(\sup_{\omega} \|\phi_\omega\|_{\ell^d_2}/\|\phi_\omega\|_{\ell^2}\), which will play a role in our subsequent discussion, achieves its minimum \(\frac{1}{\sqrt{d}}\). Sensing matrix of this kind arise in practical applications of compressed sensing such as the
multi-coset sampling and spectrum-blind recovery of multiband signals at sub-
Nyquist rates [7, 26, 27]. Similar random matrices also arise in more recent
studies on compressed sensing of analog signals [29–32].

Example 1.1.4. One author of this paper proposed the compressive ac-
quision of signals in Fourier imaging systems [2, 8, 33], which is one of
the works that invented the notion of compressed sensing. This idea has
been applied with refinements to various modalities such as magnetic re-
sonance imaging (MRI) [24, 34], photo-acoustic tomography [35], radar [36],
radar imaging [37, 38], and astronomical imaging [39], etc. The sensing
matrix $A$ for compressed sensing in Fourier imaging systems is a random
partial Fourier transform matrix with continuous-valued frequencies (con-
tinuous random partial Fourier matrix, henceforth), which is obtained similarly
to the previous example. Let $\phi_\omega \triangleq [1, e^{-j2\pi \omega}, \ldots, e^{-j2\pi (d-1)\omega}]^T$ be defined for
$\omega \in \Omega \triangleq [-\frac{1}{2}, \frac{1}{2})$. The frame $(\phi_\omega)_{\omega \in \Omega}$ in this example is a continuous tight frame, and the quantity $\sup_\omega \|\phi_\omega\|_{\ell^\infty} / \|\phi_\omega\|_{\ell^2}$ achieves its minimum $\frac{1}{\sqrt{d}}$.

Example 1.1.5. In MRI, the Fourier measurements are usually modeled
as obtained from the input signal modified by pointwise multiplication with
a mask $\lambda \in \mathbb{K}^d$, representing the receiving coil sensitivity profile. Let $\Lambda = \text{diag}(\lambda)$ denote the diagonal matrix with the elements of $\lambda$ on the diagonal.
Let $\phi_\omega \triangleq \Lambda^*[1, e^{-j2\pi \omega}, \ldots, e^{-j2\pi (d-1)\omega}]^T$ be defined for $\omega \in \Omega \triangleq [-\frac{1}{2}, \frac{1}{2})$. If $\lambda$ has no zero element, then $(\phi_\omega)_{\omega \in \Omega}$ is a frame that spans $\mathbb{K}^d$. Otherwise,
$(\phi_\omega)_{\omega \in \Omega}$ is a frame for the subspace $S$ of $\mathbb{K}^d$ spanned by the standard basis vectors corresponding to the nonzero elements of $\lambda$. In the latter case, letting the signal space be $S$ instead of $\mathbb{K}^d$, we modify the inverse problem so that $A$ constructed by (1.1.4) is a map from $S$ to $\mathbb{K}^m$. Note that each vector in the frame is multiplied from the right by $\Lambda$ compared to that in Example 1.1.4. In this example, unless the nonzero elements of $\lambda$ have the same magnitudes,
$(\phi_\omega)_{\omega \in \Omega}$ does not satisfy the two properties coming from a Fourier system
(tightness and minimal $\sup_\omega \|\phi_\omega\|_{\ell^\infty} / \|\phi_\omega\|_{\ell^2}$). Therefore, we do not restrict
our interest to the Fourier case and consider a general frame $(\phi_\omega)_{\omega \in \Omega}$.

Because random frame matrices are so ubiquitous in compressed sens-
ing, the analysis of their RIP is of major interest. Although random frame

\footnote{This was the invention of compressed sensing of analog signals. See [28] for a survey of this early work.}
matrices do not satisfy the strong concentration property, other tools are available for the analysis of their RIP. In particular, the RIP of a partial Fourier matrix has been studied using noncommutative probability theory [18, 19]. The extension of this analysis to the RIP of a random frame matrix [20] enables handling a more general class of sensing matrices. Notably, all known analyses [18–20, 40] focused on the case where \( \mathbf{D} \) corresponds to an orthonormal basis. These analyses also assumed either the exact isotropy property, \( \mathbb{E}\mathbf{A}^*\mathbf{A} = \mathbb{I} \) [18–20], or the so-called near isotropy property, \( \| \mathbb{E}\mathbf{A}^*\mathbf{A} - \mathbb{I} \| = O\left( \frac{1}{\sqrt{n}} \right) \) [40]. There is no alternative sufficient condition that does not require these properties. In fact though, these RIP analyses further extend to the following Theorem 1.1.6 (proved in Section 2.3), which addresses the case of \( \Psi = \mathbf{A}\mathbf{D} \), where \( \mathbf{A} \) is a random frame matrix and \( \mathbf{D} \) is not necessarily an orthonormal basis, and furthermore, allows a non-vanishing deviation from isotropy.

**Theorem 1.1.6.** Let \( \mathbf{A} \in \mathbb{K}^{m \times d} \) be a random matrix constructed from a frame \( (\phi_\omega)_{\omega \in \Omega} \) by (1.1.4) and let \( \mathbf{D} = [d_1, \ldots, d_n] \in \mathbb{K}^{d \times n} \) satisfy \( \delta_s(\mathbf{D}) < 1 \). Suppose that \( \sup_{\omega} \max_j |\langle \phi_\omega, d_j \rangle| \leq K \). Let \( \Psi = \mathbf{A}\mathbf{D} \). Then, \( \delta_s(\Psi) < \delta + \| \mathbb{E}\mathbf{A}^*\mathbf{A} - \mathbb{I} \| + \delta_s(\mathbf{D}) + \| \mathbb{E}\mathbf{A}^*\mathbf{A} - \mathbb{I} \| \delta_s(\mathbf{D}) \) holds with high probability for \( m = O(\delta_s^{-2} s \ln^4 n) \).

The inequality in Theorem 1.1.6 indicates that \( \delta_{ks}(\Psi) < c \) holds with high probability for \( m = O(s \ln^4 n) \) if \( \mathbf{D} \) satisfies \( \delta_{ks}(\mathbf{D}) \leq \frac{\epsilon}{4} \), and \( \mathbf{A} \) satisfies \( \| \mathbb{E}\mathbf{A}^*\mathbf{A} - \mathbb{I} \| \leq \frac{\epsilon}{4} \). Combined with the aforementioned RIP-based guarantees, this result again enables compressive sensing, when the conditions given in Theorem 1.1.6 are satisfied.

1.1.2 Motivation: Failure of Guarantees in Practical Applications

While the RIP is essential for all existing performance guarantees for compressed sensing with random frame sensing matrices, it turns out that this property is satisfied only under certain nonrealistic assumptions. Most notably, although compressed sensing has been proposed to accelerate the acquisition in imaging systems [2, 4, 24] and some of the most widely studied applications of compressed sensing to date are in such systems, the RIP has not been shown to hold for the associated sensing matrices in a realistic
setup. More specifically, $|\|\mathbb{E}A^*A - I_d\||$ is not negligible, which makes even the upper bound on $\delta_s(\Psi)$ given by Theorem 1.1.6, which is the most relaxed condition on deviation from isotropy known to date, too conservative to be used for RIP-based recovery guarantees.

One reason for the increase $|\|\mathbb{E}A^*A - I_d\|$ from the ideal case is the use of a nonuniform distribution in the construction of $A$. In Examples 1.1.3 and 1.1.4, the sensing matrix $A$ were constructed from i.i.d. samples from a tight frame $(\phi_\omega)_{\omega \in \Omega}$. In this case, if the i.i.d. sampling is done in accordance to the uniform distribution, then $\mathbb{E}A^*A = I_d$. However, in practice, i.i.d. sampling using a nonuniform distribution is often preferred for natural signals: it is desirable to take more measurements of lower frequency components, which contain more of the signal energy. Therefore, acquisition at frequencies sampled non-uniformly with a variable density is preferred [24]. As a consequence, the exact isotropy property is violated. Depending on the probability distribution, $|\|\mathbb{E}A^*A - I_d\|$ is often not negligible, and even larger than 1, which renders the upper bound on $\delta_s(\Psi)$ in Theorem 1.1.6 useless. Therefore, no known RIP analysis applies to Fourier imaging applications.

Another reason for the increase $|\|\mathbb{E}A^*A - I_d\|$ from the ideal case is that $(\phi_\omega)_{\omega \in \Omega}$ is a not tight frame. As shown in Example 1.1.5, even in a Fourier imaging system, $(\phi_\omega)_{\omega \in \Omega}$ can be a non-tight frame due to the presence of a mask. Furthermore, the application of compressed sensing is not restricted to Fourier imaging systems. The idea of compressed sensing and recovery using sparsity also applies to other inverse problems in imaging described by non-orthogonal operators (e.g., a Fredholm integral equations of the first kind). Optical diffusion tomography [41] is a concrete example of compressed sensing with such a scenario. As another example, the sensing matrix that arises in compressed sensing in shift-invariant spaces [29] is not necessarily obtained from a tight frame.

Yet another reason for the failure of the upper bound on $\delta_s(\Psi)$ in Theorem 1.1.6 has to do with the dictionary $D$. Indeed, to achieve $\delta_s(\Psi) < c$ with $m = O(s \ln^4 n)$, it is necessary that both $|\|\mathbb{E}A^*A - I_d\|$ and $\delta_s(D)$ are less than a certain threshold. However, verification of this condition for $\delta_s(D)$ is usually computationally expensive. For the special case where $D$ has full column rank (hence, $d \geq n$), $\delta_s(D)$ is easily bounded from above by $|\|D^*D - I_n\|$. In particular, if $D$ corresponds to an orthonormal basis, then $D^*D = I_n$, which implies $\delta_s(D) = 0$. Otherwise, $\delta_s(D)$ vanishes as $D$ approaches an orthonor-
mal basis. However, it is often too restrictive to make $\|D^*D - I_n\|$ less than a small threshold below 1. Moreover, without this constraint, $D$ can provide a better sparse representation, which is also desired for stable recovery. In particular, for a data-adaptive $D$, the property that $\|D^*D - I_n\|$ is less than a given threshold is not guaranteed. In this all too common situation, all known RIP analyses break down: they only provide a conservative upper bound on $\delta_s(\Psi)$, which does enable the RIP-based recovery-guarantees.

In summary, in most practical compressed sensing applications, the effective sensing matrix $\Psi = AD$ may fail to satisfy the RIP for one or more of the following reasons: the i.i.d. sampling in the construction of $A$ does not use the uniform distribution; the frame used in the construction of $A$ is not tight; or the dictionary $D$ does not have a sufficiently small restricted isometry constant. From these observations, we conclude that none of the existing performance guarantees for recovery algorithms applies to the aforementioned applications of compressed sensing.

1.1.3 Contributions

Recall that unlike the $\ell_1$-norm-based recovery, greedy recovery algorithms were designed to exploit the property that $\Psi^*\Psi x \approx x$ for sparse $x$, explicitly. For example, in the derivation of the CoSaMP algorithm [11], the procedure of applying $\Psi^*$ to $y = \Psi x$ for $x$ sparse was called the computation of a “proxy” signal, which reveals the information about the locations of nonzero elements of $x$. The same idea was also used for deriving other iterative greedy algorithms [12–14]. Indeed, if $\Psi$ satisfies the RIP, then the use of the (transpose of) the same matrix $\Psi$ to compute a proxy is a promising approach. Otherwise, one can employ a different matrix $\tilde{\Psi}$ to get a better proxy $\tilde{\Psi}^*y$. The required property is that $\tilde{\Psi}^*\Psi x \approx x$ for sparse $x$. To improve the recovery algorithms in this direction, we first extend the RIP to a property of a pair of matrices $\Psi, \tilde{\Psi} \in \mathbb{K}^{m \times n}$ called the restricted biorthogonality property (RBOP).

**Definition 1.1.7.** The $s$-restricted biorthogonality constant $\theta_s(M)$ of $M \in$
\( K^{n \times n} \) is defined as the smallest \( \delta \) that satisfies
\[
|\langle y, Mx \rangle - \langle y, x \rangle| \leq \delta \|x\|_2 \|y\|_2, \quad \forall \text{s-sparse } x, y \text{ with common support.}
\]

\[(1.1.5)\]

The pair \((\Psi, \widetilde{\Psi})\) satisfies the RBOP of order \( s \) if \( \theta_s(\widetilde{\Psi}^* \Psi) < c \) for some constant \( c \in (0, 1) \).\(^5\) Intuitively, smaller \( \theta_s(\widetilde{\Psi}^* \Psi) \) implies that \( \widetilde{\Psi}^* \Psi x \) becomes closer to \( x \) for all \( s \)-sparse \( x \). In other words, any \( s \) columns of \( \Psi \) and \( \widetilde{\Psi} \) corresponding to the same indices behave like a biorthogonal basis. If \( \widetilde{\Psi} = \Psi \), then \( \theta_s(\widetilde{\Psi}^* \Psi) \) reduces to \( \delta_s(\Psi) \); hence, the RBOP of \((\Psi, \widetilde{\Psi})\) reduces to the RIP of \( \Psi \).

We then modify the greedy recovery algorithms so that the modified algorithms employ both \( \Psi \) and \( \widetilde{\Psi} \) and, in particular, exploit the RBOP of \((\Psi, \widetilde{\Psi})\) to provide an approximation guarantee. In fact, modified thresholding and forward greedy algorithms using a different matrix \( \widetilde{\Psi} \) have been already proposed by Schnass and Vandergheynst [42]. However, our work is different from theirs in several important respects. Schnass and Vandergheynst [42] propose to use \( \widetilde{\Psi} \) numerically optimized to minimize a version of the Babel function. However, although sufficient conditions given in terms of the Babel function are easily computable, the resulting guarantees for the recovery performance are conservative. Furthermore, their numerical algorithm to design \( \widetilde{\Psi} \) is a heuristic, and does not provide any guarant e on the value of the Babel function achieved. In contrast, we propose an explicit construction of \( \widetilde{\Psi} \) so that \( \theta_s(\widetilde{\Psi}^* \Psi) \ll 1 \) holds. To show the construction, we recall the definition of a biorthogonal frame that extends the notion of a biorthogonal basis.

**Definition 1.1.8.** Let \((\phi_\omega)_{\omega \in \Omega}\) and \((\tilde{\phi}_\omega)_{\omega \in \Omega}\) be sets of vectors in \( \mathbb{K}^d \). Let \( L_2(\Omega, \mu) \) be as defined in Definition 1.1.2. Let \( \Phi^* : \ell_2^d \to L_2(\Omega, \mu) \) be the analysis operator associated to \((\phi_\omega)_{\omega \in \Omega}\) defined in (1.1.3). Let \( \Phi : L_2(\Omega, \mu) \to \ell_2^d \) be the synthesis operator associated to \((\tilde{\phi}_\omega)_{\omega \in \Omega}\) defined similarly to (1.1.2). Then, \((\phi_\omega, \tilde{\phi}_\omega)_{\omega \in \Omega}\) is a biorthogonal frame if \( \tilde{\Phi} \Phi^* = I_d \).

Matrix \( \widetilde{\Psi} \) is then constructed as the composition \( \widetilde{\Psi} = \widetilde{A} \tilde{D} \). We construct  

\(^5\)As in the case of the RIP, the threshold value of \( c \) for which the RBOP is said to be satisfied depends on the application.
\( \tilde{A} \in \mathbb{K}^{m \times d} \) from the dual frame \((\tilde{\phi}_\omega)_{\omega \in \Omega}\) by

\[
(\tilde{A})_{k,\ell} = \frac{1}{\sqrt{m}} \left[ \frac{d\nu}{d\mu}(\omega_k) \right]^{-1} (\tilde{\phi}_{\omega_k})_{\ell}, \quad \forall k \in [m], \ell \in [d]
\]

(1.1.6)

where \((\omega_k)_{k=1}^m\) are the same indices as used to define the samples from \((\psi_\omega)_{\omega \in \Omega}\) in the construction of \(A\) in (1.1.4). Assuming \(\frac{d\nu}{d\mu}(\omega) > 0\), then, by the construction of \(A\) and \(\tilde{A}\), it follows that the pair \((A, \tilde{A})\) satisfies the dual isotropy property

\[ E\tilde{A}^*A = I_d. \]

**Remark 1.1.9.** We proposed modified greedy pursuit algorithms in Section 2.2 that use both \(\Psi = AD\) and \(\widetilde{\Psi} = \tilde{A}D\) and are guaranteed using the RBOP of \((\Psi, \widetilde{\Psi})\). Therefore, it is important to check whether \(\tilde{\Psi}^* = \tilde{D}^*\tilde{A}^*\) can be efficiently implemented. The discussion on \((D, \tilde{D})\) is deferred to the next subsections and we only discuss the computational issue with \(\tilde{A}\) here. In practice, \(A^*\) is implemented using fast algorithms without forming a dense matrix explicitly. For example, if \(A\) is a partial DFT matrix, then, \(A^*\) is implemented as the fast Fourier transform (FFT) applied to the zero padded vector. Likewise, if \(A\) is a continuous partial Fourier matrix, then, the nonuniform FFT (NUFFT) [43] can be used for fast computation. In this case, since our construction of \(\tilde{A}\) in (1.1.6) only involves row-wise rescaling of \(A\) by constant factors, \(\tilde{A}^*\) is also implemented using the same fast algorithms. In the more general biorthogonal case, once the synthesis operator \(\tilde{A}\) is implemented as a fast algorithm, \(\tilde{A}^*\) is also computed efficiently using the same algorithm. In fact, in many applications, the biorthogonal dual system is given analytically. For example, if the frame \((\phi_\omega)_{\omega \in \Omega}\) is given as a filter bank system, designing perfect reconstruction filters that provide the corresponding biorthogonal dual frame is well studied [44]. Similar arguments apply to the analysis operator of analytic frames such as overcomplete discrete cosine transform (DCT) or wavelet packets.

Regarding the construction of \(\tilde{D}\), we consider the following two cases: (i) \(D\) corresponds to a basis for \(\mathbb{K}^n\) \((d = n)\); (ii) \(D\) satisfies the RIP with certain parameter. We let \(\tilde{D} = D(D^*D)^{-1}\) for the former case and \(\tilde{D} = D\) for the latter case. The RBOP of this construction is deferred to after the exposition of new recovery algorithms.

Now, we return to the discussion of the recovery algorithms. While Schnass
and Vandergheynst [42] only replaced $\Psi$ by $\tilde{\Psi}$ in the steps of computing a proxy in forward greedy algorithms (MP and OMP), we also replace the orthogonal projection used in the update of the residual in OMP by a corresponding oblique projection obtained from $\Psi$ and $\tilde{\Psi}$. Therefore, we propose a different variation of OMP called Oblique Matching Pursuit (ObMP), which is guaranteed using the RBOP of $(\Psi, \tilde{\Psi})$. We also propose similar modifications of iterative greedy recovery algorithms and their RIP-based guarantees.

The modified algorithms are different from the original algorithms: we assign them new names, with the modifier “oblique”. For example, SP is extended to oblique subspace pursuit (ObSP). CoSaMP, IHT, and HTP are likewise extended to ObCoSaMP, ObIHT, ObHTP, respectively. We call these modified greedy algorithms based on the RBOP oblique pursuits. In the numerical experiments in this paper, in scenarios where one or more of the ideal assumptions (i.i.d. sampling according to the uniform distribution, tight frame $(\phi_\omega)_{\omega \in \Omega}$, or orthonormal basis $D$) are violated, the oblique pursuits perform better than, or at least competitively with their conventional counterparts.

Importantly, the oblique pursuits come with RBOP-based approximation guarantees. In particular, similarly to its conventional counterpart, each iterative oblique pursuit algorithm is guaranteed when $\theta_{ks}(\tilde{\Psi}^*\Psi) < c$, where $k \in \{2, 3, 4\}$ and $c \in (0, 1)$ are constants specified by the algorithms. The number of measurements required for the guarantees of oblique pursuits is also similar to that required in the ideal scenario by their conventional counterparts. When combined with the subsequent RBOP analysis of $(\Psi, \tilde{\Psi})$ for random frame sensing matrices, the recovery by the iterative oblique pursuit algorithms is guaranteed with $m = O(s \ln^4 n)$. In particular, we show that it is no longer necessary to have $\|EAA^* - I_n\| \ll 1$. Therefore, the obtained guarantees apply in realistic setups of the aforementioned CS applications.

The degrees of freedom added by the freedom to design $\tilde{\Psi}$ make the RBOP easier to satisfy under milder assumptions than the RIP. In particular, with the proposed construction of $\tilde{\Psi}$, the RBOP of $(\Psi, \tilde{\Psi})$ holds without requiring the (near) isotropy property of $A$. More specifically, depending on whether $D$ corresponds to a basis or satisfies the RIP, the RIP analysis in Theorem 1.1.6 is extended to the following theorems. Recall that we proposed different constructions of $\tilde{\Psi}$ for the two cases.

**Theorem 1.1.10.** Let $A, \tilde{A} \in \mathbb{K}^{m \times d}$ be random matrices constructed from
a biorthogonal frame $(\phi_\omega, \tilde{\phi}_\omega)_{\omega \in \Omega}$ by (1.1.4) and (2.3.22), respectively. Let $D = [d_1, \ldots, d_n]$ and $\tilde{D} \in \mathbb{K}^{d \times n}$ ($d = n$) satisfy $\tilde{D}^*D = I_d$. Let $\Psi = AD$ and $\tilde{\Psi} = \tilde{A}D$. Suppose that $\sup_\omega \max_j |\langle \phi_\omega, d_j \rangle| \leq K$. Then, $\theta_s(\tilde{\Psi}^*\tilde{\Psi}) < \delta$ holds with high probability for $m = O(\delta^{-2}s \ln^4 n)$.

**Theorem 1.1.11.** Let $A, \tilde{A} \in \mathbb{K}^{m \times d}$ be random matrices constructed from a biorthogonal frame $(\phi_\omega, \tilde{\phi}_\omega)_{\omega \in \Omega}$ by (1.1.4) and (2.3.22), respectively. Let $D = [d_1, \ldots, d_n] \in \mathbb{K}^{d \times n}$ satisfy $\delta_s(D) < 1$. Let $\Psi = AD$ and $\tilde{\Psi} = \tilde{A}D$. Suppose that $\sup_\omega \max_j |\langle \phi_\omega, d_j \rangle| \leq K$. Then, $\theta_s(\tilde{\Psi}^*\tilde{\Psi}) < \delta + \delta_s(D)$ holds with high probability for $m = O(\delta^{-2}s \ln^4 n)$.

Note that the upper bounds on $\theta_s(\tilde{\Psi}^*\tilde{\Psi})$ in Theorems 1.1.10 and 1.1.11 do not depend on $\|E A^* A - I_d\|$. Therefore, unlike the RIP, which breaks down when the ideal assumptions, such as i.i.d. sampling according to the uniform distribution and tight frame, are violated, the RBOP continues to hold even with such violations.

In summary, we introduced a new tool for the design, analysis, and performance guarantees of sparse recovery algorithms, and illustrate its application to derive new guaranteed versions of several of the most popular recovery algorithms.

### 1.2 Joint Sparse Recovery

The problem of computing a sparse approximate solution to a linear system has been studied as the subset selection problem in matrix computations [45] with applications in statistical regression [46] and signal processing [26,33,47]. The study dates back to the 1970s [48,49]. Relevant theories and algorithms were further developed in the 1980s [50,51] and in the 1990s [52–55]. Recently, this subject became more popular in the signal processing community with the name of *compressed sensing* [3]. In particular, the elegant analysis derived with modern probability theory [3, 4] provided performance guarantees for polynomial-time algorithms in terms of properties of random matrices. These might be the most important contributions in recent years.

In compressed sensing, computing a sparse solution to a linear system $Ax = y$ is interpreted as the recovery of an unknown sparse vector from the measurement vector $y$ obtained through the sensing matrix $A$. The sparse
recovery problem addresses the identification of the support, which denotes the indices of the nonzero elements of the unknown sparse vector. Once the support is determined, the recovery of the unknown sparse vector reduces to a standard overdetermined linear inverse problem.

The joint sparse recovery problem, also known as the multiple measurement vector (MMV) problem, aims to identify a common support shared by unknown sparse vectors $x_1, \ldots, x_N$ from the multiple measurement vectors $y_k = Ax_k$ for $k = 1, \ldots, N$ obtained through a common sensing matrix $A$. The common support then corresponds to the indices of the nonzero rows of the unknown matrix $X \triangleq [x_1, \ldots, x_N]$. The joint sparse recovery problem arises in numerous applications including sub-Nyquist sampling of multiband signals [7, 26–28, 30, 33, 56], estimation of sparse brain excitation [54], imaging [2, 8, 41], multivariate regression [57], and direction of arrival (DOA) estimation [47], and is often an easier problem with better performance compared to the single measurement vector (SMV) case.

Algorithms that exploit the joint sparsity have been developed for the joint sparse recovery problem. Bresler and Feng showed the similarity between the joint sparse recovery problem and the DOA estimation problem in sensor array processing, and proposed to use a version of the Multiple Signal Classification (MUSIC) algorithm [58] for the full row rank case, where the nonzero rows of the unknown matrix $X$ have full row rank [7, 26, 33]. For the rank-defective case, they proposed [7, 26, 33] methods based on a greedy search inspired by the alternating projections algorithm [59] in DOA estimation.

More recently, existing solutions to the sparse recovery problem for the SMV case were extended to the MMV case. Greedy algorithms [60–62] extend orthogonal matching pursuit (OMP) [63] to the MMV case, and convex optimization formulations with mixed norm [47, 64–66] extend the corresponding SMV solution, such as basis pursuit (BP) [67] and LASSO [68] to the MMV case. Sparse Bayesian learning (SBL) [69] has also been extended to the MMV case [70, 71].

Theories have also been developed for both the joint sparse recovery problem itself and for the guarantees of the algorithms that solve it. Bresler and Feng applied the theory for the fundamental limits on DOA estimation [72] to determine the conditions for the unique identification of the support [7, 26]. The condition has been further studied in more general set-
TINGS [60, 65, 73]. MUSIC applied to joint sparse recovery [7, 26] was the first method that was guaranteed with the tightest sufficient condition, which also coincides with a necessary condition required for the support identification by any method. However, the guarantee only applies to the full row rank case with either noise-free measurement vectors, or with noise of known covariance and an exactly known measurement covariance matrix. Performance guarantees of greedy algorithms and of convex optimization formulations for joint sparse recovery have also been studied extensively in the literature [61, 62, 64, 65, 74–77]. However, the guarantees of such methods have not been proved to be strictly better than the guarantees for the SMV case. Moreover, unlike the guarantee of MUSIC [7, 26], in the full row rank noise-free case, such methods are not guaranteed with the minimal requirements.

Performance guarantees aside, the empirical performance and computational cost of any method are of key importance and usually determine its adoption in practice. Empirically, the optimization schemes with diversity measures (e.g., the mixed norm) perform better than greedy algorithms. In particular, the rate of exact support recovery in existing greedy algorithms does not improve with increasing rank of the unknown matrix beyond a certain level. While the optimization schemes with diversity measures perform better empirically than the greedy algorithms, this improved performance comes at a much higher computational cost. In contrast, greedy algorithms and MUSIC are computationally efficient. In a summary, none of the listed methods enjoys good empirical performance and computational speed at the same time.

In view of the various drawbacks of the existing algorithms for joint sparse recovery, MUSIC, when it works, is extremely attractive. In a favorable setting (the full row rank case), MUSIC provides both good empirical performance and a guarantee with minimal requirement. Moreover, MUSIC is highly efficient computationally. However, the full row rank condition is often violated in practice. For example, if the number of measurement vectors $N$ is smaller than the sparsity level $s$, then no more than $N$ rows can be linearly independent, and the nonzero rows do not have full row rank. Even with large $N$, the submatrix of nonzero rows might be rank deficient or ill-conditioned. For example, in the DOA estimation problem, correlation between sources or multipath propagation can cause a large condition number. It is well known that MUSIC fails in this practically important “rank-defective” case
and this has motivated numerous attempts to overcome this problem, without resorting to an infeasible multi-dimensional search. However, all of these previous methods use a special structure of the linear system – such as shift invariance that enables applying so-called spatial smoothing [78]. Previous extensions of MUSIC in sensor array processing are therefore not applicable to the general joint sparse recovery problem.

The main contributions of this paper are summarized as follows. First, we propose a new class of algorithms, *subspace-augmented MUSIC* (SA-MUSIC) that overcome the limitations of existing algorithms and provide the best of both worlds: good empirical performance at all rank conditions and efficient computation. In particular, SA-MUSIC algorithms improve on MUSIC so that the support is recovered in the rank-defective case. Compared to MUSIC [26], in the presence of a rank defect, SA-MUSIC has additional steps of partial support recovery and subspace augmentation. Combined with any guaranteed algorithm for partial support recovery, SA-MUSIC provides a guaranteed algorithm for the entire recovery problem. In particular, we propose a subspace-based greedy algorithm, *orthogonal subspace matching pursuit* (OSMP). When combined with partial support recovery by OSMP, SA-MUSIC provides a computationally efficient solution to joint sparse recovery with a performance guarantee. A combination with a different new subspace-based greedy algorithm is proposed and analyzed elsewhere [79]. In fact, the computational requirements of SA-MUSIC algorithms are similar to those of greedy algorithms and of MUSIC. In empirical performance, SA-MUSIC algorithms outperform previous greedy algorithms and are at least equal to, but usually outperform existing convex relaxation methods.

Second, we derive explicit conditions that guarantee each step of SA-MUSIC for the noisy and/or rank-defective case. The performance is analyzed in terms of a property [80] that is a weaker version of the restricted isometry property (RIP) [9]. We call this property the *weak-1 restricted isometry property* (weak-1 RIP). The weak-1 RIP is satisfied by a milder condition than the conventional RIP; hence, the guarantees derived with the weak-1 RIP are less demanding. Most importantly, compared to the relevant work [81] with similar but independently developed ideas, the analysis in this paper is valid for the noisy case, is non-asymptotic, and applies to a wider class of sensing matrices that arise in real applications.

Contributions of independent interest include a new subspace-based greedy
algorithm for joint sparse recovery with performance guarantees, extension of the analysis of MUSIC for joint sparse recovery to the noisy case with imperfect subspace estimation, and non-asymptotic analysis of subspace estimation from finitely many snapshots. The latter analysis is different from previous analyses of subspace methods, which were based on the law of large numbers, asymptotic normality, or low-order expansions.

1.3 Low Rank Matrix Recovery

Recent studies in compressed sensing have shown that a sparsity prior in the representation of the unknowns can guarantee unique and stable solutions to underdetermined linear systems. The idea has been generalized to the matrix case [82] with the rank replacing sparsity to define the parsimony of the representation of the unknowns. Compressed sensing of a low-rank matrix addresses the inverse problem of reconstructing an unknown low-rank matrix $X_0 \in \mathbb{C}^{m \times n}$ from its linear measurements $b = AX_0^6$ via a given linear operator $A : \mathbb{C}^{m \times n} \to \mathbb{C}^p$. As in the vector case, the inverse problem is ill-posed in the sense that the number of measurements is much smaller than the number of unknowns. Continuing the analogy with the vector case, the remarkable fact is that the number of measurements sufficient for unique and stable recovery is roughly on the same order as the number of degrees of freedom in the unknown low rank matrix. Moreover, under certain conditions, the recovery can be accomplished by polynomial-time algorithms [83].

The problem of reconstructing a low rank matrix from limited linear measurements arises in a wide range of applications. Low-order system identification can be formulated as a low rank matrix recovery problem [84,85]. The order of a linear time-invariant system determines the rank of the Hankel matrix formed from the system output and the linear measurement corresponds to the sampled response of the system obtained from a random (but known) input. Recently, the low-rank matrix recovery has been successfully applied to the reconstruction of dynamic MRI [86,87]. The sequence of vectorized image frames can be modeled as a low-rank matrix using a partially separable model [88]. The linear measurement in this case corresponds to the samples in the so-called k-t domain. Quantum state tomography is another

\footnote{We use calligraphic font for general linear operators to distinguish them from matrices.}
application where the low-rank matrix model can be useful [89]. A particular
form of the problem, known as the matrix completion problem, involves the
recovery of a low rank matrix when only a subset of its entries is known. Col-
laborative filtering, which is also better known as the Netflix problem [90],
aims to complete the rating-matrix in order to provide a user-adaptive rec-
ommendation system. Since only a limited number of entries of the matrix
are available, the completion of the matrix is a challenging problem. By
modeling the matrix as a low-rank, which is justified since the preference
of a user for a particular item is determined by few factors, successful com-
pletion becomes possible. Global positioning from local distances is another
application of matrix completion [91]. The entries of the matrix denote the
inter-sensor distances in the sensor network. The rank is restricted to the
dimension of the configuration, i.e., for the two dimensional case, the rank is
at most 2.

One method to solve the inverse problem by exploiting the prior that $X_0$
is low-rank is to solve the rank minimization problem $P_1$, to minimize the
rank within the affine space defined by $b$ and $A$:

$$P_1: \min_{X \in S_2} \text{rank}(X) \quad \text{subject to} \quad AX = b.$$ 

In practice, in the presence of measurement noise or modeling error, a more
appropriate measurement model is $b = AX_0 + \nu$ where the perturbation $\nu$
has bounded Euclidean norm, $\|\nu\|_{\ell_2} \leq \eta$. In this case, the rank minimization
problem is written as

$$P_1': \min_{X \in S_2} \text{rank}(X) \quad \text{subject to} \quad \|AX - b\|_{\ell_2} \leq \eta$$

with an ellipsoidal constraint where $\|\cdot\|_{\ell_2}$ denotes the $\ell_2$-norm of a $p$-tuple
vector. In fact, rank minimization has been studied in the more general
setting where the feasible set is not necessarily restricted to either an affine
space or an ellipsoid. However, due to the non-convexity of the rank function,
rank minimization is NP-hard even when the feasible set is convex. Fazel et
al. [84] proposed a convex relaxation of the rank minimization problem by
introducing a convex surrogate of $\text{rank}(X)$, which is known as nuclear norm
and denotes the sum of all singular values of matrix $X$. The nuclear norm of $X$ coincides with the Schatten 1-norm of $X$ and will be denoted as $\|X\|_{S_1}$.

Recht et al. [83] studied rank minimization in the framework of compressed sensing and showed that rank minimization for the matrix case is analogous to $\ell_0$-norm (number of nonzero elements) minimization for the vector case. They provided an analogy between the two problems and their respective solutions by convex relaxation. In the analogy, $\ell_1$-norm minimization for the $\ell_0$-norm minimization problem is analogous to nuclear norm minimization for rank minimization. Both are efficient algorithms, with guaranteed performance under certain conditions, to solve NP-hard problems: $\ell_0$-norm minimization and rank minimization, respectively. The respective conditions are given by the sparsity-restricted isometry property [10] and the rank-restricted isometry property [82,83], respectively. However, whereas $\ell_1$-norm minimization corresponds to a linear program (or a quadratically constrained linear program for the noisy case), nuclear norm minimization is formulated as a convex semidefinite program (SDP). Although there exist polynomial time algorithms to solve SDP, in practice they do not scale well to large SDP problem instances.

Several authors proposed methods for solving large scale SDP derived from rank minimization. These include interior point methods for SDP, projected subgradient methods, and low-rank parametrization [83] combined with a customized interior point method [85]. These methods can solve larger rank minimization problems, which the general purpose SDP solvers cannot. However, the dimension of the problem is still restricted and some of these methods do not guarantee convergence to the global minimum. Other approaches are more direct, avoiding the SDP formulation. Cai et al. [92] proposed singular value thresholding (SVT), which penalizes the objective of nuclear norm minimization by the squared Frobenius norm weighted by a parameter $\tau$. They solved the Lagrangian dual of the penalized problem by a projected subgradient method. For this particular problem, each iteration of the projected subgradient method can be computed efficiently by the truncated singular value decomposition. The sequence of solutions to the parameterized problems converges to the solution of the nuclear norm minimization problem as the penalty parameter $\tau$ goes to infinity. However, an analysis of the convergence rate is missing, and hence the quality of the solution by SVT with a particular $\tau$ is not guaranteed. Furthermore, the
efficiency of SVT is restricted to the noiseless case where the constraint is affine (i.e., linear equality). Ma et al. [93] proposed a formulation of nuclear norm minimization using the Bregman divergence and solved the problem by an efficient fixed point algorithm based on the singular value decomposition. However, they did not provide a convergence rate analysis and like SVT the efficiency of this algorithm too is restricted to the noiseless, affine constraint case. Finally, Meka et al. [94] used multiplicative updates and online convex programming to provide an approximate solution to rank minimization. However, their result depends on the (unverified) existence of an oracle that provides the solution to the rank minimization problem with a single linear constraint in constant time.

An alternative formulation of the inverse problem of compressed sensing of a matrix is minimum rank approximation,

\[
P_2: \min_{X \in S_2} \|AX - b\|_2^p \quad \text{subject to} \quad \text{rank}(X) \leq r,
\]

where \( r = \text{rank}(X_0) \) denotes the minimum rank. The advantage of formulation P2 is that it can handle both the noiseless case and the noisy case in a single form. It also works for the more general case where \( X_0 \) is not exactly low-rank but admits an accurate approximation by a low-rank matrix. When the minimum rank \( r \) is unknown, an incremental search over \( r \) will increase the complexity of the solution by at most factor \( r \). If an upper bound on \( r \) is available, then a bisection search over \( r \) can be used because the minimum of P2 is monotone decreasing in \( r \). Hence, the factor reduces to \( \log r \). Indeed, this is not an issue in many applications where the rank is assumed to be a small constant.

Recently, several algorithms have been proposed to solve P2. Haldar and Hernando [95] proposed an alternating least squares approach by exploiting the explicit factorization of a rank-r matrix. Their algorithm is computationally efficient but does not provide any performance guarantee. Keshavan et al. [96] proposed an algorithm based on optimization over a Grassmann manifold. Their algorithm first finds a good starting point by an operation called trimming, and subsequently minimizes the objective of P2 using a line search and gradient descent over a Grassmann manifold. They provide a performance guarantee only for the matrix completion problem where the
linear operator $A$ takes a few entries from $X_0$. Moreover, the performance guarantee is restricted to the noiseless case.

For a fresh look at the problem, recall that minimum rank approximation, or rank-$r$ approximation for the matrix case, is analogous to $s$-term approximation for the vector case. Like rank-$r$ matrix approximation, $s$-term vector approximation is a way to find the sparsest solution of an ill-posed inverse problem in compressed sensing. For $s$-term approximation, besides efficient greedy heuristics such as Matching Pursuit (MP) [97] and Orthogonal Matching Pursuit (OMP) [98], there are recent algorithms, which are more efficient than convex relaxation and also have performance guarantees. These include Compressive Sampling Matching Pursuit (CoSaMP) [11] and Subspace Pursuit (SP) [12]. To date, no such algorithms have been available for the matrix case.

In Chapter 4, we propose an iterative algorithm for the rank minimization problem, which is a generalization of the CoSaMP algorithm to the matrix case. We call this algorithm “Atomic Decomposition for Minimum Rank Approximation,” abbreviated as ADMiRA. ADMiRA is computationally efficient in the sense that the core computation consists of least squares and truncated singular value decompositions, which are both basic linear algebra problems and admit efficient algorithms. Importantly, ADMiRA is the first algorithm that provides a performance guarantee for the minimum rank approximation problem. Furthermore, ADMiRA provides a strong performance guarantee for P2 that covers the general case where $X_0$ is only approximately low-rank and $b$ contains noise. The strong performance guarantee of ADMiRA is comparable to that of nuclear norm minimization in [82]. In the noiseless case, SVT [92] may be considered a competitor to ADMiRA. However, for the noisy case, SVT involves more than the simple singular value thresholding operation.

Matrix completion is a special case of low-rank matrix approximation from linear measurements where the linear operator takes a few random entries.

---

7There is another generalization of CoSaMP, namely model-based CoSaMP [99]. However, this generalization addresses a completely different and unrelated problem: sparse vector approximation subject to a special (e.g., tree) structure. Furthermore, the extensions of CoSaMP to model-based CoSaMP and to ADMiRA are independent: neither one follows from the other, and neither one is a special case of the other.

8ADMiRA [100] was followed by the algorithm by Keshavan et al. [96]. The short version [100] was presented at ISIT’09.
of the unknown matrix. It has received considerable attention owing to its important applications such as collaborative filtering. However, the linear operator in matrix completion does not satisfy the rank-restricted isometry property [101]. Therefore, at the present time, ADMiRA does not have a guarantee for matrix completion. Nonetheless, empirical performance on matrix completion is better than SVT (for the experiments in Chapter 4).
CHAPTER 2

OBLIQUE PURSUITS FOR COMPRESSED SENSING

2.1 Notation

Symbol \( \mathbb{N} \) is the set of natural numbers (excluding zero), and \([n]\) denotes the set \(\{1, \ldots, n\}\) for \(n \in \mathbb{N}\). Symbol \( \mathbb{K} \) denotes a scalar field, which is either the real field \(\mathbb{R}\) or the complex field \(\mathbb{C}\). The vector space of \(d\)-tuples over \(\mathbb{K}\) is denoted by \(\mathbb{K}^d\) for \(d \in \mathbb{N}\). Similarly, for \(m, n \in \mathbb{N}\), the vector space of \(m \times n\) matrices over \(\mathbb{K}\) is denoted by \(\mathbb{K}^{m \times n}\).

We will use various notations on a matrix \(A \in \mathbb{K}^{m \times n}\). The range space spanned by the columns of \(A\) will be denoted by \(\mathcal{R}(A)\). The adjoint operator of \(A\) will be denoted by \(A^*\). This notation is also used for the adjoint of a linear operator that is not necessarily a finite matrix. The \(j\)th column of \(A\) is denoted by \(a_j\) and the submatrix of \(A\) with columns indexed by \(J \subset [n]\) is denoted by \(A_J\). The \(k\)th row of \(A\) is denoted by \(a_k\), and the submatrix of \(A\) with rows indexed by \(K \subset [m]\) is denoted by \(A_K\). Symbol \(e_k\) will denote the \(k\)th standard basis vector of \(\mathbb{K}^d\), where \(d\) is implicitly determined for compatibility. The \(k\)th element of \(d\)-tuple \(x \in \mathbb{K}^d\) is denoted by \((x)_j\). The \(k\)th largest singular value of \(A\) will be denoted by \(\sigma_k(A)\). For Hermitian symmetric \(A\), \(\lambda_k(A)\) will denote the \(k\)th largest eigenvalue of \(A\). The Frobenius norm and the spectral norm of \(A\) are denoted by \(\|A\|_F\) and \(\|A\|\), respectively. The inner product is denoted by \(\langle \cdot, \cdot \rangle\). The embedding Hilbert space, where the inner product is defined, is not explicitly mentioned when it is obvious from the context. For a subspace \(\mathcal{S}\) of \(\mathbb{K}^d\), matrices \(P_\mathcal{S} \in \mathbb{K}^{d \times d}\) and \(P_\mathcal{S}^\perp \in \mathbb{K}^{d \times d}\) denote the orthogonal projectors onto \(\mathcal{S}\) and its orthogonal complement \(\mathcal{S}^\perp\), respectively. For \(J \subset [n]\), the coordinate
projection $\Pi_J : \mathbb{K}^n \to \mathbb{K}^n$ is defined by

$$
(\Pi_J x)_k = \begin{cases} 
(x)_k & \text{if } k \in J \\
0 & \text{else.}
\end{cases}
$$

(2.1.1)

Symbols $\mathbb{P}$ and $\mathbb{E}$ will denote the probability and the expectation with respect to a certain distribution. Unless otherwise mentioned, the distribution shall be obvious from the context.

2.2 Oblique Pursuit Algorithms

In this section, we propose modified greedy pursuit algorithms that use both $\Psi$ and $\tilde{\Psi}$, and show that they are guaranteed by the RBOP of $(\Psi, \tilde{\Psi})$ similarly to the way that the corresponding conventional pursuit algorithms are guaranteed by the RIP of $\Psi$. The modified greedy pursuit algorithms will be called oblique pursuit algorithms, because they involve oblique projections instead of the orthogonal projections in the conventional algorithms.

Recall that greedy pursuit algorithms seek an approximation of signal $f$ that is exactly sparse over dictionary $D$. Let $x^* \in \mathbb{K}^n$ be an $s$-sparse vector such that

$$
x^* = \arg \min_{x \in \mathbb{K}^n} \left\{ \| f - Dx \|_2 : \| x \|_0 \leq s \right\}.
$$

We assume that the approximation error $f - Dx^*$ is small compared to $\| f \|_2$.

The measurement vector $y \in \mathbb{K}^m$ is then given by

$$
y = A(Dx^*) + z
$$

where the distortion term $z$ includes both the approximation error $A(f - Dx^*)$ in modeling $f$ as an $s$-sparse signal over $D$, and additive noise $w$,

$$
z = A(f - Dx^*) + w.
$$

Let $D\hat{x}$ be an estimate of $f$ given by a greedy pursuit algorithm such that
\( \hat{x} \) is exactly \( s \)-sparse. Then,

\[
\| f - D\hat{x} \|_2 \leq \| f - Dx^* \|_2 + \| D(x^* - \hat{x}) \|_2 \\
\leq \| f - Dx^* \|_2 + \sqrt{1 + \delta_{2s}(D)} \| \hat{x} - x^* \|_2.
\]

Since the first term \( \| f - Dx^* \|_2 \) corresponds to a fundamental limit for any greedy algorithm, we will focus in the remainder of this section on bounding \( \| \hat{x} - x^* \|_2 \).

To describe both the original greedy pursuit algorithms and our modifications, we recall the definition of the hard thresholding operator that makes a given vector exactly \( s \)-sparse by zeroing the elements except the \( s \)-largest. Formally, \( H_s : \mathbb{K}^n \to \mathbb{K}^n \) is defined by

\[
H_s(x) \triangleq \arg \min_w \{ \| x - w \| : \| w \|_0 \leq s \}.
\]

**Remark 2.2.1.** All algorithms that appear in this section extend straightforwardly to the versions that exploit the structure of the support, a.k.a. recovery algorithms for model-based compressed sensing [102]. The only task required in this modification is to replace the hard thresholding operator by a projection onto \( s \)-sparse vectors with supports satisfying certain structure (e.g., tree). The extension to model-based CS explicitly depends on the support and is only available for the greedy algorithms. To focus on the main contribution of this paper, we will not pursue the details in this direction here.

### 2.2.1 Oblique Thresholding

We start with a modification of the simple thresholding algorithm. The thresholding algorithm computes an estimate of the support \( J \) as the indices of the \( s \) largest entries of \( \Psi^*y \), which is the support of \( H_s(\Psi^*y) \).

Let us consider a special case, where \( \Psi \) has full column rank and \( y = \Psi x^* \) is noise free. While exact support recovery by naive thresholding of \( \Psi^*y \) is not guaranteed, thresholding of \( \tilde{\Psi}^*y \) with the biorthogonal dual \( \tilde{\Psi} = (\Psi^\dagger)^* \) is guaranteed to provide exact support recovery. This example leaves room to improve thresholding using another properly designed matrix \( \tilde{\Psi} \). In compressed sensing, we are interested in an underdetermined system given by \( \Psi \); hence, \( \Psi \) cannot have full column rank. In this setting, the use of the
canonical dual $\tilde{\Psi} = (\Psi^\dagger)^*$ is not necessarily a good choice of $\tilde{\Psi}$.

Schnass and Vandergheynst [42] proposed a version of the thresholding algorithm that uses another matrix $\tilde{\Psi}$ different from $\Psi$. We call this algorithm Oblique Thresholding (ObThres), as an example of the oblique pursuit algorithms that will appear in the sequel.

**Algorithm 1: Oblique Thresholding (ObThres)**

$$\hat{J} \leftarrow \text{supp} \left( H_s(\tilde{\Psi}^*y) \right);$$

Schnass and Vandergheynst [42, Theorem 3] showed a sufficient condition for exact support recovery by ObThres in the noiseless case ($z = 0$), given by

$$\frac{\tilde{\mu}_1(s, \Psi, \tilde{\Psi})}{\min_j |\tilde{\psi}_j^*\psi_j|} < \min_{j \in J^*} \|x^*\|_j \frac{1}{2\|x^*\|_\infty}$$

(2.2.1)

where the cross Babel function $\tilde{\mu}_1(s, \Psi, \tilde{\Psi})$ is defined by

$$\tilde{\mu}_1(s, \Psi, \tilde{\Psi}) \triangleq \max_k \max_{|J| = s} \sum_{j \in J} |\tilde{\psi}_j^*\psi_j|.$$

Since the left-hand side of (2.2.1) is easily computed for given $\Psi$ and $\tilde{\Psi}$, Schnass and Vandergheynst [42] proposed a numerical algorithm that designs $\tilde{\Psi}$ to minimize the left-hand side of (2.2.1). However, the minimization problem is not convex and there is no guarantee for the quality of the resulting $\tilde{\Psi}$. Moreover, their optimality criterion for $\tilde{\Psi}$ is based on the sufficient condition in (2.2.1), which is conservative (see [42, Fig. 1]). In particular, unlike the RBOP, there is no known analysis of the (cross) Babel function of random frame matrices.

Instead, we derive an alternative sufficient condition for exact support recovery by ObThres, given in terms of the RBOP of $(\Psi, \tilde{\Psi})$.

**Theorem 2.2.2 (ObThres)**. Let $x^* \in \mathbb{K}^n$ be $s$-sparse with support $J^* \subset [n]$. Let $y = \Psi x^* + z$. Suppose that $\Psi$ and $\tilde{\Psi}$ satisfy

$$\min_{j \in J^*} |(x^*)_j| > 2\theta_{s+1}(\tilde{\Psi}^*\Psi)\|x^*\|_2 + 2\max_j \|\tilde{\psi}_j\|_2 \|z\|_2.$$  \hspace{1cm} (2.2.2)

Then, ObThres will identify $J^*$ exactly.
Compared to the numerical construction of \( \tilde{\Psi} \) by Schnass and Vandergheynst [42], our construction of \( \tilde{\Psi} \) in (1.1.6) for a random frame matrix \( \Psi \) has two advantages: it is analytic; and it guarantees the RBOP of \((\Psi, \tilde{\Psi})\). Therefore, with this construction, the computation of \( \theta_{s+1}(\tilde{\Psi}^*\Psi) \) for given \( \Psi \) and \( \tilde{\Psi} \), which involves a combinatorial search, is not needed.

For the noiseless case \((z = 0)\), the sufficient condition in (2.2.2) reduces to

\[
\theta_{s+1}(\tilde{\Psi}^*\Psi) < \frac{\min_{j \in J^*} \left| \left( x^* \right)_j \right|}{2\|x^*\|_2}.
\] (2.2.3)

Even in this case though, the upper bound in (2.2.3) depends on both the dynamic range of \( x^* \) and the sparsity level \( s \). Therefore, compared to the guarantees of the iterative greedy pursuit algorithms in Section 2.2.3, the guarantee of ObThres is rather weak. In fact, the other algorithms in Section 2.2.3 outperform ObThres empirically too. However, ObThres will serve as a building block for the iterative greedy pursuit algorithms.

### 2.2.2 Oblique Matching Pursuit

Matching Pursuit (MP) and Orthogonal Matching Pursuit (OMP) are forward greedy pursuit algorithms. Unlike thresholding, which selects the support elements by a single step of hard thresholding, (O)MP increments an estimate \( \hat{J} \) of the support \( J^* \) by adding one element per step chosen by a greedy criterion:

\[
k^* = \arg \max_k \left| \left( \Psi^* (y - \Psi \hat{x}) \right)_k \right|
\] (2.2.4)

where \( y - \Psi \hat{x} \) is the residual vector computed with the estimate \( \hat{x} \) of \( x^* \) spanned by \( \Psi_j \).

Given the estimated support \( \hat{J} \), OMP updates the estimate \( \hat{x} \) optimally in the sense that \( \hat{x} \) satisfies

\[
\hat{x} = \arg \min_x \{ \| y - \Psi x \|_2 : \text{supp} (x) \subset \hat{J} \}.
\] (2.2.5)

Therefore, the criterion in (2.2.4) for OMP reduces to

\[
k^* = \arg \max_k \left| \left( \Psi^* P_{\mathcal{R}(\Psi_j)} y \right)_k \right|
\]

\[
= \arg \max_k \left| \langle P_{\mathcal{R}(\Psi_j)}^\perp \psi_k, \quad P_{\mathcal{R}(\Psi_j)} y \rangle \right|,
\] (2.2.6)
which clearly describes the idea of “orthogonal matching”.

Schnass and Vandergheynst [42] proposed variations of MP and OMP that, using $\tilde{\Psi}$, replace (2.2.4) by

$$k^* = \arg \max_k |(\tilde{\Psi}^* (y - \Psi \tilde{x}))_k|$$

(2.2.7)

and provided the following sufficient condition [42, Theorem 4] for exact support recovery by the OMP using (2.2.7)

$$\frac{\tilde{\mu}_1(s, \Psi, \tilde{\Psi})}{\min_j |\tilde{\psi}_j^* \psi_j|} < \frac{1}{2},$$

(2.2.8)

As for ObThres, they proposed to use a numerically designed $\tilde{\Psi}$ that minimizes the left-hand side of (2.2.8) (the same criterion as in their analysis of ObThres).

As discussed in the previous subsection, while easily computable for given $\Psi$ and $\tilde{\Psi}$, this sufficient condition is conservative and is not likely to be satisfied even when $\tilde{\Psi}$ is numerically optimized. Thus, the resulting algorithm will have no guarantee. Another weakness of the sufficient condition in (2.2.8) is that it has been derived without considering the orthogonal matching in OMP, and thus ignores the improvement of OMP over MP. Indeed, the same condition provides a partial guarantee of MP that each step of MP will select an element of the support $J^*$, which is not necessarily different from the previously selected ones.

In view of the weaknesses of the approach based on coherence, we turn instead to the RIP. Davies and Wakin [76] provided a sufficient condition for exact support recovery by OMP in terms of the RIP, which has been refined in the setting of joint sparsity by Lee et al. [79, Proposition 7.11]. These analyses explicitly reflect the “orthogonal matching”. In particular, one key property required for the RIP-based sufficient conditions is that the RIP is preserved under the orthogonal projection with respect to a few columns of $\Psi$, i.e., for all $\tilde{J} \subset [n]$ satisfying $|\tilde{J}| < s$,

$$\delta_s(P_{\tilde{R}(\Psi, \tilde{\Psi})}^{\perp} \Psi_{[n] \setminus \tilde{J}}) \leq \delta_s(\Psi).$$

(2.2.9)

This condition is an improvement on [76, Lemma 3.2] and was shown [79, Proof of Proposition 7.11] using the interlacing eigenvalues property of the
Schur complement [79, Lemma A.2].

The objective function in the orthogonal matching in (2.2.6) can be rewritten as

\[
|\langle P_{\mathcal{R}(\Psi_j)}^\perp \psi_k, P_{\mathcal{R}(\Psi_j)}^\perp y \rangle | = \left| \sum_{j \in J \setminus \hat{J}} \langle P_{\mathcal{R}(\Psi_j)}^\perp \psi_k, P_{\mathcal{R}(\Psi_j)}^\perp \psi_j \rangle (x^*)_j + \langle P_{\mathcal{R}(\Psi_j)}^\perp \psi_k, z \rangle \right|. \tag{2.2.10}
\]

The RIP of \(\Psi\) together with (2.2.9) imply that the left-hand side of (2.2.10) is close to \(|(\Pi_{J \setminus \hat{J}} x^*)_k|\), with the perturbation bounded as a function the RIC of \(\Psi\). Then, orthogonal matching will choose \(k^*\) as

\[
k^* = \arg \max_{k \in J \setminus \hat{J}} |(x^*)_k|.
\]

This explains why orthogonal matching is a good strategy when \(\Psi\) satisfies the RIP.

The OMP using (2.2.7) by Schnass and Vanderheynst [42] still employs the orthogonal matching. However, we are interested in the scenario where \(\Psi\) does not satisfy the RIP but instead satisfies the RBOP with a certain \(\tilde{\Psi}\). Unfortunately, unlike the RIP of \(\Psi\), the RBOP of \((\Psi, \tilde{\Psi})\) is no longer valid when the orthogonal projection \(P_{\mathcal{R}(\Psi_j)}^\perp\) is applied to both matrices. Instead, we show that the RBOP of \((\Psi, \tilde{\Psi})\) is preserved under an oblique projection, which is analogous to the RIP result in (2.2.9). To this end, we recall the definition of an oblique projection.

**Definition 2.2.3 (Oblique projection).** Let \(\mathcal{V}, \mathcal{W} \subset \mathcal{H}\) be two subspaces such that \(\mathcal{V} \oplus \mathcal{W}^\perp = \mathcal{H}\). The oblique projection onto \(\mathcal{V}\) along \(\mathcal{W}^\perp\), denoted by \(E_{\mathcal{V},\mathcal{W}^\perp}\), is defined as a linear map \(E_{\mathcal{V},\mathcal{W}^\perp}: \mathcal{H} \to \mathcal{H}\) that satisfies

1. \((E_{\mathcal{V},\mathcal{W}^\perp})x = x, \ \forall x \in \mathcal{V}\).
2. \((E_{\mathcal{V},\mathcal{W}^\perp})x = 0, \ \forall x \in \mathcal{W}^\perp\).

By the definition of the oblique projection, it follows that

\[
I_{\mathcal{H}} - E_{\mathcal{V},\mathcal{W}^\perp} = E_{\mathcal{W}^\perp,\mathcal{V}} \text{ and } E_{\mathcal{V},\mathcal{W}^\perp}^* = E_{\mathcal{W},\mathcal{V}^\perp}.
\]

When \(\mathcal{V} = \mathcal{W}\), the oblique projection reduces to the orthogonal projection \(P_\mathcal{V}\) onto \(\mathcal{V}\).
Lemma 2.2.4. Suppose that $M, \tilde{M} \in \mathbb{K}^{m \times k}$ for $k \leq m$ satisfy that $\tilde{M}^*M$ has full rank. Then, $\mathcal{R}(M)$ and $\mathcal{R}(\tilde{M})^\perp$ are complementary, i.e., $\mathcal{R}(M) \cap \mathcal{R}(\tilde{M})^\perp = \{0\}$.

Proof of Lemma 2.2.4. Assume that there is a nonzero $x \in \mathcal{R}(M) \cap \mathcal{R}(\tilde{M})^\perp$. Then, $x = My$ for some $y \in \mathbb{K}^k$ and $\tilde{M}^*My = 0$ since $x \in \mathcal{R}(\tilde{M})^\perp = \mathcal{N}(\tilde{M}^*)$. Since $\tilde{M}^*M$ is invertible, it follows that $y = 0$, which is a contradiction. \qed

The RBOP of $(\Psi, \tilde{\Psi})$ implies that $\tilde{\Psi}^*\tilde{\Psi} \tilde{j}$ is invertible. Furthermore, $\mathcal{R}(\tilde{\Psi} \tilde{j})$ and $\mathcal{R}(\tilde{\Psi} \tilde{j})$ are complementary by Lemma 2.2.4. Therefore, $\Psi \tilde{j}(\tilde{\Psi}^*\tilde{\Psi} \tilde{j})^{-1}\tilde{\Psi}^* \tilde{j}$ is an oblique projection onto $\mathcal{R}(\tilde{\Psi} \tilde{j})$ along $\mathcal{R}(\tilde{\Psi} \tilde{j})^\perp$. It follows that $E = I_{[\tilde{j}]} - \Psi \tilde{j}(\tilde{\Psi}^*\tilde{\Psi} \tilde{j})^{-1}\tilde{\Psi}^* \tilde{j}$ is an oblique projection onto $\mathcal{R}(\tilde{\Psi} \tilde{j})^\perp$ along $\mathcal{R}(\tilde{\Psi} \tilde{j})$.

Lemma 2.2.5. Suppose that $\Psi, \tilde{\Psi} \in \mathbb{K}^{m \times n}$ satisfy

$$\theta_s(\tilde{\Psi}^*\Psi) < 1.$$ 

Let $\tilde{j} \subset [n]$. Let $E = I_{[\tilde{j}]} - \Psi \tilde{j}(\tilde{\Psi}^*\tilde{\Psi} \tilde{j})^{-1}\tilde{\Psi}^* \tilde{j}$. Then,

$$\theta_s(\tilde{\Psi}^*E_{[n]\setminus\tilde{j}}) \leq \theta_s(\tilde{\Psi}^*\Psi).$$

Remark 2.2.6. When $\tilde{\Psi} = \Psi$, Lemma 2.2.5 reduces to (2.2.9).

Proof of Lemma 2.2.5. Follows directly from Lemma A.1.6 in the Appendix. \qed

Lemma 2.2.5 suggests that if $\Psi$ does not satisfy the RIP but $\Psi$ and $\tilde{\Psi}$ satisfy the RBOP, then it might better to replace the orthogonal matching by the “oblique matching” given by

$$k^* = \arg \max_k \left| \langle E^*\tilde{\psi}_k, Ey \rangle \right|,$$ (2.2.11)

where $E$ is an oblique projector defined as

$$E = I_{[\tilde{j}]} - \Psi \tilde{j}(\tilde{\Psi}^*\tilde{\Psi} \tilde{j})^{-1}\tilde{\Psi}^* \tilde{j}.$$ 

To affect the appropriate modification in OMP, recall that orthogonal matching in (2.2.6) corresponds to matching each column of $\tilde{\Psi}$ with the
residual $y - \Psi \hat{x}$ computed with a solution $\hat{x}$ to the least square problem in (2.2.5). Similarly, oblique matching is obtained by replacing the least square problem in (2.2.5) by the following weighted least square problem:

$$\hat{x} = \arg \min_x \{ \| \tilde{\Psi}^*_j (y - \Psi x) \|_2 : \text{supp} (x) \subset \tilde{J} \}.$$  

We call the resulting forward greedy pursuit algorithm with the oblique matching pursuit (ObMP). ObMP is summarized in Algorithm 2. In particular, when $\tilde{\Psi} = \Psi$, ObMP reduces to the conventional OMP. Like OMP, ObMP does not select the same support element more than once. This is guaranteed since the selected columns are within the null space of the oblique projection associated with the oblique matching.

**Algorithm 2: Oblique Matching Pursuit (ObMP)**

\[
\begin{align*}
\tilde{J} &\leftarrow \emptyset; \hat{x} \leftarrow 0; \\
\text{while } |\tilde{J}| < s &\text{ do} \\
&\quad k^* \leftarrow \arg \max_{k \in [n] \setminus \tilde{J}} | (\tilde{\Psi}^* (y - \Psi \hat{x}))_k |; \\
&\quad \tilde{J} \leftarrow \tilde{J} \cup \{ k^* \}; \\
&\quad \hat{x} \leftarrow \arg \min_x \{ \| \tilde{\Psi}^*_j (y - \Psi x) \|_2 : \text{supp} (x) \subset \tilde{J} \}; \\
\text{end}
\end{align*}
\]

Next, we present a guarantee of ObMP in terms of the RBOP.

**Proposition 2.2.7** (A Single Step of ObMP). Let $x^* \in K^n$ be $s$-sparse with support $J^* \subset [n]$. Let $y = \Psi x^* + z$ and $J \subsetneq J^*$. Suppose that $\Psi$ and $\tilde{\Psi}$ satisfy

$$\| \Pi_{J^* \setminus J} x^* \|_\infty - 2\theta_{s+1}(\tilde{\Psi}^* \Psi) \| \Pi_{J^* \setminus J} x^* \|_2 > \left( \frac{\| \Psi_{J^*} \| \| \tilde{\Psi}_{J^*} \|_2}{1 - \theta_{s+1}(\tilde{\Psi}^* \Psi)} \right) \frac{2 \max_j \| \tilde{\psi}_j \|_2 }{\| z \|_2}. $$

(2.2.12)

where the coordinate projection $\Pi_{J^* \setminus J}$ is defined in (2.1.1). Then, the next step of ObMP given $J$ will identify an element of $J^* \setminus J$.

The following theorem is a direct consequence of Proposition 2.2.7.

**Theorem 2.2.8** (ObMP). Let $x^* \in K^n$ be $s$-sparse with support $J^* \subset [n]$. 

32
Let \( y = \Psi^* \star x + z \). Suppose that \( \Psi \) and \( \tilde{\Psi} \) satisfy
\[
\min_{j \in J^*} |(x^*)_j| \left( \min_{J \subset J^*, J \neq \emptyset} \frac{\|\Pi_J x^*\|_\infty}{\|\Pi_J x^*\|_2} - 2\theta_{s+1}(\tilde{\Psi}^* \Psi) \right) > \left( \frac{\|\Psi^* J^*\|_1}{1 - \theta_{s+1}(\tilde{\Psi}^* \Psi)} \right) 2 \max_j \|\tilde{\psi}_j\|_2 \|z\|_2.
\] (2.2.13)

Then, ObMP will identify \( J^* \) exactly.

If \( \tilde{\Psi} = \Psi \), then ObMP reduces to OMP; hence, Proposition 2.2.7 reduces to the single measurement vector case of [79, Proposition 7.11], with the requirement on \( \Psi \) in (2.2.12) reduced to
\[
\|\Pi_{J^* \setminus J^*} x^*\|_\infty - 2\delta_{s+1}(\Psi)\|\Pi_{J^* \setminus J^*} x^*\|_2 > 2 \max_j \|\tilde{\psi}_j\|_2 \|z\|_2.
\] (2.2.14)

In fact, the proof of Proposition 2.2.7 in the Appendix is carried out by modifying that of [79, Proposition 7.11] so that the non-Hermitian case is appropriately managed. Similarly, the guarantee of ObMP in Theorem 2.2.8 reduces to that of OMP given by
\[
\min_{j \in J^*} |(x^*)_j| \left( \min_{J \subset J^*, J \neq \emptyset} \frac{\|\Pi_J x^*\|_\infty}{\|\Pi_J x^*\|_2} - 2\delta_{s+1}(\Psi) \right) > 2 \max_j \|\tilde{\psi}_j\|_2 \|z\|_2.
\] (2.2.15)

To satisfy the condition in (2.2.15), it is required that \( \delta_{s+1}(\Psi) < c \) for some \( c \in (0, 1) \) that depends on \( x^* \). As will be shown in Section 2.3, this RIP condition is often not satisfied in a typical scenario of practical applications. In contrast, \( \theta_{s+1}(\tilde{\Psi}^* \Psi) \) is still satisfied with a properly designed \( \tilde{\Psi} \) in the same scenario. Therefore, the guarantee of ObMP in Theorem 2.2.8 is less demanding than the corresponding guarantee of OMP.

We observe that the bound on the noise amplification in ObMP is larger by the factor \( \frac{\|\Psi^* J^*\|_1}{1 - \delta_{s+1}(\Psi, \Psi)} \) than in OMP. This factor is an upper bound on the spectral norm of the oblique projection onto \( \mathcal{R}(\Psi_J) \) along \( \mathcal{R}(\tilde{\Psi}_J)^\perp \). The analogous operator in OMP is an orthogonal projector and the spectral norm is trivially bounded from above by 1. However, when oblique matching is used instead of orthogonal matching, this is no longer valid. The spectral norm of the oblique projection is the reciprocal of the cosine of the angle between the two subspaces \( \mathcal{R}(\Psi_J) \) and \( \mathcal{R}(\tilde{\Psi}_J) \). This result is consistent with
the known analysis of oblique projections.\footnote{In a general context, unrelated to CS, it has been shown \cite{103} that oblique projectors are suboptimal in terms of minimizing the projection residual, which is however bounded within factor $\frac{1}{\cos \theta}$ of the optimal error.}

For the noiseless case ($z = 0$), the sufficient condition in (2.2.13) reduces to

$$\theta_{s+1}(\tilde{\Psi}^*\Psi) < \min_{J \subset J^*, J \neq \emptyset} \frac{\|\Pi_J x^*\|_\infty}{\frac{1}{2} \|\Pi_J x^*\|_2}. \quad (2.2.16)$$

Compared to the sufficient condition for ObThres in (2.2.3), where depending on the dynamic range of $x^*$, the upper bound on the RBOC can be arbitrary small, the right-hand side in (2.2.16) is no smaller than $\frac{1}{2} \sqrt{\frac{1}{s\pi}}$ for any $x^*$. Although ObMP is guaranteed under a milder RBOP condition than ObThres, the corresponding sufficient condition is still demanding compared to those of iterative greedy pursuit algorithms.

However, ObThres and ObMP are important, since they provide basic building blocks for the iterative greedy pursuit algorithms. The thresholding and OMP algorithms have been modified to ObThres and ObMP by replacing two basic blocks, “$\Psi^*$ followed by hardthresholding”, and “orthogonal matching”, to “$\Psi^*$ followed by hardthresholding”, and “oblique matching”, respectively. The modifications of these two basic blocks will similarly alter the other greedy pursuit algorithms and their RIP-based guarantees.

In the next section, we present the oblique versions of some iterative greedy pursuit algorithms (CoSaMP, SP, IHT, and HTP). However, the conversion to the oblique version of both algorithm and guarantee is not restricted to these examples. It applies to any other greedy pursuit algorithm that builds on these basic blocks (e.g., Fast Nesterov’s Iterative Hard Thresholding (FNIHT) \cite{104}).

\section*{2.2.3 Iterative Oblique Greedy Pursuit Algorithms}

Compressive Sampling Matching Pursuit (CoSaMP) \cite{11} and Subspace Pursuit (SP) \cite{12} are more sophisticated greedy pursuit algorithms that iteratively update the $s$-sparse estimate of $x^*$. At a high level, both CoSaMP and SP update the estimate of the true support using the following procedure:

\begin{enumerate}
\item Augment the estimated set by adding more indices that might include the missing elements of the true support.
\end{enumerate}
2. Refine the augmented set to a subset with \( s \) elements.

The two algorithms differ in the size of the increment in the augmentation. More important, SP completes each iteration by updating the residual using an orthogonal projection, which is similar to that of OMP. CoSaMP and SP provide RIP-based guarantees, which are comparable to those of \( \ell_1 \)-based solutions such as BP.

Both algorithms use the basic building blocks of correlation maximization by hard thresholding and least squares problems. Therefore, following the same approach we used to modify thresholding and OMP to ObThres and ObMP, we modify CoSaMP and SP to their oblique versions called Oblique CoSaMP (ObCoSaMP) and Oblique SP (ObSP), respectively. ObCoSaMP and ObSP are summarized in Algorithm 3 and Algorithm 4.

**Algorithm 3: Oblique Compressive Matching Pursuit (ObCoSaMP)**

```
while stop condition not satisfied do
  \( \tilde{J}_{t+1} \leftarrow \text{supp}(x_t) \cup \text{supp} \left( H_{2s}(\tilde{\Psi}^*(y - \Psi x_t)) \right) \);
  \( \tilde{x} \leftarrow \arg \min_x \left\{ \| \tilde{\Psi}^*_{\tilde{J}_{t+1}} (y - \Psi x) \|_2 : \text{supp}(x) \subset \tilde{J}_{t+1} \right\} \);
  \( x_{t+1} \leftarrow H_s(\tilde{x}) \);
  \( t \leftarrow t + 1 \);
end
```

**Algorithm 4: Oblique Subspace Pursuit (ObSP)**

```
while stop condition not satisfied do
  \( \tilde{J}_{t+1} \leftarrow \text{supp}(x_t) \cup \text{supp} \left( H_{s}(\tilde{\Psi}^*(y - \Psi x_t)) \right) \);
  \( \tilde{x} \leftarrow \arg \min_x \left\{ \| \tilde{\Psi}^*_{\tilde{J}_{t+1}} (y - \Psi x) \|_2 : \text{supp}(x) \subset \tilde{J}_{t+1} \right\} \);
  \( J_{t+1} \leftarrow \text{supp}(H_s(\tilde{x})) \);
  \( x_{t+1} \leftarrow \arg \min_x \left\{ \| \tilde{\Psi}^*_{J_{t+1}} (y - \Psi x) \|_2 : \text{supp}(x) \subset J_{t+1} \right\} \);
  \( t \leftarrow t + 1 \);
end
```
Iterative Hard Thresholding (IHT) [13] and Hard Threshold Pursuit (HTP) [14] are two other greedy pursuit algorithms with RIP-based guarantees. HTP is a modified version of IHT, which updates the residual using orthogonal projection like SP. Since both IHT and HTP use the same basic building blocks used in the other greedy pursuit algorithms, they too admit the oblique versions. We name these modified versions **Oblique IHT** (ObIHT) and **Oblique HTP** (ObHTP). ObIHT and ObHTP are summarized in Algorithm 5 and Algorithm 6. Note that these iterative oblique greedy pursuit algorithms reduce to their conventional counterparts when $\tilde{\Psi} = \Psi$.

Algorithm 5: Oblique Iterative Hard Thresholding (ObIHT)

```plaintext
while stop condition not satisfied do
    $x_{t+1} \leftarrow H_s(x_t + \tilde{\Psi}^*(y - \Psi x_t))$;
    $t \leftarrow t + 1$;
end
```

Algorithm 6: Oblique Hard Thresholding Pursuit (ObHTP)

```plaintext
while stop condition not satisfied do
    $J_{t+1} \leftarrow \text{supp} \left( H_s(x_t + \tilde{\Psi}^*(y - \Psi x_t)) \right)$;
    $x_{t+1} \leftarrow \arg \min_x \left\{ \| \tilde{\Psi}^*_{J_{t+1}}(y - \Psi x) \|_2 : \text{supp} (x) \subset J_{t+1} \right\}$;
    $t \leftarrow t + 1$;
end
```

We briefly review the currently available RIP-based guarantees of the original algorithms. The guarantees of the iterative greedy pursuit algorithms were provided in their original papers [11–14]. In particular, Needell and Tropp, in their technical report on CoSaMP [105], showed that CoSaMP (with exact arithmetic) converges within a finite number of iterations, which is at most $O(s)$ for the worst case and can be as small as $O(\ln s)$. We will show that the same analysis applies to SP, HTP, and their oblique versions. The guarantees of the iterative greedy pursuit algorithms are provided by sufficient conditions given in a common form $\delta_{ks}(\Psi) < c$, where the condition

\[ \delta_{ks}(\Psi) = \max_{\Phi \in \mathcal{E}_s} \frac{\| \Phi \|_{\infty}}{\min_{\phi \in \Phi} \langle \Phi, \Psi \rangle} \]

with $\mathcal{E}_s$ the set of $s$-sparse vectors.
becomes more demanding for larger $k$ and smaller $c$. Recently, Foucart [106] refined the guarantees of CoSaMP and IHT by increasing required $c$. We will show that the guarantee of SP is similarly improved using similar techniques and replacing triangle inequalities by the Pythagorean theorem when applicable.\footnote{As an aside, inspired by the existing RIP analysis that $\delta_{ks}(\Psi) < c$ holds with $m = O(kc^{-2}\ln^4 n)$, Foucart [106] proposed to compare sufficient conditions by comparing the values of $kc^{-2}$. Nevertheless, this comparison is heuristic and only relies on sufficient conditions for the worst case guarantee. Therefore, it is not necessarily true that an algorithm with smaller $kc^{-2}$ performs better.}

Next, we show that the RIP-based guarantees of the iterative greedy pursuit algorithms are replaced by similar guarantees of the corresponding oblique pursuit greedy algorithms, in terms of the RBOP. In fact, the modification of the guarantees is rather straightforward, as was the modification of the algorithms. We only provide the full derivation for the RBOP-based guarantee of ObSP. Replacing $\Psi$ by $\Psi$ in the result and the derivation will provide an RIP-based guarantee for SP. The guarantees of the other iterative oblique pursuit algorithms (ObCoSaMP, ObIHT, and ObHTP) are obtained by similarly modifying the corresponding results [14, 106]. Therefore, we do not repeat the derivations but only state the results.

**Theorem 2.2.9.** Let $\text{Alg} \in \{\text{ObSP, ObCoSaMP, ObIHT, ObHTP}\}$. Let $(x_t)_{t \in \mathbb{N}}$ be the sequence generated by algorithm $\text{Alg}$. Then

$$\|x_{t+1} - x^*\|_2 \leq \rho\|x_t - x^*\|_2 + \tau\|z\|_2$$

(2.2.17)

where $\rho$ and $\tau$ are positive constants depending on $\text{Alg}$, given as explicit functions of $\theta_{ks}(\Psi^*\Psi)$, $\delta_{ks}(\Psi)$, and $\delta_{ks}(\Psi)$. Moreover, $\rho$, which only depends on $\theta_{ks}(\Psi^*\Psi)$, is less than 1, provided that the condition in Table 2.1 specified by $\text{Alg}$ is satisfied.

**Proof of Theorem 2.2.9.** We only provide the proof for ObSP in Appendix A.4. The formulae for $\rho$ and $\tau$ are provided for all listed algorithms. \qed
Table 2.1: The RBOP condition required for linear convergence in Theorem 2.2.9.

<table>
<thead>
<tr>
<th>Alg</th>
<th>ObCoSaMP</th>
<th>ObSP</th>
<th>ObIHT</th>
<th>ObHTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{ks}(\Psi^*\Psi) &lt; c$</td>
<td>$\theta_{4s}(\Psi^*\Psi) &lt; 0.384$</td>
<td>$\theta_{3s}(\Psi^*\Psi) &lt; 0.325$</td>
<td>$\theta_{3s}(\Psi^*\Psi) &lt; 0.5$</td>
<td>$\theta_{3s}(\Psi^*\Psi) &lt; 0.577$</td>
</tr>
</tbody>
</table>
For $\rho < 1$, (2.2.17) implies that in the noiseless case the iteration converges linearly at rate $\rho$ to the true solution, whereas in the noisy case the error at convergence is $\|x_\infty - x^*\|_2 = \tau/(1 - \rho)\|z\|_2$.

Unlike ObIHT, the other algorithms (ObCoSaMP, ObSP, and ObHTP) involve the step of updating the estimate by solving a least squares problem. This additional step provides the property in Lemma 2.2.10, which enables the finite convergence of the algorithms.

**Lemma 2.2.10.** Let $\text{Alg} \in \{\text{ObSP}, \text{ObCoSaMP}, \text{ObHTP}\}$. Let $(x_t)_{t \in \mathbb{N}}$ be the sequence generated by $\text{Alg}$. Then, the approximation error $\|x_t - x^*\|_2$ is less than the $\ell_2$ norm of the missed components of $x^*$ to within a constant factor $\bar{\rho}$ plus the noise term, i.e.,

$$
\|x_{t+1} - x^*\|_2 \leq \bar{\rho}\|\Pi_{\mathcal{J}_{t+1}} x^*\|_2 + \bar{\tau}\|z\|_2
$$

(2.2.18)

where $\bar{\rho}$ and $\bar{\tau}$ are positive constants given as explicit functions (depending on $\text{Alg}$) of $\theta_{ks}(\tilde{\Psi}^* \Psi)$, $\delta_{ks}(\Psi)$, and $\delta_{ks}(\tilde{\Psi})$.

**Proof of Lemma 2.2.10.** Lemma 2.2.10 is an intermediate step for proving Theorem 2.2.9. For example, for ObSP, it corresponds to Lemma A.4.1 in Appendix A.4. For the other algorithms, we only provide the formulae for $\bar{\rho}$ and $\bar{\tau}$ in Appendix A.4. 

Needell and Tropp [105] showed finite convergence of CoSaMP. The same analysis also applies to ObCoSaMP, ObSP, and ObHTP. To show this, let us recall the relevant definitions from the technical report on CoSaMP [105]. The **component bands** $(\mathcal{B}_j)$ of $x^*$ are by

$$
\mathcal{B}_j \triangleq \{i : 2^{-(j+1)}\|x^*\|_2^2 < |(x^*)_i| \leq 2^{-j}\|x^*\|_2^2\}, \quad \forall j \in \mathbb{Z} \cup \{0\}.
$$

Then, the **profile** of $x^*$ is defined as the number of nonempty component bands. By definition, the profile of $x^*$ is not greater than the sparsity level of $x^*$.

**Lemma 2.2.11** (A Paraphrase of [105, Theorem B.1]). Let $p$ be the profile of $x^*$. Suppose that $(x_t)_{t \in \mathbb{N}}$ satisfies eqs. (2.2.17) and (2.2.18). Then, for

$$
t > L + p \ln \left(1 + 2 \left[\frac{\bar{\rho} + \frac{\tau}{\bar{\tau}}(1 - \rho - \eta)}{\bar{\tau}}\right]^{\frac{s}{p}} \ln \left(\frac{1}{1 - \eta}\right)\right)^{-1}, \quad (2.2.19)
$$

39
it holds that
\[ \| x_t - x^* \|_2 \leq \left[ \rho^{L} \left( \frac{\tau \bar{\rho}}{1 - \rho - \eta} + \bar{\tau} \right) + \left( \frac{1 - \rho^{L}}{1 - \rho} \right) \tau \right] \| z \|_2. \]

The minimal number of iterations for the convergence (the right-hand side of (2.2.19)) is maximized when \( p = s \) [105]. The following theorem is a direct consequence of Theorem 2.2.9, Lemma 2.2.10, and Lemma 2.2.11.

**Theorem 2.2.12.** Let \( \text{Alg} \in \{ \text{ObSP}, \text{ObCoSaMP}, \text{ObHTP} \} \). Suppose that \( \theta_{ks}(\bar{\Psi}^* \Psi) < c \) holds depending on \( \text{Alg} \) as in Table 2.1. After \( t_{\max} = C_1(s + 1) \) iterations, \( \text{Alg} \) provides an estimate \( \hat{x} \) satisfying \( \| \hat{x} - x^* \|_2 \leq C_2 \| z \|_2 \). Here, \( k, c, C_1, \) and \( C_2 \) are constants, specified by \( \text{Alg} \).

The fast convergence of iterative greedy pursuit algorithms that involve the least square steps is important. When the problem is large (e.g., in CS imaging, the image size is typically \( 512 \times 512 \) pixels), solving the least squares problems is the most computationally demanding step of the recovery algorithms. Empirically, as the theory suggests, the iterative algorithms (ObCoSaMP, ObSP, and ObHTP) converge at most within \( O(s) \) iterations, and are even more computationally efficient than the non-iterative ObMP.

**Remark 2.2.13.** The extension of greedy pursuit algorithms and their RIP-based guarantees to those based on the RBOP is not restricted to the aforementioned algorithms. For example, Fast Nesterov’s Iterative Hard Thresholding (FNIHT) [104] is another promising algorithm with an RIP-based guarantee, which will extend likewise.

### 2.3 Restricted Biorthogonality Property

In this section, we show that the RBOP-based guarantees of oblique pursuits apply to realistic models of compressed sensing systems in practice. For example, when applied to random frame matrices, the guarantees remain valid even though the i.i.d. sampling is done according to a nonuniform distribution. Recall that the guarantees of oblique pursuits in Section 2.2 required \( \theta_{ks}(\bar{\Psi}^* \Psi) < c \) where \( k \in \{ 2, 3, 4 \} \) and \( c \in (0, 1) \) are constants specified by the
algorithm in question. The noise amplification in the reconstruction for these guarantees also depend on $\delta_{ks}(\Psi)$ and $\delta_{ks}(\Psi)$. However, unlike $\theta_{ks}(\tilde{\Psi}^\star \Psi)$, the RICs $\delta_{ks}(\Psi)$ and $\delta_{ks}(\Psi)$ need not be less than 1 to provide the guarantees. In fact, as discussed later, reasonable upper bounds on $\delta_{ks}(\Psi)$ and $\delta_{ks}(\Psi)$ (possibly larger than 1) are obtained with no additional conditions whenever $\theta_{ks}(\tilde{\Psi}^\star \Psi) < c$ is achieved. Therefore, we may focus on the condition $\theta_{ks}(\tilde{\Psi}^\star \Psi) < c$. Also recall that the guarantees for the corresponding conventional pursuit algorithms require $\delta_{ks}(\Psi) < c$, for $k \in \{2, 3, 4\}$, $c \in (0, 1)$, with the same $k$ and $c$ as the corresponding oblique pursuits. To compare the guarantees of the oblique vs. the conventional pursuit algorithms, assuming $k \in \{2, 3, 4\}$ and $c \in (0, 1)$ arbitrarily fixed constants, we compare the difficulty in achieving the respective bounds on $\delta_{ks}(\Psi)$ and $\theta_{ks}(\tilde{\Psi}^\star \Psi)$. While both properties are guaranteed when $m = O(s \ln n)$, $\theta_{ks}(\tilde{\Psi}^\star \Psi) < c$ is achieved without additional conditions required for achieving $\delta_{ks}(\Psi) < c$, which are often violated in practical compressed sensing.

2.3.1 General Estimate

We extend [20, Theorem 8.4] to the following theorem, so that it provides an upper bound on $\theta_s(\tilde{\Psi}^\star \Psi)$.

Theorem 2.3.1. Let $\Psi, \tilde{\Psi} \in \mathbb{K}^{m \times n}$ be random matrices not necessarily mutually independent, each with i.i.d. rows with elements bounded in magnitude as

$$\max_{k,\ell} |(\Psi)_{k,\ell}| \leq \frac{K}{\sqrt{m}} \quad \text{and} \quad \max_{k,\ell} |(\tilde{\Psi})_{k,\ell}| \leq \frac{\tilde{K}}{\sqrt{m}} \quad (2.3.1)$$

for $K, \tilde{K} \geq 1$. Then, $\theta_s(\tilde{\Psi}^\star \Psi) < \delta + \theta_s(\mathbb{E}\tilde{\Psi}^\star \Psi)$ holds with probability $1 - \eta$ provided that

$$m \geq C_1 \delta^{-2} \left( K \sqrt{2 + \theta_s(\mathbb{E}\Psi^\star \Psi)} + \tilde{K} \sqrt{2 + \theta_s(\mathbb{E}\tilde{\Psi}^\star \tilde{\Psi})} \right)^2 s (\ln s)^2 \ln n \ln m, \quad (2.3.2)$$

$$m \geq C_2 \delta^{-2} \tilde{K} \max(K, \tilde{K}) s \ln(\eta^{-1}) \quad (2.3.3)$$

for universal constants $C_1$ and $C_2$.

Proof of Theorem 2.3.1. See Appendix A.5.
Letting $\tilde{\Psi} = \Psi$ in Theorem 2.3.1 provides the following corollary.$^3$

**Corollary 2.3.2.** Let $\Psi \in \mathbb{K}^{m \times n}$ be a random matrix with i.i.d. rows with elements bounded in magnitude as $\max_{k,\ell}|(\Psi)_{k,\ell}| \leq \frac{K}{\sqrt{m}}$ for $K \geq 1$. Then, $\delta_s(\Psi) < \delta + \theta_s(\mathbb{E}\Psi^*\Psi)$ holds with probability $1 - \eta$ provided that

$$m \geq C_1\delta^{-2}K^24[2 + \theta_s(\mathbb{E}\Psi^*\Psi)]s(\ln s)^2 \ln n \ln m,$$

$$m \geq C_2\delta^{-2}K^2s\ln(\eta^{-1})$$

for universal constants $C_1$ and $C_2$.

The following corollary is obtained by combining Theorem 2.3.1 and Corollary 2.3.2 applied to $\Psi$ and to $\tilde{\Psi}$, respectively. Corollary 2.3.3 aims to provide an upper bound on $\theta_s(\tilde{\Psi}^*\tilde{\Psi})$. It also provides upper bounds on both $\delta_s(\Psi)$ and $\delta_s(\tilde{\Psi})$.

**Corollary 2.3.3.** Let $\Psi, \tilde{\Psi} \in \mathbb{K}^{m \times n}$ be random matrices with i.i.d. rows with elements bounded in magnitude as $\max_{k,\ell}|(\Psi)_{k,\ell}| \leq \frac{K}{\sqrt{m}}$ and $\max_{k,\ell}|(\tilde{\Psi})_{k,\ell}| \leq \frac{\tilde{K}}{\sqrt{m}}$ for $K, \tilde{K} \geq 1$. Then, $\theta_s(\tilde{\Psi}^*\tilde{\Psi}) < \delta + \theta_s(\mathbb{E}\tilde{\Psi}^*\tilde{\Psi})$, $\delta_s(\Psi) < \delta + \theta_s(\mathbb{E}\Psi^*\Psi)$, and $\delta_s(\tilde{\Psi}) < \delta + \theta_s(\mathbb{E}\tilde{\Psi}^*\tilde{\Psi})$ hold with probability $1 - \eta$ provided that

$$m \geq C_1\delta^{-2}\max(K^2, \tilde{K}^2)4[2 + \max\left(\theta_s(\mathbb{E}\Psi^*\Psi), \theta_s(\mathbb{E}\tilde{\Psi}^*\tilde{\Psi})\right)]s(\ln s)^2 \ln n \ln m,$$

$$m \geq C_2\delta^{-2}\max(K^2, \tilde{K}^2)s\ln(\eta^{-1})$$

for universal constants $C_1$ and $C_2$.

Corollary 2.3.2 and Corollary 2.3.3 have very different implications. Corollary 2.3.2 guarantees that $\delta_{ks}(\Psi) < c$ holds with high probability when $m = O(s\ln^4 n)$ if $\max_{k,\ell}|(\Psi)_{k,\ell}| = O(\frac{1}{\sqrt{m}})$ and $\theta_{ks}(\mathbb{E}\Psi^*\Psi) < 0.5c$. The former condition implies that the rows of $\Psi$ are incoherent to the standard basis vectors and is called the incoherence property. As will be discussed in later subsections, the latter condition, $\theta_{ks}(\mathbb{E}\Psi^*\Psi) < 0.5c$, is often difficult to satisfy for small $c \in (0, 1)$, in particular, in practical settings of compressed sensing. Although this condition has not been shown to be a necessary condition for $\delta_{ks}(\Psi) < c$, no alternative analysis is available for random frame

---

$^3$A direct derivation of Corollary 2.3.2 might provide better constants, but we do not attempt to optimize the universal constants.
matrices. In contrast, $\theta_{ks}(E \bar{\Psi}^* \Psi)$ can be made small by an appropriate choice of $\bar{\Psi}$, which by Corollary 2.3.3 suffices to make $\theta_{ks}(E \bar{\Psi}^* \Psi) < c$. In fact, it is often the case that $\bar{\Psi}$ can be chosen to make $\theta_{ks}(E \bar{\Psi}^* \Psi)$ much smaller than $\theta_{ks}(E \Psi^* \Psi)$, or even zero, and to satisfy the incoherence property at the same time. In this case, $\theta_{ks}(\bar{\Psi}^* \Psi) < c$ is guaranteed, whereas $\delta_{ks}(\Psi)$ is not guaranteed so. This key difference in the guarantees in Corollaries 2.3.2 and 2.3.3 establishes the advertised result that the RBOP-based guarantees of oblique pursuits apply to more general cases, in which the RIP-based guarantees of the corresponding conventional pursuits fail.

In the next subsections, we elaborate the comparison of the two different approaches: oblique pursuits with RBOP-based guarantees vs. conventional pursuits with RIP-based guarantees (per Corollaries 2.3.2 and 2.3.3) in more concrete scenarios in which $\Psi$ is given as the composition of the sensing matrix $A$ obtained from a frame and the dictionary $D$ with certain properties.

2.3.2 Case I: Sampled Frame $A$ and Nonredundant $D$ of Full Rank

We first consider the case of $\Psi = AD$, where the sensing matrix $A$ is constructed from a frame $(\phi_\omega)_{\omega \in \Omega}$ by (1.1.4) using a probability measure $\nu$, and the sparsifying dictionary $D$ is nonredundant ($n \leq d$) with full column rank.

Using the isotropy property, $EA^*A = I_d$, conventional RIP analysis [20, Theorem 8.4] showed that $\delta_s(\Psi) < \delta$ holds with high probability for $m = O(\delta^{-2}s \ln^4 n)$ under the following ideal assumptions:

(AI-1) $(\phi_\omega)_{\omega \in \Omega}$ is a tight frame, i.e., $\Phi \Phi^* = I_d$ where $\Phi, \Phi^*$ denotes the associated synthesis and the analysis operators.

(AI-2) $\nu$ is the uniform measure.

(AI-3) $D^*D = I_n$.

Corollary 2.3.2 generalizes [20, Theorem 8.4], so that the same RIP result continues holds when the ideal assumptions are “slightly” violated. To quantify this statement, we introduce the following metrics that measure the deviation from the ideal assumptions.
• Nonuniform distribution \( \nu \): We additionally assume that \( \nu \) is absolutely continuous with respect to \( \mu \).\(^4\) Define

\[
\nu_{\min} \triangleq \text{ess inf}_{\omega \in \Omega} \frac{d\nu}{d\mu}(\omega) \quad \text{and} \quad \nu_{\max} \triangleq \text{ess sup}_{\omega \in \Omega} \frac{d\nu}{d\mu}(\omega) \quad (2.3.8)
\]

where the essential infimum and supremum are w.r.t. to the measure \( \nu \). If \( \Omega \) is a finite set, then \( \frac{d\nu}{d\mu}(\omega) \) reduces to the probability that \( \omega \in \Omega \) will be chosen, multiplied by the cardinality of \( \Omega \). By their definitions, \( \nu_{\min} \) and \( \nu_{\max} \) satisfy \( \nu_{\min} \leq 1 \leq \nu_{\max} \). Note that \( \nu_{\min} \) and \( \nu_{\max} \) measure how different \( \nu \) is from the uniform measure \( \mu \). In particular, \( \nu_{\min} = \nu_{\max} = 1 \) if \( \nu \) coincides with \( \mu \).

• Non-tight frame \((\phi_\omega)_{\omega \in \Omega}\): Multiplying \( \Psi \) and \( y \) by a common scalar does not modify the inverse problem \( \Psi x = y \). Therefore, replacing \( \Phi \Phi^* \) by the same matrix multiplied by an appropriate scalar, we assume without loss of generality that

\[
\lambda_1(\Phi \Phi^*) = 1 + \frac{\kappa(\Phi \Phi^*) - 1}{\kappa(\Phi \Phi^*) + 1} \quad (2.3.9)
\]

and

\[
\lambda_d(\Phi \Phi^*) = 1 - \frac{\kappa(\Phi \Phi^*) - 1}{\kappa(\Phi \Phi^*) + 1} \quad (2.3.10)
\]

where \( \kappa(\Phi \Phi^*) \) denotes the condition number of \( \Phi \Phi^* \). Equations (2.3.9) and (2.3.10) imply

\[
\theta_d(\Phi \Phi^*) = ||\Phi \Phi^* - I_d|| = \frac{\kappa(\Phi \Phi^*) - 1}{\kappa(\Phi \Phi^*) + 1}
\]

where the first identity follows from the definition of \( \theta_d \). Note that \( \theta_d(\Phi \Phi^*) = 0 \) if \( \Phi \Phi^* = I_d \).

• Non-orthonormal \( D \): Similarly, for nonredundant \( D \), we assume without loss of generality that

\[
\lambda_1(D^*D) = 1 + \frac{\kappa(D^*D) - 1}{\kappa(D^*D) + 1} \quad (2.3.11)
\]

\(^4\)If \( \Omega \) is a finite set, then \( \mu \) is the counting measure and any probability measure \( \nu \) is absolutely continuous.
and
\[ \lambda_n(D^*D) = 1 - \frac{\kappa(D^*D) - 1}{\kappa(D^*D) + 1} \] (2.3.12)

where \( \kappa(D^*D) \) denotes the condition number of \( D^*D \). Equations (2.3.11) and (2.3.12) imply
\[ \theta_n(D^*D) = \|D^*D - I_n\| = \frac{\kappa(D^*D) - 1}{\kappa(D^*D) + 1}. \]

Note that \( \theta_n(D^*D) = 0 \) if \( D \) corresponds to an orthonormal basis, i.e., \( D^*D = I_n \).

Now, invoking Corollary 2.3.2 with the above metrics, we obtain the following Theorem 2.3.4, of which Theorem 1.1.6 is a simplified version. Under the ideal assumptions, \( K_0 \) vanishes and Theorem 2.3.4 reduces to [20, Theorem 8.4].

**Theorem 2.3.4.** Let \( (\phi_\omega)_\omega \in \Omega \) and \( D = [d_1, \ldots, d_n] \in \mathbb{K}^{d \times n} \) satisfy
\[ \sup_{\omega} \max_j |\langle \phi_\omega, d_j \rangle| \leq K \text{ for some } K \geq 1. \]
Let \( A \in \mathbb{K}^{m \times d} \) be constructed from \( (\phi_\omega)_\omega \in \Omega \) by (1.1.4) using a probability measure \( \nu \), and let \( \Psi = AD \). Let \( \nu_{\min} \) and \( \nu_{\max} \) be defined in (2.3.8). Then, \( \delta_s(\Psi) < \delta + K_0 \) holds with probability \( 1 - \eta \) provided that \( m \geq C_1(1 + K_0)^2K^2\delta^{-2}s(\ln s)^2 \ln m \ln n m \) and \( m \geq C_2K^2\delta^{-2}s \ln(\eta^{-1}) \) for universal constants \( C_1 \) and \( C_2 \) where \( K_0 \) is given in terms of \( \nu_{\min}, \nu_{\max}, \delta_s(D), \) and \( \theta_d(\Phi \Phi^*) \) by
\[ K_0 = \max(1 - \nu_{\min}, \nu_{\max} - 1) + \nu_{\max}[\delta_s(D) + \theta_d(\Phi \Phi^*) + \delta_s(D) \cdot \theta_d(\Phi \Phi^*)]. \] (2.3.13)

**Proof.** See Appendix A.6.

Theorem 2.3.4 shows that the ideal assumptions (AI-1) - (AI-3) for achieving the RIP of \( \Psi \) can be relaxed to a certain extent. However, even the relaxed assumptions are still too demanding to be satisfied in many practical applications of compressed sensing. When the ideal assumptions are not all satisfied, each deviation increases \( K_0 \) and the obtained upper bound on \( \delta_s(\Psi) \) also increases. For example, when \( \Phi \Phi^* = I_d \) and \( D^*D = I_n \), depending on \( \nu \), the upper bound on \( \delta_s(\Psi) \) may turn out to be even larger than 1, which fails to provide an RIP-based guarantee. As another example, when \( \nu = \mu \) and
\[ \Phi \Phi^* = I_d \] (the rows of \( A \) are obtained from i.i.d. samples from a tight frame according to the uniform distribution), \( \delta_s(D) \) determines the quality of the upper bound. Although, in general, computation of \( \delta_s(D) \) is NP hard, an easy upper bound on \( \delta_s(D) \) is given as \( \delta_n(D) = \| D^* D - I_n \| \). Now, note that \( \delta_n(D) \geq 0.6 \) for \( \kappa(D) \geq 2 \).

Therefore, considering that the RIP-based guarantee of HTP [14] requires \( \delta_{3s}(\Psi) < 0.57 \), which is the largest upper bound on \( \delta_{3s}(\Psi) \) among all sufficient conditions for known RIP-based guarantees. This suggests that even when the other ideal assumptions are satisfied, \( D \) needs to be near ideally conditioned. This strong requirement on \( D \) is often too restrictive, in particular, for learning a data-adaptive dictionary \( D \).

Next, we show that \( \theta_s(\widetilde{\Psi}^* \Psi) < c \) is achieved more easily, without the aforementioned restriction on \( \Phi, \nu, \) or \( D \). To this end, we would like to use Corollary 2.3.3; however, the \( \widetilde{K} \) parameter in Corollary 2.3.3 requires further attention. While the incoherence parameter \( K \) is determined by the inverse problem, the other incoherence parameter \( \widetilde{K} \) is determined by our own choice of \( \widetilde{A} \) and \( \widetilde{D} \). Recall the construction of \( \widetilde{\Psi} = \widetilde{A} \widetilde{D} \): matrix \( \widetilde{A} \in \mathbb{K}^{m \times d} \) is constructed from the dual frame \( (\tilde{\phi}_\omega)_{\omega \in \Omega} \) by (1.1.6) using the same probability measure \( \nu \) used to construct \( A \) per (1.1.4), whereas \( \widetilde{D} \) is given as \( \widetilde{D} = D(D^* D)^{-1} \), so that \( \widetilde{D}^* D = I_n \). It follows that \( \widetilde{K} \) is related to \( \Phi \) and \( D \), and thus to \( K \). By deriving an upper bound on \( \widetilde{K} \) in terms of \( K \) and using it in Corollary 2.3.3, we obtain the following theorem.

**Theorem 2.3.5.** Let \( (\phi_\omega)_{\omega \in \Omega} \) and \( D = [d_1, \ldots, d_n] \in \mathbb{K}^{d \times n} \) satisfy

\[
\sup_\omega \max_j |\langle \phi_\omega, d_j \rangle| \leq K \text{ for } K \geq 1.
\]

Let \( \nu \) be a probability measure on \( \Omega \) such that its derivative is strictly positive. Let \( A, \widetilde{A} \in \mathbb{K}^{m \times n} \) be random matrices constructed from a biorthogonal frame \( (\phi_\omega, \tilde{\phi}_\omega)_{\omega \in \Omega} \) by (1.1.4) and (1.1.6), respectively using \( \nu \). Let \( \Psi = AD \) and \( \widetilde{\Psi} = \widetilde{A} \widetilde{D} \) where \( \widetilde{D} = D(D^* D)^{-1} \). Let \( \nu_{\text{min}} \) and \( \nu_{\text{max}} \) be defined in (2.3.8). Then, \( \theta_s(\widetilde{\Psi}^* \Psi) < \delta, \delta_s(\Psi) < \delta + K_1, \) and \( \delta_s(\widetilde{\Psi}) < \delta + K_1 \) hold with probability \( 1 - \eta \) provided that

\[
m \geq C_1 (1 + K_1)^2 K_2^2 \delta^{-2} s (\ln s)^2 \ln n \ln m, \tag{2.3.14}
\]
\[
m \geq C_2 K_2^2 \delta^{-2} s \ln(\eta^{-1}) \tag{2.3.15}
\]

for universal constants \( C_1 \) and \( C_2 \), where \( K_1 \) and \( K_2 \) are given in terms of
\[ K_1 = \max(1 - \nu_{\max}^{-1}, \nu_{\min}^{-1} - 1) + \max(\nu_{\max}, \nu_{\min}^{-1}) \]
\[ \cdot \left\{ 1 + \frac{\delta_n(D)}{1 - \delta_n(D)} + \frac{\theta_d(\Phi^*)}{1 - \theta_d(\Phi^*)} + \frac{\delta_n(D) \theta_d(\Phi^*)}{[1 - \delta_n(D)][1 - \theta_d(\Phi^*)]} \right\} \]
\[(2.3.16)\]

and
\[ K_2 = \frac{\| (D^*D)^{-1} e_i^1 \|_{\ell_1^d}}{\nu_{\min}^2} \left[ K + \left( \sup_{\omega \in \Omega} \| \phi_\omega \|_{\ell_2^d} \right) \cdot \frac{\theta_d(\Phi^*)}{1 - \theta_d(\Phi^*)} \cdot \left( \max_{j \in [n]} \| d_j \|_{\ell_2^d} \right) \right]. \]
\[(2.3.17)\]

**Proof.** See Appendix A.7. \(\Box\)

With any significant violation of the ideal assumptions (AI-1) – (AI-3), Theorem 2.3.4 fails to provide \(\delta_{ks}(\Psi) < c\), whereas Theorem 2.3.5 still provides \(\theta_{ks}(\Psi^*\Psi) < c\). Therefore, the RBOP-based guarantee of recovery by oblique pursuits is a significant improvement over the conventional RIP-based guarantees, in the sense that the former applies to a practical setup (subset selection with a nonuniform distribution, non-tight frame, and non-orthonormal dictionary) while the latter does not. This is because violation of the ideal assumptions does not affect the upper bound on \(\theta_{s}(\Psi^*\Psi)\) in Theorem 2.3.5. Instead, it increases the upper bounds on \(\delta_{s}(\Psi)\) and \(\delta_{s}(\Psi^*)\). However, in the guarantees of oblique pursuits, unlike \(\theta_{s}(\Psi^*\Psi)\), the restricted isometry constants \(\delta_{s}(\Psi)\) and \(\delta_{s}(\Psi^*)\) need not be bounded from above by a certain threshold.

**Example 2.3.6.** We show the implication of Theorem 2.3.5 in a 2D Fourier imaging example. The corresponding numerical results for this scenario can be found in Section 2.4. The measurements are taken over random frequencies sampled i.i.d. from the uniform 2D lattice grid \(\Omega\) with a nonuniform measure \(\nu\). The signal of interest is sparse over a data-adaptive dictionary \(D\), which is invertible \((n = d)\) and has block diagonal structure.

More specifically, \(D\) in this example is constructed as follows. Recently, Ravishankar and Bresler [107] proposed an efficient algorithm that learns a data-adaptive square transform \(T\) with a regularizer on its condition number.
When the condition number of $T$ is reasonably small, $D$ given by $D = T^{-1}$ serves as a good dictionary for sparse representation. In particular, they designed a patch-based transform $T$ that applies to each patch of the image. When the patches are nonoverlapping, $T$ and $D$ have block diagonal structure; hence, applying $D$ and $D^*$ is computationally efficient. Furthermore, when the patches are much smaller than the image, each atom in $D$ is sparse and has low mutual coherence to the Fourier transform that applies to the entire image. For example, $D \in \mathbb{C}^{512 \times 512}$ used in the numerical experiment in Section 2.4 was designed so that it applies to $8 \times 8$ pixel patches. It has condition number 1.99, which implies $\delta_n(D) = 0.60$. We also observed that $D$ satisfies $\| (D^*D)^{-1} \|_{\ell_2 \to \ell_2} = 2.13$.

Since $(\phi_\omega)_{\omega \in \Omega}$ corresponding to the 2D DFT is tight, it follows that $\theta_d(\Phi \Phi^*) = 0$. Therefore, the expressions for $K_1$ and $K_2$ in eqs. (2.3.16) and (2.3.17) reduce to

$$K_1 = \max(1 - \nu_{\max}^{-1}, \nu_{\min}^{-1} - 1) + 2.5 \max(\nu_{\max}, \nu_{\min}^{-1})$$ (2.3.18)

and

$$K_2 = \frac{2.13}{\nu_{\min}^2} K.$$ (2.3.19)

Recall that $\nu_{\min}$ and $\nu_{\max}$ in this scenario correspond to the minimum and maximum probability that a measurement is taken at a certain frequency component. The simplified expressions of $K_1$ and $K_2$ in (2.3.18) and (2.3.19) show quantitatively how the use of nonuniform distribution for the i.i.d. sampling in the construction of a random frame matrix increases the required number of measurements.

2.3.3 Case II: Sampled Frame $A$ and Overcomplete $D$ with the RIP

The analysis in the previous section focused on the case where the dictionary $D$ is not redundant. In fact though, the analysis extends to certain cases of redundant/overcomplete $D$. One such case is when $D$ is, like $A$, a random frame matrix. Then, using a construction similar to our construction of $\tilde{A}$ will produce a matrix $\tilde{D}$ with $\mathbb{E}\tilde{D}^*D = I_n$, which combined with $\mathbb{E}\tilde{A}^*A = I_d$ provides $\mathbb{E}\tilde{\Psi}^*\Psi = I_n$. However, usually, $D$ is given as a deterministic matrix.
(e.g., concatenation of analytic bases, analytic frame, data-adaptive dictionary, etc). Therefore, in the general redundant $D$ case, using the biorthogonal dual of $D$ as $\tilde{D}$ is not a promising approach. Instead, we focus in the remainder of this subsection on the case where $D$ satisfies the RIP with small $\delta_s(D)$. Using $\tilde{\Psi} = \tilde{A}D$, we show the RBOP of $(\Psi, \tilde{\Psi})$ in this case.

**Theorem 2.3.7.** Let $(\phi_\omega)_{\omega \in \Omega}$ and $D = [d_1, \ldots, d_n] \in \mathbb{K}^{d \times n}$ satisfy $\sup_\omega \max_j |\langle \phi_\omega, d_j \rangle| \leq K$ for $K \geq 1$. Let $A, \tilde{A} \in \mathbb{K}^{m \times n}$ be random matrices constructed from a biorthogonal frame $(\phi_\omega, \tilde{\phi}_\omega)_{\omega \in \Omega}$ by (1.1.4) and (1.1.6), respectively using a probability measure $\nu$. Suppose that $\delta_s(D) < 1$. Let $\Psi = AD$, and $\tilde{\Psi} = \tilde{A}D$. Let $\nu_{\min}$ and $\nu_{\max}$ be defined in (2.3.8). Then, $\theta_s(\tilde{\Psi}^* \Psi) < \delta + \delta_s(D)$, $\delta_s(\Psi) < \delta + K_1$, and $\delta_s(\tilde{\Psi}) < \delta + K_1$ hold with probability $1 - \eta$ provided that

\begin{align}
    m &\geq C_1 (1 + K_1)^2 K_2^2 \delta^{-2} s (\ln s)^2 \ln n \ln m, \quad (2.3.20) \\
    m &\geq C_2 K_2^2 \delta^{-2} s \ln(\eta^{-1}) \quad (2.3.21)
\end{align}

for universal constants $C_1$ and $C_2$, where $K_1$ and $K_2$ are given in terms of $K, \nu_{\min}, \nu_{\max}, \delta_s(D)$, and $\theta_d(\Phi^*)$ by

$$
K_1 = \max(1 - \nu_{\max}^{-1}, \nu_{\min}^{-1} - 1) + \max(\nu_{\max}^{-1}, \nu_{\min}^{-1}) \\
\cdot \left(1 + \delta_s(D) + \frac{\theta_d(\Phi^*)}{1 - \theta_d(\Phi^*)} + \frac{\delta_s(D) \theta_d(\Phi^*)}{1 - \theta_d(\Phi^*)}\right).
$$

and

$$
K_2 = \frac{1}{\nu_{\min}} \left[ K + \left(\sup_{\omega \in \Omega} \|\phi_\omega\|_{\ell_2^d}\right) \cdot \frac{\theta_d(\Phi^*)}{1 - \theta_d(\Phi^*)} \cdot \left(\max_{j \in [n]} \|d_j\|_{\ell_2^d}\right)\right].
$$

**Proof.** See Appendix A.8. \hfill \Box

### 2.3.4 Case III: Sampled Tight Frame $A$ and Orthonormal Basis $D$ / RIP Matrix $D$

In the special case where the use of a nonuniform distribution for the i.i.d. sampling in the construction of $A$ is the only cause for the resulting failure of the exact/near isotropy property, the failure of the conventional RIP analy-
sis can be fixed differently. Recall that the construction of \( \tilde{A} \) in (1.1.6) only involves the weighting of rows of a matrix obtained from the biorthogonal dual frame \((\tilde{\phi}_\omega)_{\omega \in \Omega}\), with sampling at the same indices as used for the construction of \( A \) from the frame \((\phi_\omega)_{\omega \in \Omega}\). Therefore, for the special case when \((\phi_\omega)_{\omega \in \Omega}\) is a tight frame and \(D^*D = I_n\), it is possible to derive the RIP of a preconditioned version of \( \Psi \).

We construct a preconditioned sensing matrix \( \tilde{A} \) as

\[
(\tilde{A})_{k,\ell} = \frac{1}{\sqrt{m}} \left[ \frac{d\nu}{d\mu}(\omega_k) \right]^{-1/2} (\phi_{\omega_k})_{\ell}, \quad \forall k \in [m], \ \ell \in [d]
\]

(2.3.22)

where \((\omega_k)_{k=1}^m\) are the same sampling points used in the construction of \( A \) in (1.1.4). Then, by construction, \( \tilde{A} \) satisfies the isotropy property \( \mathbb{E} \tilde{A}^* \tilde{A} = I_d \). Furthermore, if \( \sup_\omega \max_j |\langle \phi_\omega, d_j \rangle| \leq K \), then \( \max_{k,\ell} |(\tilde{\Psi})_{k,\ell}| \leq \sqrt{\frac{\nu_{\min}^{-1/2} K}{m}} \) holds.

In this case, it suffices to invoke [20, Theorem 8.4] to show the RIP of \( \tilde{\Psi} \).

Invoking instead Theorem 2.3.4, this approach extends in a straightforward way to the case where \( D \) satisfies the RIP. In the case of tight frame \( A \) and \( D \) that is an orthobasis or an RIP matrix, these results provide an alternative (and equivalent) approach to obtain guaranteed algorithms, without invoking RBOP. In particular, defining \( \Lambda \) as the diagonal matrix given by \((\Lambda)_{j,j} = [(d\nu/d\mu)(\omega_k)]^{-1/2} \) for \( j \in [m] \), conventional recovery algorithms with an RIP-based guarantee can be used to solve the modified inverse problem \( \Lambda \Psi = \Lambda y \).

As discussed earlier, non-tight frame and/or non-orthonormal or non-RIP dictionaries arise in applications of compressed sensing, and in these instances too the conventional RIP analysis fails. We are currently investigating whether, and if so how, the above approach to “preconditioned” \( \tilde{\Psi} \) may be extended in general beyond the aforementioned cases.

### 2.4 Numerical Results

We performed two experiments to compare the oblique pursuits to their conventional counterparts and to other methods.

In the first experiment, we tested the algorithms on a generic data set. Synthesis operators \( \Phi \) and \( \tilde{\Phi} \) for a random biorthogonal frame \((\phi_\omega,\tilde{\phi}_\omega)_{\omega \in \Omega}\) were generated using random unitary matrices \( U, V \in \mathbb{R}^{n \times n} \) and a fixed diagonal matrix \( \Sigma \) as \( \Phi = U\Sigma V^* \) and \( \tilde{\Phi} = U\Sigma^{-1}V^* \). The diagonal entries of
Σ increase linearly from $\sqrt{\frac{2}{3}}$ to $\sqrt{\frac{4}{3}}$. Sensing matrix $A \in \mathbb{R}^{m \times n}$ was formed by $m$ random rows of $\Phi$ scaled by $\frac{1}{\sqrt{m}}$, where the row selection was done with respect to the uniform distribution. Then, the condition number of $EA^*A$ is 2 and the isotropy property is not satisfied. In this setting the oblique pursuit algorithms are different from their conventional counterparts. Signal $x^* \in \mathbb{K}^n$ is exactly $s$-sparse in the standard basis vectors ($D = I_n$) and the nonzero elements have unit magnitude and random signs. The success of each algorithm is defined as the exact recovery of the support.

Figure 2.1 shows the empirical phase transition of each algorithm as a function of $m/n$ and $s/n$. The results were averaged over 100 repetitions. Oblique versions of thresholding and IHT showed dramatic improvement in performance while the performance of the other algorithms is almost the same. While the oblique pursuit algorithms can be guaranteed without $A$ satisfying the isotropy property, the modification of the algorithms at least do not result in the degradation of the performance.
Figure 2.1: Phase transition of support recovery by various greedy pursuit algorithms (the horizontal and vertical axes denote the ratio \( m/n \) of number of measurements to number of unknowns and ratio \( s/m \) of sparsity level to number of measurements, respectively): signal \( x^\star \) is exactly \( s \)-sparse with nonzero entries that are \( \pm 1 \) with random sign. \( n = 1024, \) SNR = 30dB, \( \kappa = 2. \)
In the second experiment, we tested the algorithms on a CS Fourier imaging system. The partial DFT sensing matrix $A$ used in this experiment was constructed using the variable density suggested by Lustig et al. [24]. We used a data-adaptive square dictionary $D$ that applies to non-overlapping patches. Dictionary $D$ was learned from the fully sampled complex valued brain image using the algorithm proposed by Ravishankar and Bresler [107] (See Example 2.3.6 for more detail). The resulting $D$ was well conditioned with condition number $\kappa(D) = 1.99$. The Oblique pursuit algorithms use $\Psi = \widetilde{A} \widetilde{D}$, where $\widetilde{D}$ is given as the biorthogonal dual $\widetilde{D} = (D^{-1})^*$. Since the patches are non-overlapping, applying $D$, $\widetilde{D}$, and their adjoint operators are patch-wise operations, and are computed efficiently.

The input image was a phantom image obtained by $s$-sparse approximation over the dictionary $D$ of an original brain image with sparsity ratio $s/n = 0.125$. Our goal in this experiment is not to compete with the state of the art of recovery algorithms in CS imaging system; rather, we want to check whether the oblique pursuit algorithms perform competitively with their conventional counterparts in a setting where the RBOP of $(\Psi, \widetilde{\Psi})$ is guaranteed. This motivates our choice of a simplified test scenario. We also compare the oblique pursuit algorithms to simple zero filling, and to NESTA [108] that solves the $\ell_1$ analysis formulation [23]. In fact, when the original brain image is used as the input image, all sparsity-based reconstruction algorithms, including NESTA, performed worse than zero filling. To get a meaningful result in this setting, we replaced the input image by an exactly $s$-sparse phantom obtained by the $s$-sparse approximation of the original brain image.

\footnote{To achieve good performance on the original image requires a more sophisticated recovery algorithm with overlapping patches, and adaptive sparsity level [109].}
Table 2.2: Quality (PSNR in dB) of images reconstructed from noisy variable density Fourier samples with measurement SNR = 30 (dB). Results averaged over 100 random sampling patterns.

<table>
<thead>
<tr>
<th></th>
<th>Thres</th>
<th>CoSaMP</th>
<th>SP</th>
<th>IHT</th>
<th>HTP</th>
<th>$\ell_1$-Analysis</th>
<th>Zero Filling</th>
</tr>
</thead>
<tbody>
<tr>
<td>conventional</td>
<td>14.48</td>
<td>42.93</td>
<td>45.24</td>
<td>9.06</td>
<td>40.02</td>
<td>34.04</td>
<td>34.73</td>
</tr>
<tr>
<td>oblique</td>
<td>37.59</td>
<td>43.30</td>
<td>44.79</td>
<td>44.96</td>
<td>45.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) Downsample by 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>conventional</td>
<td>9.34</td>
<td>29.46</td>
<td>34.74</td>
<td>9.34</td>
<td>31.58</td>
<td>30.96</td>
<td>31.55</td>
</tr>
<tr>
<td>oblique</td>
<td>31.13</td>
<td>32.21</td>
<td>36.17</td>
<td>31.10</td>
<td>36.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) Downsample by 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.2: Error images and PSNR for recovery by various algorithms from noisy measurements (the maximum intensity of the input image is normalized as 1): SNR = 30dB, downsample by 3.
Table 2.2 shows the PSNR of the reconstructed images using the various algorithms with different downsampling ratio. The error images truncated at the maximum magnitude of the input image divided by 10 are shown in Fig. 2.2. Downsampling by factors of 2 and 3 is presented, but the results for larger downsampling factor are qualitatively the same.

In most cases, the oblique pursuit algorithms performed better than the conventional counterparts. In the few exceptions, the difference in performance is not significant. In particular, ObSP and ObHTP performed significantly better than zero filling. We observed that thresholding and IHT totally failed in this experiment. In this experiment, the step sizes of IHT was fixed as 1 for its RIP-based guarantees. By employing an empirically tuned step size, the performance of IHT might be improved. In contrast, ObIHT provided a reasonable performance with a fixed step size.

Fig. 2.2 also shows that the error in the reconstruction include blocky artifacts that are more severe in the reconstruction by the $\ell_1$ analysis formulation. This issue can be resolved by replacing the non-overlapping patches by overlapping patches. Furthermore, sparse representation of overlapping patches allows more redundancy, which helps reduce the sparse approximation error. In this case, applying the inverse and the biorthogonal dual of the sparsifying transform are no longer patch-wise operations, but the inverse operation might be still efficiently computed by solving a structured inverse problem. More generally, the sparsifying dictionary might be replaced by any redundant dictionary.

However, we do not pursue the various possible improvements of the reconstruction performance in this paper. As mentioned earlier, the purpose of the numerical results in this section is just to confirm that the modification made in the oblique pursuit algorithms from the original ones does not degrade their empirical performance. It turned out fortuitously that the oblique pursuit algorithms, designed to provide guarantees in terms of the RBOP, also show significant improvement in empirical performance.
CHAPTER 3

SUBSPACE METHODS FOR JOINT SPARSE RECOVERY

3.1 Notation

Symbol \( \mathbb{N} \) is the set of natural numbers (excluding zero), and \([n]\) denotes the set \(\{1, \ldots, n\}\) for \(n \in \mathbb{N}\). Symbol \( \mathbb{K} \) denotes a scalar field, which is either the real field \(\mathbb{R}\) or the complex field \(\mathbb{C}\). The vector space of \(d\)-tuples over \(\mathbb{K}\) is denoted by \(\mathbb{K}^d\) for \(d \in \mathbb{N}\). Similarly, for \(m, n \in \mathbb{N}\), the vector space of \(m \times n\) matrices over \(\mathbb{K}\) is denoted by \(\mathbb{K}^{m \times n}\).

We will use various notations on a matrix \(A \in \mathbb{K}^{m \times n}\). The range space spanned by the columns of \(A\) will be denoted by \(\mathcal{R}(A)\). The Hermitian transpose of \(A\) will be denoted by \(A^*\). The \(j\)th column of \(A\) is denoted by \(a_j\) and the submatrix of \(A\) with columns indexed by \(J \subset [n]\) is denoted by \(A_J\). The \(k\)th row of \(A\) is denoted by \(a_k\), and the submatrix of \(A\) with rows indexed by \(K \subset [m]\) is denoted by \(A_K\). Symbol \(e_k\) will denote the \(k\)th standard basis vector of \(\mathbb{K}^d\), where \(d\) is implicitly determined for compatibility. The \(k\)th largest singular value of \(A\) will be denoted by \(\sigma_k(A)\). For Hermitian symmetric \(A\), \(\lambda_k(A)\) will denote the \(k\)th largest eigenvalue of \(A\). The Frobenius norm and the spectral norm of \(A\) are denoted by \(\|A\|_F\) and \(\|A\|\), respectively. For \(p, q \in [1, \infty]\), the mixed \(\ell_{p,q}\) norm of \(A\) is defined by \(\|A\|_{p,q} \triangleq \left(\sum_{k=1}^m \|a_k\|_p^q\right)^{\frac{1}{q}}\) for \(q < \infty\) and \(\|A\|_{p,\infty} \triangleq \max_{k \in [m]} \|a_k\|_p\). The inner product is denoted by \(\langle \cdot, \cdot \rangle\). The embedding Hilbert space, where the inner product is defined, is not explicitly mentioned when it is obvious from the context. For a subspace \(S\) of \(\mathbb{K}^d\), matrices \(P_S \in \mathbb{K}^{d \times d}\) and \(P_S^\perp \in \mathbb{K}^{d \times d}\) denote the orthogonal projectors onto \(S\) and its orthogonal complement \(S^\perp\), respectively. For two matrices \(A\) and \(B\) of the same dimension, \(A \geq B\) if \(A - B\) is positive semidefinite. Symbols \(\mathbb{P}\) and \(\mathbb{E}\) will denote the probability and the expectation with respect to a certain distribution. Unless otherwise mentioned, the distribution shall be obvious from the context.
3.2 Problem Statement

A vector $x \in \mathbb{K}^n$ is $s$-sparse if it has at most $s$ nonzero components. The support of $x$ and its sparsity level denote, respectively, the set of the indices of nonzero components of $x$ and their number. Let $\{x_i\}_{i=1}^N \subset \mathbb{K}^n$ be jointly $s$-sparse (i.e., the union of the supports of $x_i$ for $i = 1, \ldots, N$ has at most $s$ elements). Then, $X_0 = [x_1, \ldots, x_N] \in \mathbb{K}^{n \times N}$ has no more than $s$ nonzero rows and is called row $s$-sparse. The support of $X_0$ denotes the union of the supports of the columns of $X_0$, or equivalently, the set of indices of nonzero rows. The joint sparse recovery problem is to find the support of the unknown signal matrix $X_0$ from the matrix $Y \in \mathbb{K}^{m \times N}$ with multiple measurement vectors (columns of $Y$) given by

$$Y = AX_0 + W$$

with a common sensing matrix $A \in \mathbb{K}^{m \times n}$ and with perturbation $W \in \mathbb{K}^{m \times N}$.

For the noiseless case ($W = 0$), the recovery of $X_0$ and the recovery of its support are equivalent. Conditions for the unique identification of the support from noiseless measurements have been well studied [7, 60, 65, 73]. However, for the noisy case, computing a good estimate of $X_0$ does not necessarily involve the recovery of the support of $X_0$. For example, the estimates of $X_0$ produced by the mixed-norm-based approaches are usually not exactly row $s$-sparse, but close to $X_0$ in the Hilbert-Schmidt norm. In some applications, the support has important physical meaning; hence, the identification of the support is explicitly required. For example, in imaging applications of compressed sensing, the support corresponds to the location of the target object, and in sparse linear regression, the most contributing variables are identified by the support (cf. [110]). In such applications, unless the solution obtained to the sparse recovery problem is exactly $s$-sparse, a step of thresholding the obtained solution to the nearest $s$-sparse vector is necessary. For this reason, in this chapter, we are interested in the exact identification of the support of $X_0$.

Let us now return to the main problem. An important parameter in the problem will be the rank of the unknown signal matrix $X_0$, which will be assumed unknown as well. Let $J_0$ denote the support of $X_0$. Since $X_0$ is row $s$-sparse, $X_0^{J_0}$, the submatrix with the nonzero rows of $X_0$, satisfies
rank($X_0$) = rank($X_0^{J_0}$). When $X_0^{J_0}$ has full row rank, rank($X_0^{J_0}$) assumes its maximum value, rank($X_0^{J_0}$) = $s$, and we will refer to this as the full row rank case. This is the case preferred by the algorithms in this chapter. Otherwise, rank($X_0^{J_0}$) < $s$, considered as violation of the full row rank case, will be called the rank-defective case.

3.3 MUSIC Revisited

In this section, we review the version of MUSIC for joint sparse recovery [26], on which our new algorithm in Section 3.4 improves. We also elaborate the guarantees of MUSIC in [7, 26] to hold without ideal assumptions.

3.3.1 MUSIC for the Joint Sparse Recovery Problem

Revisited

As in the original MUSIC algorithm in sensor array processing [58], the first step of MUSIC for joint sparse recovery [7, 26] is to estimate the so-called signal subspace $S$ defined by

$$S \triangleq \mathcal{R}(AX_0) = \mathcal{R}(A_{J_0}X_0^{J_0})$$

from the snapshot matrix

$$Y = A_{J_0}X_0^{J_0} + W \in \mathbb{K}^{m \times N}.$$  

In sensor array processing, MUSIC [58] estimates $S$ using the eigenvalue decomposition (EVD) of $\frac{YY^*}{N}$. The same method is applied to the joint sparse recovery problem and the exact estimation of $S$ is guaranteed asymptotically, as $N \to \infty$, under the assumption that

$$\lim_{N \to \infty} \frac{YY^*}{N} = \lim_{N \to \infty} \frac{A_{J_0}X_0^{J_0}(X_0^{J_0})^*A_{J_0}^*}{N} + \sigma_w^2 I,$$  

(3.3.1)

which is achieved with statistical assumptions on $A_{J_0}X_0^{J_0}$ and $W$ [7].\footnote{We adopt the terminology from the sensor array processing literature. To emphasize the analogy between the joint sparse recovery problem and DOA estimation, we also call each of the $N$ columns of $Y$ a snapshot. Then, $N$ will denote the number of snapshots.} It is also assumed that $A_{J_0} \in \mathbb{K}^{m \times s}$ has full column rank and $X_0^{J_0} \in \mathbb{K}^{s \times N}$.

\footnote{Depending on the statistical assumptions, the limits in (3.3.1) are either in probability or almost surely.

59
has full row rank for all \( N \geq s \). Then, the dimension of \( \mathcal{S} \) coincides with \( \text{rank}(X_0^{J_0}) = s \). The assumed relation (3.3.1) implies that the multiplicity of the smallest eigenvalue of \( \frac{YY^*}{N} \) converges to \( m - s \) as \( N \to \infty \) (in the same limit sense as in (3.3.1)); hence, the dimension of \( \mathcal{S} \) is exactly determined. The signal subspace \( \mathcal{S} \) is then exactly computed as the invariant subspace spanned by the \( s \) dominant eigenvectors of \( \lim_{N \to \infty} \frac{YY^*}{N} \) since the noise part \( \sigma^2 \mathbf{I} \) in (3.3.1) only shifts the eigenvalues of \( \lim_{N \to \infty} A_{J_0} X_0^{J_0} (X_0^{J_0})^* A_{J_0}^* N \). The joint sparse recovery problem then reduces to the case where the subspace estimation is error-free; hence, the subsequent analysis of the support recovery in [7, 26, 33] considered this error-free case.\(^3\)

Given the signal subspace \( \mathcal{S} \), MUSIC for joint sparse recovery identifies the row-support \( J_0 \) as the set of the indices \( k \) of columns of \( A \) such that \( P_{\mathcal{S}} a_k = 0 \). In other words, MUSIC accepts the index \( k \) as an element of the support if \( a_k \in \mathcal{S} \) and rejects it otherwise. In the remainder of this chapter, the acronym MUSIC will denote the version for joint sparse recovery [26] rather than the original MUSIC algorithm for sensor array processing [58].

A sufficient condition that guarantees the success of MUSIC for the special case where \( X_0^{J_0} \) has full row rank, is given in terms of the Kruskal rank [111] defined below.

**Definition 3.3.1.** The Kruskal rank of a matrix \( A \), denoted by \( \text{krank}(A) \), is the maximal number \( k \) such that any \( k \) columns of \( A \) are linearly independent.

**Proposition 3.3.2** ([7, 26]). Let \( J_0 \) be an arbitrary subset of \([n]\) with \( s \) elements. Let \( X_0 \in \mathbb{K}^{n \times N} \) be row \( s \)-sparse with support \( J_0 \). Suppose that \( \text{rank}(X_0^{J_0}) = s \) and \( A \) satisfies

\[
\text{krank}(A) > s. \tag{3.3.2}
\]

Then

\[
P_{\mathcal{S}} a_k = 0 \tag{3.3.3}
\]

for \( \mathcal{S} \triangleq \mathcal{R}(A_{J_0} X_0^{J_0}) \) if and only if \( k \in J_0 \).

The following result follows directly from Proposition 3.3.2.

**Corollary 3.3.3.** Under the conditions on \( A \) and \( X_0^{J_0} \) in Proposition 3.3.2, given the exact signal subspace \( \mathcal{S} \), MUSIC is guaranteed to recover \( J_0 \).

\(^3\)Obviously, for the noiseless case, the subspace estimation is error-free without relying on the asymptotic in \( N \).
MUSIC, with its performance guarantee in Proposition 3.3.2, is remarkable in the context of compressive sensing for the following reasons. First, MUSIC is a polynomial-time algorithm with a performance guarantee under a condition that coincides with the necessary condition \( m > s \) for unique recovery (by any algorithm, no matter how complex). Second, MUSIC is simple and cheap, involving little more than a single EVD of the data covariance matrix. In fact, efficient methods for partial EVD or other rank-revealing decompositions can further reduce the cost.

Unfortunately, as is well known in the sensor array processing literature [58, 116], and also demonstrated by numerical experiments later in Section 3.8, MUSIC is prone to failure when \( X_0 \) does not have full row rank, or when it is ill-conditioned in the presence of noise. In sensor array processing, this is known as the “source coherence problem” [117] and (with the exception of the case of a Vandermonde matrix \( A \)) no general solution to this problem is known. This motivates our work to propose a new subspace-based method for joint sparse recovery that improves on MUSIC.

For the noisy case, the analysis of signal subspace estimation based on the asymptotic in \( N \) is not practical. In particular, from a perspective of compressed sensing (with joint sparsity), recovery of the support from a finite (often small) number of snapshots is desired. In the following subsection, we propose a subspace estimation scheme that works with finitely many snapshots, which will be applied to both MUSIC and the new subspace-based methods in this chapter. In Section 3.3.3, we formalize the MUSIC algorithm for support recovery in the presence of a perturbation in the estimate of \( S \). This will lay the ground for the subsequent analysis of MUSIC in the noisy scenario, and for its extension in the same scenario to SA-MUSIC.

---

4Since the mid-1990s, when these results (for what became known later as compressive sampling) [7, 26, 33] were published, until recently, MUSIC was the only polynomial-time algorithm with such a guarantee. For the SMV case, algebraic algorithms with guarantee by the minimal requirements have been later proposed [112, 113]. However, these guarantees are restricted to the case where \( A \) is a Vandermonde matrix and the measurements are noise-free. For the MMV case, a recent result [73] showed another (greedy) algorithm with the same guarantee as the guarantee [7, 26] for MUSIC, in the noise-free case.

5The rank revealing decompositions can be computed by the Lanczos method [45] for the truncated singular value decomposition (SVD). If the matrix is large, randomized algorithms [114, 115] can be also used to compute the truncated SVD.
3.3.2 Signal Subspace Estimation from Finitely Many Snapshots

We study the problem of signal subspace estimation from finitely many snapshots in this subsection. For later use in other subspace-based algorithms in Section 3.4, we weaken the assumptions in the previous section. In particular, we assume finite $N$ and no longer assume that $X_{0}^{J_0}$ has full row rank. We also assume that the columns of the noise matrix $W$ are i.i.d. random vectors with white spectrum (i.e., $EWW^* = \sigma^2_w I_m$). (Otherwise, the standard prewhitening schemes developed in sensor array processing may be applied prior to subspace estimation.)

When $A_{J_0}X_{0}^{J_0}$ is ill-conditioned, the last few singular values of $A_{J_0}X_{0}^{J_0}$ are small, making the estimation of $S \triangleq \mathcal{R}(A_{J_0}X_{0}^{J_0})$ highly sensitive to the noise in the snapshots. To improve the robustness against noise, instead of estimating the entire subspace $S$, we only estimate an $r$-dimensional subspace of $S$ for $r < s$. The dimension $r$ is determined so that the gap between $\sigma_r(A_{J_0}X_{0}^{J_0})$ and $\sigma_{r+1}(A_{J_0}X_{0}^{J_0})$ is significant. The obtained $r$ is an underestimate of $\text{rank}(X_{0}^{J_0})$.

We propose and analyze a simple scheme that determines the dimension $r$ by thresholding the eigenvalues of $\frac{YY^*}{N}$ so that the estimated signal subspace $\widehat{S}$ spanned by the $r$ dominant eigenvectors of $\frac{YY^*}{N}$ is close to an $r$-dimensional subspace of $S$. The procedure for estimating the signal subspace and its dimension is described next.

Given the snapshot matrix $Y = A_{J_0}X_{0}^{J_0} + W$, we compute its sample covariance matrix $\Gamma_Y$ defined by

$$\Gamma_Y \triangleq \frac{YY^*}{N}.$$ 

Then, we compute a biased matrix $\widehat{\Gamma}$ by

$$\widehat{\Gamma} \triangleq \Gamma_Y - \lambda_m(\Gamma_Y)I_m.$$ 

Note that $\widehat{\Gamma}$ and $\Gamma_Y$ have the same eigenvectors. Recall that our goal is to find an $r$-dimensional subspace $\widehat{S}$ for some $r \leq s$ from $\widehat{\Gamma}$ such that there exists an $r$-dimensional subspace $\mathcal{S}$ of the signal subspace $S \triangleq \mathcal{R}(A_{J_0}X_{0}^{J_0})$ satisfying $\|P_{\mathcal{S}} - P_{\widehat{S}}\| \leq \eta$ for small $\eta$. For better performance of support recovery by algorithms in Section 3.4, larger $r$ is preferred.
For an ideal case where (3.3.1) holds, $\hat{\Gamma}$ reduces to $\Gamma_S$ defined by

$$\Gamma_S \triangleq \frac{A_{J_0}X_0^T(X_0^T)^*A_{J_0}^*}{N}.$$ 

Since $S = \mathcal{R}(\Gamma_S)$, by setting $r$ to $\text{rank}(\Gamma_S)$, we compute $S$ itself rather than a proper subspace of $S$.

For finite $N$, usually, the cross-correlation terms in the sample covariance matrix $\Gamma_Y$ between the noise term $W$ and the signal term $A_{J_0}X_0^T$ are smaller than the autocorrelation terms. Since the autocorrelation term of $W$ is nearly removed in $\hat{\Gamma}$, we will show that it is likely that $D \triangleq \hat{\Gamma} - \Gamma_S$ is small in the spectral norm. In particular, let $\hat{S}$ be the subspace spanned by the $r$ dominant eigenvectors of $\hat{\Gamma}$; if $\|D\|$ is small compared to the gap between $\lambda_r(\hat{\Gamma})$ and $\lambda_{r+1}(\hat{\Gamma})$, then there exists an $r$-dimensional subspace $\bar{S}$ of $S$ with small $\|P_S - P_{\bar{S}}\|$. Since $\|D\|$ is not available, we determine the dimension $r$ by simple thresholding. More specifically, given a threshold $\tau > 0$, the dimension of $\hat{S}$ is determined as the maximal number $r$ satisfying

$$\frac{\lambda_r(\hat{\Gamma}) - \lambda_{r+1}(\hat{\Gamma})}{\lambda_1(\hat{\Gamma})} \geq \tau > \frac{\lambda_k(\hat{\Gamma}) - \lambda_{k+1}(\hat{\Gamma})}{\lambda_1(\hat{\Gamma})}, \quad \forall k > r. \quad (3.3.4)$$

When $A_{J_0}X_0^T$ (and hence, $\Gamma_S$) is ill-conditioned, it is likely that the gap between the consecutive eigenvalues $\lambda_{r_0}(\hat{\Gamma})$ and $\lambda_{r_0+1}(\hat{\Gamma})$ where $r_0 = \text{rank}(\Gamma_S)$ is small compared to $\|D\|$, which is roughly proportional to $\sigma_w^2/\lambda_1(\hat{\Gamma})$. The aforementioned increased robustness to noise is provided by choosing $r < r_0$ so that the gap $\lambda_r(\hat{\Gamma}) - \lambda_{r+1}(\hat{\Gamma})$ is large, which will, in turn, reduce the estimated subspace dimension $r$. More sophisticated methods for determining $r$ are possible, but this simple method suffices for our purposes and is amenable to analysis (see Section 3.7). This algorithm for estimating the signal subspace and its dimension is summarized as Algorithm 7.
**Algorithm 7: Signal Subspace Estimation**

input: \( Y \in \mathbb{K}^{m \times N} \) and \( \tau > 0 \)

output: \( r \in \mathbb{N} \) and \( P_S \in \mathbb{K}^{m \times m} \)

1. \( \Gamma_Y \leftarrow \frac{YY^*}{N} \);
2. \( \hat{\Gamma} \leftarrow \Gamma_Y - \lambda_m(\Gamma_Y)I_m \);
3. \( r \leftarrow m - 1 \);
4. while \( \lambda_r(\hat{\Gamma}) - \lambda_{r+1}(\hat{\Gamma}) < \tau \lambda_1(\hat{\Gamma}) \) do
5. \( r \leftarrow r - 1 \);
6. end
7. \( U \leftarrow r \) dominant eigenvectors of \( \hat{\Gamma} \);
8. \( P_S \leftarrow UU^* \);
9. return \( r, P_S \);

### 3.3.3 MUSIC Applied to an Inaccurate Estimate of the Signal Subspace

In the presence of a perturbation in the estimated signal subspace \( \hat{S} \), MUSIC finds the set \( \hat{J} \) of \( s \) indices that satisfy

\[
\min_{k \in \hat{J}} \frac{\|P_S a_k\|_2}{\|a_k\|_2} > \max_{k \in [n] \setminus \hat{J}} \frac{\|P_S a_k\|_2}{\|a_k\|_2}.
\tag{3.3.5}
\]

The corresponding algorithm is summarized in Algorithm 8. To provide an intuition for the selection criterion in (3.3.5), we use the notion of the “angle function” [118].

**Definition 3.3.4 ([118]).** The angle function between two subspaces \( S_1 \) and \( S_2 \) is defined by

\[
\angle_2(S_1, S_2) \triangleq \sin^{-1} \left( \min \{ \| P_{S_1} P_{S_2} \|, \| P_{S_2} P_{S_1} \| \} \right).
\tag{3.3.6}
\]

**Remark 3.3.5.** The angle function \( \angle_2(S_1, S_2) \) is different from the largest principal angle between \( S_1 \) and \( S_2 \) [45]. Unlike the largest principal angle, the angle function satisfies the metric properties even when \( S_1 \) and \( S_2 \) have different dimensions. In particular, when \( \dim(S_1) \geq \dim(S_2) \), the expression
on the right-hand side of (3.3.6) reduces to
\[ \langle \angle_2(S_1, S_2) = \sin^{-1}\left(\|P_{S_1}P_{S_2}\|\right). \] (3.3.7)

By (3.3.7), the criterion in (3.3.5) is equivalent to
\[ \max_{k \in J} \langle \angle_2(\hat{S}, \mathcal{R}(a_k)) \rangle < \min_{k \in [n] \setminus \hat{J}} \langle \angle_2(\hat{S}, \mathcal{R}(a_k)) \rangle. \] (3.3.8)

That is, MUSIC finds, among all subspaces spanned by a single column of \( A \), \( s \) subspaces nearest to \( S \) (in the angle function metric).

**Remark 3.3.6.** The previous work [7] on MUSIC considered specific applications (spectrum blind sampling and DOA estimation) where the columns of \( A \) have the same \( \ell_2 \) norm. Under these conditions, (3.3.5) reduces to the criterion of the algorithm in [7]. However, for general \( A \), employing the normalization makes a nontrivial difference. We consider the version of MUSIC with this normalization, which is a valid subspace method as discussed before.

**Algorithm 8: MUSIC**

```
input : Y ∈ \( \mathbb{K}^{m \times N} \), A ∈ \( \mathbb{K}^{m \times n} \), s ∈ \( \mathbb{N} \)
output: J ⊂ [n]
1 \( P_S \in \mathbb{K}^{m \times m}, r \in \mathbb{N} \leftarrow \) estimate signal subspace from \( Y \);
2 \( J \leftarrow \emptyset \);
3 for \( \ell = 1, \ldots, n \) do
4 \( \zeta_\ell \leftarrow \|P_S a_\ell\|_2/\|a_\ell\|_2; \)
5 end
6 \( J \leftarrow \) indices of the s-largest \( \zeta_\ell \)’s;
7 return \( J \);
```
3.4 Subspace-Augmented MUSIC

3.4.1 MUSIC Applied to an Augmented Signal Subspace

When \(X_{J_0}^0\) has full row rank, the signal subspace \(S \triangleq \mathcal{R}(A_{J_0}X_{J_0}^0)\) coincides with \(\mathcal{R}(A_{J_0})\). In this case, given the exact \(S\), MUSIC is guaranteed to recover the support \(J_0\) because: (1) \(a_k \in S\) for all \(k \in J_0\); and (2) \(a_k \notin S\) for all \(k \in [n] \setminus J_0\), which is implied by \(\text{krank}(A) = s + 1\) (Proposition 3.3.2).

However, in the rank-defective case, when \(X_{J_0}^0\) does not have full row rank (i.e., when \(\text{rank}(X_{J_0}^0)\) is strictly smaller than the sparsity level \(s\)), we have \(\dim(S) \leq \text{rank}(X_{J_0}^0) < s = \dim(\mathcal{R}(A_{J_0}))\). Therefore, \(S\) is a proper subspace of \(\mathcal{R}(A_{J_0})\) and \(a_k \notin S\) may occur for some (or all) \(k \in J_0\). This will cause MUSIC to miss valid components of \(J_0\). Since, in the presence of noise (imperfect subspace estimate), MUSIC selects, by (3.3.8), the \(s\) indices \(k \in [n]\) for which \(\mathcal{R}(a_k)\) is closest (in the sense of the angle function) to \(S\), this may result in the selection of spurious indices into the estimate of \(J_0\). This explains the well-known fact that in the rank-defective case, MUSIC is prone to failure.

Subspace-augmented MUSIC overcomes the limitation of MUSIC to the full row rank case by capitalizing on the following observation: MUSIC fails in the rank-defective case because \(S\) is a proper subspace of \(\mathcal{R}(A_{J_0})\); however, if another subspace \(T\) that complements \(S\) is given so that \(S + T = \mathcal{R}(A_{J_0})\), then MUSIC applied to the augmented subspace \(S + T\) will be successful.

Unfortunately, in general, finding such an oracle subspace is not feasible. The search procedure cannot even be enumerated. However, if \(X_{J_0}^0\), the matrix of nonzero rows of \(X_0\), or more generally, the subspace \(\mathcal{R}(X_{J_0}^0)\) satisfies a mild condition, then the search may be restricted without loss of generality to subspaces spanned by \(s - r\) columns of \(A\). The following proposition states this result.

**Definition 3.4.1.** Matrix \(X\) is row-nondegenerate if

\[
\text{krank}(X^*) = \text{rank}(X).
\] (3.4.1)

**Remark 3.4.2.** Condition (3.4.1) says that every \(k\)-row subset of \(X\) is linearly independent for \(k \leq \text{rank}(X)\). This is satisfied by \(X\) that is generic in the set of full rank matrices of the same size (equivalently, for \(X\), whose rows
are in general position). In fact, an even weaker requirement on $X$ suffices, as shown by the next argument, which reduces the requirement to $\mathcal{R}(X)$.

**Remark 3.4.3.** Condition (3.4.1) is invariant to multiplication of $X$ on the right by any full row rank matrix of compatible size. In particular, Condition (3.4.1) holds if and only if

$$\text{krank}(Q^*) = \text{rank}(X)$$

(3.4.2)

for any orthonormal basis $Q$ of $\mathcal{R}(X)$. It follows that (3.4.1) is a property of the subspace $\mathcal{R}(X)$. Furthermore, (3.4.1) also implies that any $\tilde{Q}$ such that $\mathcal{R}(\tilde{Q}) \subset \mathcal{R}(X)$ is also row-nondegenerate (Lemma B.2.1).

**Proposition 3.4.4** (Subspace Augmentation). Let $X_0 \in \mathbb{K}^{n \times N}$ be row $s$-sparse with support $J_0 \subset [n]$. Let $\mathcal{S}$ be an arbitrary $r$-dimensional subspace of $\mathcal{R}(A_{J_0}X_0^{J_0})$ where $r < s$. Let $J_1$ be an arbitrary subset of $J_0$ with $s - r$ elements. Suppose that $X_0^{J_0}$ is row-nondegenerate and $A_{J_0}$ has full column rank. Then

$$\mathcal{S} + \mathcal{R}(A_{J_1}) = \mathcal{R}(A_{J_0}).$$

(3.4.3)

**Proof.** See Appendix B.2. □

**Remark 3.4.5.** Note that $\dim(\mathcal{R}(A_{J_1})) \geq \dim(\mathcal{R}(A_{J_0})) - \dim(\mathcal{S}) = s - r$ is a necessary condition for (3.4.3). Therefore, $J_1$ should be a subset of $J_0$ with at least $s - r$ elements for the success of the subspace augmentation.

**Remark 3.4.6.** The row-nondegeneracy condition on $X_0^{J_0}$ is a necessary condition to guarantee (3.4.3) for an arbitrary subset $J_1$ of $J_0$ with $s - r$ elements. Suppose that $X_0^{J_0}$ fails to satisfy the row-nondegeneracy condition (i.e., $\text{krank}((X_0^{J_0})^*) < \text{rank}(X_0^{J_0})$). By the assumption on $\mathcal{S}$, there exists a row $s$-sparse matrix $U \in \mathbb{K}^{n \times r}$ with support $J_0$ such that $\mathcal{S} = \mathcal{R}(A_{J_0}U^{J_0})$. Since $\mathcal{S}$ was an arbitrary $r$-dimensional subspace of $\mathcal{S}$, without loss of generality, we may assume that $\text{krank}((U^{J_0})^*) < r$. By the projection update formula, it follows that $P_{\mathcal{R}(A_{J_1})+\mathcal{S}} = P_{\mathcal{R}(A_{J_1})} + P_{\mathcal{P}_{\mathcal{R}(A_{J_1})}^\perp \mathcal{S}}$; hence, it suffices to show $\dim(P_{\mathcal{P}_{\mathcal{R}(A_{J_1})}^\perp} \mathcal{S}) < r$ for the failure of (3.4.3). Since $\text{krank}((U^{J_0})^*) < r$,
there exists $J_1 \subset J_0$ of size $s - r$ such that \( \text{rank}(U^{J_0 \setminus J_1}) < r \). Then,

\[
\dim(P_{\mathcal{R}(A_{J_1})}^\perp \mathcal{S}) = \text{rank}(P_{\mathcal{R}(A_{J_1})}^\perp A_{J_0} U^{J_0}) \\
= \text{rank}(P_{\mathcal{R}(A_{J_1})}^\perp A_{J_0 \setminus J_1} U^{J_0 \setminus J_1}) \\
\leq \text{rank}(U^{J_0 \setminus J_1}) < r.
\]

It follows that (3.4.3) fails for this specific $J_1$. Therefore, the row-nondegeneracy condition on $X^{J_0}$ is essential. Furthermore, by Remarks 3.4.2 and 3.4.3, the row-nondegeneracy is a mild condition; it will be assumed to hold henceforth.

Let $X^{J_0}$ be row-nondegenerate and suppose that an error-free estimate of $\mathcal{S}$ is available. In this case, Proposition 3.4.4 implies that, given a correct partial support $J_1$ of size $s - r$, MUSIC applied to the augmented subspace $\mathcal{S} + \mathcal{R}(A_{J_1})$ enjoys the same guarantee as MUSIC for the full row rank case (Proposition 3.3.2). We will see in Section 3.6 that a similar statement applies even with an imperfect estimate $\hat{\mathcal{S}}$.

Based on the aforementioned result, we propose a class of methods for joint sparse recovery called Subspace-Augmented MUSIC (SA-MUSIC) consisting of the following steps:

**Step 1)** Signal subspace estimation: Compute an estimate $\hat{\mathcal{S}}$ of the signal subspace $\mathcal{S} \triangleq \mathcal{R}(A_{J_0} X^{J_0}) = \mathcal{R}(AX_0)$.

**Step 2)** Partial support recovery: Compute a partial support $J_1$ of size $s - r$ from $\hat{\mathcal{S}}$ and $A$, where $r = \dim(\hat{\mathcal{S}})$ and $s$ is the sparsity level known a priori.

**Step 3)** Augment signal subspace: Compute the augmented subspace $\bar{\mathcal{S}}$

\[
\bar{\mathcal{S}} = \hat{\mathcal{S}} + \mathcal{R}(A_{J_1}).
\]

**Step 4)** Support completion: Complete $J_1$ to produce $J_0 \supset J_1$, by adding $r$ more support elements obtained by applying “MUSIC” to $\bar{\mathcal{S}}$, that is, finding $\tilde{J} \subset [n]$ satisfying

\[
\min_{k \in J \setminus J_1} \frac{\|P_S a_k\|_2}{\|a_k\|_2} > \max_{k \in [n] \setminus \tilde{J}} \frac{\|P_S a_k\|_2}{\|a_k\|_2}. \tag{3.4.4}
\]

68
The general SA-MUSIC algorithm is summarized as Algorithm 9. An actual implementation might use orthonormal bases \( \hat{Q} \) and \( \tilde{Q} \) for the subspaces \( \hat{S} \) and \( \tilde{S} \) instead of constructing the projection operators \( P_{\hat{S}} \) and \( P_{\tilde{S}} \). Step 3 could then be performed by a QR decomposition of matrix \( [\hat{Q}, A_{J_1}] \).

**Algorithm 9: SA-MUSIC**

```latex
\textbf{input} : Y \in \mathbb{K}^{m \times N}, A \in \mathbb{K}^{m \times n}, s \in \mathbb{N}
\textbf{output} : J \subset [n]
1 \quad P_{\hat{S}} \in \mathbb{R}^{m \times m}, r \in \mathbb{N} \leftarrow \text{estimate signal subspace from } Y;
2 \quad J_1 \leftarrow \text{partial support recovery of size } s - r;
3 \quad P_{\tilde{S}} \leftarrow P_{\hat{S}} + P_{\mathcal{R}(P_{\hat{S}}^\perp A_{J_1})};
4 \quad \text{for } \ell \in [n] \setminus J_1 \text{ do}
5 \quad \quad \zeta_\ell \leftarrow \|P_{\tilde{S}}a_\ell\|_2/\|a_\ell\|_2;
6 \quad \text{end}
7 \quad J \leftarrow J_1 \cup \{\text{indices of the } r\text{-largest } \zeta_\ell\text{'s}\};
8 \quad \text{return } J;
```

A particular instance of the SA-MUSIC algorithm is specified by the particular methods used for the steps of signal subspace estimation and partial support recovery. For subspace estimation, in the analysis in Sections 3.6 and 3.7, we will consider the EVD-based scheme in Section 3.3.2. However, as mentioned earlier, the subspace estimation scheme is not restricted to the given method. For example, if the noise \( W \) is also sparse, then robust principal component analysis (RPCA) \cite{RPCA} will provide a better estimate of \( S \) than the usual SVD.

The choice of method for partial support recovery is discussed in the next subsection. Here, we note some special cases. As \( r \) increases, the size of the partial support required in SA-MUSIC decreases. For small \( s - r \), we can use an exhaustive search over \( J_1 \), the computational cost of which also decreases in \( r \). In particular, for the special case where \( r = s \), the step of partial support recovery is eliminated, and SA-MUSIC reduces to MUSIC \cite{MUSIC}.
3.4.2 Partial Support Recovery with Practical Algorithms

When there is a “rank defect” in $X_0^J$ (i.e., $\text{rank}(X_0^J) < s$), SA-MUSIC requires partial support recovery of size $s - r$. In addition to computational efficiency, a key desirable property of an algorithm to accomplish this is that it solves the partial support recovery problem more easily than solving the full support recovery problem.

From this perspective, greedy algorithms for the joint sparse recovery problem are attractive candidates. Both empirical observations and the performance guarantees in the sequel suggest that the first few steps of greedy algorithms are more likely to succeed than the entire greedy algorithms. In other words, greedy algorithms take advantage of the reduction to partial support recovery when they are combined with SA-MUSIC.

Any of the known greedy algorithms for joint sparse recovery may be used in SA-MUSIC, producing a different SA-MUSIC algorithm. We propose Orthogonal Subspace Matching Pursuit (OSMP) for the partial support recovery step in SA-MUSIC. OSMP is a modification of another greedy algorithm, rank-aware order recursive matching pursuit (RA-ORMP) [73]. OSMP replaces the snapshot matrix $Y$ in RA-ORMP by the orthogonal projector $P_S$ onto the estimated signal subspace or, equivalently, by an orthogonal basis matrix for the estimated signal subspace. Given the estimated support $J$ from the previous steps, RA-ORMP updates $J$ by adding $k$ selected by

$$k = \arg \max_{\ell \in [n] \setminus J} \| (P_{R(A_J)}^\perp Y) a_\ell \|_2 / \| P_{R(A_J)}^\perp a_\ell \|_2. \tag{3.4.5}$$

Similarly, OSMP updates the support by adding $k$ selected by

$$k = \arg \max_{\ell \in [n] \setminus J} \| (P_{R(A_J)}^\perp P_S) a_\ell \|_2 / \| P_{R(A_J)}^\perp a_\ell \|_2. \tag{3.4.6}$$

In general, the columns of $P_{R(A_J)}^\perp Y$ and $P_{R(A_J)}^\perp P_S$ span different subspaces; hence, the two projection operators used in (3.4.5) and (3.4.6) are different. The two projection operators coincide only for the special case when there is

---

6. The name “rank-aware ORMP” proposed for this algorithm appears to be a misnomer. RA-ORMP, as originally proposed [73], does not have any feature to determine rank. It computes an orthonormal basis for $Y$. However, whereas in the ideal, noiseless case this basis will have dimension $r$ equal to the rank of $X_0^J$, with any noise present $Y$ will have full rank equal to $\min\{m, N\}$, and this will also be the dimension of the computed orthonormal basis. Hence, the algorithm does not seem to have any built-in rank-awareness.
We interpret (3.4.6) using the angle function between two subspaces, that is, (3.4.6) is equivalent to

\[ k = \arg\min_{\ell \in [n] \setminus J} \angle_2(P_{R(A_J)}^\perp \hat{S}, P_{R(A_J)}^\perp R(a_\ell)). \] (3.4.7)

Given the subspace \( R(A_J) \), which is spanned by the columns of \( A \) corresponding to support elements \( J \) determined in the preceding steps of the algorithm, the selection rule finds the nearest subspace to \( P_{R(A_J)}^\perp \hat{S} \) among all subspaces, each of which is spanned by \( P_{R(A_J)}^\perp a_\ell \) for \( \ell \in [n] \setminus J \). The name “orthogonal subspace matching pursuit” is intended to distinguish the matching using a subspace metric in OSMP from that of OMP and its other variations.

Again, we use a partial run of \( s - r \) steps of OSMP for partial support recovery and switch to MUSIC applied to the augmented subspace. The complete OSMP is summarized as Algorithm 10.

**Algorithm 10: OSMP**

<table>
<thead>
<tr>
<th>input</th>
<th>( Y \in \mathbb{K}^{m \times N} ), ( A \in \mathbb{K}^{m \times n} ), ( s \in \mathbb{N} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>output</td>
<td>( J \subset [n] )</td>
</tr>
<tr>
<td>1</td>
<td>( P_S \in \mathbb{R}^{m \times m} ), ( r \in \mathbb{N} \leftarrow \text{estimate signal subspace from } Y; )</td>
</tr>
<tr>
<td>2</td>
<td>( J \leftarrow \emptyset; )</td>
</tr>
<tr>
<td>3 while (</td>
<td>J</td>
</tr>
<tr>
<td>4</td>
<td>( k \leftarrow \arg\max_{\ell \in [n] \setminus J} | (P_{R(A_J)}^\perp \hat{S}) a_\ell |<em>2/| P</em>{R(A_J)}^\perp a_\ell |_2; )</td>
</tr>
<tr>
<td>5</td>
<td>( J \leftarrow J \cup {k}; )</td>
</tr>
<tr>
<td>6</td>
<td>end</td>
</tr>
<tr>
<td>7</td>
<td>return ( J; )</td>
</tr>
</tbody>
</table>

### 3.4.3 Stopping Conditions for Unknown Sparsity Level

In most analyses of greedy algorithms, the sparsity level \( s \) is assumed to be known \emph{a priori}. In fact, this assumption is only for simplicity of the analysis and not a limitation of the greedy algorithms. We run the SA-
MUSIC algorithms until the angle function \( \sphericalangle_2(\mathcal{S}, \mathcal{R}(A_J)) = \| P_{\mathcal{R}(A_J)}^\perp P_S \| \) falls below a certain threshold \( \eta \) that depends on the noise level (or vanishes, in the noiseless case). For the noiseless case, \( \eta \) is set to 0, whereas for the noisy case, \( \eta \) is set to an estimate of \( \sphericalangle_2(\mathcal{S}, \mathcal{R}(A_{J_0})) \). The SA-MUSIC algorithm for unknown sparsity level \( s \) is summarized in Algorithm 11. The update criterion in Step 9 of Algorithm 11 determines a specific SA-MUSIC algorithm. For example, SA-MUSIC with OSMP (SA-MUSIC+SS-OSMP henceforth) uses the condition in (3.4.6). For simplicity, the analyses in Section 3.6 will assume that the sparsity level is known.

### Algorithm 11: SA-MUSIC for unknown \( s \)

**input**: \( Y \in \mathbb{K}^{m \times N}, A \in \mathbb{K}^{m \times n}, \eta > 0 \)

**output**: \( J \subset [n] \)

1. \( P_S \in \mathbb{K}^{m \times m}, r \in \mathbb{N} \leftarrow \text{estimate signal subspace from } Y; \)
2. \( J \leftarrow \emptyset; \)
3. **for** \( \ell = 1, \ldots, n \) **do**
   4. \( \zeta_\ell \leftarrow \| P_S a_\ell \|_2 / \| a_\ell \|_2; \)
5. **end**
6. \( J \leftarrow \text{indices of the } r\text{-largest } \zeta_k 's; \)
7. \( J_1 \leftarrow \emptyset; \)
8. **while** \( \| P_{\mathcal{R}(A_J)}^\perp P_S \| > \eta \) **do**
   9. Select \( k \) by an update criterion;
10. \( J_1 \leftarrow J_1 \cup \{ k \}; \)
11. \( P_S \leftarrow P_S + P_{\mathcal{R}(P_S^\perp A_{J_1})}; \)
12. **for** \( \ell \in [n] \setminus J_1 \) **do**
13. \( \zeta_\ell \leftarrow \| P_S a_\ell \|_2 / \| a_\ell \|_2; \)
14. **end**
15. \( J \leftarrow J_1 \cup \{ \text{indices of the } r\text{-largest } \zeta_\ell 's \}; \)
16. **end**
17. return \( J; \)
3.5 Weak Restricted Isometry Property

3.5.1 Uniform Restricted Isometry Property

The \textit{restricted isometry property} (RIP) has been proposed in the study of the reconstruction of sparse vectors by the $\ell_1$ norm minimization [9]. Matrix $A \in \mathbb{K}^{m \times n}$ satisfies the RIP of order $s$ if there exists a constant $\delta \in (0, 1)$ such that

\[
(1 - \delta) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta) \|x\|_2^2, \quad \forall x, \|x\|_0 \leq s. \tag{3.5.1}
\]

The smallest $\delta$ that satisfies (3.5.1) is called the \textit{restricted isometry constant} (RIC) of order $s$ and is denoted by $\delta_s(A)$. Note that (3.5.1) is equivalent to

\[
(1 - \delta)I_s \leq A_J^* A_J \leq (1 + \delta)I_s, \quad \forall J \subset [n], |J| = s; \tag{3.5.2}
\]

Hence, $\delta_s(A)$ satisfies

\[
\delta_s(A) = \max_{|J| = s} \|A_J^* A_J - I_s\|.
\]

The RIP of order $s$ implies that all submatrices of $A$ with $s$ columns are uniformly well conditioned. We therefore refer to this RIP as the \textit{uniform RIP}.

3.5.2 Weak Restricted Isometry Property

In many analyses of sparse signal recovery, the uniform RIP is unnecessarily strong and requires a demanding condition on $A$ that is not satisfied by the matrices that arise in applications. Therefore, weaker versions of RIP have been proposed, tailored to specific analyses [40,80].

Matrix $A \in \mathbb{K}^{m \times n}$ satisfies the \textit{weak restricted isometry property} (weak RIP) [40] with parameter $(J, s, t, \delta)$, where $s, t \in \mathbb{N}$, $J \subset [n]$ with $|J| = s$, and $\delta \in (0, 1)$, if

\[
(1 - \delta)I_{s+t} \leq A_K^* A_K \leq (1 + \delta)I_{s+t}, \quad \forall K \supset J, |K| = s + t. \tag{3.5.3}
\]
The corresponding **weak restricted isometry constant** (weak RIC) is given by

\[
\delta_{s+t}^\text{weak}(A; J) = \max_{K \supset J, |K| = s+t} \|A_K^* A_K - I_{s+t}\|.
\]

The special case of the weak RIP with \(t = 1\) has been previously proposed [80] to derive an average case analysis of the solution of the MMV problem by the mixed \(\ell_{2,1}\) norm minimization, also known as MMV basis pursuit (M-BP) [65]. This specific case of the weak RIP with \(t = 1\), which we call the **weak-1 RIP**, is useful for the analysis in this chapter. By the definitions, \(\delta_{s+1}^\text{weak}(A; J) \leq \delta_{s+1}(A)\); hence, the weak-1 RIP is satisfied by a less stringent condition on \(A\) than the uniform RIP of the same order.

Recall that RIC provides the upper and the lower bounds in (3.5.1), which are symmetrically deviated from 1. Tighter bounds can be obtained without this symmetry in the following form:

\[
\alpha \|x\|_2 \leq \|Ax\|_2 \leq \beta \|x\|_2, \quad \forall x, \|x\|_0 \leq s. \tag{3.5.4}
\]

The **asymmetric RIP** [120,121] is another variation of the uniform RIP, which instead of (3.5.1), is defined by (3.5.4). Again, the weak-1 version of the asymmetric RIP suffices for the analyses in this chapter. Similarly to the weak-1 RIP, matrix \(A\) satisfies the **weak-1 asymmetric RIP** if there exist \(\alpha, \beta > 0\) such that

\[
\alpha \leq \sigma_{s+1}(A_{J \cup \{j\}}) \leq \sigma_1(A_{J \cup \{j\}}) \leq \beta, \quad \forall j \in [n] \setminus J.
\]

The corresponding **weak-1 asymmetric RICs** are defined as follows:

\[
\alpha_{s+1}^\text{weak}(A; J) \triangleq \min_{j \in [n] \setminus J} \sigma_{s+1}(A_{J \cup \{j\}}), \\
\beta_{s+1}^\text{weak}(A; J) \triangleq \max_{j \in [n] \setminus J} \sigma_1(A_{J \cup \{j\}}).
\]

They are related to the weak-1 RIC by

\[
\delta_{s+1}^\text{weak}(A; J) = \max \left\{ 1 - \left[ \alpha_{s+1}^\text{weak}(A; J) \right]^2, \left[ \beta_{s+1}^\text{weak}(A; J) \right]^2 - 1 \right\}.
\]

Furthermore, \(\delta_{s+1}^\text{weak}(A; J) < 1 - \alpha^2\) is a sufficient condition for \(\alpha_{s+1}^\text{weak}(A; J) > \alpha\).

In the following text, we list some matrices that satisfy the weak-1 RIP
along with the required conditions. We show that, compared to the uniform RIP, the requirement on the number of measurements to satisfy the weak-1 RIP is reduced by large factors, ranging between 200 to thousands fold. Importantly, in addition to Gaussian matrices, which lend themselves to relatively easy analysis, the list of matrices satisfying the weak-1 RIP includes partial Fourier matrices that arise in applications, and provides reasonable constants for them.

3.5.3 Gaussian Matrix

Eldar and Rauhut derived a condition for the weak-1 RIP of an i.i.d. Gaussian matrix [80, Proposition 5.3]. Their proof starts with the concentration of the quadratic form \( \|Gx\|_2^2 \) around its expectation, where \( G \) is an i.i.d. Gaussian matrix, to bound the singular values of \( G \). This approach was originally proposed for the uniform RIP in [17]. We provide an alternative and much tighter condition for the weak-1 RIP of \( A \) directly using the concentration of the singular values of an i.i.d. Gaussian matrix [122].

**Proposition 3.5.1.** Given \( m, n \in \mathbb{N} \) with \( m \leq n \), let \( A \in \mathbb{R}^{m \times n} \) be an i.i.d. Gaussian matrix whose entries follow \( \mathcal{N}(0, \frac{1}{m}) \). Suppose that \( J \) is a subset of \([n]\) with \( s \) elements and

\[
\sqrt{m} \geq \frac{\sqrt{s + 1} + \sqrt{2 \ln \left( \frac{2(n-s)}{\epsilon} \right)}}{\sqrt{1 + \delta} - 1} \tag{3.5.5}
\]

for \( \epsilon, \delta \in (0, 1) \). Then

\[
P(\delta_{s+1}^{\text{weak}}(A; J) \geq \delta) \leq \epsilon. \tag{3.5.6}
\]

**Proof.** See Appendix B.3. \( \square \)

**Remark 3.5.2.** For large \( s \) such that the log term is negligible, Condition (3.5.5) reduces to \( m/s \geq (\sqrt{1 + \delta} - 1)^{-2} \).

**Remark 3.5.3.** Using the concavity of the square root function, it follows that

\[
m \geq \frac{2}{(\sqrt{1 + \delta} - 1)^2} \left\{ s + 2 \ln \left( \frac{2(n-s)}{\epsilon} \right) + 1 \right\} \tag{3.5.7}
\]
is a sufficient condition for (3.5.5); hence, it also guarantees (3.5.6).

Remark 3.5.4. By slightly modifying the proof of Proposition 3.5.1, we obtain the uniform RIP of $A$: $\delta_s(A) < \delta$ holds with probability at least $1 - \epsilon$ provided that

$$m \geq \frac{2}{(\sqrt{1 + \delta} - 1)^2} \left\{ \left[ 3 + \ln \left( \frac{n}{s} \right) \right] s + 2 \ln \left( \frac{2}{\epsilon} \right) + 1 \right\}. \quad (3.5.8)$$

Compared to Condition (3.5.7) required for the weak-1 RIP, in Condition (3.5.8) for the uniform RIP, the required oversampling factor $\frac{m}{s}$ has been increased roughly by the factor $3 + 2 \ln \left( \frac{n}{s} \right)$.

The proof of Proposition 3.5.1 reveals that a submatrix $A_J$ of $A$ in Proposition 3.5.1 has extreme singular values symmetrically deviated from 1. However, the extreme eigenvalues of $A_J^* A_J$, which are the squared extreme singular values of $A_J$, are not symmetrically deviated from 1. Therefore, for an i.i.d. Gaussian $A$, we can obtain much sharper estimates of the weak-1 asymmetric RICs compared to that of the symmetric weak-1 RIC. In this case, the weak-1 asymmetric RIP is satisfied by a much milder condition on $A$ even compared to the weak-1 RIP.

Proposition 3.5.5. Given $m, n \in \mathbb{N}$ with $m \leq n$, let $A \in \mathbb{R}^{m \times n}$ be an i.i.d. Gaussian matrix whose entries follow $\mathcal{N}(0, \frac{1}{m})$. Suppose that $J$ is a subset of $[n]$ with $s$ elements and

$$\sqrt{m} \geq \sqrt{s + 1} + \sqrt{2 \ln \left( \frac{n-s}{\epsilon} \right)} \frac{\gamma}{\gamma}, \quad (3.5.9)$$

for $\epsilon, \gamma \in (0, 1)$. Then

$$\mathbb{P}\left( [\alpha_{s+1}^{\text{weak}}(A; J) \leq 1 - \gamma] \lor [\beta_{s+1}^{\text{weak}}(A; J) \geq 1 + \gamma] \right) \leq \epsilon. \quad (3.5.10)$$

Proof. See Appendix B.4. \qed

Remark 3.5.6. For large $s$ such that the log term is negligible, Condition (3.5.9) reduces to $\frac{m}{s} \geq \frac{1}{\gamma^2}$.
Remark 3.5.7. Using the concavity of the square root function, it follows that
\[ m \geq \frac{2}{\gamma^2} \left\{ s + 2 \ln \left( \frac{2(n-s)}{\epsilon} \right) + 1 \right\} \tag{3.5.11} \]
is a sufficient condition for (3.5.9); hence, it also guarantees (3.5.10).

Remark 3.5.8. Proposition 3.5.5 provides a sufficient condition, which is not necessarily tightest, in particular, in the limit when \( \gamma \to 1 \). Indeed, an i.i.d. Gaussian \( A \) satisfies \( \alpha_{s+1}^{\text{weak}}(A;J) > 0 \) with probability 1 if \( m > s \), but Condition (3.5.9) does not converge to \( m > s \) as \( \delta \) approaches 1. Nevertheless, this gap vanishes if \( s \) goes to infinity with \( n = o(e^s) \), that is, \( s \) grows faster than \( \ln n \).

3.5.4 Random Partial Fourier Matrix

Candes and Plan [40] showed that a matrix \( A \) composed of randomly selected rows of a discrete Fourier transform (DFT) matrix satisfies the following local restricted isometry property under a certain mild condition: \( \|A_J^*A_J - I_s\| \leq \delta \) with high probability for a fixed \( J \subset [n] \) of size \( s \) ([40, Lemma 2.1]). It is not difficult to derive the weak-1 RIP of such a matrix from its local RIP. We only need to consider the union of the events corresponding to all subset of \([n]\) that include the support \( J \) and one more element outside \( J \).

Proposition 3.5.9 (Corollary to [40, Lemma 2.1]). Let \( k_1, \ldots, k_m \in [n] \) be i.i.d. with respect to the uniform distribution. For \( j = 1, \ldots, m \), let the \( j \)th row of \( A \) be the \( k_j \)th row of the \( n \times n \) DFT matrix divided by \( \sqrt{m} \). Suppose that \( J \) is a fixed subset of \([n]\) of cardinality \( s \), and
\[ m \geq \left( \frac{2(3+\delta)}{3\delta^2} \right) \left\{ \ln \left( \frac{2(n-s)}{\epsilon} \right) + \ln(s+1) \right\} (s+1) \tag{3.5.12} \]
for \( \epsilon, \delta \in (0,1) \). Then
\[ \mathbb{P}(\delta_{s+1}^{\text{weak}}(A;J) \geq \delta) \leq \epsilon. \tag{3.5.13} \]

Proof. See Appendix B.5. \( \square \)

\footnote{In fact, the result in [40], and hence, our argument derived from their result apply to a wider class of matrices.}

77
Remark 3.5.10. The uniform RIP of a random partial Fourier matrix has been studied before [18,20,123]. In particular, Rauhut [20, Theor. 8.4] showed a sufficient condition with explicit constants given by

\[
\frac{m}{\ln(10m)} \geq C \delta^{-2} \ln^2(100s) \ln(4n)s,
\]
\[
m \geq D \delta^{-2} \ln(\epsilon^{-1})s,
\]

where \( C \leq 17,190 \) and \( D \leq 456 \), which is a much more demanding condition than (3.5.12). In particular, when \( \frac{s}{n} \) is a constant, these conditions imply that \( m \geq C \delta^{-2}(\ln^4 n)s \), which is worse than (3.5.12) in the order, not only in constants. Hence, the weak-1 RIP for the partial Fourier matrix case reduces the required oversampling factor \( \frac{m}{s} \) not only by factors of thousands, but also in the order, compared to the uniform RIP.

3.6 Performance Guarantees

3.6.1 MUSIC for the Full Row Rank Case

With an imperfect estimate of the signal subspace, the support recovery by MUSIC is no longer guaranteed by an algebraic property of \( A \). Instead, in the following proposition, we provide a new guarantee.

Theorem 3.6.1 (MUSIC, Noisy, Full Row Rank Case). Let \( X_0 \in \mathbb{K}^{n \times N} \) be row \( s \)-sparse with support \( J_0 \subset [n] \). Let \( \tilde{S} \) be an \( s \)-dimensional subspace in \( \mathbb{K}^m \) such that \( \|P_{\tilde{S}} - P_{R(A_{J_0}X_0^*J_0)}\| \leq \eta \) for \( \eta \in (0,0.5) \). Suppose that \( X_0^\dagger_{J_0} \) has full row rank and \( A \) satisfies

\[
\alpha_{s+1}^\text{weak}(A; J_0) > \alpha 
\]

for

\[
\alpha \geq 2\sqrt{\eta(1-\eta)\|A^*\|_{2,\infty}}.
\]

Then MUSIC applied to \( P_{\tilde{S}} \) will identify \( J_0 \).

Proof. See Appendix B.6.

Remark 3.6.2. The columns of the sensing matrix \( A \) are often normalized in the \( \ell_2 \) norm (e.g., for a partial Fourier matrix) or their \( \ell_2 \) norms are highly
concentrated around 1 (e.g., for an i.i.d. Gaussian matrix). We therefore consider the quantity \( \|A^*\|_{2,\infty} \) to be 1 or close to 1. In particular, we assume that all columns of \( A \) are normalized in the \( \ell_2 \) norm (“\( A \) is normalized,” in short) in all of the figures and the numerical experiments in this chapter.

**Remark 3.6.3.** With normalized \( A \), we have \( \|A^*\|_{2,\infty} = 1 \), and \( \delta_{s+1}^{\text{weak}}(A; J_0) < 1 - \alpha^2 \) is a sufficient condition for (3.6.1).

**Remark 3.6.4.** When the signal subspace estimation is perfect (i.e., in the noiseless case), Conditions (3.6.1) and (3.6.2) reduce to \( \alpha_{s+1}^{\text{weak}}(A; J_0) > 0 \), which is an algebraic condition implying \( \text{rank}([A_{J_0}, a_j]) = s + 1 \) for all \( j \in [n] \setminus J_0 \). This algebraic condition is implied by a much milder condition than the weak-1 asymmetric RIP, which is an analytic condition. For example, an i.i.d. Gaussian \( A \) with \( m \geq s + 1 \) satisfies this with probability 1.

**Remark 3.6.5.** Theorem 3.6.1 guarantees that MUSIC recovers a fixed support \( J_0 \). Replacing Condition (3.6.1) by its uniform analog, \( \sigma_{s+1}(A_J) > \alpha \) for all \( J \subset [n] \) with \( |J| = s + 1 \), provides a uniform guarantee that MUSIC recovers an arbitrary support of size \( s \). With perfect subspace estimation, the uniform guarantee reduces to Proposition 3.3.2.

### 3.6.2 SA-MUSIC with Given Partial Support

SA-MUSIC finds the support by using the augmented subspace \( \tilde{S} \) constructed as \( \tilde{S} = \mathcal{R}(A_{J_1}) + \hat{S} \), where \( J_1 \) is a subset of \( J_0 \) of size \( s - r \) and \( \hat{S} \) is the estimated signal subspace of dimension \( r \). For SA-MUSIC to be successful, the estimate \( \hat{S} \) needs to be good enough. Here, we quantify this by assuming that there exists an \( r \)-dimensional subspace \( \tilde{S} \) of the signal subspace \( S \triangleq \mathcal{R}(A_{J_0} X_{J_0}^H) \) satisfying \( \|P_{\tilde{S}} - P_{\hat{S}}\| \leq \eta \). Section 3.7 shows that if the number of snapshots is large enough relative to the noise variance, then Algorithm 7 computes \( \hat{S} \) with the property that such an \( \tilde{S} \) exists. Recall (by Proposition 3.4.4) that assuming row-nondegenerate \( X_{J_0}^H \) implies \( \mathcal{R}(A_{J_1}) + \hat{S} = \mathcal{R}(A_{J_0}) \), which is desired by the MUSIC step in SA-MUSIC. However, since \( \hat{S} \) is constructed using \( \hat{S} \) rather than \( \tilde{S} \), which is not available, to show noise robustness of support recovery, we need to bound \( \|P_{\tilde{S}} - P_{\mathcal{R}(A_{J_0})}\| \). Now, by the projection update formula, it follows that

\[
P_{\tilde{S}} - P_{\mathcal{R}(A_{J_0})} = P_{P_{\mathcal{R}(A_{J_1})}\tilde{S}} - P_{P_{\mathcal{R}(A_{J_1})}\hat{S}}.
\]
Therefore, we need to consider the distance between the two subspaces \( \tilde{S} \) and \( \hat{S} \) as projected onto \( \mathcal{R}(A_{J_1})^\perp \). However, in general, projecting onto another subspace can either increase or decrease the distance between subspaces arbitrarily. In our specific case, the distance is bounded depending on the condition number of \( A_{J_0} \). We state the result in a formal way in the following proposition.

**Proposition 3.6.6.** Let \( X_0 \in \mathbb{K}^{n \times N} \) be row \( s \)-sparse with support \( J_0 \subset [n] \). Let \( J \) be a proper subset of \( J_0 \). Let \( \hat{S} \) be an \( r \)-dimensional subspace in \( \mathbb{K}^m \) such that there exists an \( r \)-dimensional subspace \( \bar{S} \) of \( \mathcal{R}(A_{J_0}X_0^{J_0}) \) satisfying \( \|P_{\bar{S}} - P_{\hat{S}}\| \leq \eta \). Suppose that \( X_0^{J_0} \) is row-nondegenerate and \( A \) satisfies

\[
\eta \cdot \kappa(A_{J_0}) < 1 \tag{3.6.3}
\]

where \( \kappa(A_{J_0}) \) denotes the condition number of \( A_{J_0} \). Then

\[
\|P_{P_{\mathcal{R}(A_{J_1})} \hat{S}} - P_{P_{\mathcal{R}(A_{J_1})} \bar{S}}\| \leq \frac{1}{1/[\eta \cdot \kappa(A_{J_0})] - 1}. \tag{3.6.4}
\]

**Proof.** See Appendix B.7.

We are now ready to state one of the main results of this chapter.

**Theorem 3.6.7 (SA-MUSIC with Correct Partial Support).** Let \( X_0 \in \mathbb{K}^{n \times N} \) be row \( s \)-sparse with support \( J_0 \subset [n] \). Let \( J_1 \) be an arbitrary subset of \( J_0 \) with \( s - r \) elements. Let \( \hat{S} \) be an \( r \)-dimensional subspace in \( \mathbb{K}^m \) such that there exists an \( r \)-dimensional subspace \( \bar{S} \) of \( \mathcal{R}(A_{J_0}X_0^{J_0}) \) satisfying \( \|P_{\bar{S}} - P_{\hat{S}}\| \leq \eta \). Suppose that \( X_0^{J_0} \) is row-nondegenerate and \( A \) satisfies

\[
\alpha < \alpha_{s+1}^{\text{weak}}(A; J_0) \leq \beta_{s+1}^{\text{weak}}(A; J_0) < \beta \tag{3.6.5}
\]

for \( \alpha \) and \( \beta \) satisfying

\[
1 - \sqrt{1 - \frac{\alpha^2}{\|A^*\|_2^2}} \geq \frac{2\eta\beta}{\alpha - \eta\beta}. \tag{3.6.6}
\]

Then MUSIC applied to \( \tilde{S} = \hat{S} + \mathcal{R}(A_{J_1}) \) will identify \( J_0 \setminus J_1 \).

**Proof.** See Appendix B.8.
Remark 3.6.8. With normalized $A$, the condition in (3.6.5) and (3.6.6) is implied by $\delta_{s+1}^{\text{weak}}(A; J_0) < \delta$ for

$$\eta \leq \sqrt{\frac{1 - \delta}{1 + \delta} \cdot \frac{1 - \sqrt{\delta}}{3 - \sqrt{\delta}}},$$

(3.6.7)

Remark 3.6.9. Compared to the guarantee on MUSIC (full row rank case, Theorem 3.6.1), the guarantee for SA-MUSIC (for the rank-defective case, Theorem 3.6.7) additionally requires $X_{j_0}^0$ to be row-nondegenerate. In the noiseless case, with normalized $A$, both Theorems 3.6.1 and 3.6.7 require only a mild algebraic condition $\alpha_{s+1}^{\text{weak}}(A; J_0) > 0$. However, they differ in the noisy case. As shown in Fig. 3.1, even with known partial support, SA-MUSIC in the rank-defective case suffers more from the perturbation in the subspace estimate than does MUSIC in the full row rank case. This difference in the sensitivity to the subspace estimation error is due to the reduced dimension $r < s$ of the estimated signal subspace, which, in turn, is due to the rank defect of $X_{j_0}^0$.

Theorem 3.6.7 provides a performance guarantee for SA-MUSIC with an oracle providing a correct partial support estimate, but with a noisy subspace estimate. In this scenario, SA-MUSIC provides its best performance. How realistic is this assumption? In the following sections, we will show that if the error in the subspace estimate is small enough, a suboptimal greedy algorithm is indeed guaranteed to recover the partial support exactly. In particular, when $r/s$ is large, SA-MUSIC, combined with partial support recovery by a greedy algorithm, provides a guarantee comparable to that given in Theorem 3.6.7 for SA-MUSIC with correct partial support estimate.

3.6.3 OSMP

We first consider the OSMP algorithm by itself, and provide a performance guarantee for it. This result, which covers the noisy and rank-defective case, is of interest in its own right.

**Theorem 3.6.10 (OSMP, Rank-Defective Case).** Let $X_0 \in \mathbb{K}^{n \times N}$ be row $s$-sparse with support $J_0 \subset [n]$. Let $J$ be a proper subset of $J_0$. Let $\hat{S}$ be an $r$-dimensional subspace in $\mathbb{K}^m$ such that there exists an $r$-dimensional
Figure 3.1: Comparison of MUSIC for full row rank case versus SA-MUSIC with known partial support (SA-MUSIC+Oracle) in the rank-defective case: tradeoff between parameter $\delta$ (for the weak-1 RIP) and $\eta$ (for the subspace estimate perturbation). The region below the curve provides a guarantee.

subspace $\tilde{S}$ of $\mathcal{R}(A_{J_0}X_{J_0}^0)$ satisfying $\|P_{\tilde{S}} - P_S\| \leq \eta$. Suppose that $X_{J_0}^0$ is row-nondegenerate and $A$ satisfies

$$\alpha < \alpha_{s+1}^{weak}(A; J_0) \leq \beta_{s+1}^{weak}(A; J_0) < \beta$$

for $\alpha$ and $\beta$ satisfying

$$\sqrt{\min\left(1, \frac{r}{s-|J|}\right)} \cdot \frac{\alpha}{\|A^*\|_{2,\infty}} - \sqrt{1 - \frac{\alpha^2}{\|A^*\|_{2,\infty}^2}} \geq \frac{2\eta\beta}{\alpha - \eta\beta}. \tag{3.6.9}$$

Then a single step of OSMP given $J$ will identify an element of $J_0 \setminus J$. For the special case when $r = s$, Condition (3.6.9) is replaced by the weaker condition

$$1 - \sqrt{1 - \frac{\alpha^2}{\|A^*\|_{2,\infty}^2}} \geq \frac{2\eta\beta}{\alpha - \eta\beta}. \tag{3.6.10}$$

Proof. See Appendix B.9

In the full row rank case, $X_{J_0}^0$ is trivially row-nondegenerate. In the noiseless case, we have $\eta = 0$ and with normalized $A$, Condition (3.6.10) reduces to $\alpha_{s+1}^{weak}(A; J_0) > 0$. Therefore, for the full row rank and noiseless case, OSMP is guaranteed by $\alpha_{s+1}^{weak}(A; J_0) > 0$, which coincides with the condition for the
guarantee of MUSIC in the same scenario (Proposition 3.3.2). In fact, in this noiseless and full row rank case, OSMP is equivalent to the corresponding data-domain algorithm, RA-ORMP. The coincidence of the guarantees of MUSIC and of RA-ORMP in this special case has been shown before [73]. (Unlike the analysis of RA-ORMP [73] though, Theorem 3.6.10 also applies to the noisy and/or rank-defective cases.)

3.6.4 SA-MUSIC with Partial Support Recovery by OSMP

Combining Theorem 3.6.10 and Theorem 3.6.7, we obtain another main result of this chapter: a guarantee for SA-MUSIC+OSMP.

**Theorem 3.6.11** (SA-MUSIC+OSMP, Rank-Defective Case). Let $X_0 \in \mathbb{K}^{n \times N}$ be row $s$-sparse with support $J_0 \subset [n]$. Let $\hat{S}$ be an $r$-dimensional subspace in $\mathbb{R}^m$ such that there exists an $r$-dimensional subspace $\bar{S}$ of $\mathcal{R}(A_{J_0};X_0^{J_0})$ satisfying $\|P_{\bar{S}} - P_{\hat{S}}\| \leq \eta$. Suppose that $X_0^{J_0}$ is row-nondegenerate and $A$ satisfies

$$\alpha < \alpha_{s+1}^{\text{weak}}(A; J_0) \leq \beta_{s+1}^{\text{weak}}(A; J_0) < \beta$$  \hspace{1cm} (3.6.11)

for $\alpha$ and $\beta$ satisfying

$$1 - \sqrt{1 - \frac{\alpha^2}{\|A^*\|_{2,\infty}^2}} \geq \frac{2\eta \beta}{\alpha - \eta \beta}$$  \hspace{1cm} (3.6.12)

and

$$\sqrt{\frac{r}{s}} \cdot \frac{\alpha}{\|A^*\|_{2,\infty}} - \sqrt{1 - \frac{\alpha^2}{\|A^*\|_{2,\infty}^2}} \geq \frac{2\eta \beta}{\alpha - \eta \beta}.$$  \hspace{1cm} (3.6.13)

Then SA-MUSIC+OSMP applied to $\hat{S}$ will identify $J_0$.

**Remark 3.6.12.** With normalized $A$, Condition (3.6.12) is implied by Condition (3.6.13), which means that partial support recovery by OSMP requires more stringent conditions than the subsequent MUSIC step in the guarantee of SA-MUSIC+OSMP. This results in the same guarantee for SA-MUSIC+OSMP as for OSMP alone for recovery of up to $s - r$ components. However, in the numerical experiments in Section 3.8, SA-MUSIC+OSMP performed substantially better than OSMP alone. To interpret this, we compare SA-MUSIC and OSMP conditioned on the event that a correct partial
support of size $s - r$ has been found. Theorem 3.6.10 implies that with known partial support, the remaining steps of OSMP are guaranteed by

$$\frac{\alpha}{\|A^\ast\|_{2,\infty}} - \sqrt{1 - \frac{\alpha^2}{\|A^\ast\|_{2,\infty}^2}} \geq \frac{2\eta\beta}{\alpha - \eta\beta}.$$ (3.6.14)

With normalized $A$, the asymmetric weak-1 RIP of $A$ in (3.6.11) for $\alpha$ and $\beta$ satisfying (3.6.14) is implied by $\delta_{s+1}^{\text{weak}}(A; J_0) < \delta$ for $\delta$ satisfying

$$\eta \leq \sqrt{\frac{1 - \delta}{1 + \delta}} \cdot \frac{\sqrt{1 - \delta} - \sqrt{\delta}}{2 + \sqrt{1 - \delta} - \sqrt{\delta}}.$$ (3.6.15)

In contrast, as shown in Fig. 3.2, when a partial support of size $s - r$ is given, the condition in (3.6.7) for the guaranteed success of SA-MUSIC is substantially milder than (3.6.15). Note though that neither (3.6.7) nor (3.6.15) depend on $r$, since an oracle partial support is given.

Figure 3.2: Comparison of SA-MUSIC and OSMP when a partial support of size $s - r$ is given: tradeoff between parameter $\delta$ (for the weak-1 RIP) and $\eta$ (for the subspace estimate perturbation). The region below the curve provides a guarantee.

**Remark 3.6.13.** With normalized $A$, the condition in (3.6.11), (3.6.12), and (3.6.13) is implied by $\delta_{s+1}^{\text{weak}}(A; J_0) < \delta$ for $\delta$ satisfying

$$\eta \leq \sqrt{\frac{1 - \delta}{1 + \delta}} \cdot \frac{\sqrt{r/s} \sqrt{1 - \delta} - \sqrt{\delta}}{2 + \sqrt{r/s} \sqrt{1 - \delta} - \sqrt{\delta}}.$$ (3.6.16)
Furthermore, if we assume \( \eta = 0 \), then the guarantee for SA-MUSIC+OSMP only requires

\[
\delta_{s+1}(A; J_0) < \frac{r}{r + s}
\]  

(3.6.17)

which, as shown in Fig. 3.3, becomes less demanding as \( r/s \) increases.

Figure 3.3: Required \( \delta \) (for the weak-1 RIP) for the guarantee of SA-MUSIC+OSMP in Theorem 3.6.11 in the noiseless case (\( \eta = 0 \)). The dot at the top right of the plot corresponds to \( \alpha_{s+1}^{\text{weak}}(A; J_0) > 0 \), or \( \delta_{s+1}^{\text{weak}}(A; J_0) < 1 \), required in the full row rank case, \( r = s \) (in this case SA-MUSIC+OSMP reduces to MUSIC).

For the noisy case, Condition (3.6.16) provides a tradeoff between the parameters \( \delta \) and \( \eta \) for the guarantee of SA-MUSIC+OSMP, which is visualized in Fig. 3.4. As \( r/s \) increases, the admissible \( \delta \) for the guarantee becomes larger for a fixed subspace estimate perturbation level \( \eta \). In other words, the guarantee of SA-MUSIC+OSMP benefits from the higher dimension of the estimated signal subspace \( \hat{S} \).

3.6.5 Implication of the Guarantees in Terms of the Weak-1 RIP to the Oversampling Factor

The results in the previous subsections were stated in terms of the weak-1 RIP. Given an upper bound on \( \delta_{s+1}^{\text{weak}}(A; J_0) \) (or bounds on \( \alpha_{s+1}^{\text{weak}}(A; J_0) \) and \( \beta_{s+1}^{\text{weak}}(A; J_0) \)), Section 3.5 then provides explicit conditions on the parameters \( n, m, s \) that provide the weak-1 RIP for the matrices \( A \) discussed there.

Example 1 (i.i.d. Gaussian Matrix \( A \), Asymptotic Case): In the first example, we consider asymptotic analysis with an i.i.d. Gaussian \( A \). By
Figure 3.4: Tradeoff between parameters $\delta$ (for the weak-1 RIP) and $\eta$ (for the subspace estimate perturbation) for the guarantee of SA-MUSIC+OSMP in Theorem 3.6.11 in the noisy case. Values $(\eta, \delta)$ in the region below the curve provide a guarantee.

Proposition 3.5.5, for $\gamma \in (0, 1)$,

$$1 - \gamma < \alpha_{s+1}^{\text{weak}}(A; J_0) \leq \beta_{s+1}^{\text{weak}}(A; J_0) < 1 + \gamma$$

holds with probability 1, provided that $n$ and $s$ go to infinity while satisfying $n = o(e^s)$ (i.e., $s$ grows faster than $\ln n$) and

$$m > \left( \frac{1}{\gamma^2} \right) s.$$ 

Furthermore, in this asymptotic, $\|A^*\|_{2,\infty} = 1$ with probability 1. Suppose that the estimated signal subspace $\hat{S}$ is error-free ($\eta = 0$). Then, for the full row rank case (i.e., $\dim(\hat{S}) = s$), all SA-MUSIC algorithms reduce to MUSIC without partial support recovery and are guaranteed by $m > s$, which is also a necessary condition for the support recovery. On the other hand, if $\dim(\hat{S}) = r < s$, then, by Theorem 3.6.11, SA-MUSIC+OSMP is guaranteed by $\alpha_{s+1}^{\text{weak}}(A; J_0) \leq \sqrt{s + r}$, which is implied by

$$m > \left( 1 + \frac{s}{r} \right)^2 s.$$ 

Unfortunately, this condition does not converge to $m > s$ as $r/s \to 1$. The discontinuity at $r/s = 1$ is a limitation of the current analysis in this chapter. However, it provides a valid upper bound on the oversampling factor $m/s$ in
the given asymptotic.

**Example 2 (Partial Fourier Matrix $A$, Non-asymptotic Case):** In the second example, we perform the analysis with a partial Fourier matrix that arises in practical applications (e.g., spectrum blind sampling or DOA estimation). Note that a partial Fourier matrix $A$ has normalized columns (i.e., $\|A^*\|_{2,\infty} = 1$). Given $\delta$, by Proposition 3.5.9, $\delta^\text{weak}_{s+1}(A; J_0) < \delta$ holds with probability $1 - \epsilon$, provided that

$$m \geq C_{\delta} \left[ \ln(n-s) + \ln(s+1) + \ln(2\epsilon^{-1}) \right] (s+1) \quad (3.6.18)$$

where the constant factor $C_{\delta}$ is given by $C_{\delta} = \frac{2(3+\delta)}{3\delta^2}$.

Recall that the admissible $\delta$ for given $\eta$ and $r/s$ is computed by (3.6.16). Substituting $\delta$ into these expressions determines the explicit scaling of $m$ versus $s$ and the other problem parameters that will provide guaranteed recovery.

It follows from (3.6.16) that, as $r/s$ increases, $\delta$ increases, and hence, the constant factor $C_{\delta}$ in (3.6.18) decreases. In particular, in the noiseless and full row rank case ($\eta = 0$, $r = s$), SA-MUSIC reduces to MUSIC without need of any partial support recovery and hence, is guaranteed by $e^\text{weak}_{s+1}(A; J_0) > 0$, which corresponds to $m > s$ without the $\ln n$ factor. Fig. 3.5 visualizes $C_{\delta}$ as a function of $\eta$ for varying $r/s$. Recall that the values shown in Fig. 3.5 for $C_{\delta}$ are conservative, because they have been derived from a sufficient condition for the guarantee. Nonetheless, when $r/s \geq 0.7$ and $\eta \leq 0.05$, $C_{\delta}$ has a modest value of less than 10. In contrast, as discussed in Section 3.5.4, for the same case of a partial Fourier matrix, the uniform RIP requires $m \geq C\delta^{-2} s \ln^4 n$ with $C \leq 17, 190$. Therefore, the guarantee of SA-MUSIC+OSMP is better than other guarantees given in terms of the uniform RIP (for example, the guarantee of $\ell_1$ minimization in the SMV case) not only in the order, but also in the constant factor. The relation between $\eta$ and the number of snapshots $N$ is investigated in the next section.

### 3.7 Analysis of Signal Subspace Estimation

Unlike the previous work in sensor array processing (cf. [116]), which relies on asymptotics, we analyze the perturbation in the estimate of the signal...
Figure 3.5: Constant factor $C_\delta$ in (3.6.18) for the guarantee of SA-MUSIC+OSMP when $A$ is a partial Fourier matrix. Values $(\eta, C_\delta)$ in the region above the curve provide a guarantee.

subspace with finitely many observations. Combined with the results in Section 3.6, this analysis provides non-asymptotic guarantees in the noisy case for the new proposed algorithms directly in terms of the measurement noise. The results also enable us to extend the previous performance guarantees [7, 26, 33] of MUSIC to the noisy and finite snapshot case, which was missing before.

**Assumption 1 (A1: Measurement Model and Noise):** Given row $s$-sparse $X_0 \in \mathbb{K}^{n \times N}$ with support $J_0$, the snapshot matrix $Y \in \mathbb{K}^{m \times N}$ is obtained with the sensing matrix $A \in \mathbb{K}^{m \times n}$ as

$$Y = A_{J_0}X_0^{J_0} + W$$

where $A_{J_0}$ has full column rank, and the columns of $W$ are independent realizations of Gaussian vector $w \in \mathbb{K}^m$ with $\mathbb{E}w = 0$ and $\mathbb{E}ww^* = \sigma_w^2I_m$. For the complex field case, we assume a circular Gaussian distribution.

**Assumption 2 (A2: Number of Snapshots):** We assume that the number of snapshots $N$ is large but finite, more specifically, $N$ satisfies $N \geq m$. It is also assumed that $m > s$, which is required for support recovery by any method.

Our assumption on the number of snapshots is motivated by the following considerations. In compressed sensing, the goal is usually to minimize the number of expensive measurements. Now, in certain applications of
joint sparse recovery, taking many snapshots is a rather trivial task compared to acquiring many measurements in a single snapshot. For example, in spectrum-blind sampling [26], the number of measurements per snapshot, $m$, determines the sampling rate, the increase of which is usually expensive and limited by hardware. In contrast, taking many snapshots only results in delay in the support recovery and is usually less expensive than raising the sampling rate. Similarly, in DOA estimation [47] and in distributed sensor networks [124], increasing $m$ requires more sensors, which is expensive, whereas increasing the number of snapshots $N$ corresponds to delay in estimation, which is relatively less expensive. This motivates the setting $N \geq m$ in the analysis of this subsection.

Assumption 3 (A3: Signal): We assume that the nonzero rows of $X_0$ follow the mixed multichannel model given by:

$$X_0^{J_0} = \Psi \Lambda \Phi$$

where $\Psi \in \mathbb{K}^{s \times M}$ with $M \leq s$ is a mixing matrix that has full column rank; $\Lambda \in \mathbb{R}^{M \times M}$ is a deterministic, positive, and diagonal matrix; and the elements of $\Phi \in \mathbb{K}^{M \times N}$ are independent zero mean and unit variance Gaussian random variables. Note that $\text{rank}(X_0^{J_0}) = \text{rank}(\Phi) = M$ with probability 1. We assume that $\Phi$ is independent of $W$. The rows of $\Lambda \Phi$ correspond to realizations of $M$ statistically independent sources, where the diagonal entries of $\Lambda$ represent the magnitudes of the sources. The columns of $X_0^{J_0}$ in this model are independent realizations of Gaussian vector $x \in \mathbb{K}^s$ with $\mathbb{E}x = 0$ and $\mathbb{E}xx^* = \Psi \Lambda^2 \Psi^*$.

The mixed multichannel model generalizes the multichannel model [62] proposed for the average case analysis of various methods for joint sparse recovery. With $\Psi = I_s$, the mixed multichannel model reduces to the multichannel model. However, with a rectangular mixing matrix $\Psi \in \mathbb{K}^{s \times M}$ for $M < s$, the mixed multichannel model can describe the “rank defect”, which is due to the correlation between the mixed sources (i.e., between the rows of $X_0^{J_0}$). Such correlation, which often arises in the aforementioned applications, cannot be represented by the multichannel model, in which $X_0^{J_0}$ has full row rank with probability one for $N \geq s$.

With the mixed multichannel model, the $N$ columns of $A_{J_0}X_0^{J_0} \in \mathbb{K}^{m \times N}$ are independently distributed Gaussian vectors with zero mean and covari-
Assumption 4 \((A4: \text{Covariance Matrix})\): We assume that a significant gap exists between at least one pair of consecutive eigenvalues of \(\Gamma\), more specifically, the covariance matrix \(\Gamma\) satisfies the following conditions given by the parameters \(r \in \mathbb{N}\) and \(\tau, \nu, \theta \in (0, 1)^{90}\):

\[
(1 - \theta)\lambda_r(\Gamma) - (1 + \theta)\lambda_{r+1}(\Gamma) \geq (1 + \theta)(1 + \nu)\tau\lambda_1(\Gamma) \tag{3.7.1}
\]
\[
(1 + \theta)\lambda_k(\Gamma) - (1 - \theta)\lambda_{k+1}(\Gamma) < (1 - \theta)(1 - \nu)\tau\lambda_1(\Gamma), \quad \forall k > r. \tag{3.7.2}
\]

Condition (3.7.1) asserts that a significant gap exists between two consecutive eigenvalues \(\lambda_r(\Gamma)\) and \(\lambda_{r+1}(\Gamma)\). Condition (3.7.2) asserts that a significant gap does not exist between any two consecutive eigenvalues smaller than \(\lambda_r(\Gamma)\). Together, the two conditions imply that \(r\) is the maximal value that satisfies (3.7.1). (A gap cannot be both big enough and small enough at the same time.)

If \(\Gamma\) is well conditioned with the condition number \(\kappa(\Gamma)\) satisfying

\[
\kappa(\Gamma) \triangleq \frac{\lambda_1(\Gamma)}{\lambda_M(\Gamma)} \leq \frac{1 - \theta}{(1 + \theta)(1 + \nu)\tau}
\]

then \(r\) that satisfies (3.7.1)–(3.7.2) will assume its maximal value of \(r = \text{rank}(\Gamma) = M\). In this case \(\lambda_{r+1}(\Gamma) = 0\), and (3.7.2) is trivially satisfied. Otherwise, we assume that \(\Gamma\) is ill-conditioned with one or more insignificant eigenvalues and set \(r\) to the index of the smallest eigenvalue larger than those considered insignificant. In this case, (3.7.2) implies that \(\lambda_{r+1}(\Gamma)\) is bounded from above by

\[
\lambda_{r+1}(\Gamma) < (1 - \nu)\tau\lambda_1(\Gamma) \sum_{\ell=1}^{M-r} \left(\frac{1 - \theta}{1 + \theta}\right)^\ell.
\]

The two cases for \(\Gamma\) considered above do not cover all possibilities. However, if neither assumption holds, then it will be difficult to obtain a good estimate of the signal subspace from the noisy measurements.

**Proposition 3.7.1.** Suppose that Assumptions \(A1-A4\) hold and define

\[
C_{\eta, \nu, \theta, \tau} \triangleq (1 + \theta)\tau \min \left\{\frac{(1 + \nu)\eta}{3}, \frac{\nu}{2 + \tau}\right\}. \tag{3.7.3}
\]
Let $\bar{S}$ be the subspace spanned by the $r$ dominant eigenvectors of $\Gamma_S$ defined by

$$\Gamma_S \triangleq \frac{A_{j_0} X_{j_0}^* (X_{j_0}^j) A_{j_0}^*}{N}.$$ 

Let $\epsilon, \eta \in (0, 1)$. If the number of snapshots $N$ satisfies

1. $N > 2(m + s)$ \hspace{1cm} (3.7.4)
2. $N \geq \left(\frac{36}{\theta^2}\right) \left[s + \ln \left(\frac{8}{\epsilon}\right)\right]$ \hspace{1cm} (3.7.5)
3. $N \geq \left(\frac{144}{C_{\eta, \nu, \theta, \tau}^2}\right) \left(\frac{\sigma_w^2}{\lambda_1(\Gamma)} + 2\sqrt{\frac{\sigma_w^2}{\lambda_1(\Gamma)}}\right) \left[m + s + \ln \left(\frac{8}{\epsilon}\right)\right]$ \hspace{1cm} (3.7.6)

then with probability $1 - \epsilon$, Algorithm 7 with parameter $\tau$ computes an $r$-dimensional subspace $\bar{S}$ such that

$$\|P_S - P_{\bar{S}}\| \leq \eta.$$ \hspace{1cm} (3.7.7)

**Proof.** See Appendix B.10. \hfill \Box

We are interested in achieving small perturbation $\eta$ when the noise level is moderate (i.e., the noise variance $\sigma_w^2$ is less than $\lambda_1(\Gamma)$). As $\eta$ decreases to 0 with the model parameters $\theta$, $\nu$, and $\tau$ remaining constant, $C_{\eta, \nu, \theta, \tau}$ becomes proportional to $\eta$. In this case, Condition (3.7.6) dominates the first two and is simplified as

$$N = O \left(\frac{(\sigma_w^2/\lambda_1(\Gamma))^{1/2} [m + s + \ln(8/\epsilon)]}{\eta^2}\right)$$

so that the number of snapshots required for the guarantee scales linearly in $m$. Alternatively, in the same scenario, Proposition 3.7.1 implies that (3.7.7) holds for

$$\eta = O \left(\left(\frac{\sigma_w^2}{\lambda_1(\Gamma)}\right)^{1/4} \sqrt{\frac{m + s + \ln(8/\epsilon)}{N}}\right).$$

To help interpret this condition, define the average per-sample signal-to-noise ratio (SNR) as the ratio of the powers of the measured signal and noise

$$\text{SNR} \triangleq \frac{\mathbb{E} \| A_{j_0} X_{j_0} \|_F^2}{\mathbb{E} \| W \|_F^2} = \frac{\sum_{k=1}^M \lambda_k(\Gamma)}{m \sigma_w^2}.$$
Then, $\lambda_1(\Gamma)/\sigma_w^2$ is related to the SNR by

$$
\frac{\lambda_1(\Gamma)}{\sigma_w^2} = \left( \frac{\lambda_1(\Gamma)}{\frac{1}{m}\text{tr}(\Gamma)} \right) \cdot \text{SNR}.
$$

It follows that for fixed $\Gamma$ and SNR, $\eta$ scales proportionally to $N^{-1/2}$. With more snapshots, SA-MUSIC algorithms access a more accurate estimate of the signal subspace (smaller $\eta$); hence, as shown in Fig. 3.4 (in the case of SA-MUSIC+OSMP), the admissible $\delta$ increases, resulting in a decrease of the required oversampling factor $m/s$. Eventually, as $N$ goes to infinity, the performance converges to that in the noiseless case.

### 3.8 Numerical Experiments

We compared the performance of the proposed SA-MUSIC algorithm (SA-MUSIC+OSMP) versus MUSIC [26], M-BP,\(^8\) and OSMP. As an upper bound on the performance of SA-MUSIC, we included in the comparison SA-MUSIC with known (“oracle”) partial support. The sensing matrix $A$ was generated as randomly selected $m$ rows of the $n \times n$ DFT matrix. The snapshot matrix $Y = AX_0 + Z$ was corrupted by additive i.i.d. circular complex Gaussian noise $Z$.

The algorithms were tested on random row $s$-sparse $X_0$ of rank $r$ less than $s$, with uniformly random support $J_0$. The nonzero rows $X_{0,J_0}^r$ follow the mixed multichannel model $X_{0,J_0}^r = \Psi \Lambda \Phi$. In order to observe the effect of the rank defect rather than ill-conditioning, in the first experiment, $\Psi \in \mathbb{C}^{s \times r}$ and $\Lambda$ were set to random orthonormal columns and the identity matrix, respectively, so that the resulting covariance matrix of the mixed multichannel model is ideally conditioned. The performance is assessed by the rate of successful support recovery.\(^9\)

As shown in Fig. 3.6, MUSIC fails when $\text{rank}(X_{0,J_0}^r) < s$. The performance of SA-MUSIC+OSMP and M-BP improves with increasing rank $r$. SA-MUSIC+OSMP performed better than M-BP in this experiment, at much

---

\(^8\)In the experiment, the noise variance is given to the M-BP algorithm; hence, the performance of M-BP is no worse than that of Group LASSO.

\(^9\)M-BP does not produce an $s$-sparse solution in the presence of noise. In this case, the solution by M-BP has been approximated to the nearest $s$-sparse vector, and the support is computed as that of the $s$-sparse approximation.
Figure 3.6: Test on rank defect, $n = 128$, $s = 8$, $N = 256$, Left column noiseless data; right column SNR = 30 dB: (a),(b) rank($X_{j0}^0$) = 4. (c),(d) rank($X_{j0}^0$) = 6. (e),(f) rank($X_{j0}^0$) = 8 (full row rank).

lower computational cost. SA-MUSIC with known partial support of size $s - r$ is labeled as “SA-MUSIC+Oracle” in Fig. 3.6 and shows perfect recov-
ery when \( m > s + 1 \), which is nearly the necessary condition \( m > s \). This suggests that the success of partial support recovery is more critical than the subsequent steps and leaves room for improving SA-MUSIC by combining with a better algorithm for partial support recovery than OSMP.

The performance of SA-MUSIC+OSMP and OSMP coincides for the noiseless case (left column of Fig. 3.6). However, with noise, the performance of OSMP severely degrades even for the full row rank case (right column of Fig. 3.6). For the full row rank case, all algorithms except OSMP (noisy case) were successful in terms of nearly achieving the necessary condition \( m > s \). Again, OSMP is sensitive to the perturbation in the estimate of signal subspace in this case.

Regarding the computation, we compared the runtime of each algorithm by increasing the size of the problem. In this experiment, fixing \( n = (\text{scale factor}) \times 64 \), we set the other parameters to \( s = n/16 \), \( r = \lceil 7s/8 \rceil \), and \( m = 2s \). As shown in Fig. 3.7, SA-MUSIC+OSMP is about 100 times faster than M-BP.\(^\text{10}\)

![Figure 3.7: Comparison of runtime \( N = 256 \), SNR = 30 dB.](image)

Fig. 3.8 compares the performance of the algorithms with varying number of snapshots for the case \( r/s = 6/8 \). Although the analysis in Section 3.7 assumed that \( N \) is large, the proposed subspace methods also performed well empirically with a small number of snapshots. In particular, compared to M-BP, MUSIC (for the full row rank case) and SA-MUSIC+OSMP performed

\(^\text{10}\)For M-BP, we used an efficient implementation SPGL1 [125, 126]. On the other hand, the other methods were implemented as plain Matlab script. Therefore, the speed comparison does not unfairly favor SA-MUSIC.
better with fewer snapshots. Anyway, as expected, the performance for the noisy case improves with increasing $N$. Figs. 3.6(c) and (d) correspond to the same scenario, with $N = 256$, and show the same trend. Again, as in the previous experiment, OSMP degraded severely in the presence of noise, and MUSIC performed poorly in the rank-defective case.

![Graphs showing performance comparison](a)-(d)

Figure 3.8: Test on varying $N$, $n = 128$, $s = 8$, $r = 6$. Left column noiseless; right column SNR = 30 dB: (a),(b) $N = 16$. (c),(d) $N = 64$.

In order to see the effect of ill-conditioning, in the second experiment, we tested the algorithms on the mixed multichannel model with $\Psi$ of full row rank and $\Lambda$ having geometrically decaying eigenvalues. We set the $k$th diagonal entry of $\Lambda$ as $\kappa^{-(k-1)/(s-1)}$ for $k = 1, \ldots, s$, so that the condition number of $\Lambda$ becomes $\kappa$. The mixing matrix $\Psi$ was generated randomly as in the first experiment. Fig. 3.9 compares the performance of the algorithms for the weak noise case. We note that M-BP is sensitive to the
ill-conditioning of $\Lambda$. When $\Lambda$ is well conditioned with $\kappa = 10$, the dimension of the estimated signal subspace is equal to the row rank of $X_0^J$ ($= s$); hence, SA-MUSIC+OSMP reduces to MUSIC without the partial support recovery step. However, when $\kappa = 50$, the estimated rank $r$ by using (3.3.4) in Algorithm 7 with $\tau = 10^{-3}$ is smaller than rank($X_0^J$); as a consequence, MUSIC suffers from the rank defect while SA-MUSIC+OSMP provides consistent performance.

3.9 Discussion

3.9.1 Comparison to Compressive MUSIC

An algorithm similar to SA-MUSIC named “compressive MUSIC” (CS-MUSIC) has been independently proposed by Kim et al. [81]. Although the main ideas underlying the SA-MUSIC and CS-MUSIC algorithms are similar, in fact, the two studies differ in the following significant ways.

First, the algorithms considered are different in several respects as follow:

1. The forward greedy algorithms for partial support recovery are different. The criteria for the update of OSMP in SA-MUSIC and the update in CS-MUSIC [81] maximize $\| (P_{P_{R(A)}} \hat{a})_k \|_2 / \| P_{P_{R(A)}} a_k \|_2$ and $\| P_{P_{R(A)}} \hat{a} \|_2 / \| P_{P_{R(A)}} a_k \|_2$ and

Figure 3.9: Test on large condition number, $n = 128$, $s = 8$, rank($X_0^J$) = $s$ (full row rank), $N = 256$, SNR = 30 dB. (a) $\kappa = 10$. (b) $\kappa = 50$. 

}\end{verbatim}
\[(P_{R(A)\perp}a_k)\|_2\}, \text{ respectively. The OSMP terms introduce the normalization of } P_{R(A)\perp}a_k \text{ by its } \ell_2 \text{ norm, which is missing in the greedy algorithm in CS-MUSIC [81]. Owing to this normalization, the OSMP criterion uses a valid metric between two subspaces, while the greedy algorithm in CS-MUSIC [81] does not.}

2. Given partial support \( J_1 \subset J_0 \), like SA-MUSIC, CS-MUSIC also constructs an augmented subspace \( \tilde{S} = \tilde{S} + R(A_{J_1}) \). However, the criteria that determine the remaining support elements are different. SA-MUSIC and CS-MUSIC maximize \( \|P_{\tilde{S}}a_k\|_2/\|a_k\|_2 \) and \( \|P_{\tilde{S}}a_k\|_2 \), respectively. Again, the difference in the normalization implies that the SA-MUSIC criterion is based on the subspace metric while that of CS-MUSIC is not.\(^\text{11}\)

3. SA-MUSIC and CS-MUSIC differ in the estimation schemes of the signal subspace. In an ideal case where \( X_{J_0} \) has full column rank (hence \( N \leq s \)) and \( Y \) is noise-free, a perfect signal subspace estimate is trivially computed as the range space of \( Y \). Otherwise, a signal subspace estimate can be computed by a truncated SVD. However, to get a reliable signal subspace estimate, it is important to determine the dimension of the estimate carefully. We propose and analyze an algorithm for the signal subspace estimation in this perspective. Such an estimation scheme is missing in [81], which focused more on the scenario where \( N < s \).

Second, the analyses in the two studies are fundamentally different. The analysis of Kim et al. [81] depends heavily on the assumption that \( A \) is an i.i.d. Gaussian matrix and the size of the problem goes to infinity satisfying certain scaling laws. The authors showed that under certain conditions, the probability of failure of the support recovery converges to 0 in their “large system model”. However, since no convergence rate is shown, the analysis provides no guarantee on any finite dimensional problem. In contrast, the guarantees in this chapter are based on the weak-1 RIP and are non-asymptotic. Our guarantees provide explicit formulae for the required \( m \) as functions of \( s \) and \( n \), for various sensing matrices \( A \), including i.i.d. Gaussian

\(^\text{11}\)Unlike the CS-MUSIC algorithm, which must always use finite data, the theoretical analysis in [81] is not affected by this issue, because in the large system model \( \|a_k\|_2 = 1 \), and the two criteria coincide.
and random partial Fourier, whereas the analysis in [81] only applies to an i.i.d. Gaussian $A$.

Finally, the comparison to MUSIC [26], which is the most relevant previous work, is missing in the numerical results of [81]. In fact, in the regime where CS-MUSIC dominates other methods in [81, Fig. 5], CS-MUSIC coincides with MUSIC [26] since $X^J_0$ has full row rank. However, this is not shown. Including MUSIC in the comparison would reveal that CS-MUSIC has only a marginal advantage over MUSIC in the scenario studied there. Different scenarios would have to be studied to better characterize CS-MUSIC. In contrast, we studied the cases where MUSIC is not successful due to either rank defect and/or ill-conditioning. SA-MUSIC improves on MUSIC in the sense that SA-MUSIC performs well in such settings which are unfavorable to MUSIC.

3.9.2 Comparison to the Guarantee of M-BP with the Multichannel Model

Various practical algorithms including $p$-SOMP, $p$-thresholding, and M-BP, have been analyzed under the multichannel model [62, 80]. Although it is restricted to the noiseless case, the average case guarantee of M-BP with the multichannel model [80] has been shown to be better than the other guarantees of the same kind for other algorithms. Therefore, we compare the guarantees of SA-MUSIC algorithms to that of M-BP.

For this comparison, we too assume that the snapshots are noise-free (i.e., $Y = X^J_0 A_Jo$). Nevertheless, the guarantee of SA-MUSIC algorithms in this chapter is restricted neither to the noiseless case nor to the multichannel model.

In the noiseless case, the signal subspace estimation is perfect $\hat{S} = S \triangleq R(A_Jo X^J_0)$, with $r \triangleq \dim(\hat{S}) = \text{rank}(X^J_0)$. If $N \geq s$, where $s$ is the sparsity level, then $X^J_0$ following the multichannel model has full row rank with probability 1. In the full row rank case, any SA-MUSIC algorithm reduces to MUSIC and provides the best possible guarantee with the minimal requirement $\alpha^\text{weak}_{s+1}(A; J_0) > 0$, which reduces to $m > s$ for certain matrices such as i.i.d. Gaussian $A$. This completes the comparison in the case $N \geq s$. Therefore, to compare the performance of SA-MUSIC and M-BP in the rank-
defective case, we assume that $N < s$. The rank of $X_J^0$ is then determined by the number of snapshots (i.e., $\text{rank}(X_J^0) = N$); hence, $r = N$.

Previous work [80, Theor. 4.4] showed that M-BP is guaranteed with probability $1 - \epsilon$ if $A$ satisfies

$$\delta_{s+1}^{\text{weak}}(A; J_0) < \delta$$

for $\delta$ satisfying

$$\left(\frac{\delta}{1 - \delta}\right)^2 + 2 \ln\left(\frac{\delta}{1 - \delta}\right) \geq \frac{2 \ln(n/\epsilon)}{N} + 1.$$  

Figure 3.10: Required weak-1 RIC for the guarantees of M-BP (average case analysis with the multichannel model with error probability $\epsilon = 10^{-3}$), and SA-MUSIC (worst case analysis) for the noiseless case (a) $n = 128, s = 8$. (b) $n = 1024, s = 64$.

On the other hand, SA-MUSIC+OSMP is guaranteed by the weak-1 RIP and the row-nondegeneracy condition. In particular, when $X_J^0$ follows the multichannel model, it is row-nondegenerate with probability 1. Therefore, we need to only compare the admissible $\delta$ given by (3.6.17) for SA-MUSIC+OSMP to the one given by (3.9.2) for M-BP. Fig. 3.10 displays this comparison.

For both algorithms, as $r/s$ increases, $\delta$ required for the guarantee increases; hence, the guarantee is obtained subject to a milder condition. Fig. 3.10(a) shows that the guarantee of SA-MUSIC+OSMP is satisfied by a larger $\delta$; hence, the guarantee requires reduced oversampling factor $m/s$ compared to M-BP when the problem is small ($n = 128$). Fig. 3.10(b) shows
that SA-MUSIC+OSMP provides a better guarantee (larger $\delta$) than M-BP in the regime $r/s \geq 0.6$ when $n = 1024$.

The theoretical guarantee notwithstanding, in our simulations, the recovery rate of SA-MUSIC+OSMP was always higher than that of M-BP, and often substantially so.

3.9.3 Comparison to the Analysis of Group LASSO in High Dimension

The guarantee of Group LASSO by Obozinski et al. [66] is quite tight and, in particular, achieves the optimal guarantee by the minimal requirement ($m > s$) for certain scenarios. However, their guarantee is asymptotic and only applies to Gaussian $A$. In contrast, although our guarantee of SA-MUSIC+SS-OSMP is not as tight as that of Group LASSO [66], the guarantee is non-asymptotic (i.e., valid for any finite problem), and applies to wider class of matrices that arise in practical applications, including the partial Fourier case.

3.9.4 Comparison to Compressed Sensing with Block Sparsity

The joint sparse recovery problem can be cast as a special case of compressed sensing with block sparsity [27]. The block structure in the sparsity pattern in the latter problem has been exploited to improve the performance of sparse recovery (cf. [27, 102, 124, 127]) over the unstructured original problem. However, the reduction of the joint sparsity problem to the block sparsity problem results in a special case where the sensing matrix is block diagonal with repeated blocks. Therefore, the existing analysis [127] of the block sparsity problem, which did not cover this special case, does not apply to the joint sparsity problem.


CHAPTER 4

ADMIRA: ATOMIC DECOMPOSITION
FOR MINIMUM RANK APPROXIMATION

4.1 Vector vs. Matrix

4.1.1 Preliminaries

Let $m$, $n$, and $p$ be natural numbers. Throughout Chapter 4, we use the following two Hilbert spaces: the space of $p$-tuple vectors $\ell^p_2$, and the space of $m$ by $n$ matrices $S_2$. For $\ell^p_2$, the inner product is defined by $\langle x, y \rangle_{\ell^p_2} \triangleq y^* x$ for $x, y \in \ell^p_2$ where $y^*$ denotes the Hermitian transpose of $y$. For $S_2$, the inner product is defined by $\langle X, Y \rangle_{S_2} \triangleq \text{tr}(Y^* X)$ for $X, Y \in S_2$ where $Y^*$ denotes the Hermitian transpose of $Y$. The induced Hilbert-Schmidt norm for $S_2$ is called the Schatten 2-norm or the Frobenius norm.

4.1.2 Atomic Decomposition

Let $\Gamma$ denote the set of all nonzero rank-one matrices in $S_2$. We can refine $\Gamma$ so that any two distinct elements are not collinear. The resulting subset $\mathcal{O}$ is referred to as the set of atoms\(^1\) of $S_2$. Then the set of atomic spaces $A$ of $S_2$ is defined by $A \triangleq \{ \text{span}(\psi) : \psi \in \mathcal{O} \}$. Each subspace $\mathcal{V} \in A$ is one-dimensional and hence is irreducible in the sense that $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2$ for some $\mathcal{V}_1, \mathcal{V}_2 \in A$ implies $\mathcal{V}_1 = \mathcal{V}_2 = \mathcal{V}$. Since $\mathcal{O}$ is an uncountably infinite set in a finite dimensional space $S_2$, the elements in $\mathcal{O}$ are not linearly independent. Regardless of the choice of $\mathcal{O}$, $A$ is uniquely determined. Without loss of

---

\(^1\)The “atom” in this chapter is different from Mallat and Zhang’s “atom” [97], which is an element in a dictionary, a finite set of vectors. In our terminology, an atom is a rank-one matrix, an element in an infinite set of vectors (in the vector space $S_2$). In both cases, however, an atom denotes an irreducible quantity – a singleton subset, not representable with fewer elements. (Indeed, for each atom $\psi$, the corresponding atomic space $\text{span}(\psi)$ is irreducible.)
generality, we fix \( \mathcal{O} \) such that all elements have unit Frobenius norm.

Given a matrix \( X \in \mathcal{S}_2 \), its representation \( X = \sum_j \alpha_j \psi_j \) as a linear combination of atoms is referred to as an atomic decomposition of \( X \). Since \( \mathcal{O} \) spans \( \mathcal{S}_2 \), an atomic decomposition of \( X \) exists for all \( X \in \mathcal{S}_2 \). A subset \( \Psi \) of unit-norm and pairwise orthogonal atoms in \( \mathcal{O} \) will be called an orthonormal set of atoms.

**Definition 4.1.1.** Let \( \mathcal{O} \) be a set of atoms of \( \mathcal{S}_2 \). Given \( X \in \mathcal{S}_2 \), we define \( \text{atoms}(X) \) as the smallest set of atoms in \( \mathcal{O} \) that spans \( X \),

\[
\text{atoms}(X) \triangleq \arg \min_\Psi \{|\Psi| : \Psi \subset \mathcal{O}, X \in \text{span}(\Psi)\}. \tag{4.1.1}
\]

Note that \( \text{atoms}(X) \) is not unique.

An orthonormal set \( \text{atoms}(X) \subset \mathcal{O} \) is given by the singular value decomposition of \( X \). Let \( X = \sum_{k=1}^{\text{rank}(X)} \sigma_k u_k v_k^\ast \) denote the singular value decomposition of \( X \) with singular values in decreasing order. While \( u_k v_k^\ast \) need not be in \( \mathcal{O} \), for each \( k \), there exists \( \rho_k \in \mathbb{C} \) such that \( |\rho_k| = 1 \) and \( \rho_k u_k v_k^\ast \in \mathcal{O} \). Then an orthonormal set \( \text{atoms}(X) \subset \mathcal{O} \) is given by

\[
\text{atoms}(X) = \{\rho_k u_k v_k^\ast\}_{k=1}^{\text{rank}(X)}.
\]

**Remark 4.1.2.** \( \text{atoms}(X) \), and \( \text{rank}(X) = |\text{atoms}(X)| \) for a matrix \( X \in \mathcal{S}_2 \), are the counterparts of \( \text{supp} (x) \triangleq \{k \in \{1, \ldots, p\} : x_k \neq 0\} \) and \( \|x\|_0 = |\text{supp} (x)| \) for a vector \( x \in \ell_2^n \), respectively.

### 4.1.3 Generalized Correlation Maximization

Recht *et al.* [83] showed an analogy between rank minimization P1 and \( \ell_0 \)-norm minimization. We consider instead the rank-\( r \) matrix approximation problem P2 and its analogue – the \( s \)-term vector approximation problem

\[
P3: \min_{x \in \ell_2^n} \|Ax - b\|_2 \quad \text{subject to } \|x\|_0 \leq s.
\]

In Problem P3, variable \( x \) lives in the union of \( s \) dimensional subspaces of \( \ell_2^n \), each spanned by \( s \) elements in the finite set \( \{e_k\}_{k=1}^n \), the standard basis of \( \ell_2^n \). Thus the union contains all \( s \)-sparse vectors in \( \ell_2^n \). Importantly, finitely
many \( \frac{n!}{(n-s)!s!} \), to be precise) subspaces participate in the union. Therefore, it is not surprising that P3 can be solved exactly by exhaustive enumeration, and finite selection algorithms such as CoSaMP are applicable.

In the rank-\( r \) matrix approximation problem P2, the matrix variable \( X \) lives in the union of subspaces of \( S_2 \), each of which is spanned by \( r \) atoms in the set \( \mathcal{O} \). Indeed, if \( X \in S_2 \) is spanned by \( r \) atoms in \( \mathcal{O} \), then rank(\( X \)) \( \leq r \) by the subadditivity of the rank. Conversely, if rank(\( X \)) = \( r \), then \( X \) is a linear combination of rank-one matrices and hence there exist \( r \) atoms that span \( X \). Note that uncountably infinitely many subspaces participate in the union. Therefore, some selection rules in the greedy algorithms for \( \ell_0 \)-norm minimization and \( s \)-term vector approximation do not generalize in a straightforward way. Nonetheless, using our formulation of the rank-\( r \) matrix approximation problem in terms of an atomic decomposition, we extend the analogy between the vector and matrix cases, and propose a way to generalize these selection rules to the rank-\( r \) matrix approximation problem.

First, consider the correlation maximization in greedy algorithms for the vector case. Matching Pursuit (MP) \[97\] and Orthogonal Matching Pursuit (OMP) \[98\] choose the index \( k \in \{1, \ldots, n\} \) that maximizes the correlation \( |a_k^*(b - A\hat{x})| \) between the \( k \)th column \( a_k \) of \( A \) and the residual in each iteration, where \( \hat{x} \) is the solution of the previous iteration. Given a set \( \Psi \), let \( P_\Psi \) denote the (orthogonal) projection operator onto the subspace spanned by \( \Psi \) in the corresponding embedding space. When \( \Psi = \{\psi\} \) is a singleton set, \( P_\psi \) will denote \( P_\Psi \). For example, \( P_{e_k} \) denotes the projection operator onto the subspace in \( \ell_2^n \) spanned by \( e_k \). From

\[
|a_k^*(b - A\hat{x})| = \left| \langle A^*(b - A\hat{x}), e_k \rangle \right|_2 = \|P_{e_k}A^*(b - A\hat{x})\|_2,
\]

it follows that maximizing the correlation implies maximizing the norm of the projection of the image under \( A^* \) of the residual \( b - A\hat{x} \) onto the selected one dimensional subspace.

The following selection rule generalizes the correlation maximization to the matrix case. We maximize the norm of the projection over all one-dimensional subspaces spanned by an atom in \( \mathcal{O} \):

\[
\max_{\psi \in \mathcal{O}} \left| \langle b - A\hat{X}, A\psi \rangle_{S_2} \right| = \max_{\psi \in \mathcal{O}} \|P_\psi A^*(b - A\hat{X})\|_{S_2}, \quad (4.1.2)
\]
where $A^* : \ell_p^2 \to S_2$ denotes the adjoint operator of $A$. By the Eckart-Young Theorem, the basis of the best subspace is obtained from the singular value decomposition of $M = A^*(b - AX)$, as $\psi = u_1 v_1^*$, where $u_1$ and $v_1$ are the principal left and right singular vectors.

**Remark 4.1.3.** Applying the selection rule (4.1.2) to update $\hat{X}$ recursively leads to greedy algorithms generalizing MP and OMP to rank minimization.

Next, consider the rule in recent algorithms such as CoSaMP and SP. The selection rule chooses the subset $J$ of $\{1, \ldots, n\}$ with $|J| = s$ defined by

$$|a_k^*(b - A\hat{x})| \geq |a_j^*(b - A\hat{x})|, \quad \forall k \in J, \forall j \notin J. \quad (4.1.3)$$

This is equivalent to maximizing

$$\sum_{k \in J} |a_k^*(b - A\hat{x})|^2 = \sum_{k \in J} \|P_{e_k}A^*(b - A\hat{x})\|_{\ell_2^p}^2 = \|P_{\{e_k\}_{k \in J}}A^*(b - A\hat{x})\|_{\ell_2^p}^2.$$

In other words, selection rule (4.1.3) finds the best subspace spanned by $s$ elements in $\{e_k\}_{k=1}^n$ that maximizes the norm of the projection of $n$-tuple vector $\xi = A^*(b - A\hat{x})$ onto that $s$-dimensional subspace.

The following selection rule generalizes the selection rule (4.1.3) to the matrix case. We maximize the norm of the projection over all subspaces spanned by a subset with at most $r$ atoms in $O$:

$$\max_{\Psi \subset O} \left\{ \|P_{\Psi}A^*(b - A\hat{X})\|_{S_2} : |\Psi| \leq r \right\}$$

A basis $\Psi$ of the best subspace is again obtained from the singular value decomposition of $M = A^*(b - AX)$, as $\Psi = \{\rho_k u_k v_k^*\}_{k=1}^r$, where $u_k$ and $v_k$, $k = 1, \ldots, r$ are the $r$ principal left and right singular vectors, respectively, and for each $k$, $\rho_k \in \mathbb{C}$ satisfies $|\rho_k| = 1$.\(^2\) Note that $\Psi$ is an orthonormal set although this is not enforced as an explicit constraint in the maximization.

\(^2\)Once the best subspace is determined, it is not required to compute the constants $\rho_k$'s.
4.2 Algorithm

Algorithm 12: ADMiRA

- **input**: $A: S_2 \rightarrow \ell_p^2$, $b \in \ell_p^2$, and target rank $r \in \mathbb{N}$
- **output**: rank-$r$ solution $\hat{X}$ to P2

1. $\hat{X} \leftarrow 0$;
2. $\hat{\Psi} \leftarrow \emptyset$;
3. **while** stop criterion is false **do**
   4. $\Psi' \leftarrow \arg \max_{\Psi \subset \Omega} \left\{ \| P_{\Psi} A^* (b - A\hat{X}) \|_{S_2} : |\Psi| \leq 2r \right\}$;
   5. $\hat{\Psi} \leftarrow \Psi' \cup \hat{\Psi}$;
   6. $\hat{X} \leftarrow \arg \min_X \left\{ \| b - AX \|_{\ell_p^2} : X \in \text{span}(\hat{\Psi}) \right\}$;
   7. $\hat{\Psi} \leftarrow \arg \max_{\Psi \subset \Omega} \left\{ \| P_{\Psi} \hat{X} \|_{S_2} : |\Psi| \leq r \right\}$;
   8. $\hat{X} \leftarrow P_{\hat{\Psi}} \hat{X}$;
4. **end**
5. **return** $\hat{X}$;

Algorithm 12 describes the ADMiRA algorithm. Intuitively, ADMiRA iteratively refines the pair $(\hat{\Psi}, \hat{X}) \in \Omega \times S_2$ where $\hat{\Psi}$ is the set of $r$ atoms that spans an approximate solution $\hat{X}$ to P2. Step 4 finds a set of $2r$ atoms $\Psi'$ that spans a good approximation of $X_0 - \hat{X}$, which corresponds to the information not explained by the solution $\hat{X}$ in the previous iteration. Here ADMiRA assumes that $A$ acts like an isometry on a low-rank matrix $X_0 - \hat{X}$, which implies that $A^* A$ acts like a (scaled) identity operator on $X_0 - \hat{X}$. Under this assumption, the $2r$ leading principal components of the proxy matrix $A^* (b - A\hat{X}) = A^* A (X_0 - \hat{X})$ are a good choice for $\Psi'$. The quality of a linear approximation of $X_0 - \hat{X}$ spanned by $\Psi'$ improves as iteration goes. This will be quantitatively analyzed in the proof of the performance guarantee. If $\hat{\Psi}$ and $\Psi'$ span good approximations of $\hat{X}$ and $X_0 - \hat{X}$, respectively, then $\hat{\Psi} = \hat{\Psi} \cup \Psi'$ will span a good approximation of $X_0$. Steps 6 and 7 refine the set $\hat{\Psi}$ into a set of $r$ atoms. We first compute a rank-$3r$ approximate solution $\tilde{X}$ and then take its best rank-$r$ approximation to get a feasible solution $\hat{X}$ with rank $r$. In the process, the set $\hat{\Psi}$ of $3r$ atoms is also trimmed to the $r$
atom set $\tilde{\Psi}$ so that it can span an approximate solution $\tilde{X}$ closer to $X_0$.

ADMiRA is guaranteed to converge geometrically to the global optimum when the assumptions of ADMiRA in Section 4.3 are satisfied. However, similarly to the vector case [11], it is more difficult to verify the satisfiability of the assumptions than solve the recovery problem itself, and to date there is no known algorithm to perform this verification. Instead of relying on the theoretical bound on the number of iterations, we use an empirical stopping criterion below. If either the monotone decrease of $\|b - A\tilde{X}\|_{\ell^2_p}/\|b\|_{\ell^2_p}$ is broken or $\|b - A\tilde{X}\|_{\ell^2_p}/\|b\|_{\ell^2_p}$ falls below a given threshold, ADMiRA stops.

In terms of computation, Steps 4 and 7 involve finding a best rank-$2r$ or rank-$r$ approximation to a given matrix (e.g., by truncating the SVD), while Step 6 involves the solution of a linear least-squares problem – all standard numerical linear algebra problems. Step 5 merges two given sets of atoms in $\mathcal{O}$ by taking their union. As described in more detail in Section 4.6, these computations can be further simplified and their cost reduced by storing and operating on the low rank matrices in factored form, and taking advantage of special structure of the measurement operator $A$, such as sparsity.

Most steps of ADMiRA are similar to those of CoSaMP except Step 4 and Step 7. The common feasible set $\mathcal{O}$ of the maximization problems in Step 4 and Step 7 is infinite and not orthogonal, whereas the analogous set $\{e_k\}_{k=1}^n$ in CoSaMP is finite and orthonormal. As a result, the maximization problems over the infinite set $\mathcal{O}$ in ADMiRA are more difficult than those in the analogous steps of CoSaMP, which can be simply solved by selecting the coordinates with the largest magnitudes. Nonetheless, singular value decomposition can solve the maximization problems over the infinite set efficiently.

4.3 Main Results: Performance Guarantee

4.3.1 Rank-Restricted Isometry Property (R-RIP)

Recht et al. [83] generalized the sparsity-restricted isometry property (RIP) defined for sparse vectors to low rank matrices. They also demonstrated “nearly isometric families” satisfying this R-RIP (with overwhelming probability). These include random linear operators generated from i.i.d. Gauss-
sian, or i.i.d. symmetric Bernoulli distributions. For consistency with the analogous result in $\ell_0$-norm minimization, we modify slightly the definition \cite{83} of the restricted isometry constant to express the RIP in terms of the squares of norms, rather than the norms themselves. Given a linear operator $\mathcal{A}: \mathcal{S}_2 \to \ell_2^p$, the rank-restricted isometry constant $\delta_r(\mathcal{A})$ is defined as the minimum constant that satisfies

\[
(1 - \delta_r(\mathcal{A}))\|X\|_2^2 \leq \|\gamma \mathcal{A}X\|_2^2 \leq (1 + \delta_r(\mathcal{A}))\|X\|_2^2,
\]

for all $X \in \mathcal{S}_2$ with $\text{rank}(X) \leq r$ for some constant $\gamma > 0$. Throughout Chapter 4, we assume that the linear operator $\mathcal{A}$ is scaled appropriately so that $\gamma = 1$ in (4.3.1). If $\mathcal{A}$ has a small rank-restricted isometry constant $\delta_r(\mathcal{A}) \ll 1$, then (4.3.1) implies that $\mathcal{A}$ acts like an isometry (scaled by $\gamma$) on the matrices whose rank is equal to or less than $r$. In this case, $\mathcal{A}$ is called a rank-restricted isometry to indicate that the domain where $\mathcal{A}$ is nearly an isometry is restricted to the set of low-rank matrices.

### 4.3.2 Performance Guarantee

Subject to the R-RIP, the Atomic Decomposition for Minimum Rank Approximation Algorithm (ADMiRA) has a performance guarantee analogous to that of CoSaMP.

The following are the assumptions in ADMiRA:

**A1:** The target rank is fixed as $r$.

**A2:** The linear operator $\mathcal{A}$ satisfies $\delta_{4r}(\mathcal{A}) \leq 0.065$.

**A3:** The measurement is obtained by

\[
b = \mathcal{A}X_0 + \nu,
\]

where $\nu$ is the discrepancy between the measurement and the linear model $\mathcal{A}X_0$. No assumptions are made about the matrix $X_0$ underlying the measurement, and it can be arbitrary.

---

\( ^3 \)If $\gamma \neq 1$, then the noise term in (4.3.3) needs to be scaled accordingly.

\( ^4 \)In fact, for a geometric convergence of ADMiRA, it suffices to satisfy $\delta_{4r}(\mathcal{A}) < 0.26$. To make the constant for the geometric decay less than 0.5, we assume $\delta_{4r}(\mathcal{A}) \leq 0.065$. 

107
Assumption A2 plays a key role in deriving the performance guarantee of ADMiRA: it enforces the rank-restricted isometry property of the linear operator $A$. Although the verification of the satisfiability of A2 is as difficult as or more difficult than the recovery problem itself, as mentioned above, nearly isometric families that satisfy the condition in A2 have been demonstrated [83].

The performance guarantees are specified in terms of a measure of inherent approximation error, termed *unrecoverable energy* defined by

$$
\epsilon = \|X_0 - \Pi_r(X_0)\|_{S_2} + \frac{1}{\sqrt{r}}\|X_0 - \Pi_r(X_0)\|_{S_1} + \|\nu\|_{\ell^p_2},
$$

where $\Pi_r : S_2 \rightarrow S_2$ denotes the projection onto the set of all rank-$r$ matrices and $\Pi_r(X_0)$ denotes the best rank-$r$ approximation of $X_0$. The first two terms in $\epsilon$ define a metric of the minimum distance between the “true” matrix $X_0$ and a rank-$r$ matrix. This is analogous to the notion of a measure of compressibility of a vector in sparse vector approximation. By the Eckart-Young-Mirsky theorem [128], no rank-$r$ matrix can come closer to $X_0$ in this metric. In particular, the optimal solution to P2 cannot come closer to $X_0$ in this metric. The third term is the norm of the measurement noise, which must also limit the accuracy of the approximation provided by a solution to P2.

**Theorem 4.3.1.** Let $\hat{X}_k$ denote the estimate of $X_0$ in the $k$th iteration of ADMiRA. For each $k \geq 0$, $\hat{X}_k$ satisfies the following recursion:

$$
\|X_0 - \hat{X}_{k+1}\|_{S_2} \leq 0.5\|X_0 - \hat{X}_k\|_{S_2} + 7\epsilon,
$$

where $\epsilon$ is the unrecoverable energy. From the above relation, it follows that

$$
\|X_0 - \hat{X}_k\|_{S_2} \leq 2^{-k}\|X_0\|_{S_2} + 14\epsilon, \quad \forall k \geq 0.
$$

Theorem 4.3.1 shows the geometric convergence of ADMiRA.
4.3.3 Relationship between P1, P2, and ADMiRA

The approximation \( \hat{X} \) given by ADMiRA is a solution to P2. When there is no noise in the measurement, i.e., \( b = AX_0 \), where \( X_0 \) is the solution to P1, Theorem 4.3.1 states that if the ADMiRA assumptions are satisfied with \( r \geq \text{rank}(X_0) \), then \( \hat{X} = X_0 \). An appropriate value can be assigned to \( r \) by an incremental search over \( r \).

For the noisy measurement case, the linear constraint in P1 is replaced by a quadratic constraint and the rank minimization problem is written as:

\[
P1': \min_{X \in S_2} \text{rank}(X) \quad \text{subject to} \quad \|AX - b\|_{\ell^2} \leq \eta.
\]

Let \( X' \) denote a minimizer to P1'. In this case, the approximation \( \hat{X} \) produced by ADMiRA is not necessarily equivalent to \( X' \), but by Theorem 4.3.1 the distance between the two is bounded by \( \|X' - \hat{X}\|_{S_2} \leq 17\eta \) for all \( r \geq \text{rank}(X') \) that satisfies the ADMiRA assumptions.

4.4 Properties of the Rank-Restricted Isometry

We introduce and prove a number of properties of the rank-restricted isometry. These properties serve as key tools for proving the performance guarantees for ADMiRA in this chapter. These properties further extend the analogy between the sparse vector and the low-rank matrix approximation problems (P3 and P2, respectively), and are therefore also of interest in their own right. The proofs are contained in the Appendix.

**Proposition 4.4.1.** The rank-restricted isometry constant \( \delta_r(A) \) is nondecreasing in \( r \).

An operator satisfying the R-RIP satisfies, as a consequence, a number of other properties when composed with other orthogonal projectors defined by the atomic decomposition.

**Proposition 4.4.2.** Suppose that linear operator \( A : S_2 \to \ell^2 \) has the rank-restricted isometry constant \( \delta_r(A) \). Let \( \Psi \) be a set of atoms in \( \mathcal{O} \) such that \( |\Psi| \leq r \). Then

\[
\|\mathcal{P}_{\Psi} A^*\|_{\ell^2 \to S_2} \leq \sqrt{1 + \delta_r(A)}.
\]  

(4.4.1)
Proposition 4.4.3. Suppose that linear operator $A : \mathcal{S}_2 \rightarrow \ell^p_2$ has the rank-restricted isometry constant $\delta_r(A)$. Let $\Psi$ be a set of atoms in $\mathcal{O}$ such that $|\Psi| \leq r$. Then

$$\|P_\Psi (A^* A - I_{\mathcal{S}_2}) P_\Psi\|_{\mathcal{S}_2 \rightarrow \mathcal{S}_2} \leq \delta_r(A).$$

Corollary 4.4.4. Suppose that linear operator $A : \mathcal{S}_2 \rightarrow \ell^p_2$ has the rank-restricted isometry constant $\delta_r(A)$. Let $\Psi$ be a set of atoms in $\mathcal{O}$ such that $|\Psi| \leq r$ and let $X \in \text{span}(\Psi)$. Then

$$\|P_\Psi A^* AX\|_{\mathcal{S}_2} \geq (1 - \delta_r(A)) \|X\|_{\mathcal{S}_2}. \tag{4.4.2}$$

Finally, we relate the R-RIP to the nuclear norm, extending the analogous result [11] from the $r$-sparse vector case to the rank-$r$ matrix case.

Proposition 4.4.5. If a linear map $A : \mathcal{S}_2 \rightarrow \ell^p_2$ satisfies

$$\|AX\|_{\ell^p_2}^2 \leq (1 + \delta_r(A)) \|X\|_{\mathcal{S}_2}^2, \tag{4.4.3}$$

for all $X \in \mathcal{S}_2$ with $\text{rank}(X) \leq r$, then

$$\|AX\|_{\ell^p_2} \leq \sqrt{1 + \delta_r(A)} \left[ \|X\|_{\mathcal{S}_2} + \frac{1}{\sqrt{p}} \|X\|_{\mathcal{S}_1} \right], \tag{4.4.4}$$

for all $X \in \mathcal{S}_2$.

4.5 Proof of Theorem 4.3.1

4.5.1 Exactly Low Rank Matrix Case

Theorem 4.5.1. Assume $\text{rank}(X_0) \leq r$ in (4.3.2). Let $\hat{X}_k$ denote the estimate of $X_0$ in the $k$th iteration of ADMiRA. Then for each $k \geq 0$, $\hat{X}_k$ satisfies the following recursion:

$$\|X_0 - \hat{X}_{k+1}\|_{\mathcal{S}_2} \leq 0.5 \|X_0 - \hat{X}_k\|_{\mathcal{S}_2} + 5.5 \|\nu\|_{\ell^2_2}.$$
From the above relation, it follows that

$$\|X_0 - \hat{X}_k\|_{S_2} \leq 2^{-k}\|X_0\|_{S_2} + 11\|\nu\|_{\ell_2^p}, \quad \forall k \geq 0.$$  

Theorem 4.5.1 is proved by applying a sequence of lemmata. We generalize the proof of the performance guarantee for CoSaMP [11] to the matrix case by applying the generalized analogy proposed in this chapter. The flow and the techniques used in the proofs are similar to those in [11]. However, in the matrix case, there are additional unknowns in the form of the singular vectors. Therefore, the generalization of the proofs in [11] to the matrix case is not straightforward and the proofs are sufficiently different from those for the vector case to warrant detailed exposition. The main steps in the derivation of the performance guarantee are stated in this section and the detailed proofs are in the Appendix.

For the proof, we study the \((k+1)\)th iteration starting with the previous result in the \(k\)th iteration. Let \(X_0\) denote the true solution with rank \(r\). Matrix \(\hat{X}\) denotes \(\hat{X}_k\), which is the estimate of \(X_0\) in the \(k\)th (previous) iteration. Set \(\hat{\Psi}\) is the set of orthogonal atoms obtained in the previous iteration. From \((b - A\hat{X})\), we compute the proxy matrix \(A^*(b - A\hat{X})\). Set \(\Psi'\) is the solution of the following low rank approximation problem:

$$\Psi' \triangleq \arg \max_{\Psi} \left\{ \|P_{\Psi}A^*(b - A\hat{X})\|_{S_2} : \Psi \subset \emptyset, \ |\Psi| \leq 2r \right\}.$$  

**Lemma 4.5.2.** Let rank\((X_0) \leq r\) in \((4.3.2)\). Then

$$\|P_{\Psi'}(X_0 - \hat{X})\|_{S_2} \leq 0.493\|X_0 - \hat{X}\|_{S_2} + 3.642\|\nu\|_{\ell_2^p}.$$  

Lemma 4.5.2 shows that subject to the rank-restricted isometry property, the set \(\Psi'\) of atoms chosen in Step 4 of ADMiRA is a good set: it captures \(76\% = 1 - 0.493^2\) of the energy of the atoms in \(X_0\) that were not captured by \(\hat{X}\), and the effects of additive measurement noise are bounded by a small constant. In other words, the algorithm is guaranteed to make good progress in this step.

**Remark 4.5.3.** For the vector case, the representation of a vector \(x \in \ell_2^n\)
in terms of the standard basis \( \{ e_j \}_{j=1}^n \) of \( \ell_2^n \) determines \( \| x \|_{\ell_0} \). Let \( J_1, J_2 \subset \{1, \ldots, n\} \) be arbitrary. Then the following properties hold: (i) the projection operators \( P_{\{e_j\}_{j \in J_1}} \) and \( P_{\{e_j\}_{j \in J_2}} \) commute; and (ii) \( P_{\{e_j\}_{j \in J}}^\perp x \) is \( s \)-sparse (or sparser) if \( x \) is \( s \)-sparse. These properties follow from the orthogonality of the standard basis. \[11, \text{Lemma 4.2}\], corresponding in the vector case to our Lemma 4.5.2, requires these two properties. However, the analogues of properties (i) and (ii) do not hold for the matrix case. Indeed, for \( \Psi_1, \Psi_2 \subset \mathcal{O} \), the projection operators \( P_{\Psi_1} \) and \( P_{\Psi_2} \) do not commute in general and \( \text{rank}(P_{\Psi} X) \) can be greater than \( r \) even though \( \text{rank}(X) \leq r \). While Lemma 4.5.2 provides a weaker result with bigger constants compared to \[11, \text{Lemma 4.2}\], Lemma 4.5.2 applies to more general cases, including the matrix case, without requiring the analogues of properties (i) and (ii).

**Lemma 4.5.4.** Let \( X_0, \tilde{X} \in \mathcal{S}_2 \) and let \( \Psi', \tilde{\Psi} \) be sets of atoms in \( \mathcal{O} \) such that \( |\Psi'| \leq 2r, |\tilde{\Psi}| \leq r, \) and \( P_{\tilde{\Psi}}^\perp \tilde{X} = 0 \). Let \( \tilde{\Psi} = \Psi' \cup \tilde{\Psi} \). Then

\[
\| P_{\tilde{\Psi}}^\perp X_0 \|_{\mathcal{S}_2} \leq \| P_{\Psi'}^\perp (X_0 - \tilde{X}) \|_{\mathcal{S}_2}.
\]

Lemma 4.5.4 shows that the augmented set of atoms \( \tilde{\Psi} \) produced in Step 5 of the algorithm is at least as good in explaining the unknown \( X_0 \) as was the set \( \Psi' \) in explaining the part of \( X_0 \) not captured by the estimate \( \tilde{X} \) from the previous iteration.

**Lemma 4.5.5.** Let \( \text{rank}(X_0) \leq r \) in (4.3.2) and let \( \tilde{\Psi} \) be a set of atoms in \( \mathcal{O} \) with \( |\tilde{\Psi}| \leq 3r \). Then

\[
\tilde{X} \triangleq \text{arg min}_X \left\{ \| b - AX \|_{\ell_2^p} : X \in \text{span}(\tilde{\Psi}) \right\} \quad (4.5.1)
\]

satisfies

\[
\| \tilde{X} - X_0 \|_{\mathcal{S}_2} \leq 1.002 \| P_{\tilde{\Psi}}^\perp X_0 \|_{\mathcal{S}_2} + 1.104 \| \nu \|_{\ell_2^p}.
\]

Lemma 4.5.5 shows that the least-squares step, Step 6 of the algorithm, performs almost as well as one could do with operator \( A \) equal to an identity operator: because \( \tilde{X} \) is restricted to \( \text{span}(\tilde{\Psi}) \), it is impossible to recover components of \( X_0 \) in \( \tilde{\Psi}^\perp \). Hence, the first constant cannot be smaller than
1. A value of 1 for the second constant, the noise gain, would correspond to a perfectly conditioned system.

**Lemma 4.5.6.** Suppose \( \text{rank}(X_0) \leq r \) in (4.3.2). Then,

\[
\|\Pi_r(\tilde{X}) - X_0\|_{S_2} \leq 1.007\|\tilde{X} - X_0\|_{S_2} + 1.788\|\nu\|_{\ell_p^2}.
\]

As expected, reducing the rank of the estimate \( \tilde{X} \) from \( 3r \) to \( r \), to produce \( \tilde{X}_r \), increases the approximation error. However, Lemma 4.5.6 shows that this increase is moderate – by no more than a factor of 2.

The update \( \tilde{X}_{k+1} = \tilde{X}_r \) completes the \((k + 1)\)th iteration. Combining all the results in the lemmata provides the proof of Theorem 4.5.1.

**Proof.** (Theorem 4.5.1)

\[
\|X_0 - \tilde{X}_{k+1}\|_{S_2} = \|X_0 - \Pi_r(\tilde{X})\|_{S_2}
\leq 1.007\|X_0 - \tilde{X}\|_{S_2} + 1.788\|\nu\|_{\ell_p^2} \quad \text{(Lemma 4.5.6)}
\leq 1.009\|\mathcal{P}^\perp_\Psi X_0\|_{S_2} + 2.896\|\nu\|_{\ell_p^2} \quad \text{(Lemma 4.5.5)}
\leq 1.009\|\mathcal{P}^\perp_\Psi (X_0 - \tilde{X}_k)\|_{S_2} + 2.896\|\nu\|_{\ell_p^2} \quad \text{(Lemma 4.5.4)}
\leq 0.5\|X_0 - \tilde{X}_k\|_{S_2} + 5.5\|\nu\|_{\ell_p^2}. \quad \text{(Lemma 4.5.2)}
\]

The recursion together with the fact that \( \sum_{j=0}^{k} 2^{-j} \leq \sum_{j=0}^{\infty} 2^{-j} = 2 \) provide the final result. 

\[\square\]

4.5.2 General Matrix Case

Theorem 4.3.1 is proved by combining Theorem 4.5.1 and the following lemma, which shows how to convert the mismodeling error (deviations of \( X_0 \) from a low rank matrix) to an equivalent additive measurement noise with a quantified norm.

**Lemma 4.5.7.** Let \( X_0 \) be an arbitrary matrix in \( S_2 \). The measurement \( b = AX_0 + \nu \) is also represented as \( b = A\Pi_r(X_0) + \tilde{\nu} \) where

\[
\|\tilde{\nu}\|_{\ell_p^2} \leq 1.04 \left[ \|X_0 - \Pi_r(X_0)\|_{S_2} + \frac{1}{\sqrt{r}}\|X_0 - \Pi_r(X_0)\|_{S_1} \right] + \|\nu\|_{\ell_p^2}.
\]
Proof. Let \( \tilde{\nu} = \mathcal{A}(X_0 - \Pi_r(X_0)) + \nu \). Then \( b = \mathcal{A}\Pi_r(X_0) + \tilde{\nu} \).

\[
\| \tilde{\nu} \|_2^p \leq \| \mathcal{A}(X_0 - \Pi_r(X_0)) \|_2 + \| \nu \|_2^p \\
\leq \sqrt{1 + \delta_r(\mathcal{A})} \left[ \| X_0 - \Pi_r(X_0) \|_{S_2} + \frac{1}{\sqrt{\rho}} \| X_0 - \Pi_r(X_0) \|_{S_1} \right] + \| \nu \|_2^p,
\]

where the last inequality holds by Proposition 4.4.5. The inequality \( \delta_r(\mathcal{A}) \leq 0.065 \) implies \( \sqrt{1 + \delta_r(\mathcal{A})} \leq 1.04 \).

Proof. (Theorem 4.3.1) Let \( X \) be an arbitrary matrix in \( S_2 \). The measurement is given by \( b = \mathcal{A}\Pi_r(X_0) + \tilde{\nu} \), where \( \tilde{\nu} \) is defined in Lemma 4.5.7. By Theorem 4.5.1,

\[
\| \Pi_r(X_0) - \widehat{X}_{k+1} \|_{S_2} \leq 0.5\| \Pi_r(X_0) - \widehat{X}_k \|_{S_2} + 5.5\| \tilde{\nu} \|_2^p.
\]

Applying the triangle inequality and the above inequality,

\[
\| X_0 - \widehat{X}_{k+1} \|_{S_2} \leq \| \Pi_r(X_0) - \widehat{X}_{k+1} \|_{S_2} + \| X_0 - \Pi_r(X_0) \|_{S_2} \\
\leq 0.5\| \Pi_r(X_0) - \widehat{X}_k \|_{S_2} + 5.5\| \tilde{\nu} \|_2^p + \| X_0 - \Pi_r(X_0) \|_{S_2}.
\]

Using the upper bound on \( \| \tilde{\nu} \|_2^p \) yields

\[
\| X_0 - \widehat{X}_{k+1} \|_{S_2} \leq 0.5\| X_0 - \widehat{X}_k \|_{S_2} + 6.72\| X_0 - \Pi_r(X_0) \|_{S_2} \\
+ \frac{5.72}{\sqrt{\rho}} \| X_0 - \Pi_r(X_0) \|_{S_1} + 5.5\| \nu \|_2^p \\
< 0.5\| X_0 - \widehat{X}_k \|_{S_2} + 7\epsilon,
\]

where \( \epsilon \) is the unrecoverable energy.

4.6 Implementation and Scalability

We analyze the computational complexity of ADMiRA and will show that ADMiRA scales well to large problem instances. Each iteration of ADMiRA consists of procedures requiring the following basic operations: application of \( \mathcal{A} \) and \( \mathcal{A}^* \), singular value decompositions, and solving a least squares problem. We analyze the computational cost of the procedures in terms of
the complexity of the basic operations, which will depend on the properties of \(\mathcal{A}\). The complexity is described by the big \(O\) notation \([129]\). First note that ADMiRA keeps the matrix variables (except the proxy matrix) in factorized form through their atomic decomposition, which is advantageous for both computational efficiency and memory requirements. Furthermore, the proxy matrix is often sparse in applications such as the matrix completion problem.

**Computing the proxy matrix:** this involves the application of \(\mathcal{A}\) and \(\mathcal{A}^*\). The procedure first computes the residual \(y = b - \mathcal{A}\hat{X}\) and then computes the proxy matrix \(\mathcal{A}^*y\). Let \(\hat{X} = \sum_{k=1}^{r} \sigma_k u_k v_k^*\) denote the atomic decomposition of \(\hat{X}\). Here \(u_k v_k^*'s\) are not necessarily orthogonal. \((\mathcal{A}\hat{X})_k\) can be computed by \(\langle \hat{X}, Z_k \rangle_{S_2} = \sum_{k=1}^{r} \sigma_k v_k^* Z_k u_k, k = 1, \ldots, p,\) for an appropriate set of \(p\) matrices \(Z_k \in S_2\). Then \(\mathcal{A}^*y\) can be computed by \(\sum_{k=1}^{p} y_k Z_k\). The complexity of these operations will depend on the sparsity of \(\mathcal{A}\).

**Case 1:** \(\mathcal{A}\) is an arbitrary linear (dense) operator and the costs of computing \(\mathcal{A}\hat{X}\) and \(\mathcal{A}^*y\) are \(O(pmnr)\) and \(O(pm)\), respectively.

**Case 2:** \(\mathcal{A}\) is a sparse linear operator – so the \(Z_k\) have \(O(m + n)\) non-zero elements, and the costs of computing \(\mathcal{A}\hat{X}\) and \(\mathcal{A}^*y\) are \(O(pr(m + n))\) and \(O(p(m + n))\), respectively.

**Case 3:** \(\mathcal{A}\) is an extremely sparse linear operator (such as in the matrix completion problem), so the \(Z_k\) have \(O(1)\) nonzeros, and the costs of computing \(\mathcal{A}\hat{X}\) and \(\mathcal{A}^*y\) are \(O(pr)\) and \(O(p)\), respectively.

**Finding the 2r principal atoms of the proxy matrix:** this involves the truncated singular value decomposition with \(2r\) dominant singular triplets, which can be computed by the Lanczos method \([45]\) at a cost of \(O(mnrL)\), where \(L\) denotes the number of the Lanczos iterations per each singular value, which depends on the singular value distribution. An alternative approach is to use recent advances in low rank approximation of large matrices based on randomized algorithms (cf. \([114, 130]\), and the references therein) that compute the low-rank approximation of a given matrix in time linear in the size of the matrix. These randomized algorithms are useful when the matrix is large but the rank \(r\) remains a small constant. For example, the complexity of Har-Peled’s algorithm \([114]\) is \(O(mnr^2 \log r)\). When \(\mathcal{A}\) is sparse with \(O(1)\) nonzero elements per each \(Z_k\), the matrix-vector product \((\mathcal{A}^*y)w\) for \(w \in \ell_2^n\) can be computed as \(\sum_{k=1}^{p} y_k Z_k w\) and hence the complexity reduces to...
$O(prL)$ for the Lanczos method and $O(pr^2 \log r)$ for the randomized method, respectively.

Solving least squares problems: ADMiRA requires the solution of an overdetermined system with $p$ equations and $3r$ unknowns. The complexity is $O(pr^2)$. Similarly to CoSaMP, the Richardson iteration or the conjugate gradient method can be used to improve the complexity of this part. The convergence of the Richardson iteration is guaranteed owing to the R-RIP assumption of ADMiRA and the complexity is $O(pr)$.

Finding the $r$ principal atoms of the solution to the least squares problem: this also involves the truncated singular value decomposition of the least square solution $\tilde{X}$. In fact, this procedure can be done more efficiently by exploiting the fact that $\tilde{X}$ is available in a factorized form $\tilde{X} = UV^*$ where $U \in \mathbb{C}^{m \times 3r}, V \in \mathbb{C}^{n \times 3r}$, and $\Sigma$ is a $3r \times 3r$ diagonal matrix. Here $U, V$ do not consist of orthogonal columns in general. Let $U = Q_U R_U$ and $V = Q_V R_V$ denote the QR factorizations of $U$ and $V$, respectively. Then $Q^*_U Q_U = I_m$ and $Q^*_V Q_V = I_n$. Now let $WDZ^*$ denote the singular value decomposition of the $3r \times 3r$ matrix $R_U \Sigma R_V^*$. Then we have the desired singular value decomposition $\tilde{X} = (Q_U W) D (Q_V Z)^*$. The complexity is $O((m + n + r)r^2)$, which is negligible compared to a direct SVD of $\tilde{X}$.

Applications of $A$ and $A^*$ are the most demanding procedures of ADMiRA for a dense linear operator $A$. These operations are also required in all other algorithms for P1, P1’, or P2. To overcome this computational complexity, the linear operator $A$ should have some structure that admits efficient computation. Examples include random Toeplitz matrices and randomly subsampled Fourier measurements. For matrix completion, $A$ is sparse with $O(1)$ cost per measurement and hence these operations are dominated by the remaining operations. In this case, the computation of the truncated singular value decomposition is the most demanding procedure of ADMiRA. Equipped with the randomized low rank approximation, ADMiRA has complexity of $O(pr^2 \log r)$ per iteration. ADMiRA therefore has complexity linear in the size $p$ of the data, and it scales well to large problems.
4.7 Numerical Experiment

We tested the performance of ADMiRA with an operator \( \mathcal{A} \) generated by a Gaussian ensemble, which satisfies RIP with high probability. ADMiRA performed well in this case as predicted by our theory. Here we study reconstructions by ADMiRA with a generic matrix completion example. Note that the performance guarantee in terms of R-RIP does not apply to this case, because the linear operator in the matrix completion problem does not satisfy the RIP. Nonetheless, we want to check the empirical performance of ADMiRA in this practically important application. Our Matlab implementation uses PROPACK [131] (an implementation of the Lanczos algorithm) to compute partial SVDs in Steps 4 and 7 of ADMiRA. The test matrix \( X_0 \in \mathbb{R}^{n \times n} \) is generated as the product \( X_0 = Y_L Y_R^* \) where \( Y_L, Y_R \in \mathbb{R}^{n \times r} \) have entries following an i.i.d. Gaussian distribution. The measurement \( b \) is \( p \) randomly chosen entries of \( X \), which may be contaminated with an additive white Gaussian noise. The reconstruction error and measurement noise level are measured in terms of \( \text{SNR}_{\text{rec}} = 20 \log_{10}(\|X_0\|_2/\|X_0 - \hat{X}\|_2) \) and \( \text{SNR}_{\text{meas}} = 20 \log_{10}(\|b\|_2/\|\nu\|_2) \), respectively. Computational efficiency is measured by the number of iterations. Here we stopped the algorithm when \( \|b - \mathcal{A}\hat{X}\|_2/\|b\|_2 < 10^{-4} \). As a result, the algorithm provided \( \text{SNR}_{\text{rec}} \) around 70dB for the ideal (noiseless and exactly low-rank) case when it was successful. However, it is still possible to get higher \( \text{SNR}_{\text{rec}} \) with a few more iterations. The results in Fig. 4.1, Table 4.1, and Table 4.2 have been averaged over 20 trials.

![Figure 4.1: Completion of random matrices by ADMiRA: \( n = m = 500, r = 2 \).](image-url)
Table 4.1: Completion of random matrices by ADMiRA: $n = m$, $r = 2$, $p = 10^n^{1.2r \log_{10} n}$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p/n^2$</th>
<th>$p/d_r$</th>
<th>no noise</th>
<th>SNR_{meas} = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>SNR_{rec}</td>
<td>#iter</td>
</tr>
<tr>
<td>500</td>
<td>0.37</td>
<td>47</td>
<td>83</td>
<td>8</td>
</tr>
<tr>
<td>1000</td>
<td>0.24</td>
<td>60</td>
<td>83</td>
<td>9</td>
</tr>
<tr>
<td>1500</td>
<td>0.18</td>
<td>69</td>
<td>82</td>
<td>11</td>
</tr>
<tr>
<td>2000</td>
<td>0.15</td>
<td>76</td>
<td>81</td>
<td>12</td>
</tr>
<tr>
<td>2500</td>
<td>0.13</td>
<td>81</td>
<td>81</td>
<td>18</td>
</tr>
<tr>
<td>3000</td>
<td>0.12</td>
<td>86</td>
<td>81</td>
<td>24</td>
</tr>
<tr>
<td>3500</td>
<td>0.10</td>
<td>90</td>
<td>81</td>
<td>26</td>
</tr>
<tr>
<td>4000</td>
<td>0.09</td>
<td>95</td>
<td>80</td>
<td>32</td>
</tr>
<tr>
<td>4500</td>
<td>0.09</td>
<td>98</td>
<td>81</td>
<td>37</td>
</tr>
</tbody>
</table>

Fig. 4.1 shows that both SNR_{rec} and the number of iterations improve as $p/d_r$ increases. Here $d_r$ is the number of degrees of freedom in a real rank-$r$ matrix defined by $d_r = r(n + m - r)$ and denotes the essential number of unknowns. Fig. 4.1 suggests that we need $p/d_r \geq 20$ for $n = 500$.

Candes and Recht [101] showed that $p = O(n^{1.2r \log_{10} n})$ known entries suffice to complete an unknown $n \times n$ rank-$r$ matrix. Table 4.1 shows that ADMiRA provides nearly perfect recovery of random matrices from $p$ known entries where $p = 10^n^{1.2r \log_{10} n}$. Although SNR_{rec} in the noiseless measurement case is high enough to say that the completion is nearly perfect, the number of iterations increases as $n$ increases. We are studying whether this increase in iterations with $n$ might be an artifact of our numerical implementation of ADMiRA. In the noisy measurement case the number of iterations is low and does not increase with problem size $n$. Because in most if not all practical applications the data will be noisy, or the matrix to be recovered only approximately low rank, this low and constant number of iterations is of practical significance.

Table 4.2 shows that in most of the examples tested, ADMiRA provides slightly better performance with less computation than SVT [92]. Roughly, the computational complexity of a single iteration of ADMiRA can be compared to two times that of SVT.

Fig. 4.2 compares the phase transitions of ADMiRA and SVT. We count the number of successful matrix completions (SNR_{rec} $\geq$ 70dB) out of 10
Table 4.2: Comparison of ADMiRA and SVT: no noise, \( n = m = 1000 \).

<table>
<thead>
<tr>
<th>( r )</th>
<th>( p/n^2 )</th>
<th>( p/d_r )</th>
<th>( \text{SNR}_{\text{rec}} )</th>
<th>( #\text{iter} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ADMiRA</td>
<td>SVT</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>12.51</td>
<td>77</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>25.03</td>
<td>79</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>37.54</td>
<td>81</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>50.05</td>
<td>82</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>62.56</td>
<td>84</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>75.08</td>
<td>84</td>
<td>79</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>5.01</td>
<td>19</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>10.03</td>
<td>77</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>15.04</td>
<td>78</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>20.05</td>
<td>81</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>25.06</td>
<td>82</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>30.08</td>
<td>83</td>
<td>79</td>
</tr>
<tr>
<td>10</td>
<td>0.05</td>
<td>2.51</td>
<td>7</td>
<td>-9</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>5.03</td>
<td>30</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>7.54</td>
<td>77</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>10.05</td>
<td>79</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>12.56</td>
<td>80</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>15.08</td>
<td>80</td>
<td>78</td>
</tr>
</tbody>
</table>
Figure 4.2: Phase transition of matrix completion: $n = m = 100$.

trials for each triplet $(n, p, r)$. Brighter color implies more success. ADMiRA performed better than SVT for this example.

We emphasize that all comparisons with SVT were performed for the noiseless exactly low rank matrix case, because the current implementation [132] and theory [92] of SVT do not support the ellipsoidal constraint case. We are not aware of an efficient, scalable algorithm other than ADMiRA that supports the ellipsoidal constraint.
Previous guarantees for the reconstruction of sparse signals from compressive sensing via random frame matrices by various practical algorithms were provided in terms of the restricted isometry property (RIP) of the sensing matrix. Previous works on the RIP focused on scenarios where, to satisfy the isotropy property, the sensing matrix is constructed from i.i.d. samples from a tight frame according to the uniform distribution. However, the frame might not be tight due to the physics of the sensing procedure or due to the dictionary that provides a sparse representation. Furthermore, a non-uniform rather than the uniform distribution is often used for the i.i.d. sampling in practice in compressed sensing, especially in imaging applications, due to the signal characteristics or due to the limitation imposed by the physics of the applications. To derive guarantees without idealized assumptions, we proposed to exploit the property of biorthogonality that naturally arises in frame theory. We generalized the RIP to the restricted biorthogonality property (RBOP) that is satisfied without requiring the isotropy property. To take advantage of the new RBOP, we extended greedy pursuit algorithms with RIP-based guarantees to new variations – oblique pursuit algorithms, so that they provide RBOP-based guarantees. These guarantees apply with relaxed conditions on the sensing matrices and dictionaries, which are satisfied by practical CS imaging schemes. The extension of greedy pursuit algorithms and their RIP-based guarantees to those based on the RBOP is not restricted to the specific algorithms studied in this paper. For example, Fast Nesterov’s Iterative Hard Thresholding (FNIHT) [104] is another promising algorithm with a RIP-based guarantee, which will extend similarly. Finally, we note that although the oblique pursuit algorithms were designed to provide performance guarantees in the worst-case sense, they also perform competitively with or sometimes significantly better than their conventional counterparts empirically.
In Chapter 3, we proposed subspace-based methods for joint sparse recovery, which are computationally efficient and provide both good empirical performance and theoretical guarantees. SA-MUSIC is a new family of algorithms that improve on a previous subspace method, MUSIC, so that it can recover the support reliably under unfavorable rank-defective situations. To derive guarantees for the proposed methods, we further investigated the behavior of the sensing matrices in terms of the weak-1 restricted isometry. We also provided a non-asymptotic analysis of signal subspace estimation from finitely many snapshots, which is distinct from the conventional asymptotic studies in sensor array processing. In the analysis, we provided explicit constants for each sufficient condition for the guarantees.

The SA-MUSIC algorithm, combined with OSMP, is an example; indeed, SA-MUSIC can exploit other algorithms for the step of partial support recovery. For example, we also proposed a different algorithm, signal-subspace orthogonal matching pursuit (SS-OMP) and its combination with SA-MUSIC with the corresponding guarantees [79]. Finally, the analysis in Chapter 3 reveals that the performance of the SA-MUSIC family of algorithms is limited by the algorithm for partial support recovery, rather than by the SA-MUSIC step. Therefore, it is possible to improve the performance of the SA-MUSIC algorithms by employing a better algorithm for partial support recovery than OSMP.

In Chapter 4, we proposed a new algorithm, ADMiRA, which extends both the efficiency and the performance guarantee of the CoSaMP algorithm for \( \ell_0 \)-norm minimization to matrix rank minimization. By using the proposed generalized correlation maximization, greedy algorithms such as MP, OMP, and SP and their performance guarantees are also extended from the \( s \)-term vector approximation problem to the rank-\( r \) matrix approximation problem. ADMiRA can handle large scale rank minimization problems efficiently by using recent linear time algorithms for low rank approximation of a known matrix. Our numerical experiments demonstrate that ADMiRA is an effective algorithm even when the R-RIP is not satisfied, as in the matrix completion problem. While the performance guarantee in Chapter 4 relies on the R-RIP, it seems that a performance guarantee for ADMiRA without using the R-RIP might be possible.
A.1 Preliminaries for the Appendix A

**Definition A.1.1** (Dilation [133]). The dilation of matrix $M$ is defined by

$$\mathcal{S}(M) \triangleq \begin{bmatrix} 0 & M \\ M^* & 0 \end{bmatrix}.$$ 

By definition, $\mathcal{S}(M)$ is a Hermitian matrix and its eigenvalues satisfy

$$\lambda_i(\mathcal{S}(M)) = \begin{cases} \sigma_i(M) & \text{if } i \leq n \\ -\sigma_{n-i+1}(M) & \text{if } i > n. \end{cases}$$

**Definition A.1.2** (Schur Complement). Let $M \in \mathbb{K}^{n \times n}$ be a square matrix that can be decomposed as follows:

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

where $M_{22} \in \mathbb{K}^{q \times q}$ for $q < n$ is a minor of $M$, which is also a square matrix. The Schur complement of the block $M_{22}$ of the matrix $M$, denoted by $M/M_{22}$, is the $(n-q) \times (n-q)$ matrix defined by

$$M/M_{22} \triangleq M_{11} - M_{12}M_{22}^\dagger M_{21}.$$ 

The following lemma extends [134, Theorem 5] to the non-Hermitian case.

**Lemma A.1.3.** Let $M \in \mathbb{K}^{n \times n}$ be a nonsingular matrix and $M_{22} \in \mathbb{K}^{q \times q}$ for $q < n$ be a minor of $M$. Then,

$$\sigma_1(M) \geq \sigma_1(M/M_{22}).$$
and
\[ \sigma_j(M/M_{22}) \geq \sigma_{j+q}(M), \quad \forall j = 1, \ldots, n - q. \]

**Remark A.1.4.** The analogous result for the Hermitian case [134, Theorem 5] assumed that \( M \) is semidefinite and also showed that
\[ \sigma_j(M) \geq \sigma_j(M/M_{22}), \quad \forall j = 1, \ldots, n - q. \]

**Proof of Lemma A.1.3.** By the Cauchy interlacing theorem,
\[ \sigma_{q}(M_{22}) \geq \sigma_n(M) > 0; \] hence, \( M_{22} \) is invertible. Let
\[
M = \begin{bmatrix}
M_{11} & M_{12} \\
M_{21} & M_{22}
\end{bmatrix}
= \begin{bmatrix}
M_{11} - M_{12}M_{22}^{-1}M_{21} & 0 \\
0 & 0
\end{bmatrix}
+ \begin{bmatrix}
M_{12}M_{22}^{-1}M_{21} & M_{12} \\
M_{21} & M_{22}
\end{bmatrix} = M_1 + M_2.
\]

Let \( M_{22} = U\Sigma V^* \) be the singular value decomposition of \( M_{22} \). Then, \( M_2 \) is factorized as
\[
M_2 = \begin{bmatrix}
M_{12}V\Sigma^{-1/2} \\
U\Sigma^{1/2}
\end{bmatrix}
\begin{bmatrix}
\Sigma^{-1/2}U^*M_{21} & \Sigma^{1/2}V^*
\end{bmatrix}
\]
where the left factor has \( q \) linearly independent columns and the right factor has \( q \) linearly independent rows. Therefore, \( \text{rank}(M_2) = q \).

Now, we use Weyl’s inequalities for the eigenvalues of the sum of two Hermitian matrices [135, Theorem III.2.1]. By applying [135, Theorem III.2.1] to \( \mathcal{S}(M_1) \) and \( \mathcal{S}(M_2) \), we obtain
\[
\lambda_{j+q}(\mathcal{S}(M_1) + \mathcal{S}(M_2)) \\
\leq \lambda_j(\mathcal{S}(M_1)) + \lambda_{q+1}(\mathcal{S}(M_2)) \\
= \lambda_j(\mathcal{S}(M_1)), \quad \forall j = 1, \ldots, n - q,
\]
where we used the fact that \( \lambda_{q+1}(\mathcal{S}(M_2)) = \sigma_{q+1}(M_2) = 0 \) since \( \text{rank}(M_2) = q \). Therefore,
\[ \sigma_{j+q}(M_1 + M_2) \leq \sigma_j(M_1), \quad \forall j = 1, \ldots, n - q. \]
Since $M$ is invertible, $M/M_{22}$ is also invertible since $\sigma_{n-q}(M/M_{22}) \geq \sigma_n(M) > 0$. The Schur complement $(M/M_{22})^{-1}$ is a minor of $M^{-1}$; hence,

$$\sigma_1(M)^{-1} = \sigma_n(M^{-1}) \leq \sigma_{n-q}((M/M_{22})^{-1}) = \sigma_1(M/M_{22})^{-1}.$$ 

\[ \square \]

**Lemma A.1.5.** Let $M \in \mathbb{K}^{m \times m}$. Then,

$$\sigma_1(M - I_m) = \max (1 - \sigma_m(M), \sigma_1(M) - 1).$$

**Proof of Lemma A.1.5.** If $M$ is a Hermitian matrix, then the proof is straightforward since the eigenvalues of $M - I_m$ are the eigenvalues of $M$ shifted by 1. Otherwise, by [135, Theorem III.2.8], it follows that

$$\max_{k \in [2m]} |\lambda_k(\mathcal{F}(M)) - \lambda_k(\mathcal{F}(I_m))| \leq \|\mathcal{F}(M) - \mathcal{F}(I_m)\| \leq \max_{k \in [2m]} |\lambda_k(\mathcal{F}(M)) - \lambda_{2m-k+1}(\mathcal{F}(I_m))|,$$

where $\mathcal{F}(M)$ and $\mathcal{F}(I_m)$ are the dilations of $M$ and $I_m$, respectively.

Since

$$\lambda_k(\mathcal{F}(M)) = \begin{cases} \sigma_k(M) & k \leq m \\ -\sigma_{m-k+1}(M) & k > m \end{cases}$$

and

$$\lambda_k(\mathcal{F}(I_m)) = \begin{cases} 1 & k \leq m \\ -1 & k > m, \end{cases}$$

it follows that

$$\max_{k \in [2m]} |\lambda_k(\mathcal{F}(M)) - \lambda_k(\mathcal{F}(I_m))| \leq \max_{k \in [2m]} |\lambda_k(\mathcal{F}(M)) - \lambda_{2m-k+1}(\mathcal{F}(I_m))| = \max (1 - \sigma_m(M), \sigma_1(M) - 1);$$

125
hence,

\[
\|M - I_m\| = \|\mathcal{S}(M - I_m)\| \\
= \|\mathcal{S}(M) - \mathcal{S}(I_m)\| \\
= \max(1 - \sigma_m(M), \sigma_1(M) - 1).
\]

\[\blacksquare\]

**Lemma A.1.6.** Let \(M, \tilde{M} \in \mathbb{K}^{m \times k}\). Let \(J_1 \subseteq [k]\) and \(J_2 = [k] \setminus J_1\). Suppose \(\tilde{M}^*M\) has full rank. Then,

\[
\|\tilde{M}^*_2(I_{|J_1|} - M_{J_1}(\tilde{M}^*_1M_{J_1})^\dagger M_{J_2})M_{J_2} - I_{|J_2|}\| \leq \|\tilde{M}^*M - I_k\|.
\]

*Proof of Lemma A.1.6.* To simplify the notation, let \(E \triangleq I_{|J_1|} - M_{J_1}(\tilde{M}^*_1M_{J_1})^\dagger \tilde{M}^*_1\).

By Lemma A.1.5, it follows that

\[
\|\tilde{M}^*_2EM_{J_2} - I_{|J_2|}\| = \max \left\{1 - \sigma_{|J_2|}(\tilde{M}^*_2EM_{J_2}), \sigma_{1}(\tilde{M}^*_2EM_{J_2}) - 1\right\}. \tag{A.1.1}
\]

Furthermore, since \(\tilde{M}^*M\) has full rank, (A.1.1) is upper bounded by Lemma A.1.3 as

\[
\leq \max \left\{1 - \sigma_k(\tilde{M}^*M), \sigma_1(\tilde{M}^*M) - 1\right\} = \|\tilde{M}^*M - I_k\| \tag{A.1.2}
\]

where the last step too follows from Lemma A.1.5. \[\blacksquare\]

**Lemma A.1.7** ([136, Corollary 5.2]). Suppose that \(E \in \mathbb{K}^{n \times n}\) is idempotent \((E^2 = E)\) and is neither 0 nor \(I_n\). Then, \(\|I_n - E\| = \|E\|\).

**Lemma A.1.8.** Let \(\Psi, \tilde{\Psi} \in \mathbb{K}^{m \times n}\). Let \(P \in \mathbb{K}^{n \times n}\) be an orthogonal projector in \(\mathbb{K}^n\). Then, for all \(x, y \in \mathbb{K}^n\),

\[
\left|\left| \langle \tilde{\Psi}Px, \Psi Py \rangle \right| - \left|\langle Px, Py \rangle \right| \right| \leq \|P\tilde{\Psi}^*\Psi P - P\| \cdot \|x\|_2 \cdot \|y\|_2. \tag{A.1.3}
\]

*Proof of Lemma A.1.8.* The proof follows from the properties of an inner
where (a) follows from the triangle inequality, (b) follows since $P^2 = P$ and $P^* = P$. \hfill \Box

### A.2 Proof of Theorem 2.2.2

ObThres is guaranteed to recover $J^*$ if

\[
\min_{j \in J^*} |(\hat{\Psi}^*)_j| > \max_{j \notin J^*} |(\hat{\Psi}^*)_j|.
\]

(A.2.1)

The $j$th component of $\hat{\Psi}^*y$ is given as

\[(\hat{\Psi}^*)_j = \tilde{\psi}^*_j y = e^*_j \Pi_{\{j\}} \hat{\Psi}^* \Psi \Pi J^* x^* + \tilde{\psi}^*_j z \]

and satisfies

\[
|(\hat{\Psi}^*)_j - (x^*)_j| \\
\leq |e^*_j \Pi_{\{j\}} \hat{\Psi}^* \Psi \Pi J^* x^* - (x^*)_j| + |\tilde{\psi}^*_j z| \\
= |e^*_j (\Pi_{\{j\}} \hat{\Psi}^* \Psi \Pi J^* - \Pi_{\{j\} \cap J^*}) x^*| + |\tilde{\psi}^*_j z| \\
\leq \theta_{s+1} (\hat{\Psi}^* \Psi) \|x^*\|_2 + \max_j \|\tilde{\psi}_j\|_2 \|z\|_2
\]

(A.2.2)

where the third step follows since

\[
\|\Pi_{\{j\}} \hat{\Psi}^* \Psi \Pi J^* - \Pi_{\{j\} \cap J^*} \|
\leq \|\Pi_{\{j\} \cup J^*} \hat{\Psi}^* \Psi \Pi_{\{j\} \cup J^*} - \Pi_{\{j\} \cup J^*} \|
\leq \theta_{\{j\} \cup J^*} (\hat{\Psi}^* \Psi) \leq \theta_{s+1} (\hat{\Psi}^* \Psi).
\]
Then, (2.2.2) is obtained by applying (A.2.2) to (A.2.1).

A.3 Proof of Proposition 2.2.7

Given $J \subset J^*$, the next step of ObMP given $J$ finds an element from $J^* \setminus J$ if

$$\max_{j \in J^* \setminus J} |\tilde{\psi}_j^* (E_{\mathcal{R}\mathcal{R}(\bar{\psi}_j^+) \mathcal{R}(\Psi_j)}) y| > \max_{j \in [n] \setminus J^*} |\tilde{\psi}_j^* (E_{\mathcal{R}\mathcal{R}(\bar{\psi}_j^+) \mathcal{R}(\Psi_j)}) y|.$$  \hfill (A.3.1)

Let $E$ denote $E_{\mathcal{R}\mathcal{R}(\bar{\psi}_j^+) \mathcal{R}(\Psi_j)}$ to simplify the notation. Then, $E^* = E_{\mathcal{R}\mathcal{R}(\bar{\psi}_j^+) \mathcal{R}(\Psi_j)}$ is also an oblique projection.

To derive a sufficient condition for (A.3.1), we first derive a lower bound of the left-hand side of (A.3.1) in the following:

$$\max_{j \in J^* \setminus J} |\tilde{\psi}_j^* E y| \geq \max_{j \in J^* \setminus J} |\tilde{\psi}_j^* E \Psi \Pi_{J^* \setminus J} x^*| - |\tilde{\psi}_j^* E z|$$

$$\geq \max_{j \in J^* \setminus J} |\tilde{\psi}_j^* E \Psi \Pi_{J^* \setminus J} x^*| - \|\Psi^*\| \|E\| \|z\|.$$ \hfill (A.3.2)

The term $(\ast)$ in (A.3.2) is bounded from below by

$$(\ast) = \max_{j \in J^* \setminus J} |\tilde{\psi}_j^* E \Psi \Pi_{J^* \setminus J} x^*|$$

$$= \max_{j \in J^* \setminus J} |\langle \tilde{\psi}_j, E \Psi \Pi_{J^* \setminus J} x^* \rangle|$$

$$= \max_{j \in J^* \setminus J} |\langle \tilde{\Psi}_{\Pi_{J^* \setminus J} e_j}, E \Psi \Pi_{J^* \setminus J} x^* \rangle|$$

$$\geq (a) \max_{j \in J^* \setminus J} \|\Pi_{J^* \setminus J} e_j, \Pi_{J^* \setminus J} x^*\|$$

$$- \theta_{s+1}(\tilde{\Psi}^* \Psi) \|\Pi_{J^* \setminus J} e_j\| \|\Pi_{J^* \setminus J} x^*\|$$

$$= \max_{j \in J^* \setminus J} \|x^*\| - \theta_{s+1}(\tilde{\Psi}^* \Psi) \|\Pi_{J^* \setminus J} x^*\|$$

$$= \|\Pi_{J^* \setminus J} x^*\| - \theta_{s+1}(\tilde{\Psi}^* \Psi) \|\Pi_{J^* \setminus J} x^*\|.$$ \hfill (A.3.3)

where (a) holds by Lemma A.1.8 since it follows, by Lemma A.1.6, that

$$\|\Pi_{J^* \setminus J} \tilde{\Psi}^* E \Psi \Pi_{J^* \setminus J} - \Pi_{J^* \setminus J}\|$$
\[
\|
\tilde{\psi}_{J^*} E \Psi_{J^* \setminus J} - I_{|J^* \setminus J|} \| \\
\leq \|
\tilde{\psi}_J^* \Psi - I_s \| \leq \theta_{s+1}(\tilde{\Psi}^* \Psi).
\]

Next, we derive an upper bound on the right-hand side of \((A.3.1)\) in a similar way:

\[
\begin{align*}
\max_{j \in [n] \setminus J^*} |\tilde{\psi}_j^* E y| & \leq \max_{j \in [n] \setminus J^*} |\tilde{\psi}_j^* E \Psi_{J^*} x^*| + |\tilde{\psi}_j^* E z| \\
& \leq \max_{j \in [n] \setminus J^*} |\tilde{\psi}_j^* E \Psi_{J^* \setminus J} x^*| + \|\tilde{\Psi}^*\|_2 \|E\| \|z\|_2. 
\end{align*}
\]

The term \((\ast\ast)\) in \((A.3.4)\) is upper bounded by

\[
(\ast\ast) = \max_{j \in [n] \setminus J^*} \left| \tilde{\psi}_j^* E \Psi_{(J^* \cup \{j\}) \setminus J} x^* \right|
\]

\[
= \max_{j \in [n] \setminus J^*} \left| \tilde{\psi}_j^* E \mathbb{P}_{(J^* \cup \{j\}) \setminus J} x^* \right|
\]

\[
= \max_{j \in [n] \setminus J^*} \left| \left( \tilde{\Psi} \mathbb{P}_{(J^* \cup \{j\}) \setminus J} e_j, E \mathbb{P}_{(J^* \cup \{j\}) \setminus J} x^* \right) \right|
\]

\[
(\ast) \leq \max_{j \in [n] \setminus J^*} \left| \left( \Pi_{(J^* \cup \{j\}) \setminus J} e_j, \Pi_{(J^* \cup \{j\}) \setminus J} x^* \right) \right|
\]

\[
+ \theta_{s+1}(\tilde{\Psi}^* \Psi) \|\Pi_{(J^* \cup \{j\}) \setminus J} e_j\|_2 \|\Pi_{(J^* \cup \{j\}) \setminus J} x^*\|_2
\]

\[
= \max_{j \in [n] \setminus J^*} |(x^*)_j| + \theta_{s+1}(\tilde{\Psi}^* \Psi) \|\Pi_{(J^* \cup \{j\}) \setminus J} x^*\|_2
\]

\[
= \theta_{s+1}(\tilde{\Psi}^* \Psi) \|\Pi_{J^* \setminus J} x^*\|_2
\]

\[(A.3.5)\]

where \((\ast)\) follows by Lemma A.1.8 since it follows by Lemma A.1.6 that

\[
\|\Pi_{(J^* \cup \{j\}) \setminus J} \tilde{\Psi}^* E \Psi_{(J^* \cup \{j\}) \setminus J} - \Pi_{(J^* \cup \{j\}) \setminus J} \|
\]

\[
\leq \|\tilde{\Psi}_{J^* \setminus J} E \Psi_{(J^* \cup \{j\}) \setminus J} - I_{(J^* \cup \{j\}) \setminus J} \|
\]

\[
\leq \|\tilde{\Psi}_{J^* \setminus j} \Psi_{J^* \setminus j} - I_{s+1}\| \leq \theta_{s+1}(\tilde{\Psi}^* \Psi).
\]

Applying the bounds in \((A.3.2)\) \((A.3.5)\) to \((A.3.1)\), we conclude that, for the success of the next step, it suffices to satisfy

\[
\|\Pi_{J^* \setminus J} x^*\|_{\infty} - 2\theta_{s+1}(\tilde{\Psi}^* \Psi) \|\Pi_{J^* \setminus J} x^*\|_2 > 2\|\tilde{\Psi}^*\|_2 \|E\| \|z\|_2.
\]

129
Then, computing an upper bound on $\|E\|$ will complete the proof.

When $\tilde{\Psi} = \Psi$, $E$ reduces to an orthogonal projection and satisfies $\|E\| \leq 1$. However, since we propose to use $\tilde{\Psi} \neq \Psi$, $E$ is an oblique projection and $\|E\|$ is not necessarily bounded by 1.

Since $E$ is idempotent and $E$ is neither 0 or $I_n$, by Lemma A.1.7, it follows that

$$\|E\| = \|I_n - E\| = \|E_{\mathcal{R}(\Psi) \backslash \mathcal{R}(\tilde{\Psi})}\|$$

$$= \|\Psi_J (\tilde{\Psi}_J^* \Psi_J)^{-1} \tilde{\Psi}_J^*\|$$

$$\leq \frac{\|\Psi_J\| \|\tilde{\Psi}_J\|}{\lambda_s(\tilde{\Psi}_J^* \Psi_J)} \leq \frac{\|\Psi_J\| \|\tilde{\Psi}_J\|}{1 - \theta_{s+1}(\tilde{\Psi}^* \Psi)}.$$

### A.4 Proof of Theorem 2.2.9

The proof for the ObSP case is done by the following four steps. To simplify the notations, let

$$\theta = \theta_{3s}(\tilde{\Psi}^* \Psi), \quad \delta = \delta_s(\Psi), \quad \text{and} \quad \tilde{\delta} = \delta_{2s}(\tilde{\Psi}).$$

For $J = \{j_1, \ldots, j_\ell\} \subset [n]$, define $R_J : \mathbb{K}^n \rightarrow \mathbb{K}^\ell$ by

$$(R_Jx)_k = x_{j_k}, \quad \forall k \in J, \forall x \in \mathbb{K}^n,$$

which is the reduction map to the subvector indexed by $J$. The adjoint operator $R_J^* : \mathbb{K}^\ell \rightarrow \mathbb{K}^n$ satisfies

$$R_J^* y = \sum_{k=1}^\ell (y)_k e_{j_k}$$

where $e_k$ is the $k$th column of $I_n$.

**Lemma A.4.1** (Step 1). *Under the assumptions of Theorem 2.2.9,*

$$\|x_{t+1} - x^*\|_2 \leq \rho_1 \|\Pi_{J^* \backslash J_{t+1}} x^*\|_2 + \tau_1 \|\tilde{z}\|_2$$
where $\rho_1$ and $\tau_1$ are given by

$$
\rho_1 = \frac{1}{\sqrt{1 - \theta^2}} \quad \text{and} \quad \tau_1 = \frac{\sqrt{1 + \delta}}{1 - \theta}.
$$

**Lemma A.4.2** (Step 2). Under the assumptions of Theorem 2.2.9,

$$
\|\Pi_{J^* \setminus J_{t+1}}x^*\|_2 \leq \rho_2\|\Pi_{J^* \setminus J_{t+1}}x^*\|_2 + \tau_2\|z\|_2
$$

where $\rho_2$ and $\tau_2$ are given by

$$
\rho_2 = \frac{1 + \theta}{1 - \theta} \quad \text{and} \quad \tau_2 = \frac{2\sqrt{1 + \delta}}{1 - \theta}.
$$

**Lemma A.4.3** (Step 3). Under the assumptions of Theorem 2.2.9,

$$
\|\Pi_{J^* \setminus J_{t+1}}x^*\|_2 \leq \rho_3\|\Pi_{J^* \setminus J_{t}}x^*\|_2 + \tau_3\|z\|_2
$$

where $\rho_3$ and $\tau_3$ are given by

$$
\rho_3 = \max\left(\frac{\theta}{1 - \theta}, \frac{2\theta(1 - \theta)}{1 + 2\theta + 2\theta^2}\right)
$$

and

$$
\tau_3 = \max\left(\frac{1}{1 - \theta}, \frac{2(1 - \theta)}{1 + 2\theta + 2\theta^2}\right) \frac{2\sqrt{1 + \delta}(1 + \tilde{\delta})}{1 - \theta}
$$

(Step 4): Finally, because $\text{supp} (x_t) = J_t$,

$$
\|\Pi_{J^* \setminus J_t}x^*\|_2 = \|\Pi_{J^* \setminus J_t}(x_t - x^*)\|_2 \leq \|x_t - x^*\|_2.
$$

Then, $\rho$ and $\tau$ are given as

$$
\rho = \rho_1\rho_2\rho_3 \quad \text{and} \quad \tau = \tau_1 + \rho_1\tau_2 + \rho_1\rho_2\tau_3.
$$

If we let $\tilde{\Psi} = \Psi$, ObSP reduces to SP, and the RBOP-based guarantee for ObSP also reduces to the RIP-based guarantee of SP. However, compared to
the original guarantee [12], the guarantee of SP obtained from Theorem 2.2.9 requires a less demanding RIP condition.

The results for the other algorithms (ObCoSaMP, ObHTP, and ObIHT) are obtained from the corresponding results for the conventional algorithms (CoSaMP, HTP, and IHT) [14, 106]. We only need to replace $\Psi^*\Psi$ by $\widetilde{\Psi}^*\Psi$ in the algorithms and replace $\delta_{ks}(\Psi)$ by $\theta_{ks}(\tilde{\Psi}^*\Psi)$ in the guarantees.

Constants $\rho$ and $\tau$ are explicitly given as follows:

- **ObCoSaMP**

  \[
  \rho = \sqrt{\frac{4\theta_{4s}(\tilde{\Psi}^*\Psi)^2(1 + 3\theta_{4s}(\tilde{\Psi}^*\Psi)^2)}{1 - \theta_{4s}(\tilde{\Psi}^*\Psi)^2}},
  \]

  \[
  \tau = \left(\sqrt{\frac{2(1 + 3\theta_{4s}(\tilde{\Psi}^*\Psi)^2)}{1 - \theta_{4s}(\tilde{\Psi}^*\Psi)^2}} + \sqrt{\frac{1 + 3\theta_{4s}(\tilde{\Psi}^*\Psi)^2}{1 - \theta_{4s}(\tilde{\Psi}^*\Psi)}} + \sqrt{3}\right) \cdot \sqrt{1 + \delta_{4s}(\tilde{\Psi})}.
  \]

- **ObSP**

  \[
  \rho = \frac{\theta_{3s}(\widetilde{\Psi}^*\Psi)\sqrt{1 + \theta_{3s}(\widetilde{\Psi}^*\Psi)}}{\sqrt{1 - \theta_{3s}(\widetilde{\Psi}^*\Psi)}} \max\left\{\frac{1}{(1 - \theta_{3s}(\widetilde{\Psi}^*\Psi))^2}, \frac{2}{1 + 2\theta_{3s}(\widetilde{\Psi}^*\Psi) + 2\theta_{3s}(\tilde{\Psi}^*\Psi)^2}\right\},
  \]

  \[
  \tau = \frac{\sqrt{1 + \delta_{2s}(\widetilde{\Psi})}}{1 - \theta_{3s}(\widetilde{\Psi}^*\Psi)} + \sqrt{\frac{1 + \delta_{2s}(\widetilde{\Psi})}{(1 - \theta_{3s}(\tilde{\Psi}^*\Psi))\sqrt{1 - \theta_{3s}(\tilde{\Psi}^*\Psi)^2}}}
  \]

  \[
  + \frac{2\sqrt{1 + \delta_{s}(\Psi)(1 + \delta_{2s}(\widetilde{\Psi}))\sqrt{1 + \theta_{3s}(\widetilde{\Psi}^*\Psi)}}}{\sqrt{1 - \theta_{3s}(\tilde{\Psi}^*\Psi) \cdot (1 - \theta_{3s}(\tilde{\Psi}^*\Psi))}} \cdot \max\left\{\frac{1}{(1 - \theta_{3s}(\tilde{\Psi}^*\Psi))^2}, \frac{2}{1 + 2\theta_{3s}(\tilde{\Psi}^*\Psi) + 2\theta_{3s}(\tilde{\Psi}^*\Psi)^2}\right\}.
  \]
Lemma A.4.2 is of independent interest to provide the finite convergence
in Theorem 2.2.12. We stated Lemma A.4.2 as Lemma 2.2.10 in Section 2.2.
For ObCoSaMP and ObHTP, similar lemmata are obtained with a slight
modification from the corresponding results [14, 106]. Constants \( \bar{\rho} \) and \( \bar{\tau} \) in
Lemma 2.2.10 are explicitly given as follows:

- **ObCoSaMP**

  \[
  \bar{\rho} = \sqrt{1 + 3\theta_4s(\bar{\Psi}^*\Psi)^2} / (1 - \theta_4s(\bar{\Psi}^*\Psi)^2), \\
  \bar{\tau} = \left( \sqrt{1 + 3\theta_4s(\bar{\Psi}^*\Psi)^2} / (1 - \theta_4s(\bar{\Psi}^*\Psi)) + \sqrt{3} \right) \sqrt{1 + \delta_4s(\bar{\Psi})}.
  \]

- **ObSP**

  \[
  \bar{\rho} = 1 / \sqrt{1 - \theta_3s(\bar{\Psi}^*\Psi)^2}, \quad \bar{\tau} = \sqrt{1 + \delta_4s(\bar{\Psi})} / (1 - \theta_3s(\bar{\Psi}^*\Psi)).
  \]

- **ObHTP**

  \[
  \bar{\rho} = 1 / \sqrt{1 - \theta_3s(\bar{\Psi}^*\Psi)^2}, \quad \bar{\tau} = \sqrt{1 + \delta_3s(\bar{\Psi})} / (1 - \theta_3s(\bar{\Psi}^*\Psi)).
  \]
Proof of Lemma A.4.1. Lemma A.4.1 is an extension of the analogous result by Foucart [106] to the biorthogonal case. The modification is done by replacing some matrices and introducing the RBOP instead of the RIP. We repeat the proof with appropriate modifications as a guiding example that shows how to modify the derivations using the RBOP.

Recall that $x_{t+1}$ is given as

$$x_{t+1} = \arg \min_x \{ \| \tilde{\Psi}_{J_{t+1}}^* (y - \Psi x) \| : \text{supp } (x) \subset J_{t+1} \}.$$ 

Therefore, by the optimality condition of the least square problem, it follows that

$$(\tilde{\Psi}_{J_{t+1}}^* \Psi)^* \tilde{\Psi}_{J_{t+1}}^* (y - \Psi x_{t+1}) = 0,$$

but, by the RBOP, $\tilde{\Psi}_{J_{t+1}}^* \Psi$ has full row rank; hence,

$$\tilde{\Psi}_{J_{t+1}}^* (y - \Psi x_{t+1}) = \tilde{\Psi}_{J_{t+1}}^* (\Psi(x^* - x_{t+1}) + z) = 0,$$

which implies

$$\Pi_{J_{t+1}} \tilde{\Psi}^* \Psi (x_{t+1} - x^*) = \Pi_{J_{t+1}} \tilde{\Psi}^* z. \quad (A.4.1)$$

Now,

$$\| \Pi_{J_{t+1}} (x_{t+1} - x^*) \|_2^2$$

$$= \langle x_{t+1} - x^*, \Pi_{J_{t+1}} (x_{t+1} - x^*) \rangle$$

$$= \langle x_{t+1} - x^*, (I_n - \Psi^* \tilde{\Psi}) \Pi_{J_{t+1}} (x_{t+1} - x^*) \rangle$$

$$+ \langle x_{t+1} - x^*, \Psi^* \tilde{\Psi} \Pi_{J_{t+1}} (x_{t+1} - x^*) \rangle$$

$$= \langle \Pi_{J^* \cup J_{t+1}} (x_{t+1} - x^*), (I_n - \Psi^* \tilde{\Psi}) \Pi_{J_{t+1}} (x_{t+1} - x^*) \rangle$$

$$+ \langle \Pi_{J_{t+1}} \tilde{\Psi}^* \Psi (x_{t+1} - x^*), \Pi_{J_{t+1}} (x_{t+1} - x^*) \rangle$$

$$\leq \langle x_{t+1} - x^*, \Pi_{J^* \cup J_{t+1}} (I_n - \Psi^* \tilde{\Psi}) \Pi_{J_{t+1}} (x_{t+1} - x^*) \rangle$$

$$+ \langle \Pi_{J_{t+1}} \tilde{\Psi}^* z, \Pi_{J_{t+1}} (x_{t+1} - x^*) \rangle$$

$$\leq \theta \| x_{t+1} - x^* \|_2 \| \Pi_{J_{t+1}} (x_{t+1} - x^*) \|_2$$

$$+ \sqrt{1 + \delta} \| z \|_2 \| \Pi_{J_{t+1}} (x_{t+1} - x^*) \|_2$$

(A.4.2)
where (a) follows from (A.4.1), and (b) holds since
\[ \| \Pi_{J^* \cup J_{t+1}} (I_n - \Psi^* \tilde{\Psi}) \Pi_{J_{t+1}} \| 
\leq \| I_q - \Psi^* \tilde{\Psi}_{J^* \cup J_{t+1}} \| \leq \theta \]

where \( q = |J^* \cup J_{t+1}| \leq 2s \).

It follows from (A.4.2) that
\[ \| \Pi_{J_{t+1}} (x_{t+1} - x^*) \|_2 \leq \theta \| x_{t+1} - x^* \|_2 + \sqrt{1 + \tilde{\delta}} \| z \|_2. \]

Therefore, (A.4.2) implies
\[ \| x_{t+1} - x^* \|_2^2 
= \| \Pi_{J_{t+1}} (x_{t+1} - x^*) \|_2^2 + \| \Pi_{J^* \setminus J_{t+1}} (x_{t+1} - x^*) \|_2^2 
\leq (\theta \| x_{t+1} - x^* \|_2 + \sqrt{1 + \tilde{\delta}} \| z \|_2)^2 + \| \Pi_{J^* \setminus J_{t+1}} x^* \|_2^2, \]
hence, we have
\[ \| x_{t+1} - x^* \|_2 \]
\[ \leq \frac{\theta \sqrt{1 + \tilde{\delta}} \| z \|_2 + \sqrt{1 + \tilde{\delta}} \| x_{t+1} - x^* \|_2 + (1 + \tilde{\delta}) \| z \|_2^2}{1 - \theta^2} \]
\[ \leq \frac{1}{\sqrt{1 - \theta^2}} \| \Pi_{J^* \setminus J_{t+1}} x^* \|_2 + \frac{\sqrt{1 + \tilde{\delta}}}{1 - \theta} \| z \|_2. \]

\( \square \)

**Proof of Lemma A.4.2.** Recall that \( J_{t+1} \) is chosen as the subset of \( \tilde{J}_{t+1} \) corresponding to the \( s \) largest elements of \((\tilde{\Psi}^*_{J_{t+1}} \tilde{\Psi}_{J_{t+1}})^{-1} \tilde{\Psi}^*_{J_{t+1}} y \); hence, it satisfies
\[ \| \Pi_{J_{t+1}} R_{\tilde{J}_{t+1}}^* (\tilde{\Psi}^*_{\tilde{J}_{t+1}} \tilde{\Psi}_{\tilde{J}_{t+1}})^{-1} \tilde{\Psi}^*_{\tilde{J}_{t+1}} y \| 
\geq \| \Pi_{J^*} R_{\tilde{J}_{t+1}}^* (\tilde{\Psi}^*_{\tilde{J}_{t+1}} \tilde{\Psi}_{\tilde{J}_{t+1}})^{-1} \tilde{\Psi}^*_{\tilde{J}_{t+1}} y \|, \]
which implies
\[ \| \Pi_{\tilde{J}_{t+1} \setminus J_{t+1}} R_{\tilde{J}_{t+1}}^* (\tilde{\Psi}^*_{\tilde{J}_{t+1}} \tilde{\Psi}_{\tilde{J}_{t+1}})^{-1} \tilde{\Psi}^*_{\tilde{J}_{t+1}} y \|. \]
\[
\leq \| \Pi_{J_t+1\setminus J_t} R_{J_t+1}^* (\tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1})^{-1} \tilde{\Psi}_{J_t+1}^* y \|. \tag{A.4.3}
\]

The left-hand side of (A.4.3) is the norm of the following term:

\[
\begin{align*}
\Pi_{J_t+1\setminus J_t} R_{J_t+1}^* (\tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1})^{-1} \tilde{\Psi}_{J_t+1}^* y \\
= \Pi_{J_t+1\setminus J_t} R_{J_t+1}^* (\tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1})^{-1} \tilde{\Psi}_{J_t+1}^* (\Psi x^* + z) \\
= \Pi_{J_t+1\setminus J_t} R_{J_t+1}^* (\tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1})^{-1} \tilde{\Psi}_{J_t+1}^* \\
\cdot (\Psi \Pi_{J_t} x^* + \Psi \Pi_{J_t^* \setminus J_t} x^* + z). \tag{A.4.4}
\end{align*}
\]

The first summand in (A.4.4) is rewritten as:

\[
\begin{align*}
\Pi_{J_t+1\setminus J_t} R_{J_t+1}^* (\tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1})^{-1} \tilde{\Psi}_{J_t+1}^* \Psi \Pi_{J_t} x^* \\
= \Pi_{J_t+1\setminus J_t} R_{J_t+1}^* (\tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1})^{-1} \tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1} R_{J_t+1} x^* \\
= \Pi_{J_t+1\setminus J_t} R_{J_t+1}^* R_{J_t+1} x^* \\
= \Pi_{J_t+1\setminus J_t} \Pi_{J_t} x^* \\
= \Pi_{J_t+1\setminus J_t} x^*. \tag{A.4.5}
\end{align*}
\]

By the RBOP, the other summands in (A.4.4) are bounded from above in the $\ell_2$ norm by

\[
\begin{align*}
\| \Pi_{J_t+1\setminus J_t} R_{J_t+1}^* (\tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1})^{-1} \tilde{\Psi}_{J_t+1}^* \Psi \Pi_{J_t^* \setminus J_t} x^* \|_2 \\
\leq \| (\tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1})^{-1} \| \cdot \| \tilde{\Psi}_{J_t+1}^* \Psi \Pi_{J_t^* \setminus J_t} x^* \|_2 \\
\leq \frac{\theta}{1 - \theta} \| \Pi_{J_t^* \setminus J_t} x^* \|_2 \tag{A.4.6}
\end{align*}
\]

and by

\[
\begin{align*}
\| \Pi_{J_t+1\setminus J_t} R_{J_t+1}^* (\tilde{\Psi}_{J_t+1}^* \Psi_{J_t+1})^{-1} \tilde{\Psi}_{J_t+1}^* z \|_2 \\
\leq \frac{\sqrt{1 + \delta}}{1 - \theta} \| z \|_2. \tag{A.4.7}
\end{align*}
\]

Combining eqs. (A.4.4) to (A.4.7) implies that the left-hand side of (A.4.3) is lower bounded by

\[
\| \Pi_{J_t+1\setminus J_t} x^* \|_2 - \frac{\theta}{1 - \theta} \| \Pi_{J_t^* \setminus J_t} x^* \|_2 - \frac{\sqrt{1 + \delta}}{1 - \theta} \| z \|_2. \tag{A.4.8}
\]
The right-hand side of (A.4.3) is the norm of the following term:

\[
\Pi_{\tilde{J}_{t+1}} J^* R_{\tilde{J}_{t+1}}^* (\tilde{\Psi}_{\tilde{J}_{t+1}}^* \Psi_{\tilde{J}_{t+1}})^{-1} \tilde{\Psi}_{\tilde{J}_{t+1}}^* y
\]

\[
= \Pi_{\bar{J}_{t+1}\setminus J^*} R_{\bar{J}_{t+1}}^* (\tilde{\Psi}_{\bar{J}_{t+1}}^* \Psi_{\bar{J}_{t+1}})^{-1} \tilde{\Psi}_{\bar{J}_{t+1}}^* (\Psi x^* + z)
\]

\[
= \Pi_{\bar{J}_{t+1}\setminus J^*} R_{\bar{J}_{t+1}}^* (\tilde{\Psi}_{\bar{J}_{t+1}}^* \Psi_{\bar{J}_{t+1}})^{-1} \tilde{\Psi}_{\bar{J}_{t+1}}^* \\
\cdot (\Psi \Pi_{\tilde{J}_{t+1}} x^* + \Psi \Pi_{J^* \setminus \tilde{J}_{t+1}} x^* + z).
\] (A.4.9)

Similarly to (A.4.5), the first summand in (A.4.9) is rewritten as

\[
\Pi_{\bar{J}_{t+1}\setminus J^*} R_{\bar{J}_{t+1}}^* (\tilde{\Psi}_{\bar{J}_{t+1}}^* y^* \Psi_{\bar{J}_{t+1}})^{-1} \tilde{\Psi}_{\bar{J}_{t+1}}^* \Psi \Pi_{\tilde{J}_{t+1}} x^* = \Pi_{\bar{J}_{t+1}\setminus J^*} x^* = 0.
\] (A.4.10)

In a similar way, the other summands in (A.4.9) are bounded from above in the \(\ell_2\) norm by

\[
\|\Pi_{\bar{J}_{t+1}\setminus J^*} R_{\bar{J}_{t+1}}^* (\tilde{\Psi}_{\bar{J}_{t+1}}^* \Psi_{\bar{J}_{t+1}})^{-1} \tilde{\Psi}_{\bar{J}_{t+1}}^* \Psi \Pi_{J^* \setminus \tilde{J}_{t+1}} x^*\|_2 \leq \frac{\theta}{1-\theta} \|\Pi_{J^* \setminus \tilde{J}_{t+1}} x^*\|_2
\] (A.4.11)

and

\[
\|\Pi_{\bar{J}_{t+1}\setminus J^*} R_{\bar{J}_{t+1}}^* (\tilde{\Psi}_{\bar{J}_{t+1}}^* \Psi_{\bar{J}_{t+1}})^{-1} \tilde{\Psi}_{\bar{J}_{t+1}}^* z\|_2 \leq \frac{\sqrt{1+\delta}}{1-\theta} \|z\|_2.
\] (A.4.12)

Combining eqs. (A.4.9) to (A.4.12) implies that the right-hand side of (A.4.3) is upper bounded by

\[
\frac{\theta}{1-\theta} \|\Pi_{J^* \setminus \tilde{J}_{t+1}} x^*\|_2 + \frac{\sqrt{1+\delta}}{1-\theta} \|z\|_2.
\] (A.4.13)

Therefore, by (A.4.8) and (A.4.13), we have

\[
\|\Pi_{\tilde{J}_{t+1}\setminus J_{t+1}} x^*\|_2 \leq \frac{2\theta}{1-\theta} \|\Pi_{J^* \setminus \tilde{J}_{t+1}} x^*\|_2 + \frac{2\sqrt{1+\delta}}{1-\theta} \|z\|_2.
\] (A.4.14)

Note that \(J^* \setminus J_{t+1} = (J^* \setminus \tilde{J}_{t+1}) \cup (J^* \cap (\tilde{J}_{t+1} \setminus J_{t+1}))\) and \(J^* \setminus \tilde{J}_{t+1} \cap (\tilde{J}_{t+1} \setminus J_{t+1})\) are disjoint. Therefore, since \(x^*\) is supported on \(J^*\), it follows that

\[
\|\Pi_{J^* \setminus J_{t+1}} x^*\|_2^2 = \|\Pi_{J^* \setminus \tilde{J}_{t+1}} x^*\|_2^2 + \|\Pi_{\tilde{J}_{t+1}\setminus J_{t+1}} x^*\|_2^2.
\] (A.4.15)
Applying (A.4.15) to (A.4.14), we obtain

\[
\sqrt{\|\Pi_{J^\star \setminus J_t+1} x^*\|^2_2 - \|\Pi_{J^\star \setminus J_t+1} x^*\|^2_2} \\
\leq \frac{2\theta}{1-\theta} \|\Pi_{J^\star \setminus J_t+1} x^*\|_2 + \frac{2\sqrt{1+\delta}}{1-\theta} \|z\|_2,
\]

which implies the desired inequality after simplification using \(\sqrt{a^2 + b^2} \leq a + b\) for \(a, b \geq 0\). \(\square\)

**Proof of Lemma A.4.3.** The last step in each iteration of ObSP updates \(x_t\) by

\[\quad x_t = R^{J_t}(\tilde{\Psi}^*_{J_t} \Psi_{J_t})^{-1} \tilde{\Psi}^*_{J_t} y.\]

Since \(\Psi\) and \(\tilde{\Psi}\) satisfy the RBOP, by Lemma 2.2.4, \(\tilde{\Psi}_{J_t}^{-1}(\tilde{\Psi}^*_{J_t} \Psi_{J_t})^{-1} \tilde{\Psi}^*_{J_t} \tilde{\Psi}^*_{J_t} y\) is a valid oblique projector onto \(R(\Psi_{J_t})\) along \(R(\tilde{\Psi}_{J_t})\).

Then, \(I_n - \Psi_{J_t}(\tilde{\Psi}^*_{J_t} \Psi_{J_t})^{-1} \tilde{\Psi}^*_{J_t} \Psi_{J_t}(\tilde{\Psi}^*_{J_t} \Psi_{J_t})^{-1} \tilde{\Psi}^*_{J_t} \tilde{\Psi}^*_{J_t}\) are also oblique projectors. Let \(E\) denote the oblique projection \(I_n - \Psi_{J_t}(\tilde{\Psi}^*_{J_t} \Psi_{J_t})^{-1} \tilde{\Psi}^*_{J_t}\) to simplify the notation. Then,

\[\tilde{\Psi}^*(y - \Psi x_t) = \tilde{\Psi}^* E y.\]

Let

\[J \triangleq \text{supp} \left( H_s(\tilde{\Psi}^* E y) \right).\]

Since \(E^* \tilde{\Psi}_J = 0\) for all \(j \in J_t\), it follows that \(J\) is disjoint from \(J_t\).

By definition of \(J\), we have

\[
\|\tilde{\Psi}^*_J E y\|_2 \geq \|\tilde{\Psi}^*_J, E y\|_2.
\]

hence, it follows that

\[
\|\tilde{\Psi}^*_J, E y\|_2 \geq \|\tilde{\Psi}^*_J, E y\|_2.
\]

(A.4.16)

Since \(E b_j = 0\) for all \(j \in J_t\), the left-hand side of (A.4.16) is the norm of the following term:

\[
\tilde{\Psi}^*_J, E y = \tilde{\Psi}^*_J, E (\Psi_{J^\star \setminus J_t} x^* + z). \quad \text{(A.4.17)}
\]

The first summand in (A.4.17) is upper bounded by

\[
\|\tilde{\Psi}^*_J, E \Psi_{J^\star \setminus J_t} x^*\|_2 \leq \|\tilde{\Psi}^*_J, E \Psi_{J^\star \setminus J_t}\| \|\Pi_{J^\star \setminus J_t} x^*\|_2
\]

\(\text{(s)}\)

138
where \((\ast)\) is upper bounded by

\[
\|\Pi_{\tilde{J}\setminus J^*} \Psi E \Psi\Pi_{J^*}\| \\
= \|\Pi_{\tilde{J}\setminus J^*} (\tilde{\Psi} E \Psi - I_n)\Pi_{J^*}\| \\
\leq \|\Pi_{(\tilde{J}\cup J^*)\setminus J^*} (\tilde{\Psi} E \Psi - I_n)\Pi_{(\tilde{J}\cup J^*)\setminus J^*}\| \\
= \|\tilde{\Psi} E (\tilde{J}\cup J^*) \setminus J^* - I_{(\tilde{J}\cup J^*)\setminus J^*}\| \\
\leq \|\tilde{\Psi} E (\tilde{J}\cup J^*) \setminus J^* - I_{(\tilde{J}\cup J^*)\setminus J^*}\| \leq \theta.
\]

Therefore,

\[
\|\tilde{\Psi} E \Psi\Pi_{J^*\setminus J^*} x^*\| \leq \theta \|\Pi_{J^*\setminus J^*} x^*\|. \tag{A.4.18}
\]

The first summand in (A.4.17) is upper bounded by

\[
\|\tilde{\Psi} E \Psi\Pi_{J^*\setminus J^*} x^*\| \\
\leq \|\tilde{\Psi} E\| \|\Pi_{J^*\setminus J^*}\| \|x^*\| \\
\leq (a) \sqrt{1 + \tilde{\delta} \|I_n - E\| \|z\|} \\
= \sqrt{1 + \tilde{\delta} \|\Psi_{J^*} (\tilde{\Psi} \Psi_{J^*})^{-1} \tilde{\Psi}_{J^*}\| \|z\|} \\
\leq \frac{\sqrt{1 + \delta (1 + \tilde{\delta})}}{1 - \theta} \|z\|. \tag{A.4.19}
\]

where (a) follows from Lemma A.1.7.

The right-hand side of (A.4.16) is the norm of the following term:

\[
\tilde{\Psi} E y = \tilde{\Psi} E_{\tilde{J}\setminus J^*+1} E y \\
= \tilde{\Psi} E_{\tilde{J}\setminus J^*+1} (\Psi_{J^*\setminus J^*} x^* + z) \\
= \tilde{\Psi} E_{\tilde{J}\setminus J^*+1} (\Psi_{J^*\setminus J^*} x^* + \Psi_{J^*\cap \tilde{J}^*} x^* + z) \tag{A.4.20}
\]

where the first equality holds since \(E_{\tilde{J}^*} \tilde{\psi}_j = 0\) for all \(j \in J^*\) and the last equality holds since \(J^* \setminus J^* = (J^* \setminus \tilde{J}_{t+1}) \cup (J^* \cap \tilde{J})\), and \(J^* \setminus \tilde{J}_{t+1}\) and \(J^* \cap \tilde{J}\) are disjoint.
Therefore, where (***) is further upper bounded by

\[
\|\tilde{\Psi}^{*}_{J^{*}\setminus J_{t+1}} E\Psi \Pi_{J^{*}\setminus J_{t+1}} x^*\|_2 \\
\geq \|\Pi_{J^{*}\setminus J_{t+1}}(\tilde{\Psi}^{*}_{J^{*}\setminus J_{t+1}} E\Psi \Pi_{J^{*}\setminus J_{t+1}})\| \|\Pi_{J^{*}\setminus J_{t+1}} x^*\|_2 \\
\geq \|\Pi_{J^{*}\setminus J_{t}}(\tilde{\Psi}^{*}_{J^{*}\setminus J_{t}} E\Psi \Pi_{J^{*}\setminus J_{t}})\| \|\Pi_{J^{*}\setminus J_{t+1}} x^*\|_2 \\
\geq (1 - \theta) \|\Pi_{J^{*}\setminus J_{t+1}} x^*\|_2. \quad (A.4.21)
\]

The second term in (A.4.20) is lower bounded by

\[
\|\tilde{\Psi}^{*}_{J^{*}\setminus J_{t+1}} E\Psi \Pi_{J^{*}\cap J} x^*\|_2 \\
\leq \|\tilde{\Psi}^{*}_{J^{*}\setminus J_{t+1}} E\Psi \Pi_{J^{*}\cap J}\| \|\Pi_{J^{*}\cap J} x^*\|_2
\]

where (***) is further upper bounded by

\[
\|\tilde{\Psi}^{*}_{J^{*}\setminus J_{t+1}} E\Psi \Pi_{J^{*}\cap J}\| \\
= \|\Pi_{J^{*}\setminus J_{t+1}} \tilde{\Psi}^{*} E\Psi \Pi_{J^{*}\cap J}\| \\
= \|\Pi_{J^{*}\setminus J_{t+1}} (\tilde{\Psi}^{*} E\Psi - I_n) \Pi_{J^{*}\cap J}\| \\
\leq \|\Pi_{(J^{*}\setminus J_{t+1})\cup (J^{*}\cap J)} (\tilde{\Psi}^{*} E\Psi - I_n) \Pi_{(J^{*}\setminus J_{t+1})\cup (J^{*}\cap J)}\| \\
\leq \|\tilde{\Psi}^{*}_{(J^{*}\setminus J_{t+1})\cup (J^{*}\cap J)} E\Psi_{(J^{*}\setminus J_{t+1})\cup (J^{*}\cap J)} - I_{(J^{*}\setminus J_{t+1})\cup (J^{*}\cap J)}\| \\
\leq \|\tilde{\Psi}^{*}_{(J^{*}\setminus J_{t+1})\cup (J^{*}\cap J)} E\Psi_{(J^{*}\setminus J_{t+1})\cup (J^{*}\cap J)} - I_{(J^{*}\setminus J_{t+1})\cup (J^{*}\cap J)}\| \\
\leq \|\tilde{\Psi}^{*}_{J^{*}\setminus J_{t+1}} E\Psi_{J^{*}\setminus J_{t+1}} - I_{J^{*}\setminus J_{t+1}}\| \leq \theta.
\]

Therefore,

\[
\|\tilde{\Psi}^{*}_{J^{*}\setminus J_{t+1}} E\Psi \Pi_{J^{*}\cap J} x^*\|_2 \leq \theta \|\Pi_{J^{*}\cap J} x^*\|_2. \quad (A.4.22)
\]

The last term in (A.4.20) is upper bounded by

\[
\|\tilde{\Psi}^{*}_{J^{*}\setminus J_{t}} E z\|_2 \leq \frac{\sqrt{1 + \delta(1 + \tilde{\delta})}}{1 - \theta} \|z\|_2. \quad (A.4.23)
\]
Applying eqs. (A.4.17) to (A.4.23) to eq. (A.4.16), we obtain

\[
\theta \|\Pi_{J^* \setminus J_t} x^*\|_2 + \frac{\sqrt{1 + \delta(1 + \tilde{\delta})}}{1 - \theta} \|z\|_2 \\
\geq (1 - \theta)\|\Pi_{J^* \setminus J_{t+1}} x^*\| - \theta\|\Pi_{J^* \cap J} x^*\|_2 \\
- \frac{\sqrt{1 + \delta(1 + \tilde{\delta})}}{1 - \theta} \|z\|_2.
\] (A.4.24)

Since

\[
\|\Pi_{J^* \setminus J_t} x^*\|_2^2 = \|\Pi_{J^* \setminus J_{t+1}} x^*\|_2^2 + \|\Pi_{J^* \cap J} x^*\|_2^2,
\]

(A.4.24) implies

\[
\theta \|\Pi_{J^* \setminus J_t} x^*\|_2 + \frac{2\sqrt{1 + \delta(1 + \tilde{\delta})}}{1 - \theta} \|z\|_2 \\
\geq (1 - \theta)\|\Pi_{J^* \setminus J_{t+1}} x^*\|_2 \\
- \theta \sqrt{\|\Pi_{J^* \setminus J_t} x^*\|_2^2 - \|\Pi_{J^* \setminus J_{t+1}} x^*\|_2^2}.
\] (A.4.25)

The final result is obtained by simplifying (A.4.25).

To simplify the notation, let

\[
a = \|\Pi_{J^* \setminus J_t} x^*\|_2 \\
b = \|\Pi_{J^* \setminus J_{t+1}} x^*\|_2 \\
c = \frac{2\sqrt{1 + \delta(1 + \tilde{\delta})}}{1 - \theta} \|z\|_2.
\]

Then, (A.4.25) reduces to

\[
\theta b + c \geq (1 - \theta)a - \theta \sqrt{b^2 - a^2},
\]

which is equivalent to

\[
\theta \sqrt{b^2 - a^2} \geq (1 - \theta)a - (\theta b + c).
\]

If \((1 - \theta)a \leq \theta b + c\), then

\[
a \leq \frac{\theta}{1 - \theta} b + \frac{1}{1 - \theta} c.
\] (A.4.26)
Otherwise, if \((1 - \theta)a > \theta b + c\), we have

\[
\theta^2(b^2 - a^2) \geq ((1 - \theta)a - \theta b - c)^2,
\]

which implies

\[
(2\theta^2 + 2\theta + 1)a^2 - 2(1 - \theta)(\theta b + c)a + (\theta b + c)^2 - \theta^2 b^2 \leq 0.
\]

Therefore,

\[
a \leq \frac{2\theta(1 - \theta)}{2\theta^2 + 2\theta + 1}b + \frac{2(1 - \theta)}{2\theta^2 + 2\theta + 1}c.
\]

(A.4.27)

Combining eqs. (A.4.26) and (A.4.27) completes the proof.

A.5 Proof of Theorem 2.3.1

Let \(X\) be a random variable defined as

\[
X \triangleq \max_{|J|=s} \left\| \Pi_J \left( \tilde{\Psi}^* \Psi - E \tilde{\Psi}^* \Psi \right) \Pi_J \right\|.
\]

Let \(\xi_k\) and \(\zeta_k\) be the transposed \(k\)th row of \(\sqrt{m} \Psi\) and \(\sqrt{m} \tilde{\Psi}\), respectively, for all \(k \in [n]\). By the assumption, \((\xi_k)_{k=1}^m\) and \((\zeta_k)_{k=1}^m\) are sequences of independent random vectors such that

\[
E \xi_k \xi_k^* = E \Psi^* \Psi \quad \text{and} \quad E \zeta_k \zeta_k^* = E \tilde{\Psi}^* \tilde{\Psi}
\]

for all \(k \in [m]\). Then, \(X\) is rewritten as

\[
X = \max_{|J|=s} \left\| \Pi_J \left( \frac{1}{m} \sum_{k=1}^m (\xi_k \xi_k^* - E \xi_k \xi_k^*) \right) \Pi_J \right\|.
\]

Like the RIP analysis for the Hermitian case [18, 20], the first step is to show

\[
E X \leq \frac{8\delta}{9}.
\]
By symmetrization \[20, \text{Lemma 6.7}\], \( \mathbb{E}X \) is bounded from above by

\[
\mathbb{E}X \leq 2 \mathbb{E} \max_{|J|=s} \left\| \Pi_J \left( \frac{1}{m} \sum_{k=1}^{m} \epsilon_k \zeta_k \xi_k^* \right) \Pi_J \right\|
\]

\[
= \frac{2}{m} \mathbb{E} \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^{m} \epsilon_k \zeta_k \xi_k^* \right) \Pi_J \right\| \quad (A.5.1)
\]

where \((\epsilon_k)_{k=1}^{m}\) is a Rademacher sequence independent of \((\xi_k)_{k=1}^{m}\) and \((\zeta_k)_{k=1}^{m}\).

Define random variables \(X_1\) and \(X_2\) by

\[
X_1 \triangleq \max_{|J|=s} \left\| \Pi_J \left( \Psi^* \Psi - \mathbb{E} \Psi^* \Psi \right) \Pi_J \right\|
\]

\[
X_2 \triangleq \max_{|J|=s} \left\| \Pi_J \left( \tilde{\Psi}^* \tilde{\Psi} - \mathbb{E} \tilde{\Psi}^* \tilde{\Psi} \right) \Pi_J \right\|.
\]

Then, \(X_1\) and \(X_2\) are rewritten as

\[
X_1 = \max_{|J|=s} \left\| \Pi_J \left( \frac{1}{m} \sum_{k=1}^{m} \xi_k \xi_k^* - \mathbb{E} \xi_k \xi_k^* \right) \Pi_J \right\|
\]

\[
X_2 = \max_{|J|=s} \left\| \Pi_J \left( \frac{1}{m} \sum_{k=1}^{m} \zeta_k \zeta_k^* - \mathbb{E} \zeta_k \zeta_k^* \right) \Pi_J \right\|.
\]

By symmetrization, \(\mathbb{E}X_1\) and \(\mathbb{E}X_2\) are bounded from above by

\[
\mathbb{E}X_1 \leq \frac{2}{m} \mathbb{E} \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^{m} \epsilon_k \xi_k \xi_k^* \right) \Pi_J \right\|
\]

\[
\mathbb{E}X_2 \leq \frac{2}{m} \mathbb{E} \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^{m} \epsilon_k \zeta_k \zeta_k^* \right) \Pi_J \right\|.
\]

**Lemma A.5.1 (\[18, \text{Lemma 3.8}\]).** Let \((h_k)_{k=1}^{m}\) be vectors in \(\mathbb{K}^n\). Let \(K_h \triangleq \max_k \|h_k\|_\infty\). Then,

\[
\mathbb{E} \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^{m} \epsilon_k h_k h_k^* \right) \Pi_J \right\|
\]

143
\[
\leq C_3 \sqrt{s \ln s \ln n \ln m K} \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^{m} h_k h_k^* \right) \Pi_J \right\|^{1/2}.
\]

Since \( \max_k \|\xi_k\|_{\infty} \leq K \), by Lemma A.5.1, it follows that

\[
\mathbb{E} X_1 \leq 2C_3 \sqrt{s \ln s \ln n \ln \ln m K} \sqrt{\mathbb{E} X_1 + 1 + \theta_s(\mathbb{E} \Psi^* \Psi)}.
\]

If

\[
2C_3 \sqrt{\frac{s}{m} \ln s \ln n \ln \ln m K} \sqrt{2 + \theta_s(\mathbb{E} \Psi^* \Psi)} \leq \delta_1
\]

for some \( \delta_1 < 1 \), then \( \mathbb{E} X_1 \leq \delta_1 \) and it follows that

\[
2C_3 \sqrt{\frac{s}{m} \ln s \ln n \ln \ln m K} \max_{|J|=s} \left\| \Pi_J \left( \frac{1}{m} \sum_{k=1}^{m} \zeta_k \zeta_k^* \right) \Pi_J \right\|^{1/2}
\]

\[
\leq 2C_3 \sqrt{\frac{s}{m} \ln s \ln n \ln \ln m K} \sqrt{\mathbb{E} X_1 + 1 + \theta_s(\mathbb{E} \Psi^* \Psi)}
\]

\[
\leq 2C_3 \sqrt{\frac{s}{m} \ln s \ln n \ln \ln m K} \sqrt{\delta_1 + 1 + \theta_s(\mathbb{E} \Psi^* \Psi)}
\]

\[
\leq 2C_3 \sqrt{\frac{s}{m} \ln s \ln n \ln \ln m K} \sqrt{2 + \theta_s(\mathbb{E} \Psi^* \Psi)} \leq \delta_1.
\]

(A.5.2)

Since \( \max_k \|\zeta_k\|_{\infty} \leq \tilde{K} \), by Lemma A.5.1, it follows that

\[
\mathbb{E} X_2 \leq 2C_3 \sqrt{\frac{s}{m} \ln s \ln n \ln \ln m \tilde{K} \sqrt{\mathbb{E} X_2 + 1 + \theta_s(\mathbb{E} \tilde{\Psi}^* \tilde{\Psi})}}.
\]

Similarly, if

\[
2C_3 \sqrt{\frac{s}{m} \ln s \ln n \ln \ln m \tilde{K}} \sqrt{2 + \theta_s(\mathbb{E} \tilde{\Psi}^* \tilde{\Psi})} \leq \delta_2
\]

for some \( \delta_2 < 1 \), then \( \mathbb{E} X_2 \leq \delta_2 \); hence,

\[
2C_3 \sqrt{\frac{s}{m} \ln s \ln n \ln \ln m \tilde{K}} \max_{|J|=s} \left\| \Pi_J \left( \frac{1}{m} \sum_{k=1}^{m} \zeta_k \zeta_k^* \right) \Pi_J \right\|^{1/2} \leq \delta_2.
\]

(A.5.3)

Unlike the conventional RIP analyses \([18, 20]\), matrices \( (\zeta_k \zeta_k^*)^m_{k=1} \) are not
Hermitian symmetric. The following lemma is modified from Lemma A.5.1 to get a bound on $\mathbb{E}X$ for the non-Hermitian case.

**Lemma A.5.2.** Let $(h_k)_{k=1}^m$ and $(\tilde{h}_k)_{k=1}^m$ be vectors in $\mathbb{K}^n$. Let $K_h \triangleq \max_k \|h_k\|_\infty$ and $K_{\tilde{h}} \triangleq \max_k \|\tilde{h}_k\|_\infty$. Then,

\[ \mathbb{E}_s \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^m \epsilon_k \tilde{h}_k h_k^* \right) \Pi_J \right\| \leq C_3 \sqrt{s} \ln s \sqrt{n} \sqrt{\ln m} \cdot \left[ K_h \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^m h_k h_k^* \right) \Pi_J \right\|^{1/2} + K_{\tilde{h}} \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^m \tilde{h}_k \tilde{h}_k^* \right) \Pi_J \right\|^{1/2} \right]. \quad (A.5.4) \]

**Proof of Lemma A.5.2.** By a comparison principle [137, inequality (4.8)], the left-hand side of (A.5.4), denoted by $E_1$, is bounded from above by

\[ E_1 \leq \frac{\sqrt{\pi}}{2} \mathbb{E}_g \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^m g_k \tilde{h}_k h_k^* \right) \Pi_J \right\| = \frac{\sqrt{\pi}}{2} \mathbb{E}_g \max_{|J|=s} \max_{x,y \in B_{2}^J} \left| \sum_{k=1}^m g_k y^* \tilde{h}_k h_k^* x \right| \]

where $(g_k)_{k=1}^m$ is the standard i.i.d. Gaussian sequence and $B_{2}^J \triangleq \{ x \in \mathbb{K}^n : \|x\|_2 \leq 1, \text{ supp } (x) \subset J \}$.

Define a Gaussian process $G_{x,y}$ indexed by $(x,y) \in \mathbb{K}^n \times \mathbb{K}^n$ as

\[ G_{x,y} \triangleq \sum_{k=1}^m g_k y^* \tilde{h}_k h_k^* x. \]

By Dudley’s inequality, $E_1$ is bounded from above by

\[ E_1 \leq C_4 \int_0^\infty \left( \ln N\left( \bigcup_{|J|=s} (B_{2}^J \times B_{2}^J), d, u \right) \right)^{1/2} du \]

where $N(B, d, u)$ is the covering number of set $B$ with respect to the metric.
$d$, induced from the Gaussian process $G_{x,y}$ by

$$d((x,y),(x',y')) \triangleq (\mathbb{E}|G_{x,y} - G_{x',y'}|^2)^{1/2}.$$  

Let

$$M_1 \triangleq \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^{m} h_k h_k^* \right) \Pi_J \right\|^{1/2}$$

and

$$M_2 \triangleq \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^{m} \tilde{h}_k \tilde{h}_k^* \right) \Pi_J \right\|^{1/2}.$$  

Define

$$\|x\|_h \triangleq \max_{k \in [n]} |h_k^* x| \quad \text{and} \quad \|x\|_{\tilde{h}} \triangleq \max_{k \in [n]} |\tilde{h}_k^* x|$$

for $x \in \mathbb{K}^n$. Then, $\| \cdot \|_h$ and $\| \cdot \|_{\tilde{h}}$ are valid norms on $\mathbb{K}^n$ induced by $(h_k)_{k=1}^m$ and $(\tilde{h}_k)_{k=1}^m$, respectively.

Let $x, x', y, y'$ be arbitrary $s$-sparse vectors in $\mathbb{K}^n$. Then, $d((x,y),(x',y))$ is upper bounded by

$$d((x,y),(x',y))^2 = \mathbb{E}|G_{x,y} - G_{x',y'}|^2$$

$$= \mathbb{E} \left| \sum_{k=1}^{m} g_k y^* \tilde{h}_k h_k^* (x - x') \right|^2$$

$$\leq \max_{k \in [n]} |h_k^* (x - x')|^2 \mathbb{E} \left| \sum_{k=1}^{m} g_k y^* \tilde{h}_k \right|^2$$

$$= \max_{k \in [n]} |h_k^* (x - x')|^2 \sum_{k=1}^{m} |y^* \tilde{h}_k|^2$$

$$= \|x - x'\|_h^2 \|h_k^* (x - x')\|_h^2 \sum_{k=1}^{m} y^* \tilde{h}_k \tilde{h}_k^* y$$

$$\leq M_2^2 \|x - x'\|_h^2$$ \hspace{1cm} (A.5.5)

where the fourth step follows since $(g_k)_{k=1}^m$ is the standard i.i.d. Gaussian sequence.

Similarly, $d((x',y),(x',y'))$ is upper bounded by

$$d((x',y),(x',y')) \leq M_1 \|y - y'\|_{\tilde{h}}.$$ \hspace{1cm} (A.5.6)

Then, by the triangle inequality and eqs. (A.5.5) and (A.5.6), it follows
that
\[
d((x, y), (x', y')) \leq d((x, y), (x', y)) + d((x', y), (x', y')) \\
\leq M_2\|x - x'\|_h + M_1\|y - y'\|_{\tilde{h}};
\]

Hence,
\[
\left( \ln N \left( \bigcup_{|J|=s} (B^J_2 \times B^J_2), d, u \right) \right)^{1/2} \\
\leq \left( \ln N \left( \bigcup_{|J|=s} B^J_2, M_2 \cdot \|\cdot\|_h, u/2 \right) \right)^{1/2} \\
+ \left( \ln N \left( \bigcup_{|J|=s} B^J_2, M_1 \cdot \|\cdot\|_{\tilde{h}}, u/2 \right) \right)^{1/2} \\
\leq 2M_2\sqrt{s} \left( \ln N \left( \bigcup_{|J|=s} \frac{1}{\sqrt{s}} B^J_2, \|\cdot\|_h, u \right) \right)^{1/2} \\
+ 2M_1 \left( \ln N \left( \bigcup_{|J|=s} \frac{1}{\sqrt{s}} B^J_2, \|\cdot\|_{\tilde{h}}, u \right) \right)^{1/2}.
\]

The remaining steps are identical to the Hermitian case ([20, Lemma 8.2], [18, Lemma 3.8]) and we do not reproduce the details. We obtain the desired bound by noting
\[
\int_0^\infty \left( \ln N \left( \bigcup_{|J|=s} \frac{1}{\sqrt{s}} B^J_2, \|\cdot\|_h, u \right) \right)^{1/2} \\
\leq C_4K_h \ln s \sqrt{\ln n} \sqrt{\ln m}
\]
and
\[
\int_0^\infty \left( \ln N \left( \bigcup_{|J|=s} \frac{1}{\sqrt{s}} B^J_2, \|\cdot\|_{\tilde{h}}, u \right) \right)^{1/2} \\
\leq C_4K_{\tilde{h}} \ln s \sqrt{\ln n} \sqrt{\ln m},
\]
which have been shown in the proof of [18, Lemma 3.8].

Let
\[
\delta_1 = \frac{K \sqrt{2 + \theta_s(\mathbb{E}\Psi^*\Psi)} \cdot \frac{8\delta}{9}}{K \sqrt{2 + \theta_s(\mathbb{E}\Psi^*\Psi)} + K \sqrt{2 + \theta_s(\mathbb{E}\tilde{\Psi}^*\tilde{\Psi})}}.
\]

147
and
\[ \delta_2 = \frac{\tilde{K} \sqrt{2 \theta_s (\mathbb{E} \bar{\Psi}^* \bar{\Psi})}} {K \sqrt{2 \theta_s (\mathbb{E} \Psi^* \Psi) + \tilde{K} \sqrt{2 \theta_s (\mathbb{E} \bar{\Psi}^* \bar{\Psi})}}} \cdot \frac{8 \delta}{9}. \]

Applying Lemma A.5.2 to (A.5.1), we obtain the following bound on \( \mathbb{E}X \).

If
\[ 2C_3 \sqrt{\frac{s}{m}} \ln s \ln n \ln \ln m \leq \frac{(8/9) \delta} {K \sqrt{2 \theta_s (\mathbb{E} \Psi^* \Psi) + \tilde{K} \sqrt{2 \theta_s (\mathbb{E} \bar{\Psi}^* \bar{\Psi})}}, \]

then
\[ \mathbb{E}X \leq 2 \mathbb{E} \max_{|J|=s} \left\| \Pi_J \left( \sum_{k=1}^{m} \epsilon_k \zeta_k \xi_k^* \right) \Pi_J \right\| \]
\[ \leq 2C_3 \sqrt{\frac{s}{m}} \ln s \ln n \ln \ln m \]
\[ \cdot \mathbb{E} \left[ K \max_{|J|=s} \left\| \Pi_J \left( \frac{1}{m} \sum_{k=1}^{m} \xi_k \xi_k^* \right) \Pi_J \right\|^{1/2} \]
\[ + \tilde{K} \max_{|J|=s} \left\| \Pi_J \left( \frac{1}{m} \sum_{k=1}^{m} \zeta_k \zeta_k^* \right) \Pi_J \right\|^{1/2} \]
\[ \leq \delta_1 + \delta_2 = \frac{8 \delta}{9} \]

where the last inequality follows from (A.5.2) and (A.5.3).

The second step is to show that \( X \) is concentrated around \( \mathbb{E}X \) with high probability. The corresponding result for the Hermitian case [20, Section 8.6] has been derived using a probabilistic upper bound on a random variable defined as the supremum of an empirical process [20, Theorem 6.25]. We show that the derivation for the Hermitian case [20, Section 8.6] extends to the non-Hermitian case with slight modifications.

Let \( B_J^2 \triangleq \{ x \in \mathbb{K}^n : \|x\|_2 \leq 1, \, \text{supp}(x) \subset J \} \). Since \( B_J^2 \) is closed under
the multiplication with any scalar of unit modulus, \( mX \) is written as

\[
mX = \max_{|J|=s} \max_{x,y \in B^J_2} \left| \sum_{k=1}^{m} y^* (\zeta_k \xi_k^* - E \zeta_k \xi_k^*) x \right|
\]

\[
= \max_{|J|=s} \max_{x,y \in B^J_2} \text{Re} \left( \sum_{k=1}^{m} y^* (\zeta_k \xi_k^* - E \zeta_k \xi_k^*) x \right).
\]

Define \( f_{x,y} : \mathbb{K}^{n \times n} \to \mathbb{R} \) by

\[
f_{x,y}(Z) \triangleq \text{Re} (y^* (Z - EZ)x).
\]

Then, \( E f_{x,y}(\zeta_k \xi_k^*) = 0 \) for all \( k \in [m] \) and \( mX \) is rewritten as

\[
mX = \max_{x,y} \left\{ \sum_{k=1}^{m} f_{x,y}(\zeta_k \xi_k^*) : (x, y) \in \bigcup_{|J|=s} (B^J_2 \times B^J_2) \right\}.
\]

Let \( k \in [m] \) be fixed. Let \( x, y \in B^J_2 \). Then,

\[
|f_{x,y}(\zeta_k \xi_k^*)| \leq |y^* \Pi_J (\zeta_k \xi_k^* - E \zeta_k \xi_k^*) \Pi_J x|
\]

\[
\leq \| \Pi_J (\zeta_k \xi_k^* - E \zeta_k \xi_k^*) \Pi_J \|_{\ell^1_2 \to \ell^n_2}^\frac{1}{2}
\]

\[
\leq \| \Pi_J (\zeta_k \xi_k^* - E \zeta_k \xi_k^*) \Pi_J \|_{\ell^1_2 \to \ell^n_2}^\frac{1}{2}
\]

\[
\cdot \| \Pi_J (\zeta_k \xi_k^* - E \zeta_k \xi_k^*) \Pi_J \|_{\ell^1_2 \to \ell^n_2}^\frac{1}{2}
\]

\[
= \| \Pi_J (\zeta_k \xi_k^* - E \zeta_k \xi_k^*) \Pi_J \|_{\ell^1_2 \to \ell^n_2}^\frac{1}{2}
\]

\[
\cdot \| \Pi_J (\xi_k \xi_k^* - E \xi_k \xi_k^*) \Pi_J \|_{\ell^1_2 \to \ell^n_2}^\frac{1}{2}
\]

(A.5.7)

where the third inequality follows from Schur’s interpolation theorem [138].

We derive an upper bound on \( \| \Pi_J (\zeta_k \xi_k^* - E \zeta_k \xi_k^*) \Pi_J \|_{\ell^1 \to \ell^n_1} \) by

\[
\| \Pi_J (\zeta_k \xi_k^* - E \zeta_k \xi_k^*) \Pi_J \|_{\ell^1 \to \ell^n_1}
\]

\[
\leq \| \Pi_J \zeta_k \xi_k^* \Pi_J \|_{\ell^n_1 \to \ell^n_1} + \| \Pi_J E \zeta_k \xi_k^* \Pi_J \|_{\ell^n_2 \to \ell^n_1}
\]

\[
\leq \| \Pi_J \zeta_k \xi_k^* \Pi_J \|_{\ell^n_1 \to \ell^n_1} + \| \Pi_J \zeta_k \xi_k^* \Pi_J \|_{\ell^n_2 \to \ell^n_1}
\]

\[
\leq 2sK \bar{K}
\]

(A.5.8)

where the second inequality follows from Jensen’s inequality, and the last
step holds since
\[ \|\Pi J \zeta_k \xi_k^* \Pi J\|_{\ell_1^q - \ell_1^q} \leq s \|\zeta_k\|_\infty \|\xi_k\|_\infty \leq sK \bar{K}. \]

Similarly, we have
\[ \|\Pi J (\xi_k \zeta_k^* - \mathbb{E}\zeta_k \zeta_k^*) \|_{\ell_1^q - \ell_1^q} \leq 2sK \bar{K}. \]  \hspace{1cm} (A.5.9)

By applying eqs. (A.5.8) and (A.5.9) to (A.5.7), we obtain
\[ |f_{x,y}(\zeta_k \xi_k^*)| \leq 2sK \bar{K}. \]  \hspace{1cm} (A.5.10)

Since \( k \) was arbitrary, (A.5.10) implies that \( f_{x,y}(\zeta_k \xi_k^*) \) is uniformly bounded for all \((x, y) \in \bigcup_{|J|=s} (B_{2I}^J \times B_{2J}^I) \) and for all \( k \in [m] \).

We also verify that the second moment of \( f_{x,y}(\zeta_k \xi_k^*) \) is uniformly bounded by
\[
\mathbb{E}|f_{x,y}(\zeta_k \xi_k^*)|^2 \\
= \mathbb{E}|x^* \Pi J (\zeta_k \xi_k^* - \mathbb{E}\zeta_k \zeta_k^*) \Pi J \|_2^2 \\
\leq \mathbb{E}\|\Pi J (\zeta_k \xi_k^* - \mathbb{E}\zeta_k \zeta_k^*) \Pi J \|_2^2 \\
= \mathbb{E}[x^* \left( \|\Pi J \zeta_k \|_2^2 \Pi J \xi_k \zeta_k^* \Pi J - \Pi J (\mathbb{E}\zeta_k \zeta_k^*) \Pi J \zeta_k \xi_k^* \Pi J \\
- \Pi J \xi_k \zeta_k^* \Pi J (\mathbb{E}\zeta_k \zeta_k^*) \Pi J + \Pi J (\mathbb{E}\zeta_k \zeta_k^*) \Pi J (\mathbb{E}\zeta_k \zeta_k^*) \Pi J \right) x] \\
= x^* \left( \|\Pi J \zeta_k \|_2^2 \Pi J (\mathbb{E}\zeta_k \zeta_k^*) \Pi J - \Pi J (\mathbb{E}\zeta_k \zeta_k^*) \Pi J (\mathbb{E}\zeta_k \zeta_k^*) \Pi J \right) x \\
\leq \|\Pi J \zeta_k \|_2^2 \|\Pi J (\mathbb{E}\zeta_k \zeta_k^*) \Pi J - \Pi J (\mathbb{E}\zeta_k \zeta_k^*) \Pi J \|_2^2 \\
\leq sK^2 (1 + \theta_s(\mathbb{E}\Psi^* \Psi)) + 1 + \theta_s(\mathbb{E}\Psi^* \Psi).
\]

Then, by [20, Theorem 6.25], we obtain (A.5.11).
\[
P(X \geq \delta) \leq \mathbb{P}\left(mX \geq m\mathbb{E}X + \frac{m\delta}{9}\right) \\
\leq \exp\left(-\frac{2m}{4sK^2} \left(1 + \theta_s(\mathbb{E}\Psi^* \Psi) + \frac{1 + \theta_s(\mathbb{E}\Psi^* \Psi)}{sK^2}\right)^2 - \frac{32\delta}{9} \cdot \frac{m}{2sK} + \frac{3\delta}{27} \cdot \frac{m}{2sK}\right)
\]

150
Therefore, \( \mathbb{P}(\theta_s(\bar{\Psi}^*\Psi) \geq \delta) \leq \eta \) holds provided that \( m \) satisfies

\[
m \geq C_1 \delta^{-2} \left( K \sqrt{2 + \theta_s(\mathbb{E}\Psi^*\Psi)} + \tilde{K} \sqrt{2 + \theta_s(\mathbb{E}\bar{\Psi}^*\bar{\Psi})} \right)^2 \cdot s(\ln s)^2 \ln n \ln m
\]

and

\[
m \geq C_2 \delta^{-2} \tilde{K} \max(K, \tilde{K}) s \ln(\eta^{-1})
\]

for universal constants \( C_1 \) and \( C_2 \).

### A.6 Proof of Theorem 2.3.4

Since \( \Psi = AD \), by the construction of \( A \) from \( (\phi_\omega)_{\omega \in \Omega} \) in (1.1.4), it follows that \( \max_{k,\ell} |\langle \Psi_{k,\ell} \rangle| \leq \sup_\omega \max_j |\langle \phi_\omega, d_j \rangle| \); hence, the incoherence property of \( \Psi \) is satisfied by the assumption. To invoke Corollary 2.3.2, it remains to show \( \theta_s(\mathbb{E}\Psi^*\Psi) \leq K_0 \). By the definition of \( \theta_s \), \( \theta_s(\mathbb{E}\Psi^*\Psi) \) is rewritten as

\[
\theta_s(\mathbb{E}\Psi^*\Psi) = \max \left[ \max_{\|J\|=s} \|D^*_j \mathbb{E} A^* AD_J\| - 1, 1 - \min_{\|J\|=s} \lambda_n(D^*_j \mathbb{E} A^* AD_J) \right].
\]  

(A.6.1)

Let \( J \) be an arbitrary subset of \([n]\) with \( s \) elements. Then, it follows that

\[
\|D^*_j \mathbb{E} A^* AD_J\| \leq \|D^*_j\| \|\mathbb{E} A^*\| \|D_J\|
\]

\[
= \|\mathbb{E} A^*\| \|D_J\|^2
\]

\[
\leq \nu_{\max} \|\Phi\Phi^*\| [1 + \delta_s(D)]
\]

\[
\leq \nu_{\max} [1 + \theta_d(\Phi\Phi^*)] [1 + \delta_s(D)]
\]

(A.6.2)
and

\[
\lambda_n(D^*_j E A^* A D_j) \geq \sigma_n(D^*_j) \lambda_n(E A^* A) \sigma_n(D_j) \\
\geq \lambda_n(E A^* A) \lambda_n(D^*_j D_j) \\
\geq \nu_{\min} \lambda_n(\Phi \Phi^*) [1 - \delta_n(D)] \\
\geq \nu_{\min} [1 - \theta_n(\Phi \Phi^*)][1 - \delta_n(D)].
\] (A.6.3)

Applying (A.6.2) and (A.6.3) to (A.6.1), we verify that \( K_0 \) given in (2.3.13) is a valid upper bound on \( \theta_s(\mathbb{E} \Psi^* \Psi) \). This completes the proof.

A.7 Proof of Theorem 2.3.5

First, we note that the mutual incoherence between \((\phi_\omega)_{\omega \in \Omega}\) and \((d_j)_{j \in [n]}\) is written as an operator norm given by

\[
\sup_{\omega \in \Omega} \max_{j \in [n]} |\langle \phi_\omega, d_j \rangle| = \| \Phi^* D \|_{\ell_1^n \rightarrow L_\infty(\Omega, \mu)}.
\]

Similarly, the mutual incoherence between \((\tilde{\phi}_\omega)_{\omega \in \Omega}\) and \((\tilde{d}_j)_{j \in [n]}\) is written as

\[
\sup_{\omega \in \Omega} \max_{j \in [n]} |\langle \tilde{\phi}_\omega, \tilde{d}_j \rangle| = \| \Lambda^{-1}_\nu \tilde{\Phi}^* \tilde{D} \|_{\ell_1^n \rightarrow L_\infty(\Omega, \mu)}
\]

where \( \Lambda^{-1}_\nu : L_2(\Omega, \mu) \rightarrow L_2(\Omega, \mu) \) is a diagonal operator defined by

\[
(\Lambda^{-1}_\nu h)(\omega) = \frac{d\mu}{d\nu}(\omega) h(\omega), \quad \forall h \in L_2(\Omega, \mu).
\]

Then, \( \| \Lambda^{-1}_\nu \tilde{\Phi}^* \tilde{D} \|_{\ell_1^n \rightarrow L_\infty(\Omega, \mu)} \) is upper bounded using \( K \) as follows:

\[
\| \Lambda^{-1}_\nu \tilde{\Phi}^* \tilde{D} \|_{\ell_1^n \rightarrow L_\infty(\Omega, \mu)} \\
= \| \Lambda^{-1}_\nu \Phi^* (\Phi \Phi^*)^{-1} D (D^* D)^{-1} \|_{\ell_1^n \rightarrow L_\infty(\Omega, \mu)} \\
\leq \| \Lambda^{-1}_\nu \Phi^* D (D^* D)^{-1} \|_{\ell_1^n \rightarrow L_\infty(\Omega, \mu)} \\
+ \| \Lambda^{-1}_\nu \Phi^* [ (\Phi \Phi^*)^{-1} - I_d ] D (D^* D)^{-1} \|_{\ell_1^n \rightarrow L_\infty(\Omega, \mu)} \\
\leq \frac{1}{\nu_{\min}} \| \Phi^* D \|_{\ell_1^n \rightarrow L_\infty(\Omega, \mu)} \| (D^* D)^{-1} \|_{\ell_1^n \rightarrow \ell_1^n} \\
+ \frac{1}{\nu_{\min}} \| \Phi^* \|_{\ell_2 \rightarrow L_\infty(\Omega, \mu)} \| (\Phi \Phi^*)^{-1} - I_d \|_{\ell_2^d \rightarrow \ell_2^d}
\]

152
\[ \| D\|_{\ell_1^r \to \ell_2^r} (D^* D)^{-1}\|_{\ell_1^r \to \ell_1^r} \leq \frac{\nu_{\min}}{\nu_{\max}} \left[ K + \left( \sup_{\omega \in \Omega} \| \phi_\omega \|_{\ell_2^d} \right) \right] \cdot \frac{\theta_d(\Phi \Phi^*)}{1 - \theta_d(\Phi \Phi^*)} \cdot \left( \max_{j \in [n]} \| d_j \|_{\ell_2^d} \right). \] (A.7.1)

Let \( \tilde{K} \) be the right hand side of (A.7.1). Then, we apply the incoherence parameters \( K \) and \( \tilde{K} \) to Corollary 2.3.3. Since \( \mathbb{E} \tilde{\Psi}^* \Psi = I_n \), we have \( \theta_s(\mathbb{E} \Psi^* \Psi) = 0 \). Therefore, to obtain a condition on \( m \), it only remains to bound \( \theta_s(\mathbb{E} \Psi^* \Psi) \) and \( \theta_s(\mathbb{E} \tilde{\Psi}^* \tilde{\Psi}) \).

In the proof of Theorem 2.3.4, we derived an upper bound on \( \theta_s(\mathbb{E} \Psi^* \Psi) \) given by

\[ \theta_s(\mathbb{E} \Psi^* \Psi) \leq \max(1 - \nu_{\min}, \nu_{\max} - 1) + \nu_{\max}[\delta_n(D) + \theta_d(\Phi \Phi^*) + \delta_n(D)\theta_d(\Phi \Phi^*)]. \] (A.7.2)

This upper bound is tight in the sense that equality is achieved if \( \theta_d(\Phi \Phi^*) = \delta_n(D) = 0 \), which holds, for example, for Fourier compressed sensing with signal sparsity over an orthonormal basis \( D \).

Similarly, we derive an upper bound on \( \theta_s(\mathbb{E} \tilde{\Psi}^* \tilde{\Psi}) \). Recall that \( \tilde{D} \) is written as

\[ \tilde{D} = D(D^* D)^{-1} = D(D^* D)^{-1/2}(D^* D)^{-1/2}. \]

Therefore, it follows that

\[ \| \tilde{D} \| \leq [\lambda_n(D^* D)]^{-1/2} \leq \frac{1}{\sqrt{1 - \delta_n(D)}} \] (A.7.3)

and

\[ \sigma_n(\tilde{D}) \geq [\lambda_1(D^* D)]^{-1/2} \geq \frac{1}{\sqrt{1 + \delta_n(D)}}. \] (A.7.4)

Similarly, since \( \tilde{\Phi} \) is written as

\[ \tilde{\Phi} = (\Phi \Phi^*)^{-1} \Phi = (\Phi \Phi^*)^{-1/2}(\Phi \Phi^*)^{-1/2} \Phi, \]

it follows that

\[ \| \tilde{\Phi} \| \leq [\lambda_d(\Phi \Phi^*)]^{-1/2} \leq \frac{1}{\sqrt{1 - \theta_d(\Phi \Phi^*)}} \] (A.7.5)
Using eqs. (A.7.3) to (A.7.6), we derive upper and lower bounds on the eigenvalues of $\mathbf{EE}^*\tilde{\Psi}$ as follows:

$$
\|\mathbf{EE}^*\tilde{\Psi}\| \leq \|\bar{D}^*\mathbf{E}\bar{A}^*\bar{A}\bar{D}\|
\leq \|\bar{D}^*\|\|\mathbf{E}\bar{A}^*\bar{A}\|\|\bar{D}\|
= \|\mathbf{E}\mathbf{A}^*\mathbf{A}\|\|\bar{D}\|^2
\leq \nu_{\text{min}}^{-1}\|\bar{\Phi}\|^2[1 - \delta_n(D)]^{-1}
\leq \nu_{\text{min}}^{-1}[1 - \theta_d(\Phi\Phi^*)]^{-1}[1 - \delta_n(D)]^{-1} \tag{A.7.7}
$$

and

$$
\lambda_n(\mathbf{EE}^*\tilde{\Psi}) \geq \lambda_n(\bar{D}^*\mathbf{E}\bar{A}^*\bar{A}\bar{D})
\geq \sigma_n(\bar{D}^*)\lambda_d(\mathbf{E}\bar{A}^*\bar{A})\sigma_n(\bar{D})
\geq \lambda_d(\mathbf{E}\bar{A}^*\bar{A})[1 + \delta_n(D)]^{-1}
\geq \nu_{\text{max}}^{-1}\lambda_d(\bar{\Phi}\bar{\Phi}^*)[1 + \delta_n(D)]^{-1}
\geq \nu_{\text{max}}^{-1}[1 + \theta_d(\Phi\Phi^*)]^{-1}[1 + \delta_n(D)]^{-1}. \tag{A.7.8}
$$

Then, we derive an upper bound on $\theta_s(\mathbf{EE}^*\tilde{\Psi})$ using eqs. (A.7.7) and (A.7.8) as follows:

$$
\theta_s(\mathbf{EE}^*\tilde{\Psi}) \leq \theta_n(\mathbf{EE}^*\tilde{\Psi})
= \max[1 - \lambda_n(\mathbf{EE}^*\tilde{\Psi}), \|\mathbf{EE}^*\tilde{\Psi}\| - 1]
\leq \max\left\{1 - \nu_{\text{max}}^{-1}[1 + \delta_n(D)]^{-1}[1 + \theta_d(\Phi\Phi^*)]^{-1}, \nu_{\text{min}}^{-1}[1 - \delta_n(D)]^{-1}[1 - \theta_d(\Phi\Phi^*)]^{-1} - 1\right\}
\leq \max(1 - \nu_{\text{max}}^{-1}, \nu_{\text{min}}^{-1} - 1)
+ \nu_{\text{min}}^{-1}\left\{\frac{\delta_n(D)}{1 - \delta_n(D)} + \frac{\theta_d(\Phi\Phi^*)}{1 - \theta_d(\Phi\Phi^*)} + \frac{\delta_n(D)\theta_d(\Phi\Phi^*)}{[1 - \delta_n(D)][1 - \theta_d(\Phi\Phi^*)]}\right\}. \tag{A.7.9}
$$
Combining (A.7.2) and (A.7.9), we obtain

\[
\max(\theta_s(\mathbb{E}\Psi^*\Psi), \theta_s(\mathbb{E}\tilde{\Psi}^*\tilde{\Psi})) \\
\leq \max(1 - \nu_{\text{max}}^{-1}, \nu_{\text{min}}^{-1} - 1) \\
+ \max(\nu_{\text{max}}, \nu_{\text{min}}^{-1}) \left\{ 1 + \frac{\delta_n(D)}{1 - \delta_n(D)} \right. \\
+ \frac{\theta_d(\Phi\Phi^*)}{1 - \theta_d(\Phi\Phi^*)} + \frac{\delta_n(D)\theta_d(\Phi\Phi^*)}{[1 - \delta_n(D)][1 - \theta_d(\Phi\Phi^*)]} \left. \right\}
\]

Applying this to Corollary 2.3.3 completes the proof.

A.8 Proof of Theorem 2.3.7

The proof of Theorem 2.3.7 is almost identical to that of Theorem 2.3.5. The mutual incoherence between \((\tilde{\phi}_\omega)_{\omega \in \Omega}\) and \((d_j)_{j \in [n]}\) is bounded in terms of \(K\) by

\[
\sup_{\omega \in \Omega} \max_{j \in [n]} \frac{|\langle \tilde{\phi}_\omega, d_j \rangle|}{\|\Phi^*\|_{\ell^d_2} \rightarrow L^\infty(\Omega,\mu)} \\
= \|\Lambda^{-1}_\nu \Phi^*(\Phi\Phi^*)^{-1} D\|_{\ell^d_2 \rightarrow L^\infty(\Omega,\mu)} \\
\leq \|\Lambda^{-1}_\nu \Phi^* D\|_{\ell^d_2 \rightarrow L^\infty(\Omega,\mu)} \\
+ \|\Lambda^{-1}_\nu \Phi^*(\Phi\Phi^*)^{-1} - I_d D\|_{\ell^d_2 \rightarrow L^\infty(\Omega,\mu)} \\
\leq \frac{1}{\nu_{\text{min}}} \|\Phi^* D\|_{\ell^d_2 \rightarrow L^\infty(\Omega,\mu)} \\
+ \frac{1}{\nu_{\text{min}}} \|\Phi^*\|_{\ell^d_2 \rightarrow L^\infty(\Omega,\mu)} \|\Phi\Phi^*\|^{-1} - I_d \|_{\ell^d_2 \rightarrow \ell^d_2} \| D\|_{\ell^d_2 \rightarrow \ell^d_2} \\
\leq \frac{1}{\nu_{\text{min}}} \left[ K + \left( \sup_{\omega \in \Omega} \|\phi_\omega\|_{\ell^d_2} \right) \frac{\theta_d(\Phi\Phi^*)}{1 - \theta_d(\Phi\Phi^*)} \cdot \left( \max_{j \in [n]} \|d_j\|_{\ell^d_2} \right) \right]. \quad (A.8.1)
\]

Let \(\tilde{K}\) be the right hand side of (A.8.1). Then, we apply the incoherence parameters \(K\) and \(\tilde{K}\) to Corollary 2.3.3. It remains to bound \(\theta_s(\mathbb{E}\Psi^*\Psi)\) and \(\theta_s(\mathbb{E}\tilde{\Psi}^*\tilde{\Psi})\).

In the proof of Theorem 2.3.4, we derived an upper bound on \(\theta_s(\mathbb{E}\Psi^*\Psi)\)
given by

$$\theta_s(\mathbb{E}\Psi^*\Psi) \leq \max(1 - \nu_{\text{min}}, \nu_{\text{max}} - 1)$$

\[ + \nu_{\text{max}}(\delta_s(D) + \theta_d(\Phi\Phi^*) + \delta_s(D)\theta_d(\Phi\Phi^*)). \tag{A.8.2} \]

Similarly to the proof of Theorem 1.1.10, \(\theta_s(\mathbb{E}\tilde{\Psi}^*\tilde{\Psi})\) is bounded by

$$\theta_s(\mathbb{E}\tilde{\Psi}^*\tilde{\Psi}) \leq \max \left\{ 1 - \nu_{\text{max}}^{-1}[1 - \delta_s(D)][1 + \theta_d(\Phi\Phi^*)]^{-1}, \right.$$

\[ \nu_{\text{min}}^{-1}[1 + \delta_s(D)][1 - \theta_d(\Phi\Phi^*)]^{-1} - 1 \left. \right\} \]

\[ \leq \max(1 - \nu_{\text{max}}^{-1}, \nu_{\text{min}}^{-1} - 1) + \nu_{\text{min}}^{-1} \cdot \left[ \delta_s(D) + \frac{\theta_d(\Phi\Phi^*)}{1 - \theta_d(\Phi\Phi^*)} + \frac{\delta_s(D)\theta_d(\Phi\Phi^*)}{1 - \theta_d(\Phi\Phi^*)} \right]. \tag{A.8.3} \]

Then, (A.8.2) and (A.8.3) imply

$$\max(\theta_s(\mathbb{E}\Psi^*\Psi), \theta_s(\mathbb{E}\tilde{\Psi}^*\tilde{\Psi}))$$

\[ \leq \max(1 - \nu_{\text{max}}^{-1}, \nu_{\text{min}}^{-1} - 1) + \max(\nu_{\text{max}}, \nu_{\text{min}}^{-1}) \]

\[ \cdot \left[ 1 + \delta_s(D) + \frac{\theta_d(\Phi\Phi^*)}{1 - \theta_d(\Phi\Phi^*)} + \frac{\delta_s(D)\theta_d(\Phi\Phi^*)}{1 - \theta_d(\Phi\Phi^*)} \right]. \tag{A.8.4} \]

Applying (A.8.4) to Corollary 2.3.3 completes the proof.
APPENDIX B

PROOFS FOR CHAPTER 3

B.1 Lemmata

We list a few lemmata that will be used for the proofs in the Appendix.

**Lemma B.1.1** ([135, Corollary III.1.5]). Let $A_1 \in \mathbb{K}^{m \times n_1}$ and $A_2 \in \mathbb{K}^{m \times n_2}$. Let $A \in \mathbb{K}^{m \times n}$ be the concatenation of $A_1$ and $A_2$, that is, $A = [A_1, A_2]$, where $n = n_1 + n_2$. Then, there is an interlacing relation between the singular values of $A$ and $A_1$ given by

$$\sigma_k(A) \geq \sigma_k(A_1) \geq \sigma_{k+n_2}(A)$$

for $k = 1, \ldots, n_1$.

**Lemma B.1.2.** Let $A \in \mathbb{K}^{m \times n}$ and let $J, J_0 \subset [n]$. Then, it follows that, for $k = 1, \ldots, |J_0 \setminus J|$, 

$$\lambda_k(A_{J_0 \cup J}^* A_{J_0 \cup J}) \geq \lambda_k(A_{J_0 \setminus J}^* P_{\mathcal{R}(A)}^\perp A_{J_0 \setminus J}) \geq \lambda_{k+|J|}(A_{J_0 \cup J}^* A_{J_0 \cup J}),$$

which is equivalent to

$$\sigma_k(A_{J_0 \cup J}) \geq \sigma_k(P_{\mathcal{R}(A)}^\perp A_{J_0 \setminus J}) \geq \sigma_{k+|J|}(A_{J_0 \cup J}^* A_{J_0 \cup J}).$$

**Proof.** Note that $A_{J_0 \setminus J}^* P_{\mathcal{R}(A)}^\perp A_{J_0 \setminus J}$ is the Schur complement of the block $A_J^* A_J$ of the matrix

$$\begin{bmatrix}
A_{J_0 \setminus J}^* A_{J_0 \setminus J} & A_{J_0 \setminus J}^* A_J \\
A_J^* A_{J_0 \setminus J} & A_J^* A_J
\end{bmatrix} = \Pi^* A_{J_0 \cup J}^* A_{J_0 \cup J} \Pi$$

157
where $\Pi$ is a permutation matrix that satisfies

$$A_{J_0 \cup J} \Pi = [A_{J_0 \setminus J}, A_J].$$

The application of the interlacing relation of the eigenvalues of the Schur complement [134] completes the proof.

Lemma B.1.3. Let $A \in \mathbb{K}^{m \times n}$ and $B \in \mathbb{K}^{n \times p}$ where $m \geq n$ and $p \geq n$. Then,

$$\|AB\| \geq \sigma_{n-k+1}(A) \cdot \sigma_k(B)$$

for $k = 1, \ldots, n$.

Proof. Let $A = U_1 \Sigma_1 V_1^*$ and $A = U_2 \Sigma_2 V_2^*$ denote the extended SVD of $A$ and $B$, respectively, where $\Sigma_1, V_1, U_2, \Sigma_2 \in \mathbb{K}^{n \times n}$. Let $k \in [n]$. Then,

$$\|AB\| = \|U_1 \Sigma_1 V_1^* U_2 \Sigma_2 V_2^*\|
= \|\Sigma_1 V_1^* U_2 \Sigma_2\|
= \|V_1 \Sigma_1 V_1^* U_2 \Sigma_2 U_2^*\|
\geq \lambda_{n-k+1}(V_1 \Sigma_1 V_1^*) \cdot \lambda_k(U_2 \Sigma_2 U_2^*)
= \sigma_{n-k+1}(A) \cdot \sigma_k(B)$$

where the inequality follows from the Gel’fand-Naimark theorem [135, Theor. III.4.5].

B.2 Proof of Proposition 3.4.4

We use the following lemma to prove Proposition 3.4.4.

Lemma B.2.1. Suppose that $\Phi \in \mathbb{K}^{s \times r}$, where $r \leq s$ satisfies

$$\text{krank}(\Phi^*) = r$$

and that $\Psi \in \mathbb{K}^{s \times k}$ for $k < r$ spans a $k$-dimensional subspace of $\mathcal{R}(\Phi)$. Then,

$$\text{krank}(\Psi^*) = k.$$
Proof. There exists \( R \in \mathbb{K}^{r \times k} \) such that \( R \) has full rank and \( \Psi = \Phi R \). Let \( K \) be a set of \( k \) indices from \([s]\). Since \( \operatorname{rank}(\Phi^*) = r \) implies \( \operatorname{rank}((\Phi^*)_K) = k \), \( \operatorname{rank}((\Psi^*)_K) = \operatorname{rank}(R^*(\Phi^*)_K) = \operatorname{rank}((\Phi^*)_K) = k \). Since \( K \) was arbitrary, we have \( \operatorname{rank}(\Psi^*) = k \).

Proof of Proposition 3.4.4. By the projection update formula, it follows that

\[
P_{R(A_{J_1})+\widehat{S}} = P_{R(A_{J_1})} + P_{P^\perp_{R(A_{J_1})}\widehat{S}}.
\]

Note that \( R(A_{J_1}) \) and \( P^\perp_{R(A_{J_1})}\widehat{S} \) are orthogonal to each other and both are subspaces of \( R(A_{J_0}) \). Furthermore, by assumption, \( \operatorname{rank}(A_{J_0}) = s \); hence, \( \operatorname{rank}(A_{J_1}) = |J_1| = s - r \). Therefore, it suffices to show that

\[
\dim(P^\perp_{R(A_{J_1})}\widehat{S}) = r. \tag{B.2.1}
\]

Let \( U \in \mathbb{K}^{n \times r} \) satisfy \( S = R(A_{J_0}U_{J_0}) \) and \( U^{[n]\setminus J_0} = 0 \). Then, (B.2.1) is equivalent to

\[
\operatorname{rank}(P^\perp_{R(A_{J_1})}A_{J_0}X_{0J_0}) = r
\]

which holds since

\[
\sigma_r(P^\perp_{R(A_{J_1})}A_{J_0}U_{J_0}) = \sigma_r(P^\perp_{R(A_{J_1})}A_{J_0\setminus J_1}U_{J_0\setminus J_1}) \overset{(a)}{\ge} \sigma_r(P^\perp_{R(A_{J_1})}A_{J_0\setminus J_1})\sigma_r(U_{J_0\setminus J_1}) \overset{(b)}{\ge} \sigma_s(A_{J_0})\sigma_r(U_{J_0\setminus J_1}) > 0.
\]

Inequalities (a) and (b) follow from Lemma B.1.3, which provides a lower bound on the minimum singular value of the product, and Lemma B.1.2, respectively, and the last step follows from the assumptions that \( A_{J_0} \) has full column rank and \( \operatorname{rank}((U_{J_0}^*)^*) = r \), which hold by Lemma B.2.1 because \( X_{0J_0} \) is row-nondegenerate.

\[
\]

B.3 Proof of Proposition 3.5.1

The proof of Proposition 3.5.1 is based on the following theorem by Davidson and Szarek [122].
Theorem B.3.1 ([122, Theor. II.13]). Given \( k, m \in \mathbb{N} \) with \( k \leq m \), consider the random matrix \( G \in \mathbb{R}^{m \times k} \) whose entries are i.i.d. Gaussian following \( \mathcal{N}(0, \frac{1}{m}) \). Then, for any \( t > 0 \)

\[
\begin{align*}
\mathbb{P}(\sigma_1(G) \geq 1 + \sqrt{\frac{k}{m} + t} ) & \leq \exp \left( -\frac{mt^2}{2} \right) \\
\mathbb{P}(\sigma_k(G) \leq 1 - \sqrt{\frac{k}{m} - t} ) & \leq \exp \left( -\frac{mt^2}{2} \right).
\end{align*}
\]

Proof of Proposition 3.5.1. Note that (3.5.5) implies

\[
1 - \sqrt{1 - \delta} \geq \sqrt{1 + \delta} - 1 \geq \sqrt{s + 1} - \frac{1}{m}.
\]

Let \( j \in [n] \setminus J \). Theorem B.3.1 implies

\[
\mathbb{P}\left( \sigma_1(A_{J \cup \{j\}}) \geq \sqrt{1 + \delta} \right) \leq \exp \left( -\frac{m}{2} \left( \sqrt{1 + \delta} - 1 - \sqrt{s + 1} \right)^2 \right)
\]

and

\[
\mathbb{P}\left( \sigma_{s+1}(A_{J \cup \{j\}}) \leq \sqrt{1 - \delta} \right) \leq \exp \left( -\frac{m}{2} \left( 1 - \sqrt{1 - \delta} - \sqrt{s + 1} \right)^2 \right).
\]

Since

\[
1 - \sqrt{1 - \delta} - \sqrt{s + 1} \geq \sqrt{1 + \delta} - 1 - \sqrt{s + 1} > 0
\]

it follows that

\[
\mathbb{P}\left( \|A^{*}_{J \cup \{j\}}A_{J \cup \{j\}} - I_{s+1}\| \geq \delta \right) \leq 2 \exp \left( -\frac{m}{2} \left( \sqrt{1 + \delta} - 1 - \sqrt{s + 1} \right)^2 \right).
\]

By considering the union of the events \( \|A^{*}_{J \cup \{j\}}A_{J \cup \{j\}} - I_{s+1}\| \geq \delta \) for all
\( j \in [n] \setminus J \), we obtain

\[
\mathbb{P} (\delta_{s+1}^{\text{weak}} (A; J) \geq \delta) \leq 2(n - s) \exp \left( -\frac{m}{2} \left( \sqrt{1 + \delta - 1 - \sqrt{\frac{s + 1}{m}}} \right)^2 \right).
\]

(B.3.1)

The right-hand side of (B.3.1) is bounded from above by \( \epsilon \) if

\[
(\sqrt{1 + \delta - 1}) \sqrt{m} \geq \sqrt{s + 1} + \sqrt{2 \ln \left( \frac{2(n - s)}{\epsilon} \right)},
\]

which is implied by

\[
(\sqrt{1 + \delta - 1}) \sqrt{m} \geq \sqrt{2} \sqrt{(s + 1) + 2 \ln \left( \frac{2(n - s)}{\epsilon} \right)},
\]

(B.3.2)

where we used the concavity of the square root function.

Noting that (B.3.2) coincides with (3.5.5) completes the proof. \( \square \)

B.4 Proof of Proposition 3.5.5

Let \( j \in [n] \setminus J \). Theorem B.3.1 implies

\[
\mathbb{P} (\sigma_{s+1} (A_{J \cup \{j\}}) \leq 1 - \gamma) \leq \exp \left( -\frac{m}{2} \left( \gamma - \sqrt{\frac{s + 1}{m}} \right)^2 \right).
\]

By considering the union of the events \((\sigma_{s+1} (A_{J \cup \{j\}}) \leq 1 - \gamma)\) for all \( j \in [n] \setminus J \), we obtain

\[
\mathbb{P} (\alpha_{s+1}^{\text{weak}} (A; J) \leq 1 - \gamma) \leq (n - s) \exp \left( -\frac{m}{2} \left( \gamma - \sqrt{\frac{s + 1}{m}} \right)^2 \right). \quad (B.4.1)
\]

Condition (3.5.9) implies that the right-hand side of (B.4.1) is bounded from above by \( \epsilon \).
B.5 Proof of Proposition 3.5.9

Let \( j \in [n] \setminus J \). Then, by [40, Lemma 2.1], \( A_{J \cup \{j\}} \) satisfies

\[
\mathbb{P}(\|A^*_{J \cup \{j\}} A_{J \cup \{j\}} - I_{s+1}\| \geq \delta) \leq 2(s+1) \exp \left( -\frac{m}{s} \cdot \frac{\delta^2}{2(1+\delta/3)} \right).
\]

Therefore, by considering the union of the events \( (\|A^*_{J \cup \{j\}} A_{J \cup \{j\}} - I_{s+1}\| \geq \delta) \) for all \( j \in [n] \setminus J \), we obtain

\[
\mathbb{P}(\delta_{s+1}(A; J) \geq \delta) \leq 2(n-s)(s+1) \exp \left( -\frac{m}{s} \cdot \frac{\delta^2}{2(1+\delta/3)} \right). \tag{B.5.1}
\]

Since (3.5.12) implies that the right-hand side of (B.5.1) is less than \( \epsilon \), the proof is complete.

B.6 Proof of Theorem 3.6.1

MUSIC finds \( J_0 \) if

\[
\min_{k \in J_0} \frac{\|P_S a_k\|_2}{\|a_k\|_2} > \max_{k \in [n] \setminus J_0} \frac{\|P_S a_k\|_2}{\|a_k\|_2}. \tag{B.6.1}
\]

Since \( \alpha_{s+1}^{\text{weak}}(A; J_0) > 0 \), it follows that all columns of \( A_{J_0} \) are linearly independent. Furthermore, since \( \text{rank}(X_{J_0}^H) = s \), we have

\[
S \triangleq \mathcal{R}(A_{J_0} X_{J_0}^H) = \mathcal{R}(A_{J_0}). \tag{B.6.2}
\]

By the triangle inequality,

\[
\left| \frac{\|P_S a_k\|_2}{\|a_k\|_2} - \frac{\|P_S a_k\|_2}{\|a_k\|_2} \right| \leq \frac{\|(P_S - P_S)a_k\|_2}{\|a_k\|_2} = \|P_S - P_S\| \leq \eta. \tag{B.6.3}
\]

Then, for all \( k \in J_0 \),

\[
\frac{\|P_S a_k\|_2}{\|a_k\|_2} \geq \frac{\|P_S a_k\|_2}{\|a_k\|_2} - \eta \geq \frac{\|P_{\mathcal{R}(A_{J_0})} a_k\|_2}{\|a_k\|_2} - \eta = 1 - \eta
\]

where (a) and (b) follow from (B.6.3) and (B.6.2), respectively.
Then, we obtain a lower bound on the left-hand side of (B.6.1) given by

\[
\min_{k \in J_0} \frac{\|P^*_b S_{a_k}\|_2}{\|a_k\|_2} \geq 1 - \eta. \quad \text{(B.6.4)}
\]

Similarly, by (B.6.2) and (B.6.3), we have

\[
\frac{\|P^*_b S_{a_k}\|_2}{\|a_k\|_2} \leq \frac{\|P_{R(AJ_0)} a_k\|_2}{\|a_k\|_2} + \eta
\]

for all \( k \in [n] \setminus J_0 \), where \( \|P_{R(AJ_0)} a_k\|_2 \) is further bounded from above by

\[
\frac{\|P_{R(AJ_0)} a_k\|_2^2}{\|a_k\|_2^2} = 1 - \frac{\|P_{R(AJ_0)}^\perp a_k\|_2^2}{\|a_k\|_2^2} = 1 - \frac{\sigma^2_1(P_{R(AJ_0)}^\perp a_k)}{\|a_k\|_2^2} \leq 1 - \frac{\sigma^2_{s+1}(A_{J_0 \cup \{k\}})}{\|a_k\|_2^2} \leq 1 - \frac{\{a^\text{weak}_{s+1}(A; J_0)\}^2}{\|a_k\|_2^2} \leq 1 - \frac{\{a^\text{weak}_{s+1}(A; J_0)\}^2}{\|A^*\|_{2,\infty}^2}
\]

where (c) holds by Lemma B.1.2 and (d) follows by the definition of the weak-1 asymmetric RIP. Then, we obtain an upper bound on the right-hand side of (B.6.1) given by

\[
\max_{k \in [n] \setminus J_0} \|P^*_b S_{a_k}\|_2 \leq \left[ 1 - \frac{\{a^\text{weak}_{s+1}(A; J_0)\}^2}{\|A^*\|_{2,\infty}^2} \right]^{1/2} + \eta. \quad \text{(B.6.5)}
\]

Combining (B.6.4) and (B.6.5), we note that MUSIC is guaranteed if \( A \) satisfies

\[
a^\text{weak}_{s+1}(A; J_0) > \alpha
\]

for \( \alpha > 0 \) satisfying

\[
\alpha \geq \|A^*\|_{2,\infty} \left\{ 1 - (1 - 2\eta)^2 \right\}^{1/2}.
\]
B.7 Proof of Proposition 3.6.6

Let \( d \triangleq \dim(P^\perp_{\overline{R}(A_J)} \overline{S}) \). By the row-nondegeneracy condition on \( X_0^J_0 \) and Lemma B.2.1, it follows that \( d \geq 1 \). There exists \( Q \in \mathbb{K}^{(s-k) \times d} \) with \( k = |J| \) such that \( Q^*Q = I_d \) and \( R(A_{J_0 \setminus J}Q) \) is a subspace of \( \overline{S} \). Then, it follows that

\[
\sigma_d(P^\perp_{\overline{R}(A_J)} A_{J_0 \setminus J} Q) = \sigma_d(P^\perp_{\overline{R}(A_J)} P_S A_{J_0 \setminus J} Q)
\]

\[
\leq \sigma_d(P^\perp_{\overline{R}(A_J)} P_S) \cdot \|A_{J_0 \setminus J}Q\|
\]

\[
\leq \sigma_d(P^\perp_{\overline{R}(A_J)} P_S) \cdot \|A_{J_0}\|.
\]  \hfill (B.7.1)

By the variational characterization of the singular values, (\( \star \)) is bounded from below by

\[
(\star) \geq \sigma_{s-k}(P^\perp_{\overline{R}(A_J)} A_{J_0 \setminus J}) \geq \sigma_s(A_{J_0})
\]  \hfill (B.7.2)

where the last step follows by Lemma B.1.2.

Let \( \kappa \triangleq \sigma_1(A_{J_0})/\sigma_s(A_{J_0}) \). Combining (B.7.1) and (B.7.2), we obtain

\[
\sigma_d(P^\perp_{\overline{R}(A_J)} P_S) \geq \kappa^{-1}.
\]  \hfill (B.7.3)

Since \( \dim(\overline{S}) = \dim(\hat{S}) = r \), it follows that

\[
\|P_{\overline{S}} - P_S\| = \sup_{\overline{x} \in \overline{S}} \inf_{\|\hat{x}\|_2 = 1} \|\overline{x} - \hat{x}\|_2 = \sup_{\hat{x} \in \hat{S}} \inf_{\|\overline{x}\|_2 = 1} \|\hat{x} - \overline{x}\|_2; \]  \hfill (B.7.4)

hence, \( \|P_{\overline{S}} - P_S\| \leq \eta \) implies the followings: for all \( \overline{x} \in \overline{S} \), there exists \( \hat{x} \in \hat{S} \) such that \( \|\overline{x} - \hat{x}\|_2 \leq \eta \|\overline{x}\|_2 \). Similarly, for all \( \hat{x} \in \hat{S} \), there exists \( \overline{x} \in \overline{S} \) such that \( \|\hat{x} - \overline{x}\|_2 \leq \eta \|\hat{x}\|_2 \).

The following identity is well known (see, for example, [139]): given two subspace \( S_1 \) and \( S_2 \),

\[
\|P_{S_1} - P_{S_2}\| = \max\{\|P^\perp_{S_1} P_{S_2}\|, \|P^\perp_{S_2} P_{S_1}\|\}.
\]  \hfill (B.7.5)
Note, by (B.7.5), that

$$
\| P_{\mathbb{R}(A_j)}^\perp S - P_{\mathbb{R}(A_j)}^\perp \bar{S} \| = \max \left\{ \left\| P_{\mathbb{R}(A_j)}^\perp S P_{\mathbb{R}(A_j)}^\perp \bar{S} \right\|, \left\| P_{\mathbb{R}(A_j)}^\perp S P_{\mathbb{R}(A_j)}^\perp \bar{S} \right\| \right\}.
$$

(B.7.6)

First, we derive an upper bound on (*), which is equivalently rewritten as

$$
(*) = \sup_{z \in P_{\mathbb{R}(A_j)}^\perp \bar{S}} \inf_{y \in P_{\mathbb{R}(A_j)}^\perp \bar{S}} \| z - y \|_2.
$$

(B.7.7)

Let $z \in P_{\mathbb{R}(A_j)}^\perp \bar{S}$ satisfy $\| z \|_2 = 1$. Then, by (B.7.3), there exists $\bar{x} \in \bar{S}$ such that $P_{\mathbb{R}(A_j)}^\perp \bar{x} = z$ and $\| \bar{x} \|_2 \leq \kappa$. By the argument following (B.7.4), there exists $\hat{x} \in \bar{S}$ such that $\| \bar{x} - \hat{x} \|_2 \leq \eta \| \bar{x} \|_2 \leq \eta \kappa$. Then, it follows that

$$
\| P_{\mathbb{R}(A_j)}^\perp \bar{x} - P_{\mathbb{R}(A_j)}^\perp \hat{x} \|_2 \leq \| \bar{x} - \hat{x} \|_2 \leq \eta \kappa
$$

and hence,

$$
\inf_{y \in P_{\mathbb{R}(A_j)}^\perp \bar{S}} \| z - y \|_2 \leq \eta \kappa.
$$

Since $z$ was an arbitrary unit-norm element in $P_{\mathbb{R}(A_j)}^\perp \bar{S}$, we obtain

$$
(*) \leq \eta \kappa.
$$

(B.7.8)

Next, we derive an upper bound on (**), which is equivalently rewritten as

$$
(**) = \sup_{z \in P_{\mathbb{R}(A_j)}^\perp \bar{S}} \inf_{y \in P_{\mathbb{R}(A_j)}^\perp \bar{S}} \| z - y \|_2.
$$

(B.7.9)

Since

$$
\sigma_d(P_{\mathbb{R}(A_j)}^\perp P_S) \leq \sigma_d(P_{\mathbb{R}(A_j)}^\perp P_{\bar{S}}) + \| P_{\mathbb{R}(A_j)}^\perp (P_{\bar{S}} - P_S) \|
\leq \sigma_d(P_{\mathbb{R}(A_j)}^\perp P_{\bar{S}}) + \| P_{\bar{S}} - P_S \|
\leq \sigma_d(P_{\mathbb{R}(A_j)}^\perp P_{\bar{S}}) + \eta
$$
by (B.7.3), we obtain a lower bound on \( \sigma_d(P_{R(A_J)}^\perp P_{\tilde{S}}) \) given by

\[
\sigma_d(P_{R(A_J)}^\perp P_{\tilde{S}}) \geq \kappa^{-1} - \eta = \frac{1 - \eta \kappa}{\kappa}. \tag{B.7.10}
\]

Let \( z \in P_{R(A_J)}^\perp \tilde{S} \) satisfy \( \|z\|_2 = 1 \). Then, by (B.7.10), there exists \( \hat{x} \in \tilde{S} \) such that \( P_{R(A_J)}^\perp \hat{x} = z \) and \( \|\hat{x}\|_2 \leq \frac{\kappa}{1 - \eta \kappa} \). By the argument following (B.7.4), there exists \( \bar{x} \in \tilde{S} \) such that \( \|\hat{x} - \bar{x}\|_2 \leq \eta \|\hat{x}\|_2 \leq \frac{\eta \kappa}{1 - \eta \kappa} \). Then, it follows that

\[
\| P_{R(A_J)}^\perp \hat{x} - P_{R(A_J)}^\perp \bar{x} \|_2 \leq \| \hat{x} - \bar{x} \|_2 \leq \frac{\eta \kappa}{1 - \eta \kappa}
\]

and hence,

\[
\inf_{y \in P_{R(A_J)}^\perp \tilde{S}} \| z - y \|_2 \leq \frac{\eta \kappa}{1 - \eta \kappa}.
\]

Since \( z \) was an arbitrary unit-norm element in \( P_{R(A_J)}^\perp \tilde{S} \), we obtain

\[
(\ast \ast) \leq \frac{\eta \kappa}{1 - \eta \kappa}. \tag{B.7.11}
\]

Applying (B.7.8) and (B.7.11) to (B.7.6) completes the proof.

### B.8 Proof of Theorem 3.6.7

MUSIC applied to \( \tilde{S} \) is successful if

\[
\min_{k \in J_0 \setminus J_1} \frac{\| P_{\tilde{S}} a_k \|_2}{\| a_k \|_2} > \max_{k \in [n] \setminus J_0} \frac{\| P_{\tilde{S}} a_k \|_2}{\| a_k \|_2}. \tag{B.8.1}
\]

Since (3.6.5) implies \( \text{rank}(A_{J_0}) = s \) and \( X_{0,j}^\perp \) is row-nondegenerate, by Proposition 3.4.4, we obtain

\[
\tilde{S} + \mathcal{R}(A_{J_1}) = \mathcal{R}(A_{J_0}), \tag{B.8.2}
\]

and hence, by the projection update formula,

\[
P_{\mathcal{R}(A_{J_0})} = P_{\mathcal{R}(A_{J_1})} + P_{P_{R(A_{J_1})}^\perp \tilde{S}}. \tag{B.8.3}
\]
Since the augmented subspace is given by $\mathcal{S} = \mathcal{S} + \mathcal{R}(A_{j_1})$, by the projection update formula, we have

$$P_\mathcal{S} = P_{\mathcal{R}(A_{j_1})} + P_{P_{\mathcal{R}(A_{j_1})}^\perp}\mathcal{S}. \quad \text{(B.8.4)}$$

By the triangle inequality, it follows that

$$\left| \frac{\|P_\mathcal{S}a_k\|_2 - \|P_{\mathcal{R}(A_{j_0})}a_k\|_2}{\|a_k\|_2} \right| \leq \frac{\|(P_\mathcal{S} - P_{\mathcal{R}(A_{j_0})})a_k\|_2}{\|a_k\|_2}$$

By (B.8.2), the last expression equals

$$= \frac{\|(P_{\mathcal{R}(A_{j_1})} + \tilde{\mathcal{S}} - P_{\mathcal{R}(A_{j_1})} + \mathcal{S})a_k\|_2}{\|a_k\|_2}$$

$$\leq \frac{\|P_{P_{\mathcal{R}(A_{j_1})}^\perp}\mathcal{S} - P_{P_{\mathcal{R}(A_{j_1})}^\perp}\mathcal{S}\|a_k\|_2}{\|a_k\|_2}$$

$$\leq \frac{\eta\sigma_1(A_{j_0})}{\sigma_s(A_{j_0}) - \eta\sigma_1(A_{j_0})} \leq \frac{\eta\beta}{\alpha - \eta}\quad \text{(B.8.5)}$$

where (a) follows from (B.8.3) and (B.8.4), and (b) follows from Proposition 3.6.6 because $\sigma_s(A_{j_0}) > \eta\sigma_1(A_{j_0})$ is implied by (3.6.5) and (3.6.6).

By (B.8.5), for all $k \in J_0 \setminus J_1$ we have

$$\frac{\|P_\mathcal{S}a_k\|_2}{\|a_k\|_2} > \frac{\|P_{\mathcal{R}(A_{j_0})}a_k\|_2}{\|a_k\|_2} - \frac{\eta\beta}{\alpha - \eta\beta} = 1 - \frac{\eta\beta}{\alpha - \eta\beta}.$$ 

This yields a lower bound on the left-hand side of (B.8.1) given by

$$\min_{k \in J_0 \setminus J_1} \frac{\|P_\mathcal{S}a_k\|_2^2}{\|a_k\|_2^2} > 1 - \frac{\eta\beta}{\alpha - \eta\beta}. \quad \text{(B.8.6)}$$
Similarly, by (B.8.5), for all \( k \in [n] \setminus J_0 \) we have

\[
\frac{\| P_S a_k \|_2}{\| a_k \|_2} < \frac{\| P_{\mathcal{R}(A_{J_0})} a_k \|_2}{\| a_k \|_2} + \frac{\eta \beta}{\alpha - \eta \beta}
\]

where (*) is further bounded from above by

\[
(*) = \frac{\| a_k \|_2^2 - \| P_{\mathcal{R}(A_{J_0})} a_k \|_2^2}{\| a_k \|_2^2} = 1 - \frac{\sigma_1^2(P_{\mathcal{R}(A_{J_0})} a_k)}{\| a_k \|_2^2}
\]

\[
\leq 1 - \frac{\sigma_{s+1}^2(A_{J_0 \cup \{k\}})}{\| a_k \|_2^2}
\]

\[
\leq 1 - \frac{[\alpha^{\text{weak}}(A; J_0)]^2}{\| a_k \|_2^2}
\]

\[
< 1 - \frac{\alpha^2}{\| a_k \|_2^2}
\]

where (c) follows by Lemma B.1.2. This yields an upper bound on the right-hand side of (B.8.1) given by

\[
\max_{k \in [n] \setminus J_0} \| P_S a_k \|_2^2 < \sqrt{1 - \frac{\alpha^2}{\| A^\ast \|_2^2,\infty^2} + \frac{\eta \sigma_1^2(A_{J_0})}{\sigma_2^2(A_{J_0}) - \eta \sigma_1^2(A_{J_0})}}.
\]

Finally, by applying the bounds in (B.8.5) and (B.8.6) to (B.8.1), we note that (B.8.1) is implied by the weak-1 asymmetric RIP given by

\[
1 - \frac{\eta \beta}{\alpha - \eta \beta} \geq \sqrt{1 - \frac{\alpha^2}{\| A^\ast \|_2^2,\infty^2} + \frac{\eta \beta}{\alpha - \eta \beta}},
\]

which is equivalent to (3.6.6). This completes the proof.

### B.9 Proof of Theorem 3.6.10

Assume that \( J \nsubseteq J_0 \) with \( k = |J| < s \) is given from the previous steps.
Define
\[ q_j \triangleq \frac{P_{\tilde{R}(A_j)}a_j}{\|P_{\tilde{R}(A_j)}a_j\|_2} \] (B.9.1)
for \( j \in [n] \). Then, \( \|q_j\|_2 = 1 \) for all \( j \in [n] \).

For the guarantee of the next step of OSMP, it suffices to show
\[
\max_{j \in J_0 \setminus J} \| (P_{P_{\tilde{R}(A_j)}A_j}^\perp) q_j \| > \max_{j \in [n] \setminus J_0} \| (P_{P_{\tilde{R}(A_j)}A_j}^\perp) s_j \|.\tag{B.9.2}
\]

By the triangle inequality, it follows that
\[
\left\| (P_{P_{\tilde{R}(A_j)}A_j}^\perp) q_j - (P_{P_{\tilde{R}(A_j)}A_j}^\perp) s_j \right\|
\leq \left\| (P_{P_{\tilde{R}(A_j)}A_j}^\perp) s_j - P_{P_{\tilde{R}(A_j)}A_j}^\perp s_j \right\| \leq \frac{\eta \beta}{\alpha - \eta \beta}
\]
where the last step follows from Proposition 3.6.6, since \( X_{0j} \) is row-nondegenerate, and (3.6.9) implies \( \alpha > \sqrt{\eta \beta} \).

Then, (B.9.2) is implied by
\[
\max_{j \in J_0 \setminus J} \| (P_{P_{\tilde{R}(A_j)}A_j}^\perp) q_j \| > \max_{j \in [n] \setminus J_0} \| (P_{P_{\tilde{R}(A_j)}A_j}^\perp) s_j \| + \frac{2\eta \beta}{\alpha - \eta \beta}. \tag{B.9.3}
\]

Let \( Q = [q_1, \ldots, q_n] \in \mathbb{K}^{m \times n} \), where \( q_j \) for each \( j \in [n] \) is defined in (B.9.1). Then, \( Q_{J_0 \setminus J} \) satisfies
\[
\sigma_{s-k}(Q_{J_0 \setminus J}) \geq \frac{\sigma_{s-k}(P_{P_{\tilde{R}(A_j)}A_{0j} \setminus J})}{\max_{j \in J_0 \setminus J} \| P_{P_{\tilde{R}(A_j)}A_j}^\perp a_j \|_2}
\]
\[
\geq \frac{\sigma_s(A_{0j})}{\max_{j \in J_0 \setminus J} \| a_j \|}
\geq \frac{\sigma_s(A_{0j})}{\| A^* \|_{2,\infty}}
\geq \frac{\alpha}{\| A^* \|_{2,\infty}} \tag{B.9.4}
\]
where (a) and (b) follow by Lemma B.1.2 and Lemma B.1.1, respectively, and (c) is implied by (3.6.8).
First, we bound the left-hand side of (B.9.3) from below by

\[
\max_{j \in J_0 \setminus J} \| (P_{P_{R(A_j)}^\perp}^\perp) q_j \|
\]

\[
= \max_{j \in J_0 \setminus J} \| q_j^* P_{P_{R(A_j)}^\perp}^\perp \|
\]

\[
= \| Q_{J_0 \setminus J}^* P_{P_{R(A_j)}^\perp}^\perp \|_{2,\infty}
\]

\[
\geq \frac{\| Q_{J_0 \setminus J}^* P_{P_{R(A_j)}^\perp}^\perp \|_F}{\sqrt{s - k}}
\]

\[
\geq \left( \frac{\sum_{\ell=1}^{\dim(P_{P_{R(A_j)}^\perp}^\perp)} \sigma_{s-k+1-\ell}^2 (Q_{J_0 \setminus J})}{s - k} \right)^{1/2}
\]

(B.9.5)

where (d) follows from the variational characterization of singular values [135]. For the special case when \( r = s \), \( \tilde{S} \) is given by \( \tilde{S} = S = R(A_{J_0}) \); hence,

\[
\dim(P_{R(A_j)}^\perp \tilde{S}) = \operatorname{rank}(P_{R(A_j)}^\perp A_{J_0 \setminus J}) = s - k,
\]

(B.9.6)

which follows since \( \sigma_{s-k}(P_{R(A_j)}^\perp A_{J_0 \setminus J}) \geq \sigma_s(A_{J_0}) > 0 \) by Lemma B.1.2. In this case, the right-hand side of (B.9.5) reduces to

\[
(*) = \left( \frac{\sum_{\ell=1}^{s-k} \sigma_{s-k+1-\ell}^2 (Q_{J_0 \setminus J})}{s - k} \right)^{1/2}
\]

\[
= \left( \frac{\| Q_{J_0 \setminus J} \|_F^2}{s - k} \right)^{1/2} = \left( \sum_{\ell=1}^{s-k} \| q_j \|_2^2 \right)^{1/2} = 1.
\]

(B.9.7)

Otherwise, if \( r < s \), we derive a lower bound on the right-hand side of (B.9.5) by

\[
(*) \geq \sqrt{\frac{\dim(P_{R(A_j)}^\perp \tilde{S})}{s - k}} \cdot \sigma_{s-k}(Q_{J_0 \setminus J}) > \frac{\alpha}{\| A^* \|_{2,\infty}} \cdot \sqrt{\frac{\dim(P_{R(A_j)}^\perp \tilde{S})}{s - k}}
\]

(B.9.8)

where (e) follows from (B.9.4).

Next, we derive an upper bound on the right-hand side of (B.9.3). Assume that \( j \in [n] \setminus J_0 \). Since

\[
P_{R(A_j)}^\perp \tilde{S} \subset P_{R(A_j)}^\perp R(A_{J_0}) = R(P_{R(A_j)}^\perp A_{J_0}) = R(P_{R(A_j)}^\perp A_{J_0 \setminus J}),
\]

170
it follows that

\[ \| (P_{R(A_J)}^\perp \xi) q_j \|_2 \leq \| (P_{R(P_{R(A_J)}^\perp A_{J_0})}) q_j \|_2. \]  \hfill (B.9.9)

Since \( P_{R(A_J)}^\perp A_{J_0} \) has full column rank by (B.9.6), we can define

\[ R \triangleq [A_{J_0}^* P_{R(A_J)}^\perp A_{J_0} \] \(-1/2\)

and

\[ \Phi \triangleq P_{R(A_J)}^\perp A_{J_0} R. \]

Then, \( \Phi \) is an orthonormal basis for \( R(P_{R(A_J)}^\perp A_{J_0}) \).

Since \( A_J \) has full column rank, we can also define

\[ \Psi \triangleq A_J (A_J^* A_J)^{-1/2}. \]

Then, \( \Psi \) is an orthonormal basis for \( R(A_J) \).

We bound the right-hand side of (B.9.9) from above by

\[
\begin{align*}
\| (P_{R(P_{R(A_J)}^\perp A_{J_0})}) q_j \|_2 & = \| \Phi^* q_j \|_2 \\
& \overset{(f)}{\leq} \left\| [q_j, \Phi]^* [q_j, \Phi] - I_{s-k+1} \right\| \\
& = \left\| [q_j, \Phi]^* P_{R(A_J)}^\perp [q_j, \Phi] - I_{s-k+1} \right\| \\
& = \max \left\{ 1 - \lambda_{s-k+1} \left( [q_j, \Phi]^* P_{R(A_J)}^\perp [q_j, \Phi] \right), \left\| [q_j, \Phi]^* P_{R(A_J)}^\perp [q_j, \Phi] \right\| - 1 \right\} \\
& \overset{(g)}{\leq} \max \left\{ 1 - \lambda_s \left( [q_j, \Phi, \Psi]^* [q_j, \Phi, \Psi] \right), \left\| [q_j, \Phi, \Psi]^* [q_j, \Phi, \Psi] \right\| - 1 \right\} \\
& = \left\| [q_j, \Phi, \Psi]^* [q_j, \Phi, \Psi] - I_{s+1} \right\| 
\end{align*}
\]  \hfill (B.9.10)

where (f) and (g) follow from Lemma B.1.1 and Lemma B.1.2, respectively.

Note that \([\Phi, \Psi]\) is an orthonormal basis of \( R(A_{J_0}) \). Therefore, the last term of (B.9.10) is bounded from above by

\[
\left\| [q_j, \Phi, \Psi]^* [q_j, \Phi, \Psi] - I_{s+1} \right\|^2 \\
= \left\| \begin{bmatrix} 0 & q_j^* [\Phi, \Psi] \\
[\Phi, \Psi]^* q_j & 0 \end{bmatrix} \right\|^2
\]

171
\[ \begin{bmatrix} 0 & \|\Phi, \Psi\|^2 \end{bmatrix} \begin{bmatrix} \|\Phi, \Psi\|^2 \end{bmatrix} = 1 \]

\[ = \|P_{R(A_j)}q_j\|^2 \]

\[ = 1 - \|P_{R(A_j)}q_j\|^2 \]

\[ = 1 - \frac{\|P_{R(A_j)}q_j\|^2}{\|P_{R(A_j)}a_j\|^2} \]

\[ \leq 1 - \frac{\sigma^2_{s+1}(A_{J_0\cup j})}{\|a_j\|^2} \]

\[ \leq 1 - \frac{\alpha^2}{\|A^*\|^2_{2,\infty}} \]

where (h) follows from Lemma B.1.2 and (i) is implied by (3.6.8).

Since \( j \) was arbitrary in \([n] \setminus J_0\), (B.9.10) implies that

\[ \max_{j \in [n] \setminus J} \|P_{R(P_{R(A_j)}A_{J_0\cup j})})q_j\|_2 < \sqrt{1 - \frac{\alpha^2}{\|A^*\|^2_{2,\infty}}} \].

(B.9.11)

For the full row rank case, applying (B.9.7), and (B.9.11) to (B.9.3), we obtain Condition (3.6.10), which implies (B.9.2). For the rank-defective case, applying (B.9.8), and (B.9.11) to (B.9.3), we obtain the following sufficient condition for (B.9.2)

\[ \sqrt{\frac{\dim(P_{R(A_j)}S^c)}{s - k}} \cdot \frac{\alpha}{\|A^*\|^2_{2,\infty}} - \sqrt{1 - \frac{\alpha^2}{\|A^*\|^2_{2,\infty}}} \geq \frac{2\eta\beta}{\alpha - \eta\beta} \].

(B.9.12)

It remains to show that \( \dim(P_{R(A_j)}S^c) = \min(s - k, r) \) for (3.6.9) to imply (B.9.12).

By Remark 3.4.6, there exists a row \( s \)-sparse matrix \( U \in \mathbb{K}^{n \times r} \) with support \( J_0 \) such that \( S^c = R(A_{J_0}U_{J_0}) \). Then \( \dim(P_{R(A_j)}S^c) = \text{rank}(P_{R(A_j)}A_{J_0\cup J}U_{J_0\cup J}) \).

We have shown that \( P_{R(A_j)}A_{J_0\cup J} \) has full column rank; hence, the nullity of \( P_{R(A_j)}A_{J_0\cup J}U_{J_0\cup J} \) coincides with the nullity of \( U_{J_0\cup J} \). The row-nondegeneracy condition on \( X_0^{J_0} \) implies that \( \text{rank}(U_{J_0\cup J}) = \min(s - k, r) \); hence, the nullity of \( U_{J_0\cup J} \) is \( \max(k - s + r, 0) \). Therefore, it follows that \( \dim(P_{R(A_j)}A_{J_0\cup J}U_{J_0\cup J}) = s - k - \max(k - s + r, 0) = \min(s - k, r) \).
B.10 Proof of Proposition 3.7.1

To prove Proposition 3.7.1, we use the following lemmata. The proof of Lemma B.10.2 is deferred after the proof of Proposition 3.7.1.

**Lemma B.10.1** (Case of [135, Theor. VII.3.3]). For positive semidefinite matrices $\Gamma_1, \Gamma_2 \in \mathbb{K}^{m \times m}$ and $r \in \{1, \ldots, m-1\}$, let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the subspaces spanned by the $r$-dominant eigenvectors of $\Gamma_1$ and $\Gamma_2$, respectively. Then,

$$3\|\Gamma_1 - \Gamma_2\| \geq \| P_{\mathcal{S}_1} P_{\mathcal{S}_2} \| \cdot [\lambda_r(\Gamma_2) - \lambda_{r+1}(\Gamma_1)].$$

**Lemma B.10.2.** Given $m, n \in \mathbb{N}$ with $n > 2m$, consider the random matrix $G \in \mathbb{C}^{n \times m}$ whose entries are i.i.d. circular Gaussian variables with zero mean and variance $\frac{1}{n}$. Then,

$$\mathbb{P} \left( \|G^* G - I_m\| \geq 3\sqrt{\frac{2m}{n}} + t \right) \leq 4 \exp \left( -\frac{nt^2}{18} \right)$$

for $t > 0$ satisfying

$$\sqrt{\frac{2m}{n}} + \frac{t}{3} \leq 1.$$

**Proof of Proposition 3.7.1.** Recall that, given the sample covariance matrix $\Gamma_Y \triangleq \frac{YY^*}{N}$ of the snapshots $Y$, Algorithm 1 computes $\hat{\Gamma}$ defined by

$$\hat{\Gamma} = \Gamma_Y - \lambda_m(\Gamma_Y) I_m.$$

Also recall that $\Gamma_\mathcal{S}$ is defined by

$$\Gamma_\mathcal{S} \triangleq \frac{A_{J_0} X_{J_0} (X_{J_0}^*) A_{J_0}^*}{N} = \frac{A_{J_0} \Psi \Lambda \Phi \Phi^* \Lambda \Psi^* A_{J_0}^*}{N}$$

where $A_{J_0} X_{J_0} = \Gamma^{1/2} \Phi$ are i.i.d. Gaussian vectors with given covariance matrix $\Gamma$.

As $N$ goes to infinity, $\frac{\Phi \Phi^*}{N}$ converges to the identity matrix; hence, the sample covariance matrix $\hat{\Gamma}_\mathcal{S}$ converges to $\Gamma$. The assumptions of Proposition 3.7.1 are on $\Gamma$ rather than on $\Gamma_\mathcal{S}$ because $\Gamma$ determines the signal model.
and is independent of $N$. Since $N$ is finite in Proposition 3.7.1, we need to bound the deviation of $\frac{\Phi^*}{N}$ from the identity matrix in the following derivation.

Let $D \in \mathbb{K}^{m \times m}$ denote the distortion matrix defined by

$$D \triangleq \hat{\Gamma} - \Gamma_S.$$ 

Then, $D$ is decomposed as

$$D = D_{\text{noise}} + D_{\text{cross}} + D_{\text{bias}}$$

where

$$D_{\text{noise}} \triangleq \frac{W W^*}{N} - \sigma_w^2 I_m,$$

$$D_{\text{cross}} \triangleq \frac{A J_0 X_0^* W^*}{N} + \frac{W(X_0^* A J_0)}{N},$$

$$D_{\text{bias}} \triangleq \left[ \sigma_w^2 - \lambda_m(\Gamma_Y) \right] I_m.$$

It follows that the sample covariance matrix $\Gamma_Y$ is decomposed as $\Gamma_Y = \Gamma_S + \sigma_w^2 I_m + D_{\text{cross}} + D_{\text{noise}}$. Viewing $\Gamma_Y$ as a perturbed version of $\Gamma_S + \sigma_w^2 I_m$ with distortion $D_{\text{noise}} + D_{\text{cross}}$, we bound the perturbation in the $m$th eigenvalue by Weyl’s Theorem as

$$\left| \lambda_m(\Gamma_Y) - \lambda_m(\Gamma_S + \sigma_w^2 I_m) \right| \leq \|\Gamma_Y - \Gamma_S - \sigma_w^2 I_m\|$$

$$\geq \|D_{\text{noise}} + D_{\text{cross}}\|$$

$$\leq \|D_{\text{noise}}\| + \|D_{\text{cross}}\|.$$

Now, since $\text{rank}(\Gamma_S) \leq s$ and $s < m$, (\*) reduces to

$$\left| \sigma_w^2 - \lambda_m(\Gamma_Y) \right| = \|D_{\text{bias}}\|.$$ 

Therefore, $\|D_{\text{bias}}\|$ satisfies

$$\|D_{\text{bias}}\| \leq \|D_{\text{noise}}\| + \|D_{\text{cross}}\|.$$
For \( k \in \mathbb{N} \), where \( k < N \), let \( Z_k \in \mathbb{K}^{k \times N} \) be an i.i.d. Gaussian matrix such that \( \mathbb{E}Z_k = 0 \) and \( \mathbb{E}Z_k(Z_k)^* = I_k \). Then, we define for \( k = 1, \ldots, N - 1 \) a family of random variables

\[
\xi_k \triangleq \left\| \frac{Z_k(Z_k)^*}{N} - I_k \right\|.
\]

Using \( \xi_k \), we bound the distortion terms from above by

\[
\|D_{\text{noise}}\| = \sigma_w^2 \left\| \frac{WW^*}{\sigma_w^2 N} - I_m \right\| \sim \xi_m \sigma_w^2
\]

and

\[
\|D_{\text{cross}}\| \leq 2\sigma_w \| A J_0 \Psi \Lambda \| \left\| \frac{\Phi W^*}{\sigma_w N} \right\|
\]

\[
\leq 2\sigma_w \lambda_1^{1/2}(\Gamma) \left[ \frac{1}{N} \left[ \Phi^*, \frac{W^*}{\sigma_w} \right] \left[ \Phi^*, \frac{W^*}{\sigma_w} \right] - I_{m+M} \right]
\]

\[
\sim 2\xi_{m+M} \sigma_w \lambda_1^{1/2}(\Gamma)
\]

where \( X_1 \sim X_2 \) denotes that random variables \( X_1 \) and \( X_2 \) have the same distribution. Therefore, for any fixed \( c > 0 \), since \( \mathbb{P}(\xi_k \leq c) \) decreases in \( k \), it follows that

\[
\mathbb{P}(\|D\| \leq c) \geq \mathbb{P} \left( 2\xi_m \sigma_w^2 + 2\xi_{m+s} \sigma_w \lambda_1^{1/2}(\Gamma) \leq c \right)
\]

\[
\geq \mathbb{P} \left( 2\xi_{m+s} \sigma_w \left[ \sigma_w + 2\lambda_1^{1/2}(\Gamma) \right] \leq c \right). \tag{B.10.1}
\]

**Rank estimation:** From the sample covariance matrix \( \hat{\Gamma} \), Algorithm 1 determines the rank \( r \) as the maximal number \( k \) that satisfies

\[
\lambda_k(\hat{\Gamma}) - \lambda_{k+1}(\hat{\Gamma}) \geq \tau \lambda_1(\hat{\Gamma}).
\]

Therefore, to guarantee that Algorithm 7 stops at the desired rank \( r \), we need sufficient conditions for

\[
\lambda_r(\hat{\Gamma}) - \lambda_{r+1}(\hat{\Gamma}) \geq \tau \lambda_1(\hat{\Gamma}) \tag{B.10.2}
\]

\[
\lambda_k(\hat{\Gamma}) - \lambda_{k+1}(\hat{\Gamma}) < \tau \lambda_1(\hat{\Gamma}), \quad \forall k > r. \tag{B.10.3}
\]

175
By Weyl’s perturbation theorem [135, Corollary III.2.6], it follows that
\[
\max_{k \in [m]} \left| \lambda_k(\tilde{\Gamma}) - \lambda_k(\Gamma_S) \right| \leq \|\tilde{\Gamma} - \Gamma_S\|
\]
and hence,
\[
\frac{\lambda_r(\tilde{\Gamma}) - \lambda_{r+1}(\tilde{\Gamma})}{\lambda_1(\tilde{\Gamma})} \geq \frac{\lambda_r(\Gamma_S) - \lambda_{r+1}(\Gamma_S) - 2\|D\|}{\lambda_1(\Gamma_S) + \|D\|}.
\]
(B.10.4)

Again, by Weyl’s perturbation theorem [135, Corollary III.2.6], it follows that
\[
\left| \lambda_k \left( \Phi\Phi^* \frac{\Phi\Phi^*}{N} - 1 \right) \right| = \left| \lambda_k \left( \Phi\Phi^* \frac{\Phi\Phi^*}{N} \right) - \lambda_k(I_M) \right| \leq \left\| \frac{\Phi\Phi^*}{N} - I_M \right\| \leq \xi_M
\]
for all \( k = 1, \ldots, M \); hence,
\[
1 - \xi_M \leq \lambda_M \left( \frac{\Phi\Phi^*}{N} \right) \leq \lambda_1 \left( \frac{\Phi\Phi^*}{N} \right) \leq 1 + \xi_M.
\]
(B.10.5)

By (B.10.5), it follows that
\[
\lambda_k(\Gamma_S) \geq (1 - \xi_M)\lambda_k(\Gamma) \quad \text{(B.10.6)}
\]
\[
\lambda_k(\Gamma_S) \leq (1 + \xi_M)\lambda_k(\Gamma). \quad \text{(B.10.7)}
\]

By Lemma B.10.2, it follows that
\[
\mathbb{P} \left( \xi_k > \frac{6\sqrt{k + \ln(4/\epsilon)}}{\sqrt{N}} \right) \leq \epsilon.
\]
(B.10.8)

Therefore, by (3.7.4), (3.7.5), and (B.10.8), it follows that \( \xi_M \leq \theta \) with probability \( 1 - \epsilon/2 \). Therefore, we assume \( \xi_M \leq \theta \) in the remaining steps of the proof.

Since \( \xi_M \leq \theta \), by (B.10.4), (B.10.6), and (B.10.7) the following condition is a sufficient condition for (B.10.2):
\[
\frac{(1 - \theta)\lambda_r(\Gamma) - (1 + \theta)\lambda_{r+1}(\Gamma) - 2\|D\|}{(1 + \theta)\lambda_1(\Gamma) + \|D\|} \geq \tau
\]
which is equivalent to
\[
\|D\| \leq \frac{(1 - \theta)\lambda_r(\Gamma) - (1 + \theta)\lambda_{r+1}(\Gamma) - (1 + \theta)\tau\lambda_1(\Gamma)}{2 + \tau}.
\] (B.10.9)

Since \(\Gamma\) satisfies (3.7.1), we obtain a sufficient condition for (B.10.9) given by
\[
\frac{\|D\|}{\lambda_1(\Gamma)} \leq \frac{(1 + \theta)\nu}{2 + \tau}.
\] (B.10.10)

Similarly, since \(\Gamma\) satisfies (3.7.2), we verify that (B.10.10) is also a sufficient condition for (B.10.3).

Recall that Algorithm 1 computes \(\hat{S}\) as the subspace spanned by the \(r\) dominant eigenvectors of \(\hat{\Gamma}\). Next, we derive conditions that guarantees that \(\|P_{\hat{S}} - P_S\| \leq \eta\) where \(\hat{S}\) is the subspace spanned by the \(r\) dominant eigenvectors of \(\Gamma_S\).

We apply \(\Gamma_1 = \hat{\Gamma}\) and \(\Gamma_2 = \Gamma_S\) to Lemma B.10.1, and obtain
\[
3\|D\| \geq \|P_{\hat{S}} - P_S\| \cdot [\lambda_r(\Gamma_S) - \lambda_{r+1}(\Gamma_S)].
\] (B.10.11)

Since \(\xi_M \leq \theta\) and \(\Gamma\) satisfies (3.7.1),
\[
\lambda_r(\Gamma_S) - \lambda_{r+1}(\Gamma_S) \geq (1 - \theta)\lambda_r(\Gamma) - (1 + \theta)\lambda_{r+1}(\Gamma)
\]
\[
\geq (1 + \theta)(1 + \nu)\tau\lambda_1(\Gamma).
\] (B.10.12)

By (B.10.11) and (B.10.12), we note that \(\|P_{\hat{S}} - P_S\| \leq \eta\) is implied by
\[
\frac{3\|D\|}{\lambda_1(\Gamma)} \leq \eta(1 + \theta)(1 + \nu)\tau\lambda_1(\Gamma).
\] (B.10.13)

Recall that \(C_{\eta,\nu,\theta,\tau}\) is defined by
\[
C_{\eta,\nu,\theta,\tau} \triangleq (1 + \theta)\tau \min \left\{ \frac{(1 + \nu)\eta}{3}, \frac{\nu}{2 + \tau} \right\}.
\]

Then, by (B.10.1)
\[
\xi_{m+s} \leq \frac{C_{\eta,\nu,\theta,\tau}}{2[\sigma^2/\lambda_1(\Gamma) + 2(\sigma^2/\lambda_1(\Gamma))^{1/2}]}
\] (B.10.14)
implies (B.10.10) and (B.10.13).
Finally, we note that, by (3.7.4), (3.7.5), and (B.10.8), it follows that (B.10.14) holds with probability $1 - \epsilon/2$. □

Proof of Lemma B.10.2. We use the following lemma to prove Lemma B.10.2

Lemma B.10.3 ([140, Lemma 36]). Consider a matrix $B \in \mathbb{K}^{n \times m}$ ($m \leq n$) that satisfies

$$
\|B^*B - I\| \leq \max(\delta, \delta^2)
$$

for some $\delta > 0$. Then,

$$
1 - \delta \leq \sigma_m(B) \leq \sigma_1(B) \leq 1 + \delta. \quad (B.10.15)
$$

Conversely, if $B$ satisfies (B.10.15) for some $\delta > 0$, then $\|B^*B - I\| \leq 3\max(\delta, \delta^2)$.

We can write $G$ as

$$
G = \frac{1}{\sqrt{2}}(G_{\text{Re}} + jG_{\text{Im}})
$$

where $G_{\text{Re}}, G_{\text{Im}} \in \mathbb{R}^{n \times m}$ are mutually independent i.i.d. Gaussian matrices whose entries follow $\mathcal{N}(0, \frac{1}{n})$. Then,

$$
\|G^*G - I_s\| = \left\| \frac{1}{2} \begin{bmatrix} G_{\text{Re}} & -G_{\text{Im}} \\ G_{\text{Im}} & G_{\text{Re}} \end{bmatrix}^{*} \begin{bmatrix} G_{\text{Re}} & -G_{\text{Im}} \\ G_{\text{Im}} & G_{\text{Re}} \end{bmatrix} - I_{2s} \right\|
$$

where $(*)$ is a block matrix

$$
\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}
$$

whose block entries are given by

$$
A_{11} = \frac{1}{2} [(G^*_{\text{Re}}G_{\text{Re}} - I_s) + (G^*_{\text{Im}}G_{\text{Im}} - I_s)],
$$

$$
A_{12} = \frac{1}{2} (G^*_{\text{Im}}G_{\text{Re}} - G^*_{\text{Re}}G_{\text{Im}}),
$$

$$
A_{21} = A_{12}^*,
$$

$$
A_{22} = A_{11}.
$$
Then

\[
\|G^*G - I_s\|
\leq \frac{1}{2} \left\| \begin{bmatrix} \left(-G_\text{Re}\right)^*(-G_\text{Re}) & -G_\text{Im}^*G_\text{Im} & -I_s \\ G_\text{Im}^*(-G_\text{Re}) & G_\text{Im}^*G_\text{Im} - I_s & \end{bmatrix} \right\|
\]

\[
+ \frac{1}{2} \left\| \begin{bmatrix} G_\text{Im}^*G_\text{Im} - I_s & G_\text{Im}^*G_\text{Re} \\ \end{bmatrix} \right\|
\]

\[
= \frac{1}{2} \left\| \left[-G_\text{Re} G_\text{Im}\right]^*\left[-G_\text{Re} G_\text{Im}\right] - I_{2s} \right\|
\]

\[
+ \frac{1}{2} \left\| G_\text{Re} G_\text{Im}]^*[G_\text{Re} G_\text{Im}] - I_{2s} \right\|. 
\]  

(B.10.16)

By Theorem B.3.1 and Lemma B.10.3, we have

\[
\mathbb{P} \left( \|G_\text{Re} G_\text{Im}]^*[G_\text{Re} G_\text{Im}] - I_{2s} \| \geq 3 \sqrt{\frac{2m}{n} + t} \right) \leq 2 \exp \left( -\frac{nt^2}{18} \right)
\]

for \( t > 0 \) satisfying

\[
\sqrt{\frac{2m}{n} + \frac{t}{3}} \leq 1.
\]

By the symmetry of the Gaussian distribution, we also have

\[
\mathbb{P} \left( \|[-G_\text{Re} G_\text{Im}]^*[-G_\text{Re} G_\text{Im}] - I_{2s} \| \geq 3 \sqrt{\frac{2m}{n} + t} \right) \leq 2 \exp \left( -\frac{nt^2}{18} \right).
\]

Combining (B.10.16)–(??) completes the proof. \( \square \)
APPENDIX C

PROOFS FOR CHAPTER 4

C.1 Proof of Proposition 4.4.1

The rank-restricted isometry constant $\delta_r(A)$ is rewritten as

$$\delta_r(A) = \max \left\{ [\sigma_{r,\max}(A)]^2 - 1, 1 - [\sigma_{r,\min}(A)]^2 \right\}, \quad \text{(C.1.1)}$$

where $\sigma_{r,\max}(A)$ and $\sigma_{r,\min}(A)$ are defined respectively as

$$\sigma_{r,\max}(A) \triangleq \max_X \{ \|AX\|_{\ell^2} : \|X\|_{S^2} = 1, \ \rank(X) \leq r \} \quad \text{(C.1.2)}$$

and

$$\sigma_{r,\min}(A) \triangleq \min_X \{ \|AX\|_{\ell^2} : \|X\|_{S^2} = 1, \ \rank(X) \leq r \}. \quad \text{(C.1.3)}$$

As $r$ increases, the feasible sets of both problems in (C.1.2) and (C.1.3) increase; hence, $\sigma_{r,\max}(A)$ and $\sigma_{r,\min}(A)$ are nondecreasing and nonincreasing, respectively. Therefore, (C.1.1) implies that $\delta_r(A)$ is nondecreasing in $r$.

C.2 Proof of Proposition 4.4.2

By the R-RIP of $A$, for all $X \in S_2$, it follows that

$$\langle P_\psi A^* AP_\psi X, X \rangle_{S_2} = \langle AP_\psi X, AP_\psi X \rangle_{\ell^2_p}$$

$$= \|AP_\psi X\|_{\ell^2_p}$$

$$\leq (1 + \delta_r(A)) \|P_\psi X\|_{S_2}$$

$$\leq (1 + \delta_r(A)) \|X\|_{S_2}$$

180
where the first inequality holds since \(\text{rank}(P_X) \leq r\).

Since \(P_X A^* A P_X\) is self adjoint, we have

\[
\|P_X A^* A P_X\|_{S_2 \to S_2} = \sup_{\|X\|_{S_2} = 1} \langle P_X A^* A P_X X, X \rangle_{S_2} \leq 1 + \delta_r(A).
\]

The proof completes by noting

\[
\|P_X A^*\|_{\ell_p^2 \to S_2} = \|P_X A^* A P_X\|_{S_2 \to \ell_p^2} = \|P_X A^* A P_X\|_{S_2 \to S_2}^{1/2}.
\]

C.3 Proof of Proposition 4.4.3

By the R-RIP of \(A\), for all \(X \in S_2\), it follows that

\[
(1 - \delta_r(A))\langle P_X X, P_X X \rangle_{S_2} \leq \langle A P_X X, A P_X X \rangle_{\ell_p^2} \leq (1 + \delta_r(A))\langle P_X X, P_X X \rangle_{S_2}
\]

which implies

\[
\left| \langle P_X (A^* A - I_{S_2}) P_X X, X \rangle_{S_2} \right| \leq \delta_r(A) \left| \langle P_X X, P_X X \rangle_{S_2} \right| = \delta_r(A) \|P_X X\|_{S_2}^2 \leq \delta_r(A) \|X\|_{S_2}^2.
\]

Therefore, since \(P_X (A^* A - I_{S_2}) P_X\) is self adjoint,

\[
\|P_X (A^* A - I_{S_2}) P_X\|_{S_2 \to S_2} = \sup_{\|X\|_{S_2} \leq 1} \left| \langle P_X (A^* A - I_{S_2}) P_X X, X \rangle_{S_2} \right| \leq \delta_r(A).
\]

C.4 Proof of Corollary 4.4.4

Note that \(X = P_X X\). Therefore,

\[
\|P_X A^* A X\|_{S_2} = \|P_X A^* A P_X X\|_{S_2} = \|P_X (A^* A - I_{S_2}) + I_{S_2}) P_X X\|_{S_2}
\geq \|P_X X\|_{S_2} - \|P_X (A^* A - I_{S_2}) P_X X\|_{S_2}
\geq \|P_X X\|_{S_2} - \|P_X (A^* A - I_{S_2}) P_X\|_{S_2 \to S_2}, \|P_X X\|_{S_2}
\geq \|P_X X\|_{S_2} - \delta_r(A) \|P_X X\|_{S_2}
= (1 - \delta_r(A)) \|P_X X\|_{S_2}
\]
\[ = (1 - \delta_r(A))\|X\|_S_2 \]

where the last inequality follows from Proposition 4.4.3.

### C.5 Proof of Proposition 4.4.5

We modify the proof the analogous result for the vector case in [11] for our proposition.

For \( \Psi \subset \Omega \), the unit-ball in the subspace spanned by \( \Psi \) is defined by

\[ B^\Psi_{S_2} \triangleq \{ X \in S_2 : X \in \text{span}(\Psi), \| X \|_{S_2} \leq 1 \}. \]

Define the convex body

\[ C \triangleq \text{conv} \left\{ \bigcup_{\Psi \subset \Omega, |\Psi| \leq r} B^\Psi_{S_2} \right\}, \]

where \( \text{conv} \{G\} \) denotes the convex hull of set \( G \). By the assumption, the operator norm satisfies

\[ \| A \|_{C \rightarrow \ell_p^2} \triangleq \max_{X \in C} \| AX \|_{\ell_p^2} \leq \sqrt{1 + \delta_r(A)}. \]

Define the second convex body

\[ K \triangleq \left\{ X \in S_2 : \| X \|_{S_2} + \frac{1}{\sqrt{r}} \| X \|_{S_1} \leq 1 \right\}, \]

and consider the operator norm

\[ \| A \|_{K \rightarrow \ell_p^2} \triangleq \max_{X \in K} \| AX \|_{\ell_p^2}. \]

The claim of the proposition is equivalent to

\[ \| A \|_{K \rightarrow \ell_p^2} \leq \| A \|_{C \rightarrow \ell_p^2}. \]

It suffices to show that \( K \subset C \). Let \( X \) be an element in \( K \). Consider the
singular value decomposition of $X$,

$$X = \sum_{k=1}^{\text{rank}(X)} \sigma_k u_k v_k^*,$$

with $\sigma_{k+1} \leq \sigma_k$. Let $\sigma_k = 0$ if $k > \text{rank}(X)$ and $q = \lceil \text{rank}(X)/r \rceil - 1$, where $\lceil c \rceil$ denotes the smallest integer equal to or greater than $c$. Then we have the following decomposition

$$X = \sum_{j=0}^{q} \sum_{k=rj+1}^{r(j+1)} \sigma_k u_k v_k^* = \sum_{j=0}^{q} c_j Y_j,$$

where

$$c_j \triangleq \left\| \sum_{k=rj+1}^{r(j+1)} \sigma_k u_k v_k^* \right\|_{\mathcal{S}_2} \quad \text{and} \quad Y_j \triangleq \frac{1}{c_j} \sum_{k=rj+1}^{r(j+1)} \sigma_k u_k v_k^*.$$

For each $j \in \{1, \ldots, q\}$,

$$c_j = \sqrt{\sum_{k=rj+1}^{r(j+1)} \sigma_k^2} \leq \sqrt{r} \cdot \sigma_{rj+1} \leq \sqrt{r} \cdot \frac{1}{r} \sum_{k=r(j-1)+1}^{rj} \sigma_k.$$

Therefore

$$\sum_{j=1}^{q} c_j \leq \frac{1}{\sqrt{r}} \sum_{j=1}^{q} \sum_{k=r(j-1)+1}^{rj} \sigma_k = \frac{1}{\sqrt{r}} \sum_{k=1}^{\text{rank}(X)} \sigma_k = \frac{1}{\sqrt{r}} \left\| X \right\|_{\mathcal{S}_1}.$$

From the definition of $c_0$, it follows that $c_0 \leq \left\| X \right\|_{\mathcal{S}_2}$. Since $X \in \mathcal{K}$, we note

$$\sum_{j=0}^{q} c_j \leq \left\| X \right\|_{\mathcal{S}_2} + \frac{1}{\sqrt{r}} \left\| X \right\|_{\mathcal{S}_1} \leq 1.$$

Also note that $Y_j \in \mathcal{C}$ for all $j = 0, \ldots, q$ since $\text{rank}(Y_j) \leq r$ and $\left\| Y_j \right\|_{\mathcal{S}_2} = 1$ by construction. Therefore $X$ is the convex combination of the elements in $\mathcal{C}$. Since $\mathcal{C}$ is a convex hull, $X \in \mathcal{C}$. 

183
C.6 Proof of Lemma 4.5.2

Let $\Phi = \text{atoms}(X_0 - \tilde{X})$. Since $|\Phi| \leq \text{rank}(X_0) + \text{rank}(\tilde{X}) \leq 2r$, it follows by the selection rule of $\Psi'$ that

$$\|P_{\Phi}A^*(b - A\tilde{X})\|_{S_2} \leq \|P_{\Psi'}A^*(b - A\tilde{X})\|_{S_2}.$$  \hspace{1cm} (C.6.1)

First, we derive a lower bound on the left-hand side of (C.6.1) as

$$\|P_{\Phi}A^*(b - A\tilde{X})\|_{S_2} = \|P_{\Phi}A^*A(X_0 - \tilde{X}) + P_{\Phi}A^*\nu\|_{S_2}$$
$$\geq \|P_{\Phi}A^*A(X_0 - \tilde{X})\|_{S_2} - \|P_{\Phi}A^*\nu\|_{S_2}$$
$$= \|P_{\Phi}A^*AP_{\Phi}(X_0 - \tilde{X})\|_{S_2} - \|P_{\Phi}A^*\nu\|_{S_2}$$
$$\geq (1 - \delta_{2r}(A))\|X_0 - \tilde{X}\|_{S_2} - \sqrt{1 + \delta_{2r}(A)}\|\nu\|_{\ell^2}$$ \hspace{1cm} (C.6.2)

where the last inequality follows from Corollary 4.4.4 and Proposition 4.4.2.

Then, we derive an upper bound on the right-hand side of (C.6.1) as

$$\|P_{\Psi'}A^*(b - A\tilde{X})\|_{S_2}$$
$$= \|P_{\Psi'}A^*A(X_0 - \tilde{X}) + P_{\Psi'}A^*\nu\|_{S_2}$$
$$\leq \|P_{\Psi'}A^*A(X_0 - \tilde{X})\|_{S_2} + \|P_{\Psi'}A^*\nu\|_{S_2}$$
$$\leq \|P_{\Psi'}(A^*A - I_{S_2})(X_0 - \tilde{X})\|_{S_2} + \|P_{\Psi'}(X_0 - \tilde{X})\|_{S_2} + \|P_{\Psi'}A^*\nu\|_{S_2}$$
$$= \|P_{\Psi'}(A^*A - I_{S_2})P_{\Phi}(X_0 - \tilde{X})\|_{S_2} + \|P_{\Psi'}(X_0 - \tilde{X})\|_{S_2} + \|P_{\Psi'}A^*\nu\|_{S_2}$$
$$\leq \delta_{4r}(A)\|X_0 - \tilde{X}\|_{S_2} + \|P_{\Psi'}(X_0 - \tilde{X})\|_{S_2} + \sqrt{1 + \delta_{2r}(A)}\|\nu\|_{\ell^2}$$ \hspace{1cm} (C.6.3)

where the last inequality follows from Propositions 4.4.3 and 4.4.2.

Applying eqs. (C.6.2) and (C.6.3) to (C.6.1), we obtain

$$\|P_{\Psi'}(X_0 - \tilde{X})\|_{S_2} \geq (1 - 2\delta_{4r}(A))\|X_0 - \tilde{X}\|_{S_2} - 2\sqrt{1 + \delta_{2r}(A)}\|\nu\|_{\ell^2},$$

which implies

$$\frac{\|P_{\Psi'}(X_0 - \tilde{X})\|_{S_2}}{\|X_0 - \tilde{X}\|_{S_2}} \geq 1 - 2\delta_{4r}(A) - 2\sqrt{1 + \delta_{2r}(A)}\cdot \frac{\|\nu\|_{\ell^2}}{\|X_0 - \tilde{X}\|_{S_2}}.$$ \hspace{1cm} (C.6.4)
If $\alpha - \beta \omega < 0$, then
\[
\|P^\perp_{\Psi'}(X_0 - \hat{X})\|_{S_2} \leq \|X_0 - \hat{X}\|_{S_2} < \frac{2\sqrt{1 + \delta_{2r}(A)}}{1 - 2\delta_{4r}(A)} \|\nu\|_{\ell_2}^2. \tag{C.6.5}
\]

Since $\delta_{4r}(A) \leq 0.065$, (C.6.5) is rewritten as
\[
\|P^\perp_{\Psi'}(X_0 - \hat{X})\|_{S_2} < 2.373\|\nu\|_{\ell_2}, \tag{C.6.6}
\]
which implies the desired inequality.

It remains to consider the other case, $\alpha - \beta \omega \geq 0$. By the Pythagorean theorem, we have
\[
\frac{\|P^\perp_{\Psi'}(X_0 - \hat{X})\|_{S_2}}{\|X_0 - \hat{X}\|_{S_2}} = \sqrt{1 - \left(\frac{\|P_{\Psi'}(X_0 - \hat{X})\|_{S_2}}{\|X_0 - \hat{X}\|_{S_2}}\right)^2} \\
\leq \sqrt{1 - (\alpha - \beta \omega)^2} \\
= \sqrt{1 - \alpha^2 + 2\alpha \beta \omega - \beta^2 \omega^2}.
\]

Let $g(\omega) \triangleq \sqrt{1 - \alpha^2 + 2\alpha \beta \omega - \beta^2 \omega^2}$. The first and second order derivatives of $g$ are given respectively as
\[
\frac{dg(\omega)}{d\omega} = (1 - \alpha^2 + 2\alpha \beta \omega - \beta^2 \omega^2)^{-1/2} \beta (\alpha - \beta \omega)
\]
and
\[
\frac{d^2g(\omega)}{d\omega^2} = -(1 - \alpha^2 + 2\alpha \beta \omega - \beta^2 \omega^2)^{-3/2} \beta^2.
\]
Since $\alpha - \beta \omega \geq 0$, $g(\omega)$ is a concave function of $\omega$.

Let $\gamma \in (0, 1)$. An affine function $h(\omega)$ that is tangent to $g(\omega)$ at $\omega = \frac{\alpha}{\beta}$ is given by
\[
h(\omega) = \sqrt{1 - (1 - \gamma)^2 \alpha^2} + \frac{(1 - \gamma) \alpha \beta}{\sqrt{1 - (1 - \gamma)^2 \alpha^2}} \omega.
\]
Indeed, $h(\omega) \geq g(\omega)$ since $g(\omega)$ is a concave function.

We substitute $\delta_{2r}(A) \leq \delta_{4r}(A) \leq 0.065$ to $\alpha$ and $\beta$ and choose $\gamma \in (0, 1)$ so that $\sqrt{1 - (1 - \gamma)^2 \alpha^2} = 0.493$, which gives the constants in the final inequality.
C.7 Proof of Lemma 4.5.4

Since \( \tilde{\Psi} \subset \Psi \), \( \mathcal{P}^{-1}_{\tilde{\Psi}} \hat{X} = 0 \) implies \( \mathcal{P}^{-1}_{\Psi} \hat{X} = 0 \) and hence

\[
\| \mathcal{P}^{-1}_{\Psi} X \|_{S_2} = \| \mathcal{P}^{-1}_{\Psi} (X_0 - \hat{X}) \|_{S_2} \leq \| \mathcal{P}^{-1}_{\Psi^*} (X_0 - \hat{X}) \|_{S_2},
\]

where the inequality holds since \( \Psi^* \subset \tilde{\Psi} \) implies \( (\tilde{\Psi})^\perp \subset (\Psi^*)^\perp \).

C.8 Proof of Lemma 4.5.5

We use the following lemmata in the proof of Lemma 4.5.5. The proofs of Lemma C.8.1 and Lemma C.8.2 are deferred to after the proof of Lemma 4.5.5.

**Lemma C.8.1.** Let \( \tilde{\Psi} \) and \( \tilde{X} \) be as in Lemma 4.5.5. Then,

\[
\mathcal{P}_{\tilde{\Psi}}[A^*A(\tilde{X} - X_0)] = \mathcal{P}_{\tilde{\Psi}}(A^*\nu).
\]

**Lemma C.8.2.** Let \( \tilde{\Psi} \) and \( \tilde{X} \) be as in Lemma 4.5.5. Then,

\[
\| \mathcal{P}_{\tilde{\Psi}}(\tilde{X} - X_0) \|_{S_2} \leq \delta_{4r} \| \tilde{X} - X_0 \|_{S_2} + \sqrt{1 + \delta_{3r} \| \nu \|_{\ell_2}}.
\]

**Proof of Lemma 4.5.5.** By the Pythagorean theorem and Lemma C.8.2,

\[
\| \tilde{X} - X_0 \|_{S_2}^2 = \| (\mathcal{P}^{-1}_{\tilde{\Psi}} + \mathcal{P}_{\tilde{\Psi}})(\tilde{X} - X_0) \|_{S_2}^2 \\
= \| \mathcal{P}^{-1}_{\tilde{\Psi}}(\tilde{X} - X_0) \|_{S_2}^2 + \| \mathcal{P}_{\tilde{\Psi}}(\tilde{X} - X_0) \|_{S_2}^2 \\
= \| \mathcal{P}^{-1}_{\tilde{\Psi}} X_0 \|_{S_2}^2 + \| \mathcal{P}_{\tilde{\Psi}}(\tilde{X} - X_0) \|_{S_2}^2 \\
\leq \| \mathcal{P}^{-1}_{\tilde{\Psi}} X_0 \|_{S_2}^2 + \left( \delta_{4r}(A) \| \tilde{X} - X_0 \|_{S_2} + \sqrt{1 + \delta_{3r}(A) \| \nu \|_{\ell_2}} \right)^2.
\]

(C.8.1)

Viewing (C.8.1) as an inequality given by a quadratic polynomial of \( \| \tilde{X} - X_0 \|_{S_2} \), (C.8.1) implies that \( \| \tilde{X} - X_0 \|_{S_2} \) is less than the largest root of the polynomial, i.e.,

\[
\| \tilde{X} - X_0 \|_{S_2} \leq \frac{1}{\sqrt{1 - [\delta_{4r}(A)]^2}} \| \mathcal{P}^{-1}_{\tilde{\Psi}} X_0 \|_{S_2} + \frac{\sqrt{1 + \delta_{3r}(A)}}{1 - \delta_{4r}(A)} \| \nu \|_{\ell_2}.
\]

186
Substituting $\delta_{3r}(A) \leq \delta_{4r}(A) \leq 0.065$ gives the constants in the final inequality.

Proof of Lemma C.8.1. Let

$$f(X) \triangleq \frac{1}{2}\|b - AP_{\tilde{\psi}}X\|_{F}^{2}.$$  

Then, the gradient of $f$ is given by

$$\nabla f(X) = PA^*(AP_{\tilde{\psi}}X - b).$$

Since $\tilde{X}$ is a minimizer of $f$ and $P_{\tilde{\psi}}\tilde{X} = \tilde{X}$, it follows that

$$P_{\tilde{\psi}}[A^*(A\tilde{X} - b)] = 0. \quad (C.8.2)$$

Applying $b = AX_0 + \nu$ to eq. (C.8.2) completes the proof.

Proof of Lemma C.8.2. Let $\Psi_0 \triangleq \text{atoms}(X_0)$. We compute an upper bound of $\|P_{\tilde{\psi}}(\tilde{X} - X_0)\|_{S_2}^{2}$ by

\[
\begin{align*}
\|P_{\tilde{\psi}}(\tilde{X} - X_0)\|_{S_2}^{2} &= \langle P_{\tilde{\psi}}(\tilde{X} - X_0), P_{\tilde{\psi}}(\tilde{X} - X_0) \rangle_{S_2} \\
&= \langle P_{\tilde{\psi}}(\tilde{X} - X_0), P_{\tilde{\psi}}[(T_{S_2} - A^*A)(\tilde{X} - X_0)] \rangle_{S_2} \\
&= \langle P_{\tilde{\psi}}(\tilde{X} - X_0), P_{\tilde{\psi}}[(T_{S_2} - A^*A)(\tilde{X} - X_0)] \rangle_{S_2} \\
&+ \langle P_{\tilde{\psi}}(\tilde{X} - X_0), P_{\tilde{\psi}}[A^*A(\tilde{X} - X_0)] \rangle_{S_2} \\
&\overset{(a)}{=} \langle P_{\tilde{\psi}}(\tilde{X} - X_0), P_{\tilde{\psi}}[(T_{S_2} - A^*A)(\tilde{X} - X_0)] \rangle_{S_2} \\
&+ \langle P_{\tilde{\psi}}(\tilde{X} - X_0), P_{\tilde{\psi}}(A^*\nu) \rangle_{S_2} \\
&\overset{(b)}{\leq} \|P_{\tilde{\psi}}(A^*A - T_{S_2})\|P_{\tilde{\psi}}\|_{S_2 \rightarrow S_2} \|P_{\tilde{\psi}}(\tilde{X} - X_0)\|_{S_2} \|\tilde{X} - X_0\|_{S_2} \\
&+ \langle A^*P_{\tilde{\psi}}(\tilde{X} - X_0), \nu \rangle_{F}^{2} \\
&\overset{(c)}{\leq} \delta_{4r}(A)\|P_{\tilde{\psi}}(\tilde{X} - X_0)\|_{S_2} \|\tilde{X} - X_0\|_{S_2} \\
&+ \sqrt{1 + \delta_{3r}(A)}\|P_{\tilde{\psi}}(\tilde{X} - X_0)\|_{S_2} \|\nu\|_{F}^{2} \quad (C.8.3)
\end{align*}
\]

where (a) follows by Lemma C.8.1, (b) follows by $P_{\tilde{\psi} \cup \Psi_0}(\tilde{X} - X_0) = \tilde{X} - X_0$,
and (c) holds since
\[ \| \mathcal{P}_\Psi (A^*A - I_{S_2}) \mathcal{P}_{\Psi \cup \Psi_0} \|_{S_2 \rightarrow S_2} \leq \| \mathcal{P}_\Psi (A^*A - I_{S_2}) \mathcal{P}_{\Psi \cup \Psi_0} \|_{S_2 \rightarrow S_2} \leq \delta_{4r}(A). \]

By dividing both hand sides of (C.8.3) by \( \| \mathcal{P}_\Psi (A^*A - I_{S_2}) \mathcal{P}_{\Psi \cup \Psi_0} \|_{S_2} \), we obtain the desired inequality. \( \square \)

### C.9 Proof of Lemma 4.5.6

Note that
\[ \| \Pi_r(\tilde{X}) - X_0 \|_{S_2}^2 = \| (\tilde{X} - X_0) - (\tilde{X} - \Pi_r(\tilde{X})) \|_{S_2}^2 \]
\[ = \| \tilde{X} - X_0 \|_{S_2}^2 + \| \tilde{X} - \Pi_r(\tilde{X}) \|_{S_2}^2 - 2 \cdot \text{Re} \langle \tilde{X} - \Pi_r(\tilde{X}), \tilde{X} - X_0 \rangle_{S_2} \] (C.9.1)

where \( \text{Re} \) denotes the operation of taking the real part of a complex number.

Since \( \tilde{X} \) and \( \Pi_r(\tilde{X}) \) are spanned by \( \bar{\Psi} \), it follows that
\[ \tilde{X} - \Pi_r(\tilde{X}) = \mathcal{P}_\Psi [\tilde{X} - \Pi_r(\tilde{X})]; \quad \text{(C.9.2)} \]

Hence, we have
\[ \| \tilde{X} - \Pi_r(\tilde{X}) \|_{S_2} = \| \mathcal{P}_\Psi [\tilde{X} - \Pi_r(\tilde{X})] \|_{S_2} \leq \| \mathcal{P}_\Psi (\tilde{X} - X_0) \|_{S_2} \] (C.9.3)

where the last inequality follows from the definition of \( \Pi_r \).

Using eq. (C.9.2), we obtain a bound on the inner product term in (C.9.1) given by
\[ \langle \tilde{X} - \Pi_r(\tilde{X}), \tilde{X} - X_0 \rangle_{S_2} \]
\[ = \langle \mathcal{P}_\Psi [\tilde{X} - \Pi_r(\tilde{X})], \tilde{X} - X_0 \rangle_{S_2} \]
\[ = \langle \tilde{X} - \Pi_r(\tilde{X}), \mathcal{P}_\Psi (\tilde{X} - X_0) \rangle_{S_2} \]
\[ = \langle \tilde{X} - \Pi_r(\tilde{X}), \mathcal{P}_\Psi [(I_{S_2} - A^*A)(\tilde{X} - X_0)] \rangle_{S_2} \]
\[ = \langle \tilde{X} - \Pi_r(\tilde{X}), \mathcal{P}_\Psi [(I_{S_2} - A^*A)(\tilde{X} - X_0)] \rangle_{S_2} \]
\[ = \langle \tilde{X} - \Pi_r(\tilde{X}), \mathcal{P}_\Psi [A^*A(\tilde{X} - X_0)] \rangle_{S_2} \]
\[ = \langle \tilde{X} - \Pi_r(\tilde{X}), (I_{S_2} - A^*A)(\tilde{X} - X_0) \rangle_{S_2} \]
\[ \overset{(a)}{=} \langle \tilde{X} - \Pi_r(\tilde{X}), (I_{S_2} - A^*A)(\tilde{X} - X_0) \rangle_{S_2} \]

188
\begin{align*}
+ \langle \widetilde{X} - \Pi_r(\widetilde{X}), \mathcal{P}_\phi(A^*\nu) \rangle_{S_2} \\
= \langle \widetilde{X} - \Pi_r(\widetilde{X}), (I_{S_2} - A^*A)(\widetilde{X} - X_0) \rangle_{S_2} \\
+ \langle \mathcal{A} [\widetilde{X} - \Pi_r(\widetilde{X})], \nu \rangle_{\ell_2^p} \tag{C.9.4}
\end{align*}

where (a) follows from Lemma C.8.1.

Since |\text{atoms}(\widetilde{X} - \Pi_r(\widetilde{X})) \cup \text{atoms}(\widetilde{X} - X_0)| \leq 4r, by Proposition 4.4.3, the first term of the right-hand side of (C.9.4) satisfies

\begin{align*}
\langle \widetilde{X} - \Pi_r(\widetilde{X}), (I_{S_2} - A^*A)(\widetilde{X} - X_0) \rangle_{S_2} &\leq \delta_{4r}(A) \| \widetilde{X} - \Pi_r(\widetilde{X}) \|_{S_2} \| \widetilde{X} - X_0 \|_{S_2} 
\tag{C.9.5}
\end{align*}

Since |\text{atoms}(\widetilde{X} - \Pi_r(\widetilde{X}))| \leq 3r, by Proposition 4.4.2, the second term of the right-hand side of (C.9.4) satisfies

\begin{align*}
\langle \mathcal{A} [\widetilde{X} - \Pi_r(\widetilde{X})], \nu \rangle_{\ell_2^p} &\leq \sqrt{1 + \delta_{3r}(A)} \| \widetilde{X} - \Pi_r(\widetilde{X}) \|_{S_2} \| \nu \|_{\ell_2^p}. \tag{C.9.6}
\end{align*}

Combining Lemma C.8.2 and eqs. (C.9.1) and (C.9.3) to (C.9.6), we obtain

\begin{align*}
\| \Pi_r(\widetilde{X}) - X_0 \|_{S_2}^2 \\
\leq \| \widetilde{X} - X_0 \|_{S_2}^2 + \| \widetilde{X} - \Pi_r(\widetilde{X}) \|_{S_2}^2 \\
+ 2\delta_{4r}(A) \| \widetilde{X} - \Pi_r(\widetilde{X}) \|_{S_2} \| \widetilde{X} - X_0 \|_{S_2} \\
+ 2\sqrt{1 + \delta_{3r}(A)} \| \widetilde{X} - \Pi_r(\widetilde{X}) \|_{S_2} \| \nu \|_{\ell_2^p} \\
\leq (1 + 3[\delta_{4r}(A)]^2) \| \widetilde{X} - X_0 \|_{S_2}^2 \\
+ 6\delta_{4r}(A) \sqrt{1 + \delta_{3r}(A)} \| \widetilde{X} - X_0 \|_{S_2} \| \nu \|_{\ell_2^p} + 3(1 + \delta_{3r}(A)) \| \nu \|_{\ell_2^p}^2 \\
\leq (1 + 3[\delta_{4r}(A)]^2) \left( \| \widetilde{X} - X_0 \|_{S_2} + \sqrt{\frac{3(1 + \delta_{3r}(A))}{1 + 3[\delta_{4r}(A)]^2} \| \nu \|_{\ell_2^p}^2} \right)^2 \\
= \left( \sqrt{1 + 3[\delta_{4r}(A)]^2} \| \widetilde{X} - X_0 \|_{S_2} + \sqrt{3(1 + \delta_{3r}(A)) \| \nu \|_{\ell_2^p}^2} \right)^2.
\end{align*}

Substituting \( \delta_{3r}(A) \leq \delta_{4r}(A) \leq 0.065 \) gives the constants in the final inequality.
REFERENCES


