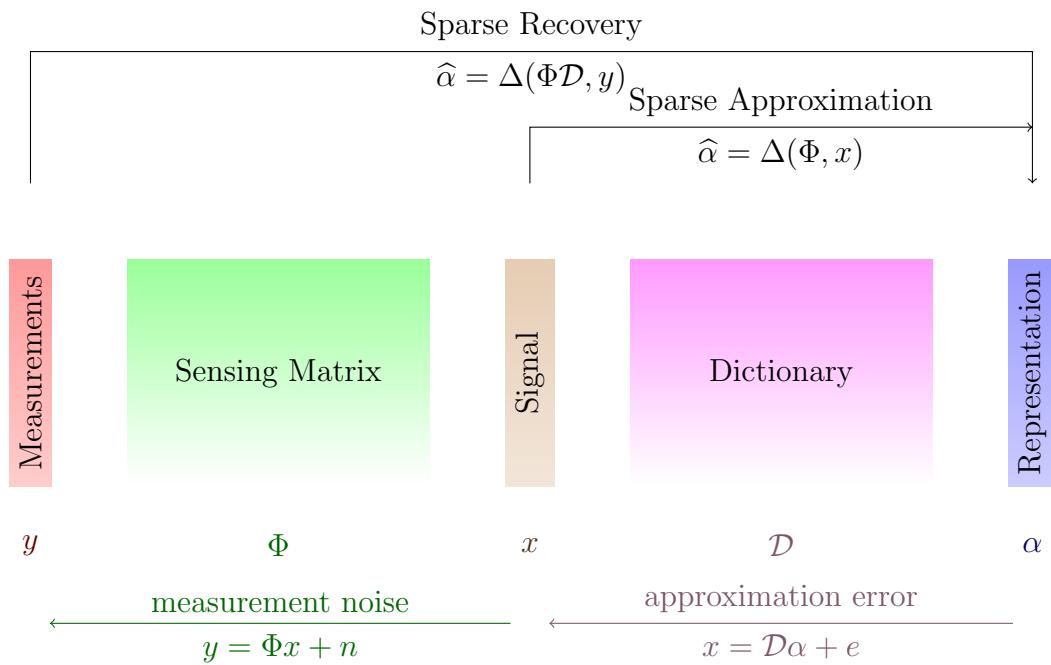


---

# Illustrated Compressed Sensing

---



Shailesh Kumar  
Electrical Engineering Department  
Indian Institute of Technology, Delhi

June 25, 2014

---

 This work is licensed under a   
Creative Commons Attribution 3.0 Unported License.

---

# Contents

List of Tables	vi
List of Figures	vii
Chapter 1. Introduction	1
1.1. Notation	2
1.2. Useful functions	5
1.3. Outline of the book	6
<b>Part 1. Sparse Representations and Compressed Sensing</b>	<b>7</b>
Chapter 2. Sparse Signal Models	8
2.1. Outline	8
2.2. Sparse solutions for under-determined linear systems	10
2.3. Sparsity in orthonormal bases	25
2.4. Sparse and redundant representations	48
2.5. p-norms and sparse signals	53
2.6. Compressible signals	58
2.7. Tools for dictionary analysis	66
2.8. Compressed sensing	90
2.9. Examples	98
2.10. Digest	106
Chapter 3. More Tools for Dictionary and Random Matrix Analysis	111
3.1. Restricted isometry property	111
3.2. Johnson Lindenstrauss theorem	146
3.3. Stable embeddings	156
3.4. Spark	160
3.5. Coherence	161
3.6. Babel function	164
3.7. Exact recovery coefficient	168
3.8. Digest	173

Chapter 4. Sensing Matrices	178
4.1. Introduction	178
4.2. Recovery of exactly sparse signals	179
4.3. Recovery of approximately sparse signals	182
4.4. Recovery in presence of measurement noise	188
4.5. The RIP and the NSP	200
4.6. Matrices satisfying RIP	204
Chapter 5. Dictionaries and Sensing Matrices	209
5.1. Dirac-DCT dictionary	209
5.2. Grassmannian frames	210
5.3. Rademacher sensing matrices	210
5.4. Gaussian sensing matrices	215
5.5. Partial Fourier sensing matrices	217
5.6. Digest	218
Chapter 6. Basis Pursuit for Sparse Recovery	219
6.1. Introduction	219
6.2. Basis Pursuit	222
6.3. Stability of sparsest solution	237
6.4. BPIC	246
6.5. $l_1$ penalty problem	249
6.6. BPIC for compressed sensing	264
6.7. Digest	280
Chapter 7. Matching Pursuit Algorithms	283
7.1. Introduction	283
7.2. Orthogonal Matching Pursuit for sparse approximation	288
7.3. Orthogonal Matching Pursuit for Compressed Sensing	306
7.4. Analysis of OMP using Restricted Isometry Property	320
7.5. Compressive Sampling Matching Pursuit	330
7.6. Digest	350
Chapter 8. Shrinkage and Thresholding Algorithms	351
8.1. Iterative hard thresholding for signal recovery	351

Chapter 9. Union of Orthonormal Bases	357
9.1. Sparse $l_p$ representations	358
9.2. Union of bases	365
9.3. Digest	379
Chapter 10. Compressed Sensing with Orthogonal Systems	381
10.1. Digest	381
<b>Part 2. Joint Recovery and Dictionary Learning Problems</b>	<b>383</b>
Chapter 11. Joint Sparsity Problems	384
11.1. Tools from matrix analysis	385
11.2. Sparse representation	388
11.3. Compressed sensing	395
11.4. Distributed compressed sensing	398
11.5. Miscellaneous results	398
11.6. Digest	399
Chapter 12. Joint Recovery Algorithms	400
12.1. Thresholding algorithm	401
12.2. Simultaneous orthogonal matching pursuit (S-OMP)	402
12.3. Thresholding recovery guarantees	406
12.4. Performance guarantees for S-OMP	410
12.5. S-OMP recovery guarantee for Exact sparse problem	411
12.6. Approximation with a sparsity bound	413
12.7. Joint $l_1$ minimization recovery guarantee	415
Chapter 13. Rank Aware and MUSIC based Algorithms for Joint Sparse Recovery	419
13.1. $l_0$ norm minimization	419
13.2. The MUSIC principle	429
13.3. MUSIC based joint recovery	432
13.4. Rank blindness in joint recovery algorithms	434
13.5. Rank aware algorithms	438
13.6. $l_1$ norm minimization	440

13.7. Digest	440
Chapter 14. Dictionary Learning	441
14.1. Introduction	441
14.2. Unique dictionary and matrix factorization	444
14.3. Digest	452
Chapter 15. Distributed Compressed Sensing	453
15.1. Introduction	453
15.2. Framework for joint sparsity	455
15.3. Theoretical bounds on measurement rates	464
15.4. Practical recover algorithms	470
<b>Part 3. Inference</b>	471
Chapter 16. Detection with Compressed Measurements	472
16.1. Binary detection theory	472
16.2. Detection with compressed measurements	490
Appendix A. Useful MATLAB Functions	499
A.1. General purpose utilities	499
A.2. Functions for generating signal patterns	499
Appendix. Bibliography	501
Appendix. Index	506

## List of Tables

1	Mutual coherence of Dirac and Fourier bases	32
2	Entries in wavelet representation of piecewise cubic polynomial signal higher than a threshold	99
1	Symbols used in this part of the book	385
1	Symbols used in this chapter	454

## List of Figures

2.1 An under-determined system $3x_1 + 4x_2 = 12$	12
2.2 Minimum $l_2$ norm solution for the under-determined system $3x_1 + 4x_2 = 12$	15
2.3 Minimum $l_1$ norm solution for the under-determined system $3x_1 + 4x_2 = 12$	18
2.4 Mutual coherence for Dirac and Fourier bases	33
2.5 A naive algorithm for computing the spark of a matrix	69
2.6 A piecewise cubic polynomials signal	100
2.7 Sparse representation of signal in wavelet basis	101
2.8 Wavelet coefficients sorted by magnitude	102
2.9 Measurement vector $y = \Phi x + e$	103
2.10 Daubechies-8 wavelet basis	104
2.11 Gaussian sensing matrix $\Phi$	104
2.12 Recovery matrix $\Phi\Psi$	105
3.1 Required subspace dimension $M$ for $K$ points for restricted isometry constant $\delta$	148
7.1 Orthogonal matching pursuit for sparse approximation	289
7.2 Orthogonal matching pursuit for sparse recovery in CS	319
7.3 Orthogonal matching pursuit [17]	320
7.4 Sketch of OMP without intermediate $\alpha^k$ computation	322
7.5 Sketch of OMP without intermediate $\alpha^k$ computation	323
7.6 CoSaMP for iterative sparse signal recovery	331

8.1 Iterative hard thresholding for sparse signal recovery	352
12. Simultaneous Orthogonal Matching Pursuit	403
14. Dictionary learning: iterative approach	443
15. Bipartite graph for DCS	466

## CHAPTER 1

### Introduction

Today we are witnessing a huge data deluge all around us. 10 megapixel cameras are now a norm while even up to 60 mega pixel cameras are available in market. Sampling process remains dominated by Nyquist formulation, thus leading to denser uniform sampling grids. This has led to two specific problems. We require more sophisticated sampling devices to be able to sample at such high resolution. Also we need much bigger storage space to store high quality signals (We will be primarily looking at audio, images, video signals).

In order to facilitate easy transfer of signals, we resort to lossy compression techniques. For most purposes there is essentially not much perceptual difference between the original signal and the lossy compressed version. This prevailing paradigm can be summarized as *SAMPLE THEN COMPRESS* paradigm.

Compression essentially is based on looking at the signal in some orthogonal basis (Fourier, DCT, Wavelet, etc.) and keeping only those coefficients in the basis which contain essential signal information. We can think of these coefficients as linear measurements on the signal samples.

Compression also leads to generation loss. Every time we go through the process of *Decompress* → *Process* → *Compress* a signal, we introduce another generation of loss to the signal. After certain generations, compression artifacts start dominating and signal becomes useless for further processing.

What if we could make such measurements directly before sampling? Consider a 10 million pixels image in which essential information is

carried in just 10 thousand wavelet coefficients. Suppose a sampling process could directly give us these 10 thousand measurements, then we could have made our sampling system much simpler as well as avoided the post sampling compression process altogether! But this looks like a pipe dream. For how would the sampling process know as to which 10 thousand coefficients are really important in a given image? And then this number could vary from image to image. For a low detail image of say a plain background, very few coefficients might be required while for a highly detailed image of a garden with variety of flowers, many more coefficients might be required. Moreover, making such specific measurements will require having elaborate analog circuitry dedicated for combining information from different samples into these measurements. Isn't there a way out?

It turns out that there is indeed a way around. Compressed sensing provides the necessary mathematical framework for working with reduced number of signal measurements rather than the huge size original signals while still retaining all necessary signal information to achieve high quality perceptual experience. This novel way of working with signals can be called as *COMPRESS THEN SAMPLE* paradigm.

In this book, we will take a tour of principles of compressed sensing and see its applications.

### 1.1. Notation

$\forall$ : for all (for each)

$\exists$ : there exists

$\implies$  : implies

$\iff$  : if and only if

$\in$ : belongs to

$\notin$ : doesn't belong to

$\subset$ : Proper subset

$\subseteq$ : Subset

$\supset$ : Proper superset

- $\supseteq$ : Superset
- $\ll$ : Much less than
- $\gg$ : Much greater than
- $\prec$ : less than (partial order)
- $\preceq$ : less than or equal to (partial order)
- $\succ$ : greater than (partial order)
- $\succeq$ : greater than or equal to (partial order)
- $\binom{N}{K}$ : Binomial coefficient  $N$  choose  $K$
- $\mathbb{E}$ : Expectation operator
- $\mathbb{N}$ : The set of natural numbers (1 onwards)
- $\mathbb{P}$ : Probability of something
- $\mathbb{Q}$ : The set of rational numbers
- $\mathbb{R}$ : The field of real numbers (a.k.a. the real line)
- $\mathbb{C}$ : The field of complex numbers
- $\mathbb{R}^N$ : The  $N$ -dimensional Euclidean space
- $\mathbb{C}^N$ : The  $N$ -dimensional complex space
- $\mathbb{R}^{M \times N}$ : The vector space of  $M \times N$  real matrices
- $\mathbb{C}^{M \times N}$ : The vector space of  $M \times N$  complex matrices
- $\mathbb{Z}$ : The set of integers
- $\mathbb{Z}^+$ : The set of positive integers
- $x$ : A signal in the signal space  $\mathbb{C}^N$
- $\|x\|_0$ :  $l_0$ -“norm” of a vector (number of non-zero entries)
- $\|x\|_2$ :  $l_2$  norm (Euclidean norm) of a vector
- $\|x\|_1$ :  $l_1$  norm (sum of absolute values) of a vector
- $\|x\|_\infty$ :  $l_\infty$  norm (maximum absolute value) of a vector
- $|x|$ : Vector of absolute values of entries in  $x$
- $x^+$ : Positive part of  $x$
- $x^-$ : Negative part of  $x$  [ $x = x^+ - x^-$ ]
- $\langle x, y \rangle$ : Inner product
- $A$ : a matrix
- $A^T$ : Transpose
- $A^H$ : Hermitian transpose
- $\overline{A}$ : Complex conjugate

$A^{-1}$ : Inverse

$A^\dagger$ : Pseudo-inverse

$A = U\Sigma V^H$ : Singular value decomposition

$A = Q\Lambda Q^H$ : Eigen value decomposition

$\det(A)$ : Determinant of a matrix

$|A|$ : Matrix of absolute values of entries of  $A$

$\lambda$ : An eigen value

$\sigma$ : A singular value

$\Lambda$ : A diagonal matrix of eigen values of  $A$

$\Sigma$ : A diagonal matrix of singular values of  $A$

$A \succ 0$ : Positive definite (p.d.)

$A \succeq 0$ : Positive semidefinite (p.s.d.)

$a_i$ :  $i$ -th column in a matrix  $A$

$\underline{a}_i$ :  $i$ -th row in a matrix  $A$

$\|A\|_F$ : Frobenius norm of a matrix

$\|A\|_S$ : Sum norm of a matrix

$\|A\|_M$ : Max norm of a matrix

$\|A\|_2$ : Spectral norm of a matrix

$\|A\|_1$ : Max column sum norm of a matrix

$\|A\|_\infty$ : Max row sum norm of a matrix

$I$ : Identity matrix

$0$ : Zero matrix

$1$ : One matrix

$N$ : Dimension of ambient signal space

$D$ : Dimension of representation space (number of atoms in a dictionary)

$M$ : Dimension of measurement space (number of measurements)

$S$ : Number of signals in an MMV problem or multi-channel recovery problem

$K$ : Sparsity level

$\Omega$ : The set of indices  $\{1, 2, \dots, N\}$

$\Gamma$ : A subset of indices  $\Gamma \subseteq \Omega$

$\mathbb{C}^\Gamma$ : The vector space of signals by restricting signals to entries indexed by  $\Gamma$  or setting entries indexed by  $\Omega \setminus \Gamma$  to 0.

$\mathcal{D}$ : A dictionary (either a set of atoms or its matrix representation)

$\Phi$ : A sensing matrix

$\phi_i$ : A column vector in a sensing matrix

$\Psi$ : An orthonormal basis (or its corresponding matrix representation)

**( $\mathcal{D}, K$ )-sparse**: A signal which is  $K$ -sparse in a dictionary  $\mathcal{D}$

$\alpha$ : A representation of a signal  $x$  in a dictionary  $\mathcal{D}$  [ $x = \mathcal{D}\alpha$ ]

$e$ : Approximation error [ $x = \mathcal{D}\alpha + e$ ] or measurement error [ $y = \Phi x + e$ ]

$\Delta$ : Sparse recovery process (algorithm)  $\hat{\alpha} = \Delta(\mathcal{D}, x)$  or  $\hat{x} = \Delta(\Phi, y)$  or  
 $\hat{x}, \hat{\alpha} = \Delta(\Phi, \mathcal{D}, y)$

$\mu$ : Dictionary coherence

$\delta$ : Restricted isometry constant

$X$ : An ensemble of signals

$Y$ : An ensemble of measurements

$E$ : An ensemble of error vectors

$\mathcal{A}$ : An ensemble of representations

$\log$ : Logarithm to base 10

$\ln$ : Natural logarithm [to base  $e$ ]

$\log_2$ : Logarithm to base 2

## 1.2. Useful functions

The function  $\text{sgn} : \mathbb{R} \rightarrow \mathbb{R}$  is defined as

$$\text{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \quad (1.2.1)$$

We define an extension  $\text{sgn} : \mathbb{R}^N \rightarrow \mathbb{R}^N$  as

$$\text{sgn}(x) = (\text{sgn}(x_1), \text{sgn}(x_2), \dots, \text{sgn}(x_N)) \quad (1.2.2)$$

$$\delta : \mathbb{N} \times \mathbb{N} \rightarrow \{0, 1\}$$

$$\delta(i, j) = \begin{cases} 1 & i = 0 \\ 0 & i \neq 0 \end{cases} \quad (1.2.3)$$

Also written as  $\delta_{ij}$  or  $\delta_{i,j}$ .

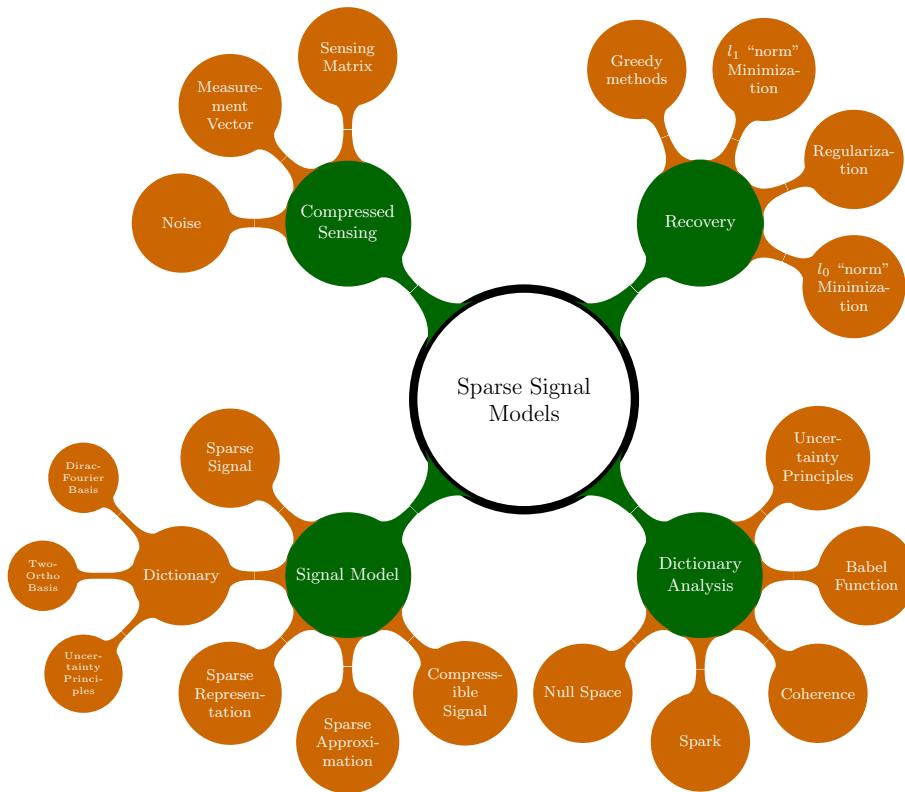
### 1.3. Outline of the book

## **Part 1**

# **Sparse Representations and Compressed Sensing**

## CHAPTER 2

# Sparse Signal Models



### 2.1. Outline

In this chapter we develop initial concepts of sparse signal models.

We begin our study with a review of solutions of under-determined systems. We build a case for solutions which promote sparsity.

We show that although the real life signals may not be sparse yet they are compressible and can be approximated with sparse signals.

We then review orthonormal bases and explain the inadequacy of those bases in exploiting the sparsity in many signals of interest. We develop an example of Dirac Fourier basis as a two ortho basis and demonstrate how it can better exploit signal sparsity compared to Dirac basis and Fourier basis individually.

We follow this with a general discussion of redundant signal dictionaries. We show how they can be used to create sparse and redundant signal representations.

We study various properties of signal dictionaries which are useful in characterizing the capabilities of a signal dictionary in exploiting signal sparsity.

In this chapter, our signals of interest will typically lie in the finite  $N$ -dimensional complex vector space  $\mathbb{C}^N$ . Sometimes we will restrict our attention to the  $N$  dimensional Euclidean space to simplify discussion.

We will be concerned with different representations of our signals of interest in  $\mathbb{C}^D$  where  $D \geq N$ . This aspect will become clearer as we go along in this chapter.

## Sparsity

We quickly define the notion of sparsity in a signal.

We recall the definition of  $l_0$ -“norm” (don’t forget the quotes) of  $x \in \mathbb{C}^N$  given by

$$\|x\|_0 = |\text{supp}(x)|$$

where  $\text{supp}(x) = \{i : x_i \neq 0\}$  denotes the support of  $x$ .

Informally we say that a signal  $x \in \mathbb{C}^N$  is **sparse** if  $\|x\|_0 \ll N$ .

More generally if  $x = \mathcal{D}\alpha$  where  $\mathcal{D} \in \mathbb{C}^{N \times D}$  with  $D > N$  is some signal dictionary (to be formally defined later), then  $x$  is sparse in dictionary  $\mathcal{D}$  if  $\|\alpha\|_0 \ll D$ .

Sometimes we simply say that  $x$  is  $K$ -sparse if  $\|x\|_0 \leq K$  where  $K < N$ . We do not specifically require that  $K \ll N$ .

An even more general definition of sparsity is the degrees of freedom a signal may have.

As an example consider all points on the surface of a unit sphere in  $\mathbb{R}^N$ . For every point  $x$  belonging to the surface  $|x|_2 = 1$ . Thus if we choose the values of  $N - 1$  components of  $x$  then the value of the remaining component is automatically fixed. Thus the number of degrees of freedom  $x$  has on the surface of the unit sphere in  $\mathbb{R}^N$  is actually  $N - 1$ . Such a surface represents a manifold in the ambient Euclidean space. Of special interest are low dimensional manifolds where the number of degrees of freedom  $K \ll N$ .

## 2.2. Sparse solutions for under-determined linear systems

The discussion in this section is largely based on chapter 1 of [21].

Consider a matrix  $\Phi \in \mathbb{C}^{M \times N}$  with  $M < N$ .

Define an under-determined system of linear equations:

$$\Phi x = y \quad (2.2.1)$$

where  $y \in \mathbb{C}^M$  is known and  $x \in \mathbb{C}^N$  is unknown.

This system has  $N$  unknowns and  $M$  linear equations. There are more unknowns than equations.

Let the columns of  $\Phi$  be given by  $\phi_1, \phi_2, \dots, \phi_N$ .

Column space of  $\Phi$  (vector space spanned by all columns of  $\Phi$ ) is denoted by  $\mathcal{C}(\Phi)$  i.e.

$$\mathcal{C}(\Phi) = \sum_{i=1}^N c_i \phi_i, \quad c_i \in \mathbb{C}.$$

We know that  $\mathcal{C}(\Phi) \subset \mathbb{C}^M$ .

Clearly  $\Phi x \in \mathcal{C}(\Phi)$  for every  $x \in \mathbb{C}^N$ . Thus if  $y \notin \mathcal{C}(\Phi)$  then we have no solution. But, if  $y \in \mathcal{C}(\Phi)$  then we have infinite number of solutions.

Let  $\mathcal{N}(\Phi)$  represent the null space of  $\Phi$  given by

$$\mathcal{N}(\Phi) = \{x \in \mathbb{C}^N : \Phi x = 0\}.$$

Let  $\hat{x}$  be a solution of  $y = \Phi x$ . And let  $z \in \mathcal{N}(\Phi)$ . Then

$$\Phi(\hat{x} + z) = \Phi\hat{x} + \Phi z = y + 0 = y.$$

Thus the set  $\hat{x} + \mathcal{N}(\Phi)$  forms the complete set of infinite solutions to the problem  $y = \Phi x$  where

$$\hat{x} + \mathcal{N}(\Phi) = \{\hat{x} + z \quad \forall z \in \mathcal{N}(\Phi)\}.$$

**Example 2.1: An under-determined system** As a running example in this section, we will consider a simple under-determined system in  $\mathbb{R}^2$ . The system is specified by

$$\Phi = \begin{bmatrix} 3 & 4 \end{bmatrix}$$

and

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

with

$$\Phi x = y = 12.$$

where  $x$  is unknown and  $y$  is known. Alternatively

$$\begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 12$$

or more simply

$$3x_1 + 4x_2 = 12.$$

The solution space of this system is a line in  $\mathbb{R}^2$  which is shown in fig. 2.1.

Specification of the under-determined system as above, doesn't give us any reason to prefer one particular point on the line as the preferred solution.

Two specific solutions are of interest

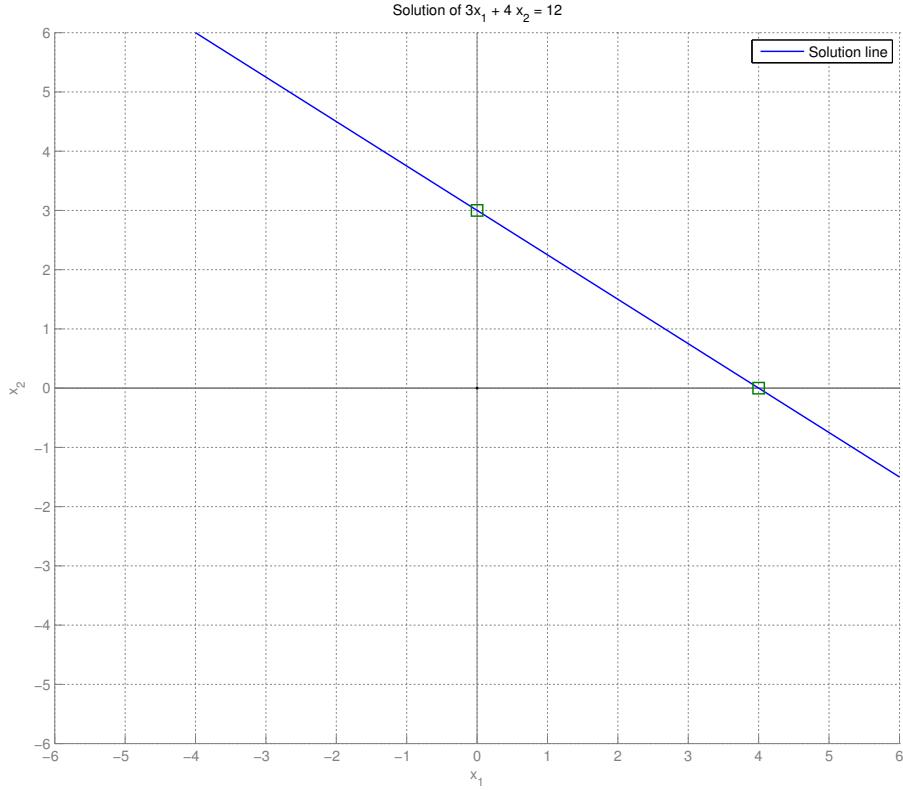


FIGURE 2.1. An under-determined system  $3x_1 + 4x_2 = 12$

- $(x_1, x_2) = (4, 0)$  lies on the  $x_1$  axis.
- $(x_1, x_2) = (0, 3)$  lies on the  $x_2$  axis.

In both of these solutions, one component is 0, thus leading these solutions to be sparse.

It is easy to visualize sparsity in this simplified 2-dimensional setup but situation becomes more difficult when we are looking at high dimensional signal spaces. We need well defined criteria to promote sparse solutions.  $\square$

### 2.2.1. Regularization

Are all these solutions equivalent or can we say that one solution is better than the other in some sense? In order to suggest that some

solution is better than other solutions, we need to define a criteria for comparing two solutions.

In optimization theory, this idea is known as **regularization**. We define a cost function  $J(x) : \mathbb{C}^N \rightarrow \mathbb{R}$  which defines the **desirability** of a given solution  $x$  out of infinitely possible solutions. The higher the cost, lower is the desirability of the solution. Thus the goal of the optimization problem is to find a desired  $x$  with minimum possible cost.

In optimization literature, the cost function is one type of **objective function**. While the objective of an optimization problem might be either minimized or maximized, cost is always minimized.

We can write this optimization problem as

$$\begin{aligned} & \underset{x}{\text{minimize}} \quad J(x) \\ & \text{subject to} \quad y = \Phi x. \end{aligned} \tag{2.2.2}$$

If  $J(x)$  is convex, then its possible to find a global minimum cost solution over the solution set.

If  $J(x)$  is not convex, then it may not be possible to find a global minimum, we may have to settle with a local minimum.

A variety of such cost function based criteria can be considered.

### 2.2.2. $l_2$ regularization

One of the most common criteria is to choose a solution with the smallest  $l_2$  norm.

The problem can then be reformulated as an optimization problem

$$\begin{aligned} & \underset{x}{\text{minimize}} \quad \|x\|_2 \\ & \text{subject to} \quad y = \Phi x. \end{aligned} \tag{P}_2$$

In fact minimizing  $\|x\|_2$  is same as minimizing its square  $\|x\|_2^2 = x^H x$ .

So an equivalent formulation is

$$\begin{aligned} & \underset{x}{\text{minimize}} && x^H x \\ & \text{subject to} && y = \Phi x. \end{aligned} \tag{P}_2$$

**Example 2.2: Minimum  $l_2$  norm solution for an under-determined system** We continue with our running example.

We can write  $x_2$  as

$$x_2 = 3 - \frac{3}{4}x_1.$$

With this definition the squared  $l_2$  norm of  $x$  becomes

$$\begin{aligned} \|x\|_2^2 &= x_1^2 + x_2^2 = x_1^2 + \left(3 - \frac{3}{4}x_1\right)^2 \\ &= \frac{25}{16}x_1^2 - \frac{9}{2}x_1 + 9. \end{aligned}$$

Minimizing  $\|x\|_2^2$  over all  $x$  is same as minimizing over all  $x_1$ .

Since  $\|x\|_2^2$  is a quadratic function of  $x_1$ , we can simply differentiate it and equate to 0 giving us

$$\frac{25}{8}x_1 - \frac{9}{2} = 0 \implies x_1 = \frac{36}{25} = 1.44.$$

This gives us

$$x_2 = \frac{48}{25} = 1.92.$$

Thus the optimal  $l_2$  norm solution is obtained at  $(x_1, x_2) = (1.44, 1.92)$ .

We note that the minimum  $l_2$  norm at this solution is

$$\|x\|_2 = \frac{12}{5} = 2.4.$$

It is instructive to note that the  $l_2$  norm cost function prefers a non-sparse solution to the optimization problem.

We can view this solution graphically by drawing  $l_2$  norm balls of different radii in fig. 2.2. The ball which just touches the solution space line (i.e. the line is tangent to the ball) gives us the optimal solution.

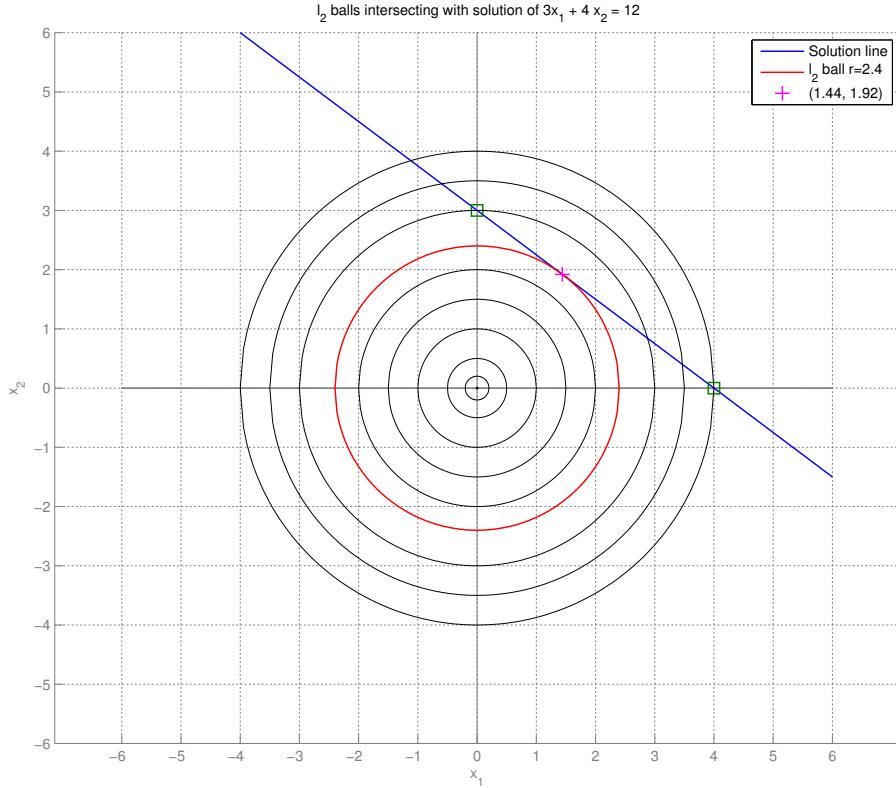


FIGURE 2.2. Minimum  $l_2$  norm solution for the under-determined system  $3x_1 + 4x_2 = 12$

All other norm balls either don't touch the solution line at all, or they cross it at exactly two points.  $\square$

A formal solution to  $l_2$  norm minimization problem can be easily obtained using Lagrange multipliers.

We define the Lagrangian

$$\mathcal{L}(x) = \|x\|_2^2 + \lambda^H(\Phi x - y) \quad (2.2.3)$$

with  $\lambda \in \mathbb{C}^M$  being the Lagrange multipliers for the (equality) constraint set.

Differentiating  $\mathcal{L}(x)$  w.r.t.  $x$  we get

$$\frac{\partial \mathcal{L}(x)}{\partial x} = 2x + \Phi^H \lambda. \quad (2.2.4)$$

By equating the derivative to 0 we obtain the optimal value of  $x$  as

$$x^* = -\frac{1}{2}\Phi^H \lambda. \quad (2.2.5)$$

Plugging this solution back into the constraint  $\Phi x = y$  gives us

$$\Phi x^* = -\frac{1}{2}(\Phi\Phi^H)\lambda = y \implies \lambda = -2(\Phi\Phi^H)^{-1}y. \quad (2.2.6)$$

In above we are implicitly assuming that  $\Phi$  is a full rank matrix thus,  $\Phi\Phi^H$  is invertible and positive definite.

Putting  $\lambda$  back in eq. (2.2.5) we obtain the well known closed form least squares solution using pseudo-inverse solution

$$x^* = \Phi^H(\Phi\Phi^H)^{-1}y = \Phi^\dagger y. \quad (2.2.7)$$

We would like to mention that there are several iterative approaches to solve the  $l_2$  norm minimization problem (like gradient descent and conjugate descent). For large systems, they are more effective than computing the pseudo-inverse.

The beauty of  $l_2$  norm minimization lies in its simplicity and availability of closed form analytical solutions. This has led to its prevalence in various fields of science and engineering. But  $l_2$  norm is by no means the only suitable cost function. Rather the simplicity of  $l_2$  norm often drives engineers away from trying other possible cost functions. In the sequel, we will look at various other possible cost functions.

**2.2.2.1. Convexity.** Convex optimization problems have a unique feature that it is possible to find the global optimal solution if such a solution exists.

The solution space  $\Omega = \{x : \Phi x = y\}$  is convex. Thus the feasible set of solutions for the optimization problem (2.2.2) is also convex. All it remains is to make sure that we choose a cost function  $J(x)$

which happens to be convex. This will ensure that a global minimum can be found through convex optimization techniques. Moreover, if  $J(x)$  is strictly convex, then it is guaranteed that the global minimum solution is *unique*. Thus even though, we may not have a nice looking closed form expression for the solution of a strictly convex cost function minimization problem, the guarantee of the existence and uniqueness of solution as well as well developed algorithms for solving the problem make it very appealing to choose cost functions which are convex.

We remind that all  $l_p$  norms with  $p \geq 1$  are convex functions. In particular  $l_\infty$  and  $l_1$  norms are very interesting and popular where

$$l_\infty(x) = \max(x_i), 1 \leq i \leq N$$

and

$$l_1(x) = \sum_{i=1}^N |x_i|.$$

In the following section we will attempt to find a unique solution to our optimization problem (2.2.2) using  $l_1$  norm.

### 2.2.3. $l_1$ regularization

In this section we will restrict our attention to the Euclidean space case where  $x \in \mathbb{R}^N$ ,  $\Phi \in \mathbb{R}^{M \times N}$  and  $y \in \mathbb{R}^M$ .

We choose our cost function  $J(x) = l_1(x)$ .

The cost minimization problem can be reformulated as

$$\begin{aligned} & \underset{x}{\text{minimize}} && \|x\|_1 \\ & \text{subject to} && \Phi x = y. \end{aligned} \tag{P}_1$$

**Example 2.3: Minimum  $l_1$  norm solution for an under-determined system** We continue with our running example.

Again we can view this solution graphically by drawing  $l_1$  norm balls of different radii in fig. 2.3. The ball which just touches the solution space line gives us the optimal solution.

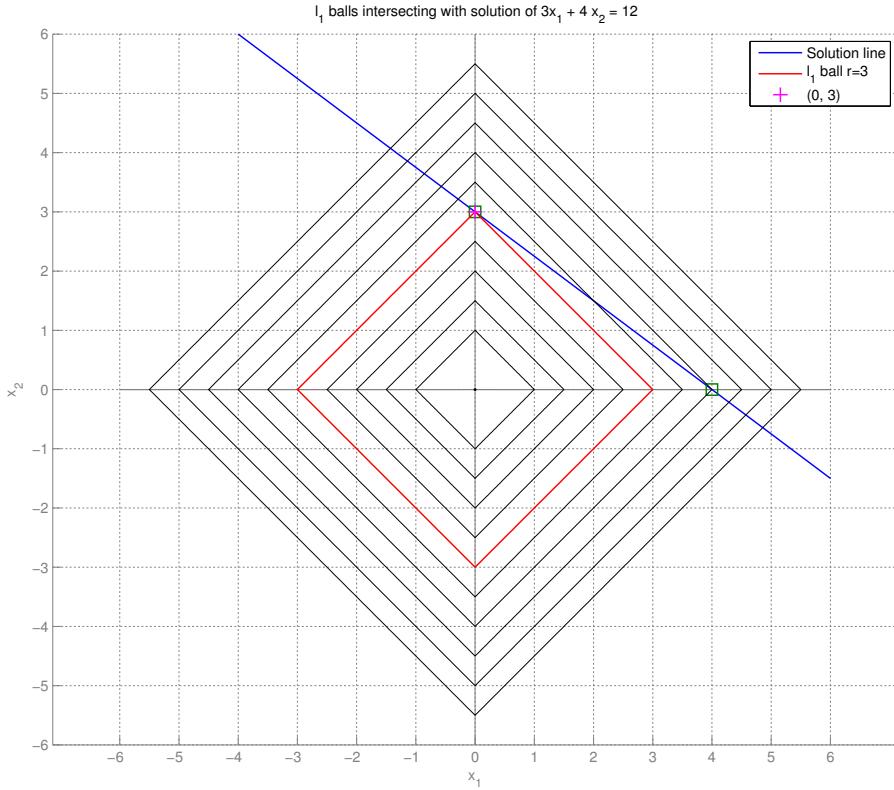


FIGURE 2.3. Minimum  $l_1$  norm solution for the under-determined system  $3x_1 + 4x_2 = 12$

As we can see from the figure the minimum  $l_1$  norm solution is given by  $(x_1, x_2) = (0, 3)$ .

It is interesting to note that  $l_1$  norm solution promotes sparser solutions while  $l_2$  norm solution promotes solutions in which signal energy is distributed amongst all of its components.  $\square$

Its time to have a closer look at our cost function  $J(x) = \|x\|_1$ . This function is convex yet not strictly convex.

**Example 2.4:  $\|x\|_1$  is not strictly convex** Consider again  $x \in \mathbb{R}^2$ . For  $x \in \mathbb{R}_+^2$  (the first quadrant),

$$\|x\|_1 = x_1 + x_2.$$

Hence for any  $c_1, c_2 \geq 0$  and  $x, y \in \mathbb{R}_+^2$ :

$$\|(c_1x + c_2y)\|_1 = (c_1x + c_2y)_1 + (c_1x + c_2y)_2 = c_1\|x\|_1 + c_2\|y\|_1.$$

Thus,  $l_1$ -norm is not strictly convex. Consequently, a unique solution may not exist for  $l_1$  norm minimization problem.

As an example consider the under-determined system

$$3x_1 + 3x_2 = 12.$$

We can easily visualize that the solution line will pass through points  $(0, 4)$  and  $(4, 0)$ . Moreover, it will be clearly parallel with  $l_1$ -norm ball of radius 4 in the first quadrant. See again fig. 2.3. This gives us infinitely possible solutions to the minimization problem ( $P_1$ ).

We can still observe that

- these solutions are gathered in a small line segment that is bounded (a bounded convex set) and
- There exist two solutions  $(4, 0)$  and  $(0, 4)$  amongst these solutions which have only 1 non-zero component.

□

For the  $l_1$  norm minimization problem since  $J(x)$  is not strictly convex, hence a unique solution may not be guaranteed. In specific cases, there may be infinitely many solutions. Yet what we can claim is

- these solutions are gathered in a set that is bounded and convex, and
- among these solutions, there exists at least one solution with at most  $M$  non-zeros (as the number of constraints in  $\Phi x = y$ ).

**Theorem 2.1** Let  $S$  denote the solution set of  $l_1$  norm minimization problem ( $P_1$ ).  $S$  contains at least one solution  $\hat{x}$  with  $\|\hat{x}\|_0 = M$ .

PROOF. We have

- $S$  is convex and bounded.
- $\Phi x^* = y \quad \forall x^* \in S$ .
- Since  $\Phi \in \mathbb{R}^{M \times N}$  is full rank and  $M < N$ , hence  $\text{rank}(\Phi) = M$ .

Let  $x^* \in S$  be an optimal solution with  $\|x^*\|_0 = L > M$ .

Consider the  $L$  columns of  $\Phi$  which correspond to  $\text{supp}(x^*)$ .

Since  $L > M$  and  $\text{rank}(\Phi) = M$  hence these columns linearly dependent.

Thus there exists a vector  $h \in \mathbb{R}^N$  with  $\text{supp}(h) \subseteq \text{supp}(x^*)$  such that

$$\Phi h = 0.$$

Note that since we are only considering those columns of  $\Phi$  which correspond to  $\text{supp}(x)$ , hence we require  $h_i = 0$  whenever  $x_i^* = 0$ .

Consider a new vector

$$x = x^* + \epsilon h$$

where  $\epsilon$  is small enough such that every element in  $x$  has the same sign as  $x^*$ .

As long as

$$|\epsilon| \leq \min_{i \in \text{supp}(x^*)} \frac{|x_i^*|}{|h_i|} = \epsilon_0$$

such an  $x$  can be constructed.

Note that  $x_i = 0$  whenever  $x_i^* = 0$ .

Clearly

$$\Phi x = \Phi(x^* + \epsilon h) = y + \epsilon 0 = y.$$

Thus  $x$  is a feasible solution to the problem ( $P_1$ ) though it need not be an optimal solution.

But since  $x^*$  is optimal hence, we must assume that  $l_1$  norm of  $x$  is greater than or equal to the  $l_1$  norm of  $x^*$

$$\|x\|_1 = \|x^* + \epsilon h\|_1 \geq \|x^*\|_1 \quad \forall |\epsilon| \leq \epsilon_0.$$

Now look at  $\|x\|_1$  as a function of  $\epsilon$  in the region  $|\epsilon| \leq \epsilon_0$ .

In this region,  $l_1$  function is continuous and differentiable since all vectors  $x^* + \epsilon h$  have the same sign pattern. If we define  $y^* = |x^*|$  (the vector of absolute values), then

$$\|x^*\|_1 = \|y^*\|_1 = \sum_{i=1}^N y_i^*.$$

Since the sign patterns don't change, hence

$$|x_i| = |x_i^* + \epsilon h_i| = y_i^* + \epsilon h_i \operatorname{sgn}(x_i^*).$$

Thus

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^N |x_i| \\ &= \sum_{i=1}^N (y_i^* + \epsilon h_i \operatorname{sgn}(x_i^*)) \\ &= \|x^*\|_1 + \epsilon \sum_{i=1}^N h_i \operatorname{sgn}(x_i^*) \\ &= \|x^*\|_1 + \epsilon h^T \operatorname{sgn}(x^*). \end{aligned}$$

The quantity  $h^T \operatorname{sgn}(x^*)$  is a constant. The inequality  $\|x\|_1 \geq \|x^*\|_1$  applies to both positive and negative values of  $\epsilon$  in the region  $|\epsilon| \leq \epsilon_0$ . This is possible only when inequality is in fact an equality.

This implies that the addition / subtraction of  $\epsilon h$  under these conditions does not change the  $l_1$  length of the solution. Thus,  $x \in S$  is also an optimal solution.

This can happen only if

$$h^T \operatorname{sgn}(x^*) = 0.$$

We now wish to tune  $\epsilon$  such that one entry in  $x^*$  gets nulled while keeping the solutions  $l_1$  length.

We choose  $i$  corresponding to  $\epsilon_0$  (defined above) and pick

$$\epsilon = \frac{-x_i^*}{h_i}.$$

Clearly for the corresponding

$$x = x^* + \epsilon h$$

the  $i$ -th entry is nulled while others keep their sign and the  $l_1$  norm is also preserved. Thus, we have got a new optimal solution with  $L - 1$  non-zeros at the most. It is possible that more than 1 entries get nulled this operation.

We can repeat this procedure till we are left with  $M$  non-zero elements.

Beyond this we may not proceed since  $\operatorname{rank}(\Phi) = M$  hence we cannot say that corresponding columns of  $\Phi$  are linearly dependent.  $\square$

We thus note that  $l_1$  norm has a tendency to prefer sparse solutions. This is a well known and fundamental property of linear programming.

#### 2.2.4. $l_1$ norm minimization problem as a linear programming problem

We now show that  $(P_1)$  in  $\mathbb{R}^N$  is in fact a linear programming problem.

Recalling the problem:

$$\begin{aligned} & \underset{x \in \mathbb{R}^N}{\text{minimize}} && \|x\|_1 \\ & \text{subject to} && y = \Phi x. \end{aligned}$$

Let us write  $x$  as  $u - v$  where  $u, v \in \mathbb{R}^N$  are both non-negative vectors such that  $u$  takes all positive entries in  $x$  while  $v$  takes all the negative entries in  $x$ .

**Example 2.5:**  $x = u - v$  Let

$$x = (-1, 0, 0, 2, 0, 0, 0, 4, 0, 0, -3, 0, 0, 0, 0, 2, 10).$$

Then

$$u = (0, 0, 0, 2, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 2, 10).$$

And

$$v = (1, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0).$$

Clearly  $x = u - v$ . □

We note here that by definition

$$\text{supp}(u) \cap \text{supp}(v) = \emptyset$$

i.e. support of  $u$  and  $v$  do not overlap.

We now construct a vector

$$z = \begin{bmatrix} u \\ v \end{bmatrix} \in \mathbb{R}^{2N}.$$

We can now verify that

$$\|x\|_1 = \|u\|_1 + \|v\|_1 = 1^T z.$$

And

$$\Phi x = \Phi(u - v) = \Phi u - \Phi v = \begin{bmatrix} \Phi & -\Phi \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \Phi & -\Phi \end{bmatrix} z$$

where  $z \succeq 0$ .

Hence the optimization problem ( $P_1$ ) can be recast as

$$\underset{z \in \mathbb{R}^{2N}}{\text{minimize}} \quad 1^T z$$

$$\text{subject to} \quad \begin{bmatrix} \Phi & -\Phi \end{bmatrix} z = y \quad (P_1(\text{LP}))$$

$$\text{and} \quad z \succeq 0.$$

This optimization problem has the classic Linear Programming structure since the objective function is affine as well as constraints are affine.

Let  $z^* = \begin{bmatrix} u^* \\ v^* \end{bmatrix}$  be an optimal solution to the problem ( $P_1(\text{LP})$ ).

In order to show that the two optimization problems are equivalent, we need to verify that our assumption about the decomposition of  $x$  into positive entries in  $u$  and negative entries in  $v$  is indeed satisfied by the optimal solution  $u^*$  and  $v^*$ . i.e. support of  $u^*$  and  $v^*$  do not overlap.

Since  $z \succeq 0$  hence  $\langle u^*, v^* \rangle \geq 0$ . If support of  $u^*$  and  $v^*$  don't overlap, then we have  $\langle u^*, v^* \rangle = 0$ . And if they overlap then  $\langle u^*, v^* \rangle > 0$ .

Now for the sake of contradiction, let us assume that support of  $u^*$  and  $v^*$  do overlap for the optimal solution  $z^*$ .

Let  $k$  be one of the indices at which both  $u_k \neq 0$  and  $v_k \neq 0$ . Since  $z \succeq 0$ , hence  $u_k > 0$  and  $v_k > 0$ .

Without loss of generality let us assume that  $u_k > v_k > 0$ .

In the equality constraint

$$\begin{bmatrix} \Phi & -\Phi \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = y$$

Both of these coefficients multiply the same column of  $\Phi$  with opposite signs giving us a term

$$\phi_k(u_k - v_k).$$

Now if we replace the two entries in  $z^*$  by

$$u'_k = u_k - v_k$$

and

$$v'_k = 0$$

to obtain an new vector  $z'$ , we see that there is no impact in the equality constraint

$$\begin{bmatrix} \Phi & -\Phi \end{bmatrix} z = y.$$

Also the positivity constraint

$$z \succeq 0$$

is satisfied. This means that  $z'$  is a feasible solution.

On the other hand the objective function  $1^T z$  value reduces by  $2v_k$  for  $z'$ . This contradicts our assumption that  $z^*$  is the optimal solution.

Hence for the optimal solution of  $(P_1(\text{LP}))$  we have

$$\text{supp}(u^*) \cap \text{supp}(v^*) = \emptyset$$

thus

$$x^* = u^* - v^*$$

is indeed the desired solution for the optimization problem  $(P_1)$ .

### 2.3. Sparsity in orthonormal bases

We start this section with a quick review of orthonormal bases and orthogonal transforms for finite dimensional signals  $x \in \mathbb{C}^N$ . We look at several examples of sparse signals in different orthonormal bases. We then demonstrate that while an orthonormal bases is a complete representation of all signals  $x \in \mathbb{C}^N$  yet, its not a good tool for exploiting the sparsity in  $x$  adequately.

We present an uncertainty principle which explains why a pair of orthonormal bases (like Dirac and Fourier basis) cannot have sparse representation of the same signal simultaneously.

We then demonstrate that a combination of two orthonormal bases can be quite useful in creating a redundant yet sparse representation of a larger class of signals which could not be sparsely represented in either of the two bases individually.

This motivates us to discuss more general over-complete signal dictionaries in the next section.

### 2.3.1. Orthonormal bases and orthogonal transforms

In DSP, we often convert a finite length time domain signal into a different domain using finite length transforms. Some of the most common transforms are *discrete Fourier transform*, the discrete cosine transform, and the *Haar transform*. They all belong to the class of transforms called orthogonal transforms.

Orthogonal transforms are characterized by a pair of equations

$$x = \Psi\alpha \quad (2.3.1)$$

and

$$\alpha = \Psi^H x \quad (2.3.2)$$

where  $\Psi$  is an **orthonormal basis** for the complex vector space  $\mathbb{C}^N$ . In particular, the columns of  $\Psi$  are unit norm, and orthogonal to each other. Thus if we write

$$\Psi = [\psi_1 \ \psi_2 \ \dots \ \psi_N]$$

then

$$\langle \psi_i, \psi_j \rangle = \delta(i - j).$$

In other way:

$$\Psi^{-1} = \Psi^H.$$

Eq. (2.3.1) is known as the synthesis equation ( $x$  is synthesized by columns of  $\Psi$ ). Eq. (2.3.2) is known as the analysis equation as we compute the coefficients in  $\alpha$  by taking the inner product of  $x$  with columns of  $\Psi$ .

$\Psi$  is known as synthesis operator while  $\Psi^H$  is known as analysis operator.

Orthogonal transforms preserve the norm of the signal:

$$\|x\|_2^2 = \|\alpha\|_2^2. \quad (2.3.3)$$

This result is commonly known as **Parseval's identity** in signal processing community.

More generally, orthogonal transforms preserve inner products

$$\langle x, y \rangle = \langle \Psi x, \Psi y \rangle \quad \forall x, y \in \mathbb{C}^N. \quad (2.3.4)$$

**Example 2.6: Dirac basis and sparse signals** The simplest orthogonal transform is the identity basis or the standard ordered basis for  $\mathbb{C}^N$ .

$$\Psi = I_N.$$

$$\Psi^{-1} = \Psi^H = I_N^H = I_N.$$

In this basis

$$x = I_N \alpha = \alpha.$$

We will drop  $N$  from suffix for convenience and refer to the matrix as  $I$  only.

This basis is also known as **Dirac basis**. The name **Dirac** comes from the Dirac delta functions used in signal analysis in continuous time domain.

The basis consists of finite length impulses denoted by  $e_i$  where

$$e_1 = (1, 0, \dots, 0),$$

$$e_2 = (0, 1, \dots, 0),$$

⋮

$$e_N = (0, 0, \dots, 1)$$

If a signal  $x$  consists of a linear combination of few  $K \ll N$  impulses, then its a sparse signal in this basis. For example the signal

$$x = (3, 4, 0, 0, -2, 0, 0, \dots, 0, 0, 0)$$

is 3-sparse in Dirac basis since

$$x = 3e_1 + 4e_2 - 2e_5$$

can be expressed as a linear combination of just 3 impulses.

In contrast if we consider a complex sinusoid in  $\mathbb{C}^N$ , there is no way we can find a sparse representation for it in the Dirac basis.  $\square$

### 2.3.2. Fourier basis and sparse signals

The most popular finite length orthogonal transform is DFT (Discrete Fourier Transform).

We define the  $N$ -th root of unity as

$$\omega = \exp\left(\frac{j2\pi}{N}\right). \quad (2.3.5)$$

Clearly

$$\omega^N = \exp(j2\pi) = 1.$$

We define the synthesis matrix of DFT as

$$\Psi = F_N = \frac{1}{\sqrt{N}} \begin{bmatrix} \omega^{kn} \end{bmatrix} \quad \forall 0 \leq k \leq N-1, 0 \leq n \leq N-1 \quad (2.3.6)$$

where

$$\omega^{kn} = \exp\left(\frac{j2\pi kn}{N}\right).$$

$k$  iterates over rows of  $F_N$  while  $n$  iterates over columns of  $F_N$ . The definition is actually symmetric. Hence

$$F_N = F_N^T.$$

Note that we have multiplied with  $\frac{1}{\sqrt{N}}$  to make sure that columns of  $F_N$  are unit norm.

The columns of  $F_N$  forms the **Fourier basis** for signals in  $\mathbb{C}^N$ .

**Example 2.7: Fourier basis for  $N = 2$**  2nd root of unity is given by

$$\omega = \exp(j\pi) = -1.$$

$$F_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} \omega^0 & \omega^0 \\ \omega^0 & \omega^1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

In this case

$$F_2^H = F_2.$$

□

**Example 2.8:** Fourier basis for  $N = 3$  3rd root of unity is given by

$$\omega = \exp\left(\frac{j2\pi}{3}\right) = -0.5 + 0.866j.$$

$$F_3 = \frac{1}{\sqrt{3}} \begin{bmatrix} \omega^0 & \omega^0 & \omega^0 \\ \omega^0 & \omega^1 & \omega^2 \\ \omega^0 & \omega^2 & \omega^4 \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -0.5 + 0.866j & -0.5 - 0.866j \\ 1 & -0.5 - 0.866j & -0.5 + 0.866j \end{bmatrix}$$

In this case

$$F_3^H = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -0.5 - 0.866j & -0.5 + 0.866j \\ 1 & -0.5 + 0.866j & -0.5 - 0.866j \end{bmatrix}$$

□

**Example 2.9:** Fourier basis for  $N = 4$  4th root of unity is given by

$$\omega = \exp\left(\frac{j2\pi}{4}\right) = j.$$

$$F_4 = \frac{1}{2} \begin{bmatrix} \omega^0 & \omega^0 & \omega^0 & \omega^0 \\ \omega^0 & \omega^1 & \omega^2 & \omega^3 \\ \omega^0 & \omega^2 & \omega^4 & \omega^6 \\ \omega^0 & \omega^3 & \omega^6 & \omega^9 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \omega^0 & \omega^0 & \omega^0 & \omega^0 \\ \omega^0 & \omega^1 & \omega^2 & \omega^3 \\ \omega^0 & \omega^2 & 1 & \omega^2 \\ \omega^0 & \omega^3 & \omega^2 & \omega^1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & -1 \\ 1 & -j & -1 & j \end{bmatrix}$$

In this case

$$\mathbf{F}_4^H = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & j \\ 1 & -1 & 1 & -1 \\ 1 & j & -1 & -j \end{bmatrix}$$

□

We drop the suffix  $N$  wherever convenient and simply refer to the synthesis matrix as  $F$ .

If a signal  $x$  is a linear combination of only a few ( $K \ll N$ ) complex sinusoids, then  $x$  has a sparse representation in the DFT basis  $F$ .

**Example 2.10: Sparse signals in  $\mathbf{F}_4$**  Consider the following signal

$$x = (-0.5 \quad 0.5 - j \quad 1.5 \quad 0.5 + j)$$

Its representation in  $\mathbf{F}_4$  is given by

$$\alpha = \mathbf{F}_4^H x = (1 \quad -2 \quad 0 \quad 0).$$

Clearly the signal is 2-sparse in  $\mathbf{F}_4$ .

Now consider a signal  $e_2$  which is sparse in the Dirac basis

$$e_2 = (0 \quad 1 \quad 0 \quad 0).$$

Its representation in  $\mathbf{F}_4$  is

$$\alpha = \mathbf{F}_4^H e_2 = (0.5 \quad 0.5j \quad -0.5 \quad 0.5j).$$

Thus we see that while  $e_2$  is sparse in  $I_4$ , it is not at all sparse in  $\mathbf{F}_4$ .

□

### 2.3.3. An uncertainty principle

As we noted, Dirac basis can give sparse representations for impulses but not for complex sinusoids. Vice versa Fourier basis can give sparse representations for complex sinusoids but not impulses.

Can we claim that a signal cannot be simultaneously represented both in time (Dirac basis) and in frequency domain (Fourier basis)?

More generally, let  $\Psi$  and  $\mathcal{X}$  be any two arbitrary orthonormal bases for  $\mathbb{C}^N$ .

For some  $x \in \mathbb{C}^N$  Let

$$x = \Psi\alpha = \mathcal{X}\beta \quad (2.3.7)$$

where  $\alpha$  and  $\beta$  are representations of  $x$  in  $\Psi$  and  $\mathcal{X}$  respectively. Can we claim something for the relationship between sparsity levels  $\|\alpha\|_0$  and  $\|\beta\|_0$ ?

The answer turns out to be yes, but it depends on how much the two bases are similar or close to each other. The results in this section were originally developed in [22].

**Definition 2.1** The **proximity** between two orthonormal bases  $\Psi$  and  $\mathcal{X}$  where

$$\Psi = [\psi_1 \ \dots \ \psi_N]$$

and

$$\mathcal{X} = [\chi_1 \ \dots \ \chi_N]$$

is defined as the maximum absolute value of inner products between the columns of these two bases:

$$\mu(\Psi, \mathcal{X}) = \max_{1 \leq i, j \leq N} |\langle \psi_i, \chi_j \rangle|. \quad (2.3.8)$$

This is also known as **mutual coherence** of the two orthonormal bases.

If the two bases are identical, then clearly  $\mu = 1$ . If any vector in  $\Psi$  is very close to some vector in  $\mathcal{X}$  then we will have a very high value of  $\mu$  close to 1.

If the vectors in  $\Psi$  and  $\mathcal{X}$  are highly dissimilar, then we will have very low value of  $\mu$  close to 0.

**Example 2.11: Proximity or mutual coherence of Dirac and Fourier bases** As an example, we consider the mutual coherence between Dirac and Fourier bases.

For some small values of  $N$  (the dimension of ambient space  $\mathbb{C}^N$ ) the values are tabulated in table 1.

TABLE 1. Mutual coherence of Dirac and Fourier bases

$N$	$\mu$
2	0.7071
4	0.5000
6	0.4082
8	0.3536

For larger values of  $N$  we can see the variation of  $\mu$  in fig. 2.4.  $\square$

We present some results related to mutual coherence of two orthonormal bases.

**Theorem 2.2** *The product of two orthonormal bases  $\Psi$  and  $\mathcal{X}$  for  $\mathbb{C}^N$  given by  $\Psi^H \mathcal{X}$  forms an orthonormal basis by itself.*

PROOF. Consider the matrix  $\Psi^H \mathcal{X}$

$$\Psi^H \mathcal{X} = \begin{bmatrix} \psi_1^H \\ \vdots \\ \psi_N^H \end{bmatrix} \begin{bmatrix} \chi_1 & \dots & \chi_N \end{bmatrix} = \begin{bmatrix} \psi_1^H \chi_1 & \psi_1^H \chi_2 & \dots & \psi_1^H \chi_N \\ \psi_2^H \chi_1 & \psi_2^H \chi_2 & \dots & \psi_2^H \chi_N \\ \vdots & \vdots & \ddots & \vdots \\ \psi_N^H \chi_1 & \psi_N^H \chi_2 & \dots & \psi_N^H \chi_N \end{bmatrix}$$

Any column of the product matrix is

$$\begin{bmatrix} \psi_1^H \chi_i \\ \psi_2^H \chi_i \\ \vdots \\ \psi_N^H \chi_i \end{bmatrix} = \Psi^H \chi_i$$

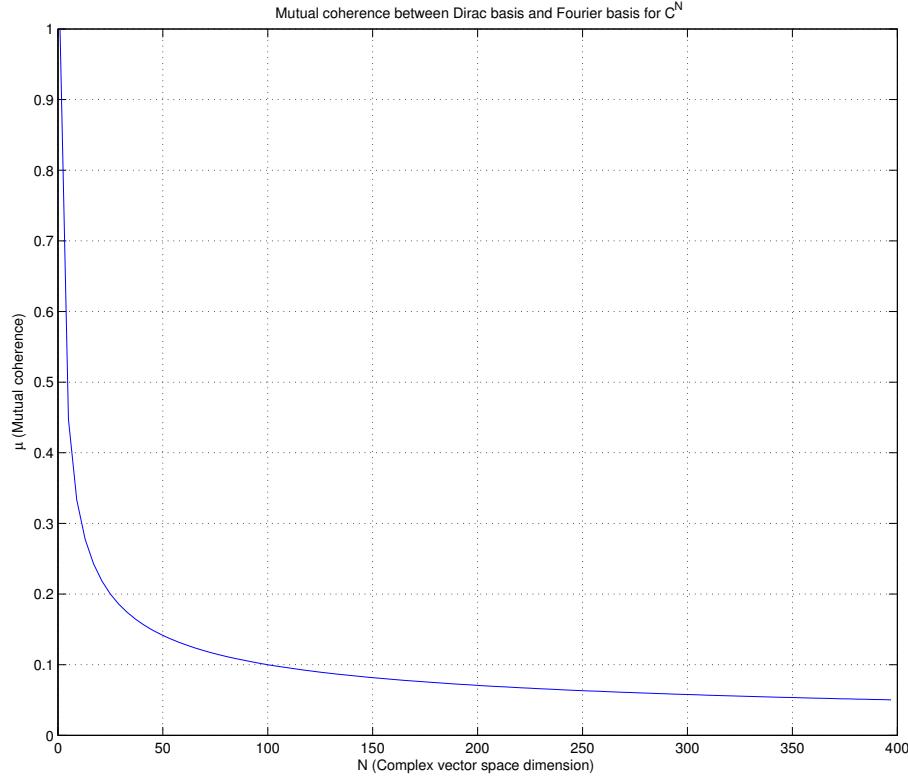


FIGURE 2.4. Mutual coherence for Dirac and Fourier bases

But then  $\Psi$  preserves norms, hence

$$\|\Psi^H \chi_i\|_2 = \|\chi_i\|_2 = 1.$$

Thus each column of the product  $\Psi^H \mathcal{X}$  is itself a unit norm vector.

Consider the inner product of two columns of  $\Psi^H \mathcal{X}$

$$\langle \Psi^H \chi_i, \Psi^H \chi_j \rangle = \chi_j^H \Psi \Psi^H \chi_i = \chi_j^H \chi_i = \delta(i - j).$$

Thus, the columns of  $\Psi^H \mathcal{X}$  are orthogonal to each other.

Hence  $\Psi^H \mathcal{X}$  forms an orthonormal basis. □

REMARK. A more general result would be to show that the set of orthonormal bases forms a group under the matrix multiplication operation. The identity element is the Dirac basis. The inverse of an orthonormal basis is also an orthonormal basis. The product of two

orthonormal bases is also an orthonormal basis. The matrix multiplication satisfies associative law.

**Theorem 2.3** *Mutual coherence of two orthonormal bases  $\Psi$  and  $\mathcal{X}$  for the complex vector space  $\mathbb{C}^N$  is bounded by*

$$\frac{1}{\sqrt{N}} \leq \mu(\Psi, \mathcal{X}) \leq 1. \quad (2.3.9)$$

**PROOF.** Since columns of both  $\Psi$  and  $\mathcal{X}$  are unit norm, hence  $|\langle \psi_i, \chi_j \rangle|$  cannot be greater than 1. Now if  $\Psi = \mathcal{X}$  then we have  $\mu(\Psi, \mathcal{X}) = 1$ . This proves the upper bound.

Now consider the matrix  $\Psi^H \mathcal{X}$  which forms an orthonormal basis by itself.

Consider any column of this matrix

$$\begin{bmatrix} \psi_1^H \chi_i \\ \psi_2^H \chi_i \\ \vdots \\ \psi_N^H \chi_i \end{bmatrix} = \Psi^H \chi_i$$

Since the column is unit norm, hence some of squares of the absolute values of entries of the column is 1. Thus the absolute value of each of the entries cannot be simultaneously less than  $\frac{1}{\sqrt{N}}$ .

Hence there exists an entry (in each column) such that

$$|\psi_j^H \chi_i| \geq \frac{1}{\sqrt{N}}.$$

Hence we get the lower bound on mutual coherence of  $\Psi$  and  $\mathcal{X}$  given by

$$\mu(\Psi, \mathcal{X}) \geq \frac{1}{\sqrt{N}}.$$

□

**Theorem 2.4** Mutual coherence of Dirac and Fourier bases is  $\frac{1}{\sqrt{N}}$ .  
i.e.

$$\mu(I, F) = \frac{1}{\sqrt{N}}. \quad (2.3.10)$$

PROOF. theorem 2.3 shows that

$$\mu(I, F) \geq \frac{1}{\sqrt{N}}.$$

We just need to show that its in fact an equality.

Consider  $i$ -th column of I as  $e_i$ .

Consider  $j$ -th column of F:

$$f_j = \frac{1}{\sqrt{N}} \begin{bmatrix} \omega^0 \\ \omega^j \\ \omega^{2j} \\ \vdots \\ \omega^{(N-1)j} \end{bmatrix}$$

where  $\omega$  is the  $N$ -th root of unity. Then

$$\langle e_i, f_j \rangle = \frac{1}{\sqrt{N}} \omega^{-(i-1)j} \implies |\langle e_i, f_j \rangle| = \frac{1}{\sqrt{N}}.$$

This doesn't depend on the choice of  $i$ -th column of I and  $j$ -th column of F.

Hence

$$\mu(I, F) = \frac{1}{\sqrt{N}}.$$

□

With basic properties of mutual coherence in place, we are now ready to state an uncertainty principle on the sparsity levels of representations of same signal  $x$  in two different orthonormal bases:

**Theorem 2.5** For any arbitrary pair of orthonormal bases  $\Psi, \mathcal{X}$  with mutual coherence  $\mu(\Psi, \mathcal{X})$ , and for any arbitrary non-zero

vector  $x \in \mathbb{C}^N$  with representations  $\alpha$  and  $\beta$  correspondingly, the following inequality holds true:

$$\|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu(\Psi, \mathcal{X})}. \quad (2.3.11)$$

Moreover for unit-length  $x$

$$\|\alpha\|_1 + \|\beta\|_1 \geq \frac{2}{\sqrt{\mu(\Psi, \mathcal{X})}}. \quad (2.3.12)$$

PROOF. Dividing  $x$  by  $\|x\|_2$  doesn't change  $l_0$  "norm" of  $\alpha$  and  $\beta$ . i.e.

$$\left\| \frac{\alpha}{\|x\|_2} \right\|_0 = \|\alpha\|_0.$$

Hence without loss of generality, we will assume that  $\|x\|_2 = 1$ .

We are given that  $x^H x = 1$ ,  $x = \Psi \alpha$  and  $x = \mathcal{X} \beta$ . Since  $\Psi$  and  $\mathcal{X}$  are orthonormal bases hence  $\|\alpha\|_2 = \|\beta\|_2 = 1$  also.

We can write as

$$\begin{aligned} 1 &= x^H x \\ &= (\Psi \alpha)^H (\mathcal{X} \beta) = \alpha^H \Psi^H \mathcal{X} \beta \\ &= \sum_{i=1}^N \sum_{j=1}^N \overline{\alpha_i} \beta_j \psi_i^H \chi_j \\ &= \left| \sum_{i=1}^N \sum_{j=1}^N \overline{\alpha_i} \beta_j \psi_i^H \chi_j \right| \\ &\leq \sum_{i=1}^N \sum_{j=1}^N |\alpha_i| |\beta_j| |\psi_i^H \chi_j| \\ &\leq \mu(\Psi, \mathcal{X}) \sum_{i=1}^N \sum_{j=1}^N |\alpha_i| |\beta_j| \\ &= \mu(\Psi, \mathcal{X}) \|\alpha\|_1 \|\beta\|_1 \end{aligned}$$

where we note that

$$\|\alpha\|_1 \|\beta\|_1 = \left( \sum_{i=1}^N |\alpha_i| \right) \left( \sum_{j=1}^N |\beta_j| \right) = \sum_{i=1}^N \sum_{j=1}^N |\alpha_i| |\beta_j|$$

and

$$\mu(\Psi, \mathcal{X}) \geq |\psi_i^H \chi_j|.$$

Hence we get the inequality

$$\mu(\Psi, \mathcal{X}) \|\alpha\|_1 \|\beta\|_1 \geq 1 \quad (2.3.13)$$

Using the inequality between algebraic mean and geometric mean

$$\sqrt{ab} \leq \frac{a+b}{2} \quad \forall a, b \geq 0$$

we get

$$\|\alpha\|_1 + \|\beta\|_1 \geq 2\sqrt{\|\alpha\|_1 \|\beta\|_1} \geq \frac{2}{\sqrt{\mu(\Psi, \mathcal{X})}}.$$

This is an uncertainty principle for the  $l_1$  norms of the two representations. We still have to get the uncertainty principle for  $l_0$  case.

We assume that

$$\|\alpha\|_0 = A$$

and

$$\|\beta\|_0 = B$$

Consider the sets

$$X_A = \{v : v = \Psi a \text{ and } \|a\|_0 = A, \|a\|_2 = 1\}$$

and

$$X_B = \{v : v = \mathcal{X} b \text{ and } \|b\|_0 = B, \|b\|_2 = 1\}.$$

Clearly

$$x \in X_A \cap X_B.$$

The representations  $a$  for vectors  $v$  in  $X_A$  have exactly  $A$  non-zero entries and are all unit norm representations. Which of them would have the longest  $l_1$  norm  $\|a\|_1$ ?

This can be written as an optimization problem of the form

$$\underset{v \in X_A}{\text{maximize}} \|a\|_1 \text{ where } a = \Psi^H v. \quad (2.3.14)$$

Let the optimal solution for this problem be  $v_a$  with corresponding representation  $a^* = \Psi^H v_a$ . Clearly

$$\|a^*\|_1 \geq \|\alpha\|_1.$$

Similarly from the set  $X_B$  let us find the vector  $v_b$  with maximum  $l_1$  norm representation in  $\mathcal{X}$

$$\underset{v \in X_B}{\text{maximize}} \|b\|_1 \text{ where } b = \mathcal{X}^H v. \quad (2.3.15)$$

Let the optimal solution for this problem be  $v_b$  with corresponding representation  $b^* = \mathcal{X}^H v_b$ . Clearly

$$\|b^*\|_1 \geq \|\beta\|_1.$$

Returning back to the inequality

$$\|\alpha\|_1 \|\beta\|_1 \geq \frac{1}{\mu(\Psi, \mathcal{X})}$$

we can write

$$\|a^*\|_1 \|b^*\|_1 \geq \frac{1}{\mu(\Psi, \mathcal{X})}.$$

An equivalent formulation of the optimization problem (2.3.14) is

$$\begin{aligned} & \underset{a}{\text{maximize}} \quad \|a\|_1 \\ & \text{subject to} \quad \|a\|_2^2 = a^H a = 1 \\ & \text{and} \quad \|a\|_0 = A. \end{aligned} \quad (2.3.16)$$

This formulation doesn't require any specific mention of the basis  $\Psi$ .

Let the optimal value for this problem be given by

$$\|a^*\|_1 = g(A) = g(\|\alpha\|_0)$$

Here we consider the optimization problem to be parameterized by the  $l_0$ -“norm” of  $a$ , i.e.  $\|a\|_0 = A$  and we write the optimal value as a function  $g$  of the parameter  $A$ .

Then by symmetry, optimal value for the problem (2.3.15) is

$$\|b^*\|_1 = g(B) = g(\|\beta\|_0)$$

Thus we can write

$$g(\|\alpha\|_0)g(\|\beta\|_0) \geq \frac{1}{\mu(\Psi, \mathcal{X})}. \quad (2.3.17)$$

This is our intended result since we have been able to write the inequality as a function of  $l_0$  “norm”s of the representations  $\alpha$  and  $\beta$ .

In order to complete the result, we need to find the solution of the optimization problem (2.3.16) given by the function  $g$ .

Without loss of generality, let us assume that the  $A$  non-zero entries in the optimization variable  $a$  appear in its first  $A$  entries and rest are zero. This is fine since changing the order of entries in  $a$  doesn’t affect any of the norms of concern  $\|a\|_0$ ,  $\|a\|_1$  and  $\|a\|_2$ .

Let us further assume that all non-zero entries of  $a$  are strictly positive real numbers. This assumption is valid since only absolute values are used in this problem. Specifically for any  $a$  with non-zero complex entries  $(a_1, a_2, \dots, a_A, 0, \dots, 0)$  there exists  $a'$  with positive entries  $(|a_1|, |a_2|, \dots, |a_A|, 0, \dots, 0)$  such that  $\|a\|_0 = \|a'\|_0$ ,  $\|a\|_1 = \|a'\|_1$  and  $\|a\|_2 = \|a'\|_2$ , hence solving the optimization problem for complex  $a$  is same as solving the optimization problem for  $a$  with strictly positive first  $A$  entries.

Using Lagrange multipliers, the  $l_0$  constraint vanishes (since the assumptions mentioned above allow us to focus on only the first  $A$  coordinates of  $a$ ), and we obtain

$$\mathcal{L}(a) = \sum_{i=1}^A a_i + \lambda \left( 1 - \sum_{i=1}^A a_i^2 \right). \quad (2.3.18)$$

Differentiating w.r.t.  $a_i$  and equating to 0 we get

$$1 - 2\lambda a_i = 0 \implies a_i = \frac{1}{2\lambda}. \quad (2.3.19)$$

The  $l_2$  constraint requires

$$\sum_{i=1}^A a_i^2 = A \frac{1}{4\lambda^2} = 1 \implies \lambda = \frac{\sqrt{A}}{2}$$

Thus

$$a_i = \frac{1}{\sqrt{A}}$$

and

$$\|\alpha\|_1 = \sum_{i=1}^A |a_i| = \sqrt{A}.$$

Thus the optimal value of the optimization problem (2.3.16) is

$$g(A) = \sqrt{A} = \sqrt{\|\alpha\|_0}.$$

Similarly

$$g(B) = \sqrt{B} = \sqrt{\|\beta\|_0}.$$

Putting back in (2.3.20) we get

$$\sqrt{\|\alpha\|_0 \|\beta\|_0} \geq \frac{1}{\mu(\Psi, \mathcal{X})}. \quad (2.3.20)$$

Applying the algebraic mean-geometric mean inequality we get the desired result

$$\|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu(\Psi, \mathcal{X})}. \quad (2.3.21)$$

□

This theorem suggests that if two orthonormal bases have low mutual coherence then

- the two representations for  $x$  cannot be jointly  $l_1$ -short and
- the two representations for  $x$  cannot be jointly sparse.

**Challenge** Can we show that the above result is sharp? i.e. For a pair of orthonormal bases  $\Psi$  and  $\mathcal{X}$ , it is always possible to find a non-zero vector  $x$  with corresponding representations  $x = \Psi\alpha$  and

$x = \mathcal{X}\beta$  which satisfies the lower bound

$$\|\alpha\|_0 + \|\beta\|_0 = \frac{2}{\mu(\Psi, \mathcal{X})} ? \quad (2.3.22)$$

**Example 2.12: Sparse representations with Dirac and Fourier bases** We showed in theorem 2.4 that

$$\mu(I, F) = \frac{1}{\sqrt{N}}.$$

Let  $x \in \mathbb{C}^N$ . Let its representation in  $F$  be given by

$$x = F\alpha.$$

Applying theorem 2.5 we have

$$\|x\|_0 + \|\alpha\|_0 \geq \frac{2}{\mu(\Psi, \mathcal{X})} = 2\sqrt{N}.$$

This tells us that a signal cannot have fewer than  $2\sqrt{N}$  non-zeros in both time and frequency domain together.

□

### 2.3.4. Linear combinations of impulses and sinusoids

What happens if a signal  $x$  is a linear combination of few complex sinusoids and few impulses?

The set of sinusoids and impulses involved in the construction of  $x$  actually specifies the degrees of freedom of  $x$ . This is the indicator of inherent sparsity of  $x$  provided this set of component signals of  $x$  is known a-priori.

In absence of prior knowledge of component signals of  $x$ , we attempt to look for a sparse representation of  $x$  in one of the well understood orthonormal bases. Here we are specifically looking at the two bases Dirac and Fourier.

While the Dirac basis can provide sparse representation for impulses, sinusoids have dense representation in Dirac basis. Vice versa, in Fourier

basis, complex sinusoids have sparse representation, yet impulses have dense representations. Thus neither of the two bases is capable of providing a sparse representation for a combination of impulses and sinusoids.

The natural question arises if there is a way to come up with a sparse representation for such signals by combining the Dirac and Fourier basis?

### 2.3.5. Dirac Fourier basis

Now we develop a representation of signals  $x \in \mathbb{C}^N$  in terms of a combination of Dirac basis I and Fourier basis F.

We define a new synthesis matrix

$$\mathcal{H} = [I \quad F] \in \mathbb{C}^{N \times 2N}. \quad (\text{DF})$$

We can write I as

$$I = [e_1 \quad \dots \quad e_N].$$

Let us write F as

$$F = [f_1 \quad \dots \quad f_N]$$

This enables us to write  $\mathcal{H}$  as

$$\mathcal{H} = [e_1 \quad \dots \quad e_N \quad f_1 \quad \dots \quad f_N]$$

We will look for a representation of  $x$  using the synthesis matrix  $\mathcal{H}$  as

$$x = \mathcal{H}\alpha \quad (2.3.23)$$

where  $\alpha \in \mathbb{C}^{2N}$ .

Since this representation is under-determined and  $\mathcal{C}(\mathcal{H}) = \mathbb{C}^N$  hence there are always infinitely many possible representations of  $x$  in  $\mathcal{H}$ .

We would prefer to choose the sparsest representation which can be stated as an optimization problem

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} \quad \|\alpha\|_0 \\ & \text{subject to} \quad x = \mathcal{H}\alpha. \end{aligned} \tag{P}_0$$

**Example 2.13: Sparse representation using Dirac Fourier Basis** Consider  $N = 4$ .

Then the Dirac Fourier basis is

$$\mathcal{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & .5 & .5 & .5 & .5 \\ 0 & 1 & 0 & 0 & .5 & .5j & -.5 & -.5j \\ 0 & 0 & 1 & 0 & .5 & -.5 & .5 & -.5 \\ 0 & 0 & 0 & 1 & .5 & -.5j & -.5 & .5j \end{bmatrix}$$

Let

$$x = 3e_1 - 2f_2 = \begin{pmatrix} 2 & -j & 1 & j \end{pmatrix}.$$

A sparse representation of  $x$  in  $\mathcal{H}$  is

$$\alpha = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & -2 & 0 & 0 \end{pmatrix}.$$

This representation is 2-sparse.

Thus we see that Dirac Fourier basis is able to provide a sparser representation of a linear combination of impulses and sinusoids compared to individual orthonormal bases (the Dirac basis and the Fourier basis).

□

This gives us motivation to consider such combination of bases which help us provide a sparse representation to a larger class of signals. This is the objective of the next section.

In due course we will revisit the Dirac Fourier basis further in several examples.

### 2.3.6. Two-ortho basis

Before we leave this section, let us define general two-ortho bases.

**Definition 2.2** Let  $\Psi$  and  $\mathcal{X}$  be any two  $\mathbb{C}^{N \times N}$  matrices whose columns vectors form orthonormal bases for the complex vector space  $\mathbb{C}^N$  individually.

We define

$$\mathcal{H} = [\Psi \quad \mathcal{X}] \in \mathbb{C}^{N \times 2N}. \quad (2.3.24)$$

Then columns of  $\mathcal{H}$  form a **two-ortho basis** for  $\mathbb{C}^N$ .

The mutual coherence of a two ortho basis is defined as

$$\mu(\mathcal{H}) = \mu(\Psi, \mathcal{X}).$$

Clearly columns of  $\mathcal{H}$  span  $\mathbb{C}^N$ .

We present a very interesting result about the null space of  $\mathcal{H}$ .

**Theorem 2.6** For a two ortho basis  $\mathcal{H} = [\Psi \quad \mathcal{X}]$  with low coherence, the non-zero vectors in the null space of  $\mathcal{H}$  are not sparse.

Concretely

$$\|v\|_0 \geq \frac{2}{\mu(\mathcal{H})} \quad \forall v \in \mathcal{N}(\mathcal{H}). \quad (2.3.25)$$

PROOF. Let  $v \in \mathcal{N}(\mathcal{H})$  be some non-zero vector.

Now let  $v_\psi$  and  $v_\chi$  be first  $N$  and last  $N$  entries of  $v$ . Then

$$[\Psi \quad \mathcal{X}] \begin{bmatrix} v_\psi \\ v_\chi \end{bmatrix} = 0 \implies \Psi v_\psi = -\mathcal{X} v_\chi = y \neq 0.$$

If  $y$  were 0 then, both  $v_\psi$  and  $v_\chi$  would have to be 0 which is not the case since by assumption  $v \neq 0$ .

We note that  $v_\psi$  and  $-v_\chi$  are representations of same vector  $y$  in two different orthonormal bases  $\Psi$  and  $\mathcal{X}$  respectively, and  $\|-v_\chi\|_0 = \|v_\chi\|_0$ .

Applying theorem 2.5 we have

$$\|v\|_0 \geq \|v_\psi\|_0 + \|v_\chi\|_0 \geq \frac{2}{\mu(\mathcal{H})}. \quad (2.3.26)$$

We also note that since the orthonormal bases preserve norm, hence

$$\|y\|_2 = \|v_\psi\|_2 = \|v_\chi\|_2.$$

This shows us that the energy of a null space vector is evenly distributed in the two components corresponding to each orthonormal basis.  $\square$

**Challenge** For a two-ortho basis  $\mathcal{H} = [\Psi \ \mathcal{X}]$  is it possible to find a null space vector  $v$  satisfying the lower bound

$$\|v\|_0 = \frac{2}{\mu(\mathcal{H})} ? \quad (2.3.27)$$

Let  $x \in \mathbb{C}^N$  be any arbitrary signal. Then its representation in  $\mathcal{H}$  is given by

$$x = \mathcal{H}\alpha. \quad (2.3.28)$$

Obviously for all  $z \in \mathcal{N}(\mathcal{H})$  ( $\alpha + z$ ) is also a representation of  $x$ .

What we are particularly interested in are sparse representations of  $x$ . A major concern for us is to ensure that a sparse representation of  $x$  in  $\mathcal{H}$  is unique. Under what conditions such is possible?

Formally, let  $\alpha$  and  $\beta$  be two different representations of  $x$  in  $\mathcal{H}$ . Can we say that if  $\alpha$  is sparse then  $\beta$  won't be sparse?

This is established in the next uncertainty principle.

**Theorem 2.7** Let  $x \in \mathbb{C}^N$  be any signal and let  $\mathcal{H}$  be a two ortho basis defined in (2.3.24). Let  $\alpha$  and  $\beta$  be two distinct representations of  $x$  in  $\mathcal{H}$  i.e.

$$x = \mathcal{H}\alpha = \mathcal{H}\beta.$$

Then the following holds

$$\|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu(\mathcal{H})}. \quad (2.3.29)$$

This is an uncertainty principle for the sparsity of distinct representations in two ortho basis.

PROOF. Let

$$e = \alpha - \beta$$

be the difference vector of representations of  $x$ .

Clearly

$$\mathcal{H}e = \mathcal{H}\alpha - \mathcal{H}\beta = x - x = 0.$$

Thus  $e \in \mathcal{N}(\mathcal{H})$ . Applying theorem 2.6, we get

$$\|e\|_0 \geq \frac{2}{\mu(\mathcal{H})}.$$

But since  $e = \alpha - \beta$  hence we have

$$\|\alpha\|_0 + \|\beta\|_0 \geq \|e\|_0 \geq \frac{2}{\mu(\mathcal{H})}.$$

□

**Challenge** For a two-ortho basis  $\mathcal{H} = [\Psi \quad \mathcal{X}]$  is it possible to find a vector  $x$  with two alternative representations  $\alpha$  and  $\beta$  satisfying the lower bound

$$\|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu(\mathcal{H})}? \quad (2.3.30)$$

This theorem suggests as that if  $\mu(\mathcal{H})$  is small (i.e. the coherence between the two orthonormal bases is small) then two representations of  $x$  cannot be simultaneously sparse.

Rather if a representation is sufficiently sparse, then all other representations of  $x$  are guaranteed to be non-sparse providing the uniqueness of sparse representation.

This is stated formally in the following **uniqueness** theorem.

**Theorem 2.8** *If a representation of  $x$  in the two ortho basis  $\mathcal{H} = [\Psi \ \mathcal{X}]$  has fewer than  $\frac{1}{\mu(\mathcal{H})}$  non-zeros, then it is necessarily the sparsest one possible, and any other representation must be denser.*

PROOF. Let  $\alpha$  be a candidate representation with

$$\|\alpha\|_0 < \frac{1}{\mu(\mathcal{H})}.$$

Let  $\beta$  be any other candidate representation. By applying theorem 2.7 we have

$$\|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu(\mathcal{H})}.$$

This gives us

$$\|\beta\|_0 \geq \frac{2}{\mu(\mathcal{H})} - \|\alpha\|_0 \implies \|\beta\|_0 > \frac{1}{\mu(\mathcal{H})}.$$

Thus we find that

$$\|\beta\|_0 > \|\alpha\|_0$$

which is true for every representation  $\beta$  of  $x$  in  $\mathcal{H}$  other than  $\alpha$ .

Hence  $\alpha$  is the sparsest possible representation of  $x$  in  $\mathcal{H}$ .  $\square$

We note here that any arbitrary choice of two bases may not be helpful in coming up with a two ortho basis which can provide us sparse representations for our signals of interest. In next few sections, we will explore this issue further in the more general context of signal dictionaries.

**Challenge** So there are signals for which a sufficiently sparse (and unique) representation doesn't exist in a given two-ortho basis. What kind relationships may exist between different (insufficiently) sparse representations of such signals?

## 2.4. Sparse and redundant representations

### 2.4.1. Dictionaries

**Definition 2.3** A **dictionary** for  $\mathbb{C}^N$  is a finite collection  $\mathcal{D}$  of unit-norm vectors which span the whole space.

The elements of a dictionary are called **atoms** and they are denoted by  $\phi_\omega$  where  $\omega$  is drawn from an index set  $\Omega$ .

The whole dictionary structure is written as

$$\mathcal{D} = \{\phi_\omega : \omega \in \Omega\} \quad (2.4.1)$$

where

$$\|\phi_\omega\|_2 = 1 \quad \forall \omega \in \Omega$$

and

$$x = \sum_{\omega \in \Omega} c_\omega \phi_\omega \quad \forall x \in \mathbb{C}^N.$$

We use the letter  $D$  to denote the number of elements in the dictionary, i.e.

$$D = |\Omega|.$$

This definition is adapted from [34].

The indices may have an interpretation, such as the time-frequency or time-scale localization of an atom, or they may simply be labels without any underlying meaning.

Note that the dictionary need not provide a unique representation for any vector  $x \in \mathbb{C}^N$ , but it provides at least one representation for each  $x \in \mathbb{C}^N$ .

When  $D = N$  we have a set of unit norm vectors which span the whole of  $\mathbb{C}^N$ . Thus we have a basis (not-necessarily an orthonormal basis). A dictionary cannot have  $D < N$ . The more interesting case is when  $D > N$ .

### 2.4.2. Redundant dictionaries and sparse signals

With  $D > N$ , clearly there are more atoms than necessary to provide a representation of signal  $x \in \mathbb{C}^N$ . Thus such a dictionary is able to provide multiple representations to same vector  $x$ . We call such dictionaries **redundant dictionaries** or **over-complete dictionaries**.

In contrast a basis with  $D = N$  is called a **complete dictionary**.

A special class of signals is those signals which have a sparse representation in a given dictionary  $\mathcal{D}$ .

**Definition 2.4** A signal  $x \in \mathbb{C}^N$  is called  $(\mathcal{D}, K)$ -sparse if it can be expressed as a linear combination of at-most  $K$  atoms from the dictionary  $\mathcal{D}$  where  $K \ll D$ .

It is usually expected that  $K \ll N$  also holds.

Let  $\Lambda \subset \Omega$  be a subset of indices with  $|\Lambda| = K$ .

Let  $x$  be any signal in  $\mathbb{C}^N$  such that  $x$  can be expressed as

$$x = \sum_{\lambda \in \Lambda} b_\lambda \phi_\lambda \quad \text{where } b_\lambda \in \mathbb{C}. \quad (2.4.2)$$

Note that this is not the only possible representation of  $x$  in  $\mathcal{D}$ . This is just one of the possible representations of  $x$ . The special thing about this representation is that it is  $K$ -sparse i.e. only at most  $K$  atoms from the dictionary are being used.

Now there are  $\binom{D}{K}$  ways in which we can choose a set of  $K$  atoms from the dictionary  $\mathcal{D}$ .

Thus the set of  $(\mathcal{D}, K)$ -sparse signals is given by

$$\Sigma_{(\mathcal{D}, K)} = \{x \in \mathbb{C}^N : x = \sum_{\lambda \in \Lambda} b_\lambda \phi_\lambda\}.$$

for some index set  $\Lambda \subset \Omega$  with  $|\Lambda| = K$ .

This set  $\Sigma_{(\mathcal{D}, K)}$  is dependent on the chosen dictionary  $\mathcal{D}$ . In the sequel, we will simply refer to it as  $\Sigma_K$ .

**Example 2.14:  $K$ -sparse signals for standard basis** For the special case where  $\mathcal{D}$  is nothing but the standard basis of  $\mathbb{C}^N$ , then

$$\Sigma_K = \{x : \|x\|_0 \leq K\}$$

i.e. the set of signals which has  $K$  or less non-zero elements.  $\square$

**Example 2.15:  $K$ -sparse signals for orthonormal basis** In contrast if we choose an orthonormal basis  $\Psi$  such that every  $x \in \mathbb{C}^N$  can be expressed as

$$x = \Psi\alpha$$

then with the dictionary  $\mathcal{D} = \Psi$ , the set of  $K$ -sparse signals is given by

$$\Sigma_K = \{x = \Psi\alpha : \|\alpha\|_0 \leq K\}.$$

$\square$

We also note that set of vectors  $\{\alpha_\lambda : \lambda \in \Lambda\}$  with  $K < N$  form a subspace of  $\mathbb{C}^N$ .

So we have  $\binom{D}{K}$   $K$ -sparse subspaces contained in the dictionary  $\mathcal{D}$ . And the  $K$ -sparse signals lie in the **union of all these subspaces**.

### 2.4.3. Sparse approximation problem

In sparse approximation problem, we attempt to express a given signal  $x \in \mathbb{C}^N$  using a linear combination of  $K$  atoms from the dictionary  $\mathcal{D}$  where  $K \ll N$  and typically  $N \ll D$  i.e. the number of atoms in a dictionary  $\mathcal{D}$  is typically much larger than the ambient signal space dimension  $N$ .

Naturally we wish to obtain a best possible sparse representation of  $x$  over the atoms  $\phi_\omega \in \mathcal{D}$  which minimizes the approximation error.

Let  $\Lambda$  denote the index set of atoms which are used to create a  $K$ -sparse representation of  $x$  where  $\Lambda \subset \Omega$  with  $|\Lambda| = K$ .

Let  $x_\Lambda$  represent an approximation of  $x$  over the set of atoms indexed by  $\Lambda$ .

Then we can write  $x_\Lambda$  as

$$x_\Lambda = \sum_{\lambda \in \Lambda} b_\lambda \phi_\lambda \quad \text{where } b_\lambda \in \mathbb{C}. \quad (2.4.3)$$

We put all complex valued coefficients  $b_\lambda$  in the sum into a list  $b$ .

The approximation error is given by

$$e = \|x - x_\Lambda\|_2. \quad (2.4.4)$$

Clearly we would like to minimize the approximation error over all possible choices of  $K$  atoms and corresponding set of coefficients  $b_\lambda$ .

Thus the sparse approximation problem can be cast as a minimization problem given by

$$\min_{|\Lambda|=K} \min_b \left\| x - \sum_{\lambda \in \Lambda} b_\lambda \phi_\lambda \right\|_2. \quad (2.4.5)$$

If we choose a particular  $\Lambda$ , then the inner minimization problem becomes a straight-forward least squares problem. But there are  $\binom{D}{K}$  possible choices of  $\Lambda$  and solving the inner least squares problem for each of them becomes prohibitively expensive.

We reemphasize here that in this formulation we are using a *fixed* dictionary  $\mathcal{D}$  while the vector  $x \in \mathbb{C}^N$  is *arbitrary*.

This problem is known as  $(\mathcal{D}, K)$ -SPARSE approximation problem.

A related problem is known as  $(\mathcal{D}, K)$ -EXACT-SPARSE problem where it is known a-priori that  $x$  is a linear combination of at-most  $K$  atoms from the given dictionary  $\mathcal{D}$  i.e.  $x$  is a  $K$ -sparse signal as defined in previous section for the dictionary  $\mathcal{D}$ .

This formulation simplifies the minimization problem (2.4.5) since it is known a priori that for  $K$ -sparse signals, a 0 approximation error can be achieved. The only problem is to find a set of subspaces from the  $\binom{D}{K}$  possible  $K$ -sparse subspaces which are able to provide a  $K$ -sparse representation of  $x$  and amongst them choose one. It is imperative to note that even the  $K$ -sparse representation need not be unique.

Clearly the EXACT-SPARSE problem is simpler than the SPARSE approximation problem. Thus if EXACT-SPARSE problem is NP-Hard then so is the harder SPARSE-approximation problem. It is expected that solving the EXACT-SPARSE problem will provide insights into solving the SPARSE problem.

In theorem 2.8 we identified conditions under which a sparse representation for a given vector  $x$  in a two-ortho-basis is unique. It would be useful to get similar conditions for general dictionaries. such conditions would help us guarantee the uniqueness of EXACT-SPARSE problem.

#### 2.4.4. Synthesis and analysis

The atoms of a dictionary  $\mathcal{D}$  can be organized into a  $N \times D$  matrix as follows:

$$\Phi \triangleq \begin{bmatrix} \phi_{\omega_1} & \phi_{\omega_2} & \dots & \phi_{\omega_D} \end{bmatrix}. \quad (2.4.6)$$

where  $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$  is the index set for the atoms of  $\mathcal{D}$ . We remind that  $\phi_\omega \in \mathbb{C}^N$ , hence they have a column vector representation in the standard basis for  $\mathbb{C}^N$ .

The order of columns doesn't matter as long as it remains fixed once chosen.

Thus in matrix terminology a representation of  $x \in \mathbb{C}^N$  in the dictionary can be written as

$$x = \Phi b \quad (2.4.7)$$

where  $b \in \mathbb{C}^D$  is a vector of coefficients to produce a superposition  $x$  from the atoms of dictionary  $\mathcal{D}$ . Clearly with  $D > N$ ,  $b$  is not unique. Rather for every vector  $z \in \mathcal{N}(\Phi)$ , we have:

$$\Phi(b + z) = \Phi b + \Phi z = x + 0 = x.$$

**Definition 2.5** The matrix  $\Phi$  is called a **synthesis matrix** since  $x$  is synthesized from the columns of  $\Phi$  with the coefficient vector  $b$ .

We can also view the synthesis matrix  $\Phi$  as a linear operator from  $\mathbb{C}^D$  to  $\mathbb{C}^N$ .

There is another way to look at  $x$  through  $\Phi$ .

**Definition 2.6** [Analysis matrix] The conjugate transpose  $\Phi^H$  of the synthesis matrix  $\Phi$  is called the **analysis matrix**. It maps a given vector  $x \in \mathbb{C}^N$  to a list of inner products with the dictionary:

$$c = \Phi^H x$$

where  $c \in \mathbb{C}^N$ .

REMARK. Note that in general  $x \neq \Phi(\Phi^H x)$  unless  $\mathcal{D}$  is an orthonormal basis.

**Definition 2.7**  $(\mathcal{D}, K)$ -EXACT-SPARSE) With the help of synthesis matrix  $\Phi$ , the  **$(\mathcal{D}, K)$ -exact-sparse** can now be written as

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} \quad \|\alpha\|_0 \\ & \text{subject to} \quad x = \Phi\alpha \\ & \text{and} \quad \|\alpha\|_0 \leq K \end{aligned} \tag{P}_0$$

**Definition 2.8**  $(\mathcal{D}, K)$ -SPARSE approximation) With the help of synthesis matrix  $\Phi$ , the  **$(\mathcal{D}, K)$ -sparse approximation** can now be written as

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} \quad \|x - \Phi\alpha\|_2 \\ & \text{subject to} \quad \|\alpha\|_0 \leq K. \end{aligned} \tag{P}_0^\epsilon$$

## 2.5. p-norms and sparse signals

### 2.5.1. $l_1$ , $l_2$ and $l_\infty$ norms

There are some simple and useful results on relationships between different  $p$ -norms listed in this section. We also discuss some interesting properties of  $l_1$ -norm specifically.

**Definition 2.9** Let  $v \in \mathbb{C}^N$ . Let the entries in  $v$  be represented as

$$v_i = r_i \exp(j\theta_i)$$

where  $r_i = |v_i|$  with the convention that  $\theta_i = 0$  whenever  $r_i = 0$ .

The sign vector for  $v$  denoted by  $\text{sgn}(v)$  is defined as

$$\text{sgn}(v) = \begin{bmatrix} \text{sgn}(v_1) \\ \vdots \\ \text{sgn}(v_N) \end{bmatrix} \quad (2.5.1)$$

where

$$\text{sgn}(v_i) = \begin{cases} \exp(j\theta_i) & \text{if } r_i \neq 0; \\ 0 & \text{if } r_i = 0. \end{cases} \quad (2.5.2)$$

**Lemma 2.9** For any  $v \in \mathbb{C}^N$ :

$$\|v\|_1 = \text{sgn}(v)^H v = \langle v, \text{sgn}(v) \rangle. \quad (2.5.3)$$

PROOF.

$$\|v\|_1 = \sum_{i=1}^N r_i = \sum_{i=1}^N [r_i e^{j\theta_i}] e^{-j\theta_i} = \sum_{i=1}^N v_i e^{-j\theta_i} = \text{sgn}(v)^H v.$$

Note that whenever  $v_i = 0$ , corresponding 0 entry in  $\text{sgn}(v)$  has no effect on the sum.  $\square$

**Lemma 2.10** Suppose  $v \in \mathbb{C}^N$ . Then

$$\|v\|_2 \leq \|v\|_1 \leq \sqrt{N} \|v\|_2. \quad (2.5.4)$$

PROOF. For the lower bound, we go as follows

$$\|v\|_2^2 = \sum_{i=1}^N |v_i|^2 \leq \left( \sum_{i=1}^N |v_i|^2 + 2 \sum_{i,j,i \neq j} |v_i||v_j| \right) = \left( \sum_{i=1}^N |v_i| \right)^2 = \|v\|_1^2.$$

This gives us

$$\|v\|_2 \leq \|v\|_1.$$

We can write  $l_1$  norm as

$$\|v\|_1 = \langle v, \text{sgn}(v) \rangle. \quad (2.5.5)$$

By Cauchy-Schwartz inequality we have

$$\langle v, \text{sgn}(v) \rangle \leq \|v\|_2 \|\text{sgn}(v)\|_2 \quad (2.5.6)$$

Since  $\text{sgn}(v)$  can have at most  $N$  non-zero values, each with magnitude 1,

$$\|\text{sgn}(v)\|_2^2 \leq N \implies \|\text{sgn}(v)\|_2 \leq \sqrt{N}.$$

Thus, we get

$$\|v\|_1 \leq \sqrt{N} \|v\|_2.$$

□

**Lemma 2.11** *Let  $v \in \mathbb{C}^N$ . Then*

$$\|v\|_2 \leq \sqrt{N} \|v\|_\infty \quad (2.5.7)$$

PROOF.

$$\|v\|_2^2 = \sum_{i=1}^N |v_i|^2 \leq N \max_{1 \leq i \leq N} (|v_i|^2) = N \|v\|_\infty^2.$$

Thus

$$\|v\|_2 \leq \sqrt{N} \|v\|_\infty.$$

□

**Lemma 2.12** *Let  $v \in \mathbb{C}^N$ . Let  $1 \leq p, q \leq \infty$ . Then*

$$\|v\|_q \leq \|v\|_p \text{ whenever } p \leq q. \quad (2.5.8)$$

PROOF. TBD

□

**Lemma 2.13** *Let  $\mathbf{1} \in \mathbb{C}^N$  be the vector of all ones i.e.  $\mathbf{1} = (1, \dots, 1)$ . Let  $v \in \mathbb{C}^N$  be some arbitrary vector. Let  $|v|$  denote the*

vector of absolute values of entries in  $v$ . i.e.  $|v|_i = |v_i| \forall 1 \leq i \leq N$ . Then

$$\|v\|_1 = \mathbf{1}^T |v| = \mathbf{1}^H |v|. \quad (2.5.9)$$

PROOF.

$$\mathbf{1}^T |v| = \sum_{i=1}^N |v|_i = \sum_{i=1}^N |v_i| = \|v\|_1.$$

Finally since  $\mathbf{1}$  consists only of real entries, hence its transpose and Hermitian transpose are same.  $\square$

**Lemma 2.14** Let  $\mathbf{1} \in \mathbb{C}^{N \times N}$  be a square matrix of all ones. Let  $v \in \mathbb{C}^N$  be some arbitrary vector. Then

$$|v|^T \mathbf{1} |v| = \|v\|_1^2. \quad (2.5.10)$$

PROOF. We know that

$$\mathbf{1} = \mathbf{1} \mathbf{1}^T$$

Thus,

$$|v|^T \mathbf{1} |v| = |v|^T \mathbf{1} \mathbf{1}^T |v| = (\mathbf{1}^T |v|)^T \mathbf{1}^T |v| = \|v\|_1 \|v\|_1 = \|v\|_1^2.$$

We used the fact that  $\|v\|_1 = \mathbf{1}^T |v|$ .  $\square$

**Theorem 2.15**  $k$ -th largest (magnitude) entry in a vector  $x \in \mathbb{C}^N$  denoted by  $x_{(k)}$  obeys

$$|x_{(k)}| \leq \frac{\|x\|_1}{k} \quad (2.5.11)$$

PROOF. Let  $n_1, n_2, \dots, n_N$  be a permutation of  $\{1, 2, \dots, N\}$  such that

$$|x_{n_1}| \geq |x_{n_2}| \geq \dots \geq |x_{n_N}|.$$

Thus, the  $k$ -th largest entry in  $x$  is  $x_{n_k}$ . It is clear that

$$\|x\|_1 = \sum_{i=1}^N |x_i| = \sum_{i=1}^N |x_{n_i}|$$

Obviously

$$|x_{n_1}| \leq \sum_{i=1}^N |x_{n_i}| = \|x\|_1.$$

Similarly

$$k|x_{n_k}| = |x_{n_k}| + \dots + |x_{n_k}| \leq |x_{n_1}| + \dots + |x_{n_k}| \leq \sum_{i=1}^N |x_{n_i}| \leq \|x\|_1.$$

Thus

$$|x_{n_k}| \leq \frac{\|x\|_1}{k}.$$

□

### 2.5.2. Sparse signals

In this section we explore some useful properties of  $\Sigma_K$ , the set of  $K$ -sparse signals in standard basis for  $\mathbb{C}^N$ .

We recall that

$$\Sigma_K = \{x \in \mathbb{C}^N : \|x\|_0 \leq K\}. \quad (2.5.12)$$

We established before that this set is a union of  $\binom{N}{K}$  subspaces of  $\mathbb{C}^N$  each of which is constructed by an index set  $\Lambda \subset \{1, \dots, N\}$  with  $|\Lambda| = K$  choosing  $K$  specific dimensions of  $\mathbb{C}^N$ .

We first present some lemmas which connect the  $l_1$ ,  $l_2$  and  $l_\infty$  norms of vectors in  $\Sigma_K$ .

**Lemma 2.16** Suppose  $u \in \Sigma_K$ . Then

$$\frac{\|u\|_1}{\sqrt{K}} \leq \|u\|_2 \leq \sqrt{K}\|u\|_\infty. \quad (2.5.13)$$

PROOF. Due to lemma 2.9, we can write  $l_1$  norm as

$$\|u\|_1 = \langle u, \text{sgn}(u) \rangle. \quad (2.5.14)$$

By Cauchy-Schwartz inequality we have

$$\langle u, \text{sgn}(u) \rangle \leq \|u\|_2 \|\text{sgn}(u)\|_2 \quad (2.5.15)$$

Since  $u \in \Sigma_K$ ,  $\text{sgn}(u)$  can have at most  $K$  non-zero values each with magnitude 1. Thus, we have

$$\|\text{sgn}(u)\|_2^2 \leq K \implies \|\text{sgn}(u)\|_2 \leq \sqrt{K} \quad (2.5.16)$$

Thus we get the lower bound

$$\|u\|_1 \leq \|u\|_2 \sqrt{K} \implies \frac{\|u\|_1}{\sqrt{K}} \leq \|u\|_2. \quad (2.5.17)$$

Now  $|u_i| \leq \max(|u_i|) = \|u\|_\infty$ . So we have

$$\|u\|_2^2 = \sum_{i=1}^N |u_i|^2 \leq K \|u\|_\infty^2$$

since there are only  $K$  non-zero terms in the expansion of  $\|u\|_2^2$ .

This establishes the upper bound:

$$\|u\|_2 \leq \sqrt{K} \|u\|_\infty \quad (2.5.18)$$

□

## 2.6. Compressible signals

In this section, we first look at some general results and definitions related to  $K$ -term approximations of arbitrary signals  $x \in \mathbb{C}^N$ . We then define the notion of a compressible signal and study properties related to it.

### 2.6.1. K-term approximation of general signals

**Definition 2.10** [Restriction of a signal on an index set] Let  $x \in \mathbb{C}^N$ . Let  $T \subset \{1, 2, \dots, N\}$  be any index set. Further let

$$T = \{t_1, t_2, \dots, t_{|T|}\}$$

such that

$$t_1 < t_2 < \dots < t_{|T|}.$$

Let  $x_T \in \mathbb{C}^{|T|}$  be defined as

$$x_T = \begin{pmatrix} x_{t_1} & x_{t_2} & \dots & x_{t_{|T|}} \end{pmatrix}. \quad (2.6.1)$$

Then  $x_T$  is a **restriction** of the signal  $x$  on the index set  $T$ .

Alternatively let  $x_T \in \mathbb{C}^N$  be defined as

$$x_T(i) = \begin{cases} x(i) & \text{if } i \in T; \\ 0 & \text{otherwise.} \end{cases} \quad (2.6.2)$$

In other words,  $x_T \in \mathbb{C}^N$  keeps the entries in  $x$  indexed by  $T$  while sets all other entries to 0. Then we say that  $x_T$  is obtained by **masking**  $x$  with  $T$ . As an abuse of notation, we will use any of the two definitions whenever we are referring to  $x_T$ . The definition being used should be obvious from the context.

**Example 2.16: Restrictions on index sets** Let

$$x = \begin{pmatrix} -1 & 5 & 8 & 0 & 0 & -3 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{C}^{10}.$$

Let

$$T = \{1, 3, 7, 8\}.$$

Then

$$x_T = \begin{pmatrix} -1 & 0 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{C}^{10}.$$

Since  $|T| = 4$ , sometimes we will also write

$$x = \begin{pmatrix} -1 & 8 & 0 & 0 \end{pmatrix} \in \mathbb{C}^4.$$

□

**Definition 2.11** [K-term approximation] Let  $x \in \mathbb{C}^N$  be an arbitrary signal. Consider any index set  $T \subset \{1, \dots, N\}$  with  $|T| = K$ . Then  $x_T$  is a **K-term approximation** of  $x$ .

Clearly for any  $x \in \mathbb{C}^N$  there are  $\binom{N}{K}$  possible  $K$ -term approximations of  $x$ .

**Example 2.17:  $K$ -term approximation** Let

$$x = \begin{pmatrix} -1 & 5 & 8 & 0 & 0 & -3 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{C}^{10}.$$

Let  $T = \{1, 6\}$ . Then

$$x_T = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & -3 & 0 & 0 & 0 & 0 \end{pmatrix}$$

is a 2-term approximation of  $x$ .

If we choose  $T = \{7, 8, 9, 10\}$ , the corresponding 4-term approximation of  $x$  is

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

□

**Definition 2.12** [Largest entries approximation] Let  $x \in \mathbb{C}^N$  be an arbitrary signal. Let  $\lambda_1, \dots, \lambda_N$  be indices of entries in  $x$  such that

$$|x_{\lambda_1}| \geq |x_{\lambda_2}| \geq \dots \geq |x_{\lambda_N}|.$$

In case of ties, the order is resolved lexicographically, i.e. if  $|x_i| = |x_j|$  and  $i < j$  then  $i$  will appear first in the sequence  $\lambda_k$ .

Consider the index set  $\Lambda_K = \{\lambda_1, \lambda_2, \dots, \lambda_K\}$ . The restriction of  $x$  on  $\Lambda_K$  given by  $x_{\Lambda_K}$  (see definition 2.10) contains the  $K$  largest entries  $x$  while setting all other entries to 0. This is known as the  **$K$  largest entries approximation** of  $x$ .

This signal is denoted henceforth as  $x|_K$ . i.e.

$$x|_K = x_{\Lambda_K} \quad (2.6.3)$$

where  $\Lambda_K$  is the index set corresponding to  $K$  largest entries in  $x$  (magnitude wise).

**Example 2.18: Largest entries approximation** Let

$$x = \begin{pmatrix} -1 & 5 & 8 & 0 & 0 & -3 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then

$$x|_1 = \begin{pmatrix} 0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

$$x|_2 = \begin{pmatrix} 0 & 5 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

$$x|_3 = \begin{pmatrix} 0 & 5 & 8 & 0 & 0 & -3 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$x|_4 = x.$$

All further  $K$  largest entries approximations are same as  $x$ .  $\square$

A pertinent question at this point is: which  $K$ -term approximation of  $x$  is the best  $K$ -term approximation? Certainly in order to compare two approximations we need some criterion. Let us choose  $l_p$  norm as the criterion. The next lemma gives an interesting result for best  $K$ -term approximations in  $l_p$  norm sense.

**Lemma 2.17** *Let  $x \in \mathbb{C}^N$ . Let the best  $K$  term approximation of  $x$  be obtained by the following optimization program:*

$$\begin{aligned} & \underset{T \subset \{1, \dots, N\}}{\text{maximize}} \quad \|x_T\|_p \\ & \text{subject to} \quad |T| = K. \end{aligned} \tag{2.6.4}$$

where  $p \in [1, \infty]$ .

Let an optimal solution for this optimization problem be denoted by  $x_{T^*}$ . Then

$$\|x_K\|_p = \|x_{T^*}\|_p.$$

i.e. the  $K$ -largest entries approximation of  $x$  is an optimal solution to eq. (2.6.4).

**PROOF.** For  $p = \infty$ , the result is obvious. In the following, we focus on  $p \in [1, \infty)$ .

We note that maximizing  $\|x_T\|_p$  is equivalent to maximizing  $\|x_T\|_p^p$ .

Let  $\lambda_1, \dots, \lambda_N$  be indices of entries in  $x$  such that

$$|x_{\lambda_1}| \geq |x_{\lambda_2}| \geq \dots \geq |x_{\lambda_N}|.$$

Further let  $\{\omega_1, \dots, \omega_N\}$  be any permutation of  $\{1, \dots, N\}$ .

Clearly

$$\|x|_K\|_p^p = \sum_{i=1}^K |x_{\lambda_i}|^p \geq \sum_{i=1}^K |x_{\omega_i}|^p.$$

Thus if  $T^*$  corresponds to an optimal solution of eq. (2.6.4) then

$$\|x|_K\|_p^p = \|x_{T^*}\|_p^p.$$

Thus  $x|_K$  is an optimal solution to eq. (2.6.4).  $\square$

This lemma helps us establish that whenever we are looking for a best  $K$ -term approximation of  $x$  under any  $l_p$  norm, all we have to do is to pickup the  $K$ -largest entries in  $x$ .

**Definition 2.13** [Restriction of a matrix on an index set] Let  $\Phi \in \mathbb{C}^{M \times N}$ . Let  $T \subset \{1, 2, \dots, N\}$  be any index set. Further let

$$T = \{t_1, t_2, \dots, t_{|T|}\}$$

such that

$$t_1 < t_2 < \dots < t_{|T|}.$$

Let  $\Phi_T \in \mathbb{C}^{M \times |T|}$  be defined as

$$\Phi_T = \begin{bmatrix} \phi_{t_1} & \phi_{t_2} & \dots & \phi_{t_{|T|}} \end{bmatrix}. \quad (2.6.5)$$

Then  $\Phi_T$  is a **restriction** of the matrix  $\Phi$  on the index set  $T$ .

Alternatively let  $\Phi_T \in \mathbb{C}^{M \times N}$  be defined as

$$(\Phi_T)_i = \begin{cases} \phi_i & \text{if } i \in T; \\ 0 & \text{otherwise.} \end{cases} \quad (2.6.6)$$

In other words,  $\Phi_T \in \mathbb{C}^{M \times N}$  keeps the columns in  $\Phi$  indexed by  $T$  while sets all other columns to 0. Then we say that  $\Phi_T$  is obtained by **masking**  $\Phi$  with  $T$ . As an abuse of notation, we will use any of the two definitions whenever we are referring to  $\Phi_T$ . The definition being used should be obvious from the context.

**Lemma 2.18** *Let  $\text{supp}(x) = \Lambda$ . Then*

$$\Phi x = \Phi_\Lambda x_\Lambda. \quad (2.6.7)$$

PROOF.

$$\Phi x = \sum_{i=1}^N x_i \phi_i = \sum_{\lambda_i \in \Lambda} x_{\lambda_i} \phi_{\lambda_i} = \Phi_\Lambda x_\Lambda.$$

□

REMARK. The lemma remains valid whether we use the restriction or the mask version of  $x_\Lambda$  notation as long as same version is used for both  $\Phi$  and  $x$ .

**Corollary 2.19.** *Let  $S$  and  $T$  be two disjoint index sets such that for some  $x \in \mathbb{C}^N$*

$$x = x_T + x_S \quad (2.6.8)$$

*using the mask version of  $x_T$  notation. Then the following holds*

$$\Phi x = \Phi_T x_T + \Phi_S x_S. \quad (2.6.9)$$

PROOF. Straightforward application of Lemma 2.18:

$$\Phi x = \Phi x_T + \Phi x_S = \Phi_T x_T + \Phi_S x_S.$$

□

**Lemma 2.20** *Let  $T$  be any index set. Let  $\Phi \in \mathbb{C}^{M \times N}$  and  $y \in \mathbb{C}^M$ .*

*Then*

$$[\Phi^H y]_T = \Phi_T^H y. \quad (2.6.10)$$

PROOF.

$$\Phi^H y = \begin{bmatrix} \langle \phi_1, y \rangle \\ \vdots \\ \langle \phi_N, y \rangle \end{bmatrix}$$

Now let

$$T = \{t_1, \dots, t_K\}.$$

Then

$$[\Phi^H y]_T = \begin{bmatrix} \langle \phi_{t_1}, y \rangle \\ \vdots \\ \langle \phi_{t_K}, y \rangle \end{bmatrix} = \Phi_T^H y.$$

□

**REMARK.** The lemma remains valid whether we use the restriction or the mask version of  $\Phi_T$  notation.

### 2.6.2. Compressible signals

We will now define the notion of a compressible signal in terms of the decay rate of magnitude of its entries when sorted in descending order.

**Definition 2.14** [ $p$ -compressible signal] Let  $x \in \mathbb{C}^N$  be an arbitrary signal. Let  $\lambda_1, \dots, \lambda_N$  be indices of entries in  $x$  such that

$$|x_{\lambda_1}| \geq |x_{\lambda_2}| \geq \dots \geq |x_{\lambda_N}|.$$

In case of ties, the order is resolved lexicographically, i.e. if  $|x_i| = |x_j|$  and  $i < j$  then  $i$  will appear first in the sequence  $\lambda_k$ . Define

$$\hat{x} = (x_{\lambda_1}, x_{\lambda_2}, \dots, x_{\lambda_N}). \quad (2.6.11)$$

The signal  $x$  is called  **$p$ -compressible** with magnitude  $R$  if there exists  $p \in (0, 1)$  such that

$$|\hat{x}_i| \leq R \cdot i^{-\frac{1}{p}} \quad \forall i = 1, 2, \dots, N. \quad (2.6.12)$$

**Lemma 2.21** Let  $x$  be  $p$ -compressible with  $p = 1$ . Then

$$\|x\|_1 \leq R(1 + \ln(N)). \quad (2.6.13)$$

**PROOF.** Recalling  $\hat{x}$  from (2.6.11) its straightforward to see that

$$\|x\|_1 = \|\hat{x}\|_1$$

since the  $l_1$  norm doesn't depend on the ordering of entries in  $x$ .

Now since  $x$  is 1-compressible, hence from (2.6.12) we have

$$|\hat{x}_i| \leq R \frac{1}{i}.$$

This gives us

$$\|\hat{x}\|_1 \leq \sum_{i=1}^N R \frac{1}{i} = R \sum_{i=1}^N \frac{1}{i}.$$

The sum on the R.H.S. is the  $N$ -th Harmonic number (sum of reciprocals of first  $N$  natural numbers). A simple upper bound on Harmonic numbers is

$$H_k \leq 1 + \ln(k).$$

This completes the proof.  $\square$

We now demonstrate how a compressible signal is well approximated by a sparse signal.

**Lemma 2.22** *Let  $x$  be a  $p$ -compressible signal and let  $x|_K$  be its best  $K$ -term approximation. Then the  $l_1$  norm of approximation error satisfies*

$$\|x - x|_K\|_1 \leq C_p \cdot R \cdot K^{1-\frac{1}{p}} \quad (2.6.14)$$

with

$$C_p = \left( \frac{1}{p} - 1 \right)^{-1}.$$

Moreover the  $l_2$  norm of approximation error satisfies

$$\|x - x|_K\|_2 \leq D_p \cdot R \cdot K^{1-\frac{1}{p}} \quad (2.6.15)$$

with

$$D_p = \left( \frac{2}{p} - 1 \right)^{-1/2}.$$

PROOF.

$$\|x - x|_K\|_1 = \sum_{i=K+1}^N |x_{\lambda_i}| \leq R \sum_{i=K+1}^N i^{-\frac{1}{p}}.$$

We now approximate the R.H.S. sum with an integral.

$$\sum_{i=K+1}^N i^{-\frac{1}{p}} \leq \int_{x=K}^N x^{-\frac{1}{p}} dx \leq \int_{x=K}^{\infty} x^{-\frac{1}{p}} dx.$$

Now

$$\int_{x=K}^{\infty} x^{-\frac{1}{p}} dx = \left[ \frac{x^{1-\frac{1}{p}}}{1-\frac{1}{p}} \right]_K^{\infty} = C_p K^{1-\frac{1}{p}}.$$

We can similarly show the result for  $l_2$  norm.  $\square$

## 2.7. Tools for dictionary analysis

In this section we review various properties associated with a dictionary  $\mathcal{D}$  which are useful in understanding the behavior and capabilities of a dictionary.

We recall that a dictionary  $\mathcal{D}$  consists of a finite number of unit norm vectors in  $\mathbb{C}^N$  called atoms which span the signal space  $\mathbb{C}^N$ . Atoms of the dictionary are indexed by an index set  $\Omega$ . i.e.

$$\mathcal{D} = \{d_\omega : \omega \in \Omega\}$$

with  $|\Omega| = D$  and  $N \leq D$  with  $\|d_\omega\|_2 = 1$  for all atoms.

The vectors  $x \in \mathbb{C}^N$  can be represented by a synthesis matrix consisting of the atoms of  $\mathcal{D}$  by a vector  $\alpha \in \mathbb{C}^D$  as

$$x = \mathcal{D}\alpha.$$

Note that we are using the same symbol  $\mathcal{D}$  to represent the dictionary as a set of atoms as well as the corresponding synthesis matrix.

We can write the matrix  $\mathcal{D}$  consisting of its columns as

$$\mathcal{D} = [d_1 \ \dots \ d_D]$$

This shouldn't be causing any confusion in the sequel. When we write the subscript as  $d_{\omega_i}$  where  $\omega_i \in \Omega$  we are referring to the atoms of the dictionary  $\mathcal{D}$  indexed by the set  $\Omega$ , while when we write the subscript as  $d_i$  we are referring to a column of corresponding synthesis matrix.

In this case,  $\Omega$  will simply mean the index set  $\{1, \dots, D\}$ . Obviously  $|\Omega| = D$  holds still.

Often, we will be working with a subset of atoms in a dictionary. Usually such a subset of atoms will be indexed by an index set  $\Lambda \subseteq \Omega$ .  $\Lambda$  will take the form of  $\Lambda \subseteq \{\omega_1, \dots, \omega_D\}$  or  $\Lambda \subseteq \{1, \dots, D\}$  depending upon whether we are talking about the subset of atoms in the dictionary or a subset of columns from the corresponding synthesis matrix.

Often we will need the notion of a sub-dictionary [37] described below.

**Definition 2.15** A sub-dictionary is a linearly independent collection of atoms. Let  $\Lambda \subset \{\omega_1, \dots, \omega_D\}$  be the index set for the atoms in the sub-dictionary. We denote the sub-dictionary as  $\mathcal{D}_\Lambda$ . We also use  $\mathcal{D}_\Lambda$  to denote the corresponding matrix with  $\Lambda \subset \{1, \dots, D\}$ .

REMARK. A subdictionary is full rank.

This is obvious since it is a collection of linearly independent atoms.

For subdictionaries often we will say  $K = |\Lambda|$  and  $G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda$  as its Gram matrix. Sometimes, we will also be considering  $G^{-1}$ .  $G^{-1}$  has a useful interpretation in terms of the **dual vectors** for the atoms in  $\mathcal{D}_\Lambda$  [35].

Let  $\{d_\lambda\}_{\lambda \in \Lambda}$  denote the atoms in  $\mathcal{D}_\Lambda$ . Let  $\{c_\lambda\}_{\lambda \in \Lambda}$  be chosen such that

$$\langle d_\lambda, c_\lambda \rangle = 1$$

and

$$\langle d_\lambda, c_\omega \rangle = 0 \text{ for } \lambda, \omega \in \Lambda, \lambda \neq \omega.$$

Each dual vector  $c_\lambda$  is orthogonal to atoms in the subdictionary at different indices and is long enough so that its inner product with  $d_\lambda$  is one. The dual system somehow inverts the sub-dictionary. In fact the

dual vectors are nothing but the columns of the matrix  $B = (\mathcal{D}_\Lambda^\dagger)^H$ .

Now, a simple calculation:

$$B^H B = (\mathcal{D}_\Lambda^\dagger)(\mathcal{D}_\Lambda^\dagger)^H = (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} = (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} = G^{-1}.$$

Therefore, the inverse Gram matrix lists the inner products between the dual vectors.

Sometimes we will be discussing tools which apply for general matrices. We will use the symbol  $\Phi$  for representing general matrices. Whenever the dictionary is an orthonormal basis, we will use the symbol  $\Psi$ .

### 2.7.1. Spark

**Definition 2.16** [Spark of a matrix] The **spark** of a given matrix  $\Phi$  is the smallest number of columns of  $\Phi$  that are linearly dependent. If all columns are linearly independent, then the spark is defined to be number of columns plus one.

Note that the definition of spark applies to all matrices (wide, tall or square). It is not restricted to the synthesis matrices for a dictionary.

Correspondingly, the spark of a dictionary is defined as the minimum number of atoms which are linearly dependent.

We recall that *rank* of a matrix is defined as the maximum number of columns which are linearly independent. Definition of spark bears remarkable resemblance yet its very hard to obtain as it requires a combinatorial search over all possible subsets of columns of  $\Phi$ .

#### Example 2.19: Spark

- Spark of the  $3 \times 3$  identity matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is 4 since all columns are linearly independent.

- Spark of the  $2 \times 4$  matrix

$$\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$

is 2 since column 1 and 3 are linearly dependent.

- If a matrix has a column with all zero entries, then the spark of such a matrix is 1. This is a trivial case and we will not consider such matrices in the sequel.
- In general for an  $N \times D$  synthesis matrix,  $\text{spark}(\mathcal{D}) \in [2, N+1]$ .

□

A naive combinatorial algorithm to calculate the spark of a matrix is given in fig. 2.5.

```

Input:  $\Phi$ 
Output:  $s = \text{Spark of } \Phi$ 
 $R \leftarrow \text{rank}(\Phi);$ 
foreach  $j \leftarrow 1, \dots, R$  do
    Identify  $\binom{D}{j}$  ways of choosing  $j$  columns from  $D$  columns of  $\Phi$ ;
    foreach choice of  $j$  columns do
        if columns are dependent then
             $s \leftarrow j$  ;
            return;
        end
    end
end
// All columns are linearly independent
 $s \leftarrow R + 1$  ;

```

FIGURE 2.5. A naive algorithm for computing the spark of a matrix

Spark is useful in characterizing the uniqueness of the solution of a  $(\mathcal{D}, K)$ -EXACT-SPARSE problem (see definition 2.7).

REMARK. The  $l_0$ -“norm” of vectors belonging to null space of a matrix  $\Phi$  is greater than or equal to  $\text{spark}(\Phi)$ :

$$\|x\|_0 \geq \text{spark}(\Phi) \quad \forall x \in \mathcal{N}(\Phi). \quad (2.7.1)$$

PROOF. If  $x \in \mathcal{N}(\Phi)$  then  $\Phi x = 0$ . Thus non-zero entries in  $x$  pick a set of columns in  $\Phi$  which are linearly dependent. Clearly  $\|x\|_0$  indicates the number of columns in the set which are linearly dependent. By definition  $\text{spark}$  of  $\Phi$  indicates the minimum number of columns which are linearly dependent hence the result.

$$\|x\|_0 \geq \text{spark}(\Phi) \quad \forall x \in \mathcal{N}(\Phi).$$

□

We now present a criteria based on  $\text{spark}$  which characterizes the uniqueness of a sparse solution to the problem  $y = \Phi x$ .

**Theorem 2.23** [Uniqueness-Spark] Consider a solution  $x^*$  to the under-determined system  $y = \Phi x$ . If  $x^*$  obeys

$$\|x^*\|_0 < \frac{\text{spark}(\Phi)}{2} \quad (2.7.2)$$

then it is necessarily the sparsest solution.

PROOF. Let  $x'$  be some other solution to the problem. Then

$$\Phi x' = \Phi x^* \implies \Phi(x' - x^*) = 0 \implies (x' - x^*) \in \mathcal{N}(\Phi).$$

Now based on previous remark we have

$$\|x' - x^*\|_0 \geq \text{spark}(\Phi).$$

Now

$$\|x'\|_0 + \|x^*\|_0 \geq \|x' - x^*\|_0 \geq \text{spark}(\Phi).$$

Hence, if  $\|x^*\|_0 < \frac{\text{spark}(\Phi)}{2}$ , then we have

$$\|x'\|_0 > \frac{\text{spark}(\Phi)}{2}$$

for all other solutions  $x'$  to the equation  $y = \Phi x$ .

Thus  $x^*$  is necessarily the sparsest possible solution.  $\square$

This result is quite useful as it establishes a global optimality criterion for the  $(\mathcal{D}, K)$ -EXACT-SPARSE problem in definition 2.7.

As long as  $K < \frac{1}{2} \text{spark}(\Phi)$  this theorem guarantees that the solution to  $(\mathcal{D}, K)$ -EXACT-SPARSE problem is unique. This is quite surprising result for a non-convex combinatorial optimization problem. We are able to guarantee a global uniqueness for the solution based on a simple check on the sparsity of the solution.

Note that we are only saying that if a sufficiently sparse solution is found then it is unique. We are not claiming that it is possible to find such a solution.

Obviously, the larger the spark, we can guarantee uniqueness for signals with higher sparsity levels. So a natural question is: *How large can spark of a dictionary be?* We consider few examples.

**Example 2.20: Spark of Gaussian dictionaries** Consider a dictionary  $\mathcal{D}$  whose atoms  $d_i$  are random vectors independently drawn from normal distribution. Since a dictionary requires all its atoms to be unit-norms, hence we divide the each of the random vectors with their norms.

We know that with probability 1 any set of  $N$  independent Gaussian random vectors is linearly independent. Also since  $d_i \in \mathbb{C}^N$  hence a set of  $N + 1$  atoms is always linearly dependent.

Thus  $\text{spark}(\mathcal{D}) = N + 1$ .

Thus, if a solution to EXACT-SPARSE problem contains  $\frac{N}{2}$  or fewer non-zero entries then it is necessarily unique with probability 1.  $\square$

**Example 2.21: Spark of Dirac Fourier basis** For

$$\mathcal{D} = \begin{bmatrix} I & F \end{bmatrix} \in \mathbb{C}^{N \times 2N}$$

it can be shown that

$$\text{spark}(\mathcal{D}) = 2\sqrt{N}.$$

In this case, the sparsity level of a unique solution must be less than  $\sqrt{N}$ .  $\square$

### 2.7.2. Coherence

Finding out the spark of a dictionary  $\mathcal{D}$  is NP-hard since it involves considering combinatorially large number of selections of columns from  $\mathcal{D}$ . In this section we consider the *coherence* of a dictionary which is computationally tractable and quite useful in characterizing the solutions of sparse approximation problems.

**Definition 2.17** [Coherence of a dictionary] The **coherence** of a dictionary  $\mathcal{D}$  is defined as the maximum absolute inner product between two distinct atoms in the dictionary:

$$\mu = \max_{j \neq k} |\langle d_{\omega_j}, d_{\omega_k} \rangle| = \max_{j \neq k} |(\mathcal{D}^H \mathcal{D})_{jk}|. \quad (2.7.3)$$

If the dictionary consists of two orthonormal bases, then coherence is also known as *mutual coherence* or *proximity*; see definition 2.1.

We note that  $d_{\omega_i}$  is the  $i$ -th column of synthesis matrix  $\mathcal{D}$ . Also  $\mathcal{D}^H \mathcal{D}$  is the **Gram matrix** for  $\mathcal{D}$  whose elements are nothing but the inner-products of columns of  $\mathcal{D}$ .

We note that by definition  $\|d_{\omega}\|_2 = 1$  hence  $\mu \leq 1$  and since absolute values are considered hence  $\mu \geq 0$ . Thus,  $0 \leq \mu \leq 1$ .

For an orthonormal basis  $\Psi$  all atoms are orthogonal to each other, hence

$$|\langle \psi_{\omega_j}, \psi_{\omega_k} \rangle| = 0 \text{ whenever } j \neq k.$$

Thus  $\mu = 0$ .

In the following, we will use the notation  $|A|$  to denote a matrix consisting of absolute values of entries in a matrix  $A$ . i.e.

$$|A|_{ij} = |A_{ij}|.$$

The off-diagonal entries of the Gram matrix are captured by the matrix  $\mathcal{D}^H \mathcal{D} - I$ . Note that all diagonal entries in  $\mathcal{D}^H \mathcal{D} - I$  are zero since atoms of  $\mathcal{D}$  are unit norm. Moreover, each of the entries in  $|\mathcal{D}^H \mathcal{D} - I|$  is dominated by  $\mu(\mathcal{D})$ .

The inner product between any two atoms  $|\langle d_{\omega_j}, d_{\omega_k} \rangle|$  is a measure of how much they look alike or how much they are correlated. Coherence just picks up the two vectors which are most alike and returns their correlation. In a way  $\mu$  is quite a blunt measure of the quality of a dictionary, yet it is quite useful.

If a dictionary is uniform in the sense that there is not much variation in  $|\langle d_{\omega_j}, d_{\omega_k} \rangle|$ , then  $\mu$  captures the behavior of the dictionary quite well.

**Definition 2.18** [Incoherent dictionaries] We say that a dictionary is **incoherent** if the coherence of the dictionary is small.

We are looking for dictionaries which are incoherent. In the sequel we will see how incoherence plays a role in sparse approximation.

**Example 2.22:** [ Two ortho bases] We established in theorem 2.3 that coherence of two ortho-bases is bounded by

$$\frac{1}{\sqrt{N}} \leq \mu \leq 1.$$

In particular we showed in theorem 2.4 that coherence of Dirac Fourier basis is  $\frac{1}{\sqrt{N}}$ . □

**Example 2.23: Coherence: Multi-ONB dictionary** A dictionary of concatenated orthonormal bases is called a multi-ONB. For some  $N$ , it is possible to build a multi-ONB which contains  $N$  or even  $N + 1$  bases yet retains the minimal coherence  $\mu = \frac{1}{\sqrt{N}}$  possible. □

**Theorem 2.24** A lower bound on the coherence of a general dictionary is given by

$$\mu \geq \sqrt{\frac{D-N}{N(D-1)}}$$

**Definition 2.19** If each atomic inner product meets this bound, the dictionary is called an **optimal Grassmannian frame**.

The definition of coherence can be extended to arbitrary matrices  $\Phi \in \mathbb{C}^{N \times D}$ .

**Definition 2.20** [Coherence of an arbitrary matrix] The **coherence** of a matrix  $\Phi \in \mathbb{C}^{N \times D}$  is defined as the maximum absolute *normalized* inner product between two distinct columns in the matrix. Let

$$\Phi = [\phi_1 \ \phi_2 \ \dots \ \phi_D].$$

Then coherence of  $\Phi$  is given by

$$\mu(\Phi) = \max_{j \neq k} \frac{|\langle \phi_j, \phi_k \rangle|}{\|\phi_j\|_2 \|\phi_k\|_2} \quad (2.7.4)$$

It is assumed that none of the columns in  $\Phi$  is a zero vector.

**2.7.2.1. Lower bounds for spark.** Coherence of a matrix is easy to compute. More interestingly it also provides a lower bound on the spark of a matrix.

**Theorem 2.25** For any matrix  $\Phi \in \mathbb{C}^{N \times D}$  (with non-zero columns) the following relationship holds

$$\text{spark}(\Phi) \geq 1 + \frac{1}{\mu(\Phi)}. \quad (2.7.5)$$

**PROOF.** We note that scaling of a column of  $\Phi$  doesn't change either the spark or coherence of  $\Phi$ . Therefore, we assume that the columns of  $\Phi$  are normalized.

We now construct the Gram matrix of  $\Phi$  given by  $G = \Phi^H\Phi$ . We note that

$$G_{kk} = 1 \quad \forall 1 \leq k \leq D$$

since each column of  $\Phi$  is unit norm.

Also

$$|G_{kj}| \leq \mu(\Phi) = \mu(\Phi) \quad \forall 1 \leq k, j \leq D, k \neq j.$$

Consider any  $p$  columns from  $\Phi$  and construct its Gram matrix. This is nothing but a leading minor of size  $p \times p$  from the matrix  $G$ .

From the Gershgorin disk theorem, if this minor is diagonally dominant, i.e. if

$$\sum_{j \neq i} |G_{ij}| < |G_{ii}| \quad \forall i$$

then this sub-matrix of  $G$  is positive definite and so corresponding  $p$  columns from  $\Phi$  are linearly independent.

But

$$|G_{ii}| = 1$$

and

$$\sum_{j \neq i} |G_{ij}| \leq (p - 1)\mu(\Phi)$$

for the minor under consideration. Hence for  $p$  columns to be linearly independent the following condition is sufficient

$$(p - 1)\mu(\Phi) < 1.$$

Thus if

$$p < 1 + \frac{1}{\mu(\Phi)},$$

then every set of  $p$  columns from  $\Phi$  is linearly independent.

Hence, the smallest possible set of linearly dependent columns must satisfy

$$p \geq 1 + \frac{1}{\mu(\Phi)}.$$

This establishes the lower bound that

$$\text{spark}(\Phi) \geq 1 + \frac{1}{\mu(\Phi)}.$$

□

This bound on spark doesn't make any assumptions on the structure of the dictionary. In fact, imposing additional structure on the dictionary can give better bounds. Let us look at an example for a two ortho-basis [20].

**Theorem 2.26** *Let  $\mathcal{D}$  be a two ortho-basis. Then*

$$\text{spark}(\mathcal{D}) \geq \frac{2}{\mu(\mathcal{D})}. \quad (2.7.6)$$

PROOF. From theorem 2.6 we know that for any vector  $v \in \mathcal{N}(\mathcal{D})$

$$\|v\|_0 \geq \frac{2}{\mu(\mathcal{D})}.$$

But

$$\text{spark}(\mathcal{D}) = \min_{v \in \mathcal{N}(\mathcal{D})} (\|v\|_0).$$

Thus

$$\text{spark}(\mathcal{D}) \geq \frac{2}{\mu(\mathcal{D})}.$$

□

For maximally incoherent two orthonormal bases, we know that  $\mu = \frac{1}{\sqrt{N}}$ . A perfect example is the pair of Dirac and Fourier bases. In this case  $\text{spark}(\mathcal{D}) \geq 2\sqrt{N}$ .

**2.7.2.2. Uniqueness-Coherence.** We can now establish a uniqueness condition for sparse solution of  $x = \Phi\alpha$ .

**Theorem 2.27** [Uniqueness-Coherence] Consider a solution  $x^*$  to the under-determined system  $y = \Phi x$ . If  $x^*$  obeys

$$\|x^*\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu(\Phi)} \right) \quad (2.7.7)$$

then it is necessarily the sparsest solution.

PROOF. This is a straightforward application of theorem 2.23 and theorem 2.25.  $\square$

It is interesting to compare the two uniqueness theorems: theorem 2.23 and theorem 2.27.

Theorem 2.23 uses spark, is sharp and is far more powerful than theorem 2.27.

Coherence can never be smaller than  $\frac{1}{\sqrt{N}}$ , therefore the bound on  $\|x^*\|_0$  in theorem 2.27 can never be larger than  $\frac{\sqrt{N}+1}{2}$ .

However, spark can be easily as large as  $N$  and then bound on  $\|x^*\|_0$  can be as large as  $\frac{N}{2}$ .

We recall from theorem 2.8 that the bound for sparsity level of sparsest solution in two-ortho basis  $\mathcal{H} = [\Psi \ \mathcal{X}]$  is given by

$$\|x^*\|_0 < \frac{1}{\mu(\mathcal{H})}$$

which is a larger bound than theorem 2.27 for general dictionaries by a factor of 2.

Thus, we note that coherence gives a weaker bound than spark for supportable sparsity levels of unique solutions. The advantage that coherence has is that it is easily computable and doesn't require any special structure on the dictionary (two ortho basis has a special structure).

### 2.7.2.3. Singular values of sub-dictionaries.

**Theorem 2.28** Let  $\mathcal{D}$  be a dictionary and  $\mathcal{D}_\Lambda$  be a sub-dictionary. Let  $\mu$  be the coherence of  $\mathcal{D}$ . Let  $K = |\Lambda|$ . Then the eigen values of  $G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda$  satisfy:

$$1 - (K - 1)\mu \leq \lambda \leq 1 + (K - 1)\mu. \quad (2.7.8)$$

Moreover, the singular values of the sub-dictionary  $\mathcal{D}_\Lambda$  satisfy

$$\sqrt{1 - (K - 1)\mu} \leq \sigma(\mathcal{D}_\Lambda) \leq \sqrt{1 + (K - 1)\mu}. \quad (2.7.9)$$

PROOF. We recall from Gershgorin's theorem that for any square matrix  $A \in \mathbb{C}^{K \times K}$ , every eigen value  $\lambda$  of  $A$  satisfies

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \text{ for some } i \in \{1, \dots, K\}.$$

Now consider the matrix  $G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda$  with diagonal elements equal to 1 and off diagonal elements bounded by a value  $\mu$ . Then

$$|\lambda - 1| \leq \sum_{j \neq i} |a_{ij}| \leq \sum_{j \neq i} \mu = (K - 1)\mu.$$

Thus,

$$-(K - 1)\mu \leq \lambda - 1 \leq (K - 1)\mu \iff 1 - (K - 1)\mu \leq \lambda \leq 1 + (K - 1)\mu$$

This gives us a lower bound on the smallest eigen value.

$$\lambda_{\min}(G) \geq 1 - (K - 1)\mu.$$

Since  $G$  is positive definite ( $\mathcal{D}_\Lambda$  is full-rank), hence its eigen values are positive. Thus, the above lower bound is useful only if

$$1 - (K - 1)\mu > 0 \iff 1 > (K - 1)\mu \iff \mu < \frac{1}{K - 1}.$$

We also get an upper bound on the eigen values of  $G$  given by

$$\lambda_{\max}(G) \leq 1 + (K - 1)\mu.$$

The bounds on singular values of  $\mathcal{D}_\Lambda$  are obtained as a straight-forward extension by taking square roots on the expressions.  $\square$

#### 2.7.2.4. Embeddings using sub-dictionaries.

**Theorem 2.29** Let  $\mathcal{D}$  be a real dictionary and  $\mathcal{D}_\Lambda$  be a sub-dictionary with  $K = |\Lambda|$ . Let  $\mu$  be the coherence of  $\mathcal{D}$ . Let  $v \in \mathbb{R}^K$  be an arbitrary vector. Then

$$|v|^T [I - \mu(\mathbf{1} - I)] |v| \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq |v|^T [I + \mu(\mathbf{1} - I)] |v| \quad (2.7.10)$$

where  $\mathbf{1}$  is a  $K \times K$  matrix of all ones. Moreover

$$(1 - (K - 1)\mu) \|v\|_2^2 \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq (1 + (K - 1)\mu) \|v\|_2^2. \quad (2.7.11)$$

PROOF. We can easily write

$$\begin{aligned} \|\mathcal{D}_\Lambda v\|_2^2 &= v^T \mathcal{D}_\Lambda^T \mathcal{D}_\Lambda v \\ v^T \mathcal{D}_\Lambda^T \mathcal{D}_\Lambda v &= \sum_{i=1}^K \sum_{j=1}^K v_i d_{\lambda_i}^T d_{\lambda_j} v_j. \end{aligned} \quad (2.7.12)$$

The terms in the R.H.S. for  $i = j$  are given by

$$v_i d_{\lambda_i}^T d_{\lambda_i} v_i = |v_i|^2.$$

Summing over  $i = 1, \dots, K$ , we get

$$\sum_{i=1}^K |v_i|^2 = \|v\|_2^2 = v^T v = |v|^T |v| = |v|^T I |v|.$$

We are now left with  $K^2 - K$  off diagonal terms. Each of these terms is bounded by

$$-\mu |v_i| |v_j| \leq v_i d_{\lambda_i}^T d_{\lambda_j} v_j \leq \mu |v_i| |v_j|.$$

Summing over the  $K^2 - K$  off-diagonal terms we get:

$$\sum_{i \neq j} |v_i| |v_j| = \sum_{i,j} |v_i| |v_j| - \sum_{i=j} |v_i| |v_j| = |v|^T (\mathbf{1} - I) |v|.$$

Thus,

$$-\mu |v|^T (\mathbf{1} - I) |v| \leq \sum_{i \neq j} v_i d_{\lambda_i}^T d_{\lambda_j} v_j \leq \mu |v|^T (\mathbf{1} - I) |v|$$

Thus,

$$|v|^T I |v| - \mu |v|^T (\mathbf{1} - I) |v| \leq v^T \mathcal{D}_\Lambda^T \mathcal{D}_\Lambda v \leq |v|^T I |v| + \mu |v|^T (\mathbf{1} - I) |v|.$$

We get the result by slight reordering of terms:

$$|v|^T [I - \mu(\mathbf{1} - I)] |v| \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq |v|^T [I + \mu(\mathbf{1} - I)] |v|$$

We note that due to lemma 2.14

$$|v|^T \mathbf{1} |v| = \|v\|_1^2.$$

Thus, the inequalities can be written as

$$(1 + \mu) \|v\|_2^2 - \mu \|v\|_1^2 \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq (1 - \mu) \|v\|_2^2 + \mu \|v\|_1^2.$$

Alternatively,

$$\|v\|_2^2 - \mu (\|v\|_1^2 - \|v\|_2^2) \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq \|v\|_2^2 + \mu (\|v\|_1^2 - \|v\|_2^2).$$

Finally, due to lemma 2.10

$$\|v\|_1^2 \leq K \|v\|_2^2 \implies \|v\|_1^2 - \|v\|_2^2 \leq (K - 1) \|v\|_2^2.$$

This gives us

$$(1 - (K - 1)\mu) \|v\|_2^2 \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq (1 + (K - 1)\mu) \|v\|_2^2.$$

□

We now present the above theorem for the complex case. The proof is based on singular values. This proof is simpler and more general than the one presented above.

**Theorem 2.30** *Let  $\mathcal{D}$  be a dictionary and  $\mathcal{D}_\Lambda$  be a sub-dictionary with  $K = |\Lambda|$ . Let  $\mu$  be the coherence of  $\mathcal{D}$ . Let  $v \in \mathbb{C}^K$  be an arbitrary vector. Then*

$$(1 - (K - 1)\mu) \|v\|_2^2 \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq (1 + (K - 1)\mu) \|v\|_2^2. \quad (2.7.13)$$

PROOF. Recall that

$$\sigma_{\min}^2(\mathcal{D}_\Lambda) \|v\|_2^2 \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq \sigma_{\max}^2(\mathcal{D}_\Lambda) \|v\|_2^2.$$

Theorem 2.28 tells us:

$$1 - (K - 1)\mu \leq \sigma^2(\mathcal{D}_\Lambda) \leq 1 + (K - 1)\mu.$$

Thus,

$$\sigma_{\min}^2(\mathcal{D}_\Lambda) \|v\|_2^2 \geq (1 - (K - 1)\mu) \|v\|_2^2$$

and

$$\sigma_{\max}^2(\mathcal{D}_\Lambda) \|v\|_2^2 \leq (1 + (K - 1)\mu) \|v\|_2^2.$$

This gives us the result

$$(1 - (K - 1)\mu) \|v\|_2^2 \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq (1 + (K - 1)\mu) \|v\|_2^2.$$

□

### 2.7.3. Babel function

Recalling the definition of coherence, we note that it reflects only the extreme correlations between atoms of dictionary. If most of the inner products are small compared to one dominating inner product, then the value of coherence is highly misleading.

In [34], Tropp introduced **Babel function**, which measures the maximum total coherence between a fixed atom and a collection of other atoms. The *Babel function* quantifies an idea as to how much the atoms of a dictionary are “speaking the same language”.

**Definition 2.21** [Babel function] The *Babel function* for a dictionary  $\mathcal{D}$  is defined by

$$\mu_1(p) \triangleq \max_{|\Lambda|=p} \max_{\psi} \sum_{\Lambda} |\langle \psi, d_{\Lambda} \rangle|, \quad (2.7.14)$$

where the vector  $\psi$  ranges over the atoms indexed by  $\Omega \setminus \Lambda$ . We define

$$\mu_1(0) = 0$$

for sparsity level  $p = 0$ .

Let us understand what is going on here. For each value of  $p$  we consider all possible  $\binom{D}{p}$  subspaces by choosing  $p$  vectors from  $\mathcal{D}$ .

Let the atoms spanning one such subspace be identified by an index set  $\Lambda \subset \Omega$ .

All other atoms are indexed by the index set  $\Gamma = \Omega \setminus \Lambda$ .

Let

$$\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$$

denote the atoms indexed by  $\Gamma$ .

We pickup a vector  $\psi \in \Psi$  and compute its inner product with all atoms indexed by  $\Lambda$ . We compute the sum of absolute value of these inner products over all  $\{d_\lambda : \lambda \in \Lambda\}$ .

We run it for all  $\psi \in \Psi$  and then pickup the maximum value of above sum over all  $\psi$ .

We finally compute the maximum over all possible  $p$ -subspaces.

This number is considered at the Babel number for sparsity level  $p$ .

We first make a few observations over the properties of Babel function.

Babel function is a generalization of coherence.

**REMARK.** For  $p = 1$  we observe that

$$\mu_1(1) = \mu(\mathcal{D})$$

the coherence of  $\mathcal{D}$ .

**REMARK.**  $\mu_1$  is a non-decreasing function of  $p$ .

**PROOF.** This is easy to see since the sum

$$\sum_{\Lambda} |\langle \psi, d_\lambda \rangle|$$

cannot decrease as  $p = |\Lambda|$  increases.

In particular for some value of  $p$  let  $\Lambda^p$  and  $\psi^p$  denote the set and vector for which the maximum in (2.7.14) is achieved. Now pick some column which is not  $\psi^p$  and is not indexed by  $\Lambda^p$  and include it for  $\Lambda^{p+1}$ . Note

that  $\Lambda^{p+1}$  and  $\psi^p$  might not be the worst case for sparsity level  $p + 1$  in (2.7.14). Clearly

$$\sum_{\Lambda^{p+1}} |\langle \psi^p, d_\lambda \rangle| \geq \sum_{\Lambda^p} |\langle \psi^p, d_\lambda \rangle|$$

$\mu_1(p+1)$  cannot be less than  $\mu_1(p)$ .

□

**Lemma 2.31** *Babel function is upper bounded by coherence as per*

$$\mu_1(p) \leq p \mu(\mathcal{D}). \quad (2.7.15)$$

PROOF.

$$\sum_{\Lambda} |\langle \psi, d_\lambda \rangle| \leq p \mu(\mathcal{D}).$$

This leads to

$$\mu_1(p) = \max_{|\Lambda|=p} \max_{\psi} \sum_{\Lambda} |\langle \psi, d_\lambda \rangle| \leq \max_{|\Lambda|=p} \max_{\psi} (p \mu(\mathcal{D})) = p \mu(\mathcal{D}).$$

□

**2.7.3.1. Computation of Babel function.** It might seem at first that computation of Babel function is combinatorial and hence prohibitively expensive. But it is not true.

We will demonstrate this through an example in this section. Our example synthesis matrix will be

$$\mathcal{D} = \begin{bmatrix} 0.5 & 0 & 0 & 0.6533 & 1 & 0.5 & -0.2706 & 0 \\ 0.5 & 1 & 0 & 0.2706 & 0 & -0.5 & 0.6533 & 0 \\ 0.5 & 0 & 1 & -0.2706 & 0 & -0.5 & -0.6533 & 0 \\ 0.5 & 0 & 0 & -0.6533 & 0 & 0.5 & 0.2706 & 1 \end{bmatrix}$$

From the synthesis matrix  $\mathcal{D}$  we first construct its Gram matrix given by

$$G = \mathcal{D}^H \mathcal{D}. \quad (2.7.16)$$

We then take absolute value of each entry in  $G$  to construct  $|G|$ .

For the running example

$$|G| = \begin{bmatrix} 1 & 0.5 & 0.5 & 0 & 0.5 & 0 & 0 & 0.5 \\ 0.5 & 1 & 0 & 0.2706 & 0 & 0.5 & 0.6533 & 0 \\ 0.5 & 0 & 1 & 0.2706 & 0 & 0.5 & 0.6533 & 0 \\ 0 & 0.2706 & 0.2706 & 1 & 0.6533 & 0 & 0 & 0.6533 \\ 0.5 & 0 & 0 & 0.6533 & 1 & 0.5 & 0.2706 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0.5 & 1 & 0 & 0.5 \\ 0 & 0.6533 & 0.6533 & 0 & 0.2706 & 0 & 1 & 0.2706 \\ 0.5 & 0 & 0 & 0.6533 & 0 & 0.5 & 0.2706 & 1 \end{bmatrix}$$

We now sort every row in descending order to obtain a new matrix  $G'$ .

$$G' = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0 & 0 & 0 \\ 1 & 0.6533 & 0.5 & 0.5 & 0.2706 & 0 & 0 & 0 \\ 1 & 0.6533 & 0.5 & 0.5 & 0.2706 & 0 & 0 & 0 \\ 1 & 0.6533 & 0.6533 & 0.2706 & 0.2706 & 0 & 0 & 0 \\ 1 & 0.6533 & 0.5 & 0.5 & 0.2706 & 0 & 0 & 0 \\ 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0 & 0 & 0 \\ 1 & 0.6533 & 0.6533 & 0.2706 & 0.2706 & 0 & 0 & 0 \\ 1 & 0.6533 & 0.5 & 0.5 & 0.2706 & 0 & 0 & 0 \end{bmatrix}$$

First entry in each row is now 1. This corresponds to  $\langle d_i, d_i \rangle$  and it doesn't appear in the calculation of  $\mu_1(p)$  hence we disregard whole of first column.

Now look at column 2 in  $G'$ . In the  $i$ -th row it is nothing but

$$\max_{j \neq i} |\langle d_i, d_j \rangle|.$$

Thus,

$$\mu(\mathcal{D}) = \mu_1(1) = \max_{1 \leq j \leq D} G'_{j,2}$$

i.e. the coherence is given by the maximum in the 2nd column of  $G'$ .

In the running example

$$\mu(\mathcal{D}) = \mu_1(1) = 0.6533.$$

Looking carefully we can note that for  $\psi = d_i$  the maximum value of sum

$$\sum_{\Lambda} |\langle \psi, d_{\lambda} \rangle|$$

while  $|\Lambda| = p$  is given by the sum over elements from 2nd to  $(p+1)$ -th columns in  $i$ -th row.

Thus

$$\mu_1(p) = \max_{1 \leq i \leq D} \sum_{j=2}^{p+1} G'_{ij}.$$

For the running example the Babel function values are given by

$$(0.6533 \ 1.3066 \ 1.6533 \ 2 \ 2 \ 2 \ 2).$$

We see that Babel function stops increasing after  $p = 4$ . Actually  $\mathcal{D}$  is constructed by shuffling the columns of two orthonormal bases. Hence many of the inner products are 0 in  $G$ .

**2.7.3.2. Babel function and spark.** We first note that *Babel function* tells something about linear independence of columns of  $\mathcal{D}$ .

**Lemma 2.32** *Let  $\mu_1$  be the Babel function for a dictionary  $\mathcal{D}$ . If*

$$\mu_1(p) < 1$$

*then all selections of  $p+1$  columns from  $\mathcal{D}$  are linearly independent.*

**PROOF.** We recall from the proof of theorem 2.25 that if

$$p + 1 < 1 + \frac{1}{\mu(\mathcal{D})} \implies p < \frac{1}{\mu(\mathcal{D})}$$

then every set of  $(p+1)$  columns from  $\mathcal{D}$  are linearly independent.

We also know from lemma 2.31 that

$$p \mu(\mathcal{D}) \geq \mu_1(p) \implies \mu(\mathcal{D}) \geq \frac{\mu_1(p)}{p} \implies \frac{1}{\mu(\mathcal{D})} \leq \frac{p}{\mu_1(p)}.$$

Thus if

$$p < \frac{p}{\mu_1(p)} \implies 1 < \frac{1}{\mu_1(p)} \implies \mu_1(p) < 1$$

then all selections of  $p+1$  columns from  $\mathcal{D}$  are linearly independent.  $\square$

This leads us to a lower bound on spark from *Babel function*.

**Lemma 2.33** [Spark lower bound from *Babel function*] *A lower bound of spark of a dictionary  $\mathcal{D}$  is given by*

$$\text{spark}(\mathcal{D}) \geq \min_{1 \leq p \leq N} \{p : \mu_1(p-1) \geq 1\}. \quad (2.7.17)$$

PROOF. For all  $j \leq p-2$  we are given that  $\mu_1(j) < 1$ . Thus all sets of  $p-1$  columns from  $\mathcal{D}$  are linearly independent (using lemma 2.32).

Finally  $\mu_1(p-1) \geq 1$ , hence we cannot say definitively whether a set of  $p$  columns from  $\mathcal{D}$  is linearly dependent or not. This establishes the lower bound on spark.  $\square$

An earlier version of this result also appeared in [20] theorem 6.

### 2.7.3.3. Babel function and singular values.

**Theorem 2.34** [36] *Let  $\mathcal{D}$  be a dictionary and  $\Lambda$  be an index set with  $|\Lambda| = K$ . The singular values of  $\mathcal{D}_\Lambda$  are bounded by*

$$1 - \mu_1(K-1) \leq \sigma^2 \leq 1 + \mu_1(K-1). \quad (2.7.18)$$

PROOF. Consider the Gram matrix

$$G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda.$$

$G$  is a  $K \times K$  square matrix.

Also let

$$\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_K\}$$

so that

$$\mathcal{D}_\Lambda = \begin{bmatrix} d_{\lambda_1} & d_{\lambda_2} & \dots & d_{\lambda_K} \end{bmatrix}.$$

The Gershgorin Disc Theorem states that every eigenvalue of  $G$  lies in one of the  $K$  discs

$$\Delta_k = \left\{ z : |z - G_{kk}| \leq \sum_{j \neq k} |G_{jk}| \right\}$$

Since  $d_i$  are unit norm, hence  $G_{kk} = 1$ .

Also we note that

$$\sum_{j \neq k} |G_{jk}| = \sum_{j \neq k} |\langle d_{\lambda_j}, d_{\lambda_k} \rangle| \leq \mu_1(K-1)$$

since there are  $K-1$  terms in sum and  $\mu_1(K-1)$  is an upper bound on all such sums.

Thus if  $z$  is an eigen value of  $G$  then we have

$$\begin{aligned} |z - 1| &\leq \mu_1(K-1) \\ \implies -\mu_1(K-1) &\leq z - 1 \leq \mu_1(K-1) \\ \implies 1 - \mu_1(K-1) &\leq z \leq 1 + \mu_1(K-1). \end{aligned} \tag{2.7.19}$$

This is OK since  $G$  is positive semi-definite, thus, the eigen values of  $G$  are real.

But the eigen values of  $G$  are nothing but the squared singular values of  $\mathcal{D}_\Lambda$ . Thus we get

$$1 - \mu_1(K-1) \leq \sigma^2 \leq 1 + \mu_1(K-1).$$

□

**Corollary 2.35.** *Let  $\mathcal{D}$  be a dictionary and  $\Lambda$  be an index set with  $|\Lambda| = K$ . If  $\mu_1(K-1) < 1$  then the squared singular values of  $\mathcal{D}_\Lambda$  exceed  $(1 - \mu_1(K-1))$ .*

PROOF. From previous theorem we have

$$1 - \mu_1(K-1) \leq \sigma^2 \leq 1 + \mu_1(K-1).$$

Since the singular values are always non-negative, the lower bound is useful only when  $\mu_1(K-1) < 1$ . When it holds we have

$$\sigma(\mathcal{D}_\Lambda) \geq \sqrt{1 - \mu_1(K-1)}.$$

□

**Theorem 2.36** [35] *Let  $\mu_1(K - 1) < 1$ . If a signal can be written as a linear combination of  $k$  atoms, then any other exact representation of the signal requires at least  $(K - k + 1)$  atoms.*

PROOF. If  $\mu_1(K - 1) < 1$ , then the singular values of any submatrix of  $K$  atoms are non-zero. Thus, the minimum number of atoms required to form a linear dependent set is  $K + 1$ . Let the number of atoms in any other exact representation of the signal be  $l$ . Then

$$k + l \geq K + 1 \implies l \geq K - k + 1.$$

□

**2.7.3.4. Babel function and gram matrix.** Let  $\Lambda$  index a sub-dictionary and let  $G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda$  denote the Gram matrix of the subdictionary  $\mathcal{D}_\Lambda$ . Assume  $K = |\Lambda|$ .

**Theorem 2.37**

$$\|G\|_\infty = \|G\|_1 \leq 1 + \mu_1(K - 1). \quad (2.7.20)$$

PROOF. Since  $G$  is Hermitian, hence the two norms are equal:

$$\|G\|_\infty = \|G^H\|_1 = \|G\|_1.$$

Now each row consists of a diagonal entry 1 and  $K - 1$  off diagonal entries. The absolute sum of all the off-diagonal entries in a row is upper bounded by  $\mu_1(K - 1)$ . Thus, the absolute sum of all the entries in a row is upper bounded by  $1 + \mu_1(K - 1)$ . Since  $\|G\|_\infty$  is nothing but the maximum  $l_1$  norm of rows of  $G$ , hence

$$\|G\|_\infty \leq 1 + \mu_1(K - 1).$$

□

**Theorem 2.38** [35] Suppose that  $\mu_1(K - 1) < 1$ . Then

$$\|G^{-1}\|_\infty = \|G^{-1}\|_1 \leq \frac{1}{1 - \mu_1(K - 1)} \quad (2.7.21)$$

PROOF. Since  $G$  is Hermitian, hence the two operator norms are equal:

$$\|G^{-1}\|_\infty = \|G^{-1}\|_1.$$

As usual we can write  $G$  as  $G = I + A$  where  $A$  consists of off-diagonal entries in  $A$  (recall that since atoms are unit norm, hence diagonal entries in  $G$  are 1).

Each row of  $A$  lists inner products between a fixed atom and  $K - 1$  other atoms (leaving the 0 at the diagonal entry). Therefore

$$\|A\|_{\infty \rightarrow \infty} \leq \mu_1(K - 1)$$

(since  $l_1$  norm of any row is upper bounded by the babel number  $\mu_1(K - 1)$ ). Now  $G^{-1}$  can be written as a Neumann series

$$G^{-1} = \sum_{k=0}^{\infty} (-A)^k.$$

Thus

$$\|G^{-1}\|_\infty = \left\| \sum_{k=0}^{\infty} (-A)^k \right\|_\infty \leq \sum_{k=0}^{\infty} \|(-A)^k\|_\infty = \sum_{k=0}^{\infty} \|A\|_\infty^k = \frac{1}{1 - \|A\|_\infty}.$$

Finally

$$\begin{aligned} \|A\|_\infty \leq \mu_1(K - 1) &\iff 1 - \|A\|_\infty \geq 1 - \mu_1(K - 1) \\ &\iff \frac{1}{1 - \|A\|_\infty} \leq \frac{1}{1 - \mu_1(K - 1)}. \end{aligned}$$

Thus

$$\|G^{-1}\|_\infty \leq \frac{1}{1 - \mu_1(K - 1)}.$$

□

### 2.7.3.5. Quasi incoherent dictionaries.

**Definition 2.22** [Quasi-incoherent dictionary] When the *Babel function* of a dictionary grows slowly, we say that the dictionary is **quasi-incoherent**.

## 2.8. Compressed sensing

In this section we formally define the problem of compressed sensing.

**Compressed sensing** refers to the idea that for sparse or compressible signals, a small number of nonadaptive measurements carries sufficient information to approximate the signal well. In the literature it is also known as **compressive sensing** and **compressive sampling**. Different authors seem to prefer different names. In this book we will stick to the name as compressed sensing.

In this section we will represent a signal dictionary as well as its synthesis matrix as  $\mathcal{D}$ .

We recall the definition of sparse signals from definition 2.4.

A signal  $x \in \mathbb{C}^N$  is  $K$ -sparse in  $\mathcal{D}$  if there exists a representation  $\alpha$  for  $x$  which has at most  $K$  non-zeros. i.e.

$$x = \mathcal{D}\alpha$$

and

$$\|\alpha\|_0 \leq K.$$

The dictionary could be standard basis, Fourier basis, wavelet basis, a wavelet packet dictionary, a multi-ONB or even a randomly generated dictionary.

Real life signals are not sparse, yet they are compressible in the sense that entries in the signal decay rapidly when sorted by magnitude. As a result compressible signals are well approximated by sparse signals. Note that we are talking about the sparsity or compressibility of the

signal in a suitable dictionary. Thus we mean that the signal  $x$  has a representation  $\alpha$  in  $\mathcal{D}$  in which the coefficients decay rapidly when sorted by magnitude.

**Definition 2.23** In compressed sensing, a **measurement** is a linear functional applied to a signal

$$y = \langle x, f \rangle.$$

The compressed sensor makes multiple such linear measurements. This can best be represented by the action of a **sensing matrix**  $\Phi$  on the signal  $x$  given by

$$y = \Phi x. \quad (2.8.1)$$

$\Phi \in \mathbb{C}^{M \times N}$  represents  $M$  different measurements made on the signal  $x$  by the sensing process. Each row of  $\Phi$  represents one linear measurement.

The vector  $y \in \mathbb{C}^M$  is known as **measurement vector**.

$\mathbb{C}^N$  forms the **signal space** while  $\mathbb{C}^M$  forms the **measurement space**.

We also note that above can be written as

$$y = \Phi x = \Phi \mathcal{D} \alpha = (\Phi \mathcal{D}) \alpha.$$

It is assumed that the signal  $x$  is  $K$ -sparse or  $K$ -compressible in  $\mathcal{D}$  and  $K \ll N$ .

The objective is to recover  $x$  from  $y$  given that  $\Phi$  and  $\mathcal{D}$  are known.

We do this by first recovering the sparse representation  $\alpha$  from  $y$  and then computing  $x = \mathcal{D} \alpha$ .

If  $M \geq N$  then the problem is a straight forward least squares problem. So we don't consider it here.

The more interesting case is when  $K < M \ll N$  i.e. the number of measurements is much less than the dimension of the ambient signal space while more than the sparsity level of signal namely  $K$ .

We note that given  $\alpha$  is found, finding  $x$  is straightforward. We therefore can remove the dictionary from our consideration and look at the simplified problem given as: recover  $x$  from  $y$  with

$$y = \Phi x$$

where  $x \in \mathbb{C}^N$  itself is assumed to be  $K$ -sparse or  $K$ -compressible and  $\Phi \in \mathbb{C}^{M \times N}$  is the sensing matrix.

### 2.8.1. The sensing matrix

There are two ways to look at the sensing matrix. First view is in terms of its columns

$$\Phi = [\phi_1 \ \phi_2 \ \dots \ \phi_N] \quad (2.8.2)$$

where  $\phi_i \in \mathbb{C}^M$  are the columns of sensing matrix. In this view we see that

$$y = \sum_{i=1}^N x_i \phi_i$$

i.e.  $y$  belongs to the column span of  $\Phi$  and one representation of  $y$  in  $\Phi$  is given by  $x$ .

This view looks very similar to a dictionary and its atoms but there is a difference. In a dictionary, we require each atom to be unit norm. We don't require columns of the sensing matrix  $\Phi$  to be unit norm.

The second view of sensing matrix  $\Phi$  is in terms of its columns. We write

$$\Phi = \begin{bmatrix} \chi_1^H \\ \chi_2^H \\ \vdots \\ \chi_M^H \end{bmatrix} \quad (2.8.3)$$

where  $\chi_i \in \mathbb{C}^N$  are conjugate transposes of rows of  $\Phi$ . This view gives us following result

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} \chi_1^H \\ \chi_2^H \\ \vdots \\ \chi_M^H \end{bmatrix} x = \begin{bmatrix} \chi_1^H x \\ \chi_2^H x \\ \vdots \\ \chi_M^H x \end{bmatrix} = \begin{bmatrix} \langle x, \chi_1 \rangle \\ \langle x, \chi_2 \rangle \\ \vdots \\ \langle x, \chi_M \rangle \end{bmatrix} \quad (2.8.4)$$

In this view  $y_i$  is a measurement given by the inner product of  $x$  with  $\chi_i$  ( $\langle x, \chi_i \rangle = \chi_i^H x$ ).

We will call  $\chi_i$  as a **sensing vector**. There are  $M$  such sensing vectors in  $\mathbb{C}^N$  comprising  $\Phi$  corresponding to  $M$  measurements in the measurement space  $\mathbb{C}^M$ .

### 2.8.2. Number of measurements

A fundamental question of compressed sensing framework is: *How many measurements are necessary to acquire  $K$ -sparse signals?* By necessary we mean that  $y$  carries enough information about  $x$  such that  $x$  can be recovered from  $y$ .

Clearly if  $M < K$  then recovery is not possible.

We further note that the sensing matrix  $\Phi$  should not map two different  $K$ -sparse signals to the same measurement vector. Thus we will need  $M \geq 2K$  and each collection of  $2K$  columns in  $\Phi$  must be non-singular.

If the  $K$ -column sub matrices of  $\Phi$  are badly conditioned, then it is possible that some sparse signals get mapped to very similar measurement vectors. Thus it is numerically unstable to recover the signal. Moreover, if noise is present, stability further degrades.

In [11] Candès and Tao showed that the geometry of sparse signals should be preserved under the action of a sensing matrix. In particular the distance between two sparse signals shouldn't change by much during sensing.

They quantified this idea in the form of a *restricted isometric constant* of a matrix  $\Phi$  as the smallest number  $\delta_K$  for which the following holds

$$(1 - \delta_K)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K)\|x\|_2^2 \quad \forall x : \|x\|_0 \leq K.$$

We will study more about this property known as restricted isometry property (RIP) in section 3.1. Here we just sketch the implications of RIP for compressed sensing.

When  $\delta_K < 1$  then the inequalities imply that every collection of  $K$  columns from  $\Phi$  is non-singular. Since we need every collection of  $2K$  columns to be non-singular, we actually need  $\delta_{2K} < 1$  which is the minimum requirement for recovery of  $K$  sparse signals.

Further if  $\delta_{2K} \ll 1$  then we note that sensing operator very nearly maintains the  $l_2$  distance between any two  $K$  sparse signals. In consequence, it is possible to invert the sensing process stably.

It is now known that many randomly generated matrices have excellent RIP behavior. One can show that if  $\delta_{2K} \leq 0.1$ , then with

$$M = \mathcal{O}(K \ln^\alpha N)$$

measurements, one can recover  $x$  with high probability.

Some of the typical random matrices which have suitable RIP properties are

- Gaussian sensing matrices
- Partial Fourier matrices
- Rademacher sensing matrices

### 2.8.3. Signal recovery

The second fundamental problem in compressed sensing is: *Given the compressed measurements  $y$  how do we recover the signal  $x$ ?* This problem is known as SPARSE-RECOVERY problem.

A simple formulation of the problem as: minimize  $\|x\|_0$  subject to  $y = \Phi x$  is hopeless since it entails a combinatorial explosion of search space.

Over the years, people have developed a number of algorithms to tackle the sparse recovery problem.

The algorithms can be broadly classified into following categories

**Greedy pursuits:** These algorithms attempt to build the approximation of the signal iteratively by making locally optimal choices at each step. Examples of such algorithms include OMP (orthogonal matching pursuit), stage-wise OMP, regularized OMP, CoSaMP (compressive sampling pursuit) and IHT (iterative hard thresholding).

**Convex relaxation:** These techniques relax the  $l_0$  “norm” minimization problem into a suitable problem which is a convex optimization problem. This relaxation is valid for a large class of signals of interest. Once the problem has been formulated as a convex optimization problem, a number of solutions are available, e.g. interior point methods, projected gradient methods and iterative thresholding.

**Combinatorial algorithms:** These methods are based on research in group testing and are specifically suited for situations where highly structured measurements of the signal are taken. This class includes algorithms like Fourier sampling, chaining pursuit, and HHS pursuit.

A major emphasis of the following chapters will be the study of these sparse recovery algorithms.

In the following we present examples of real life problems which can be modeled as compressed sensing problems.

### 2.8.4. Error correction in linear codes

The classical error correction problem was discussed in one of the seminal founding papers on compressed sensing [10].

Let  $f \in \mathbb{R}^N$  be a “plaintext” message being sent over a communication channel.

In order to make the message robust against errors in communication channel, we encode the error with an error correcting code.

We consider  $A \in \mathbb{R}^{D \times N}$  with  $D > N$  as a **linear code**.  $A$  is essentially a collection of code words given by

$$A = \begin{bmatrix} a_1 & a_2 & \dots & a_N \end{bmatrix} \quad (2.8.5)$$

where  $a_i \in \mathbb{R}^D$  are the codewords.

We construct the “ciphertext”

$$x = Af \quad (2.8.6)$$

where  $x \in \mathbb{R}^D$  is sent over the communication channel. Clearly  $x$  is a redundant representation of  $f$  which is expected to be robust against small errors during transmission.

$A$  is assumed to be full column rank. Thus  $A^T A$  is invertible and we can easily see that

$$f = A^\dagger x$$

where

$$A^\dagger = (A^T A)^{-1} A^T$$

is the left pseudo inverse of  $A$ .

But naturally the communication channel is going to add some error. What we actually receive is

$$y = x + e = Af + e \quad (2.8.7)$$

where  $e \in \mathbb{R}^D$  is the error being introduced by the channel.

The least squares solution by minimizing the error  $l_2$  norm is given by

$$f' = A^\dagger y = A^\dagger(Af + e) = f + A^\dagger e.$$

Since  $A^\dagger e$  is usually non-zero (we cannot assume that  $A^\dagger$  will annihilate  $e$ ), hence  $f'$  is not an exact replica of  $f$ .

What is needed is an exact reconstruction of  $f$ . To achieve this, a common assumption in literature is that error vector  $e$  is in fact sparse. i.e.

$$\|e\|_0 \leq K \ll D. \quad (2.8.8)$$

To reconstruct  $f$  it is sufficient to reconstruct  $e$  since once  $e$  is known we can get

$$x = y - e$$

and from there  $f$  can be faithfully reconstructed.

The question is: for a given sparsity level  $K$  for the error vector  $e$  can one reconstruct  $e$  via practical algorithms? By practical we mean algorithms which are of polynomial time w.r.t. the length of “ciphertext” ( $D$ ).

The approach in [10] is as follows.

We construct a matrix  $F \in \mathbb{R}^{M \times D}$  which can annihilate  $A$  i.e.

$$FA = 0.$$

We then apply  $F$  to  $y$  giving us

$$\tilde{y} = F(Af + e) = Fe.$$

Therefore the decoding problem is reduced to that of reconstructing a sparse vector  $e \in \mathbb{R}^D$  from the measurements  $Fe \in \mathbb{R}^M$  where we would like to have  $M \ll D$ .

With this the problem of finding  $e$  can be cast as problem of finding a sparse solution for the under-determined system given by

$$\begin{aligned} & \underset{e \in \Sigma_K}{\text{minimize}} \quad \|e\|_0 \\ & \text{subject to} \quad \tilde{y} = Fe \end{aligned} \tag{P_0}$$

This now becomes the compressed sensing problem. The natural questions are

- How many measurements  $M$  are necessary (in  $F$ ) to be able to recover  $e$  exactly?
- How should  $F$  be constructed?
- How do we recover  $e$  from  $\tilde{y}$ ?

## 2.9. Examples

In this section we will look at several examples which can be modeled using sparse and redundant representations and measured using compressed sensing techniques.

Several examples in this section have been incorporated from Sparco [6] (a testing framework for sparse reconstruction).

### 2.9.1. Piecewise cubic polynomial signal

This example was discussed in [9]. Our signal of interest is a piecewise cubic polynomial signal as shown in fig. 2.6. It has a sparse representation in a wavelet basis (fig. 2.7). We can sort the wavelet coefficients by magnitude and plot them in descending order to visualize how sparse the representation is in fig. 2.8. The chosen basis is a Daubechies wavelet basis  $\Psi$  fig. 2.10. A Gaussian random sensing matrix  $\Phi$  (fig. 2.11) is used to generate the measurement vector  $y$  (fig. 2.9). Finally the product of  $\Phi$  and  $\Psi$  given by  $\Phi\Psi$  will be used for actual recovery of sparse representation  $\alpha$  from the measurements  $y$  (fig. 2.12). Fundamental equations are:

$$x = \Psi\alpha$$

and

$$y = \Phi x + e = \Phi \Psi \alpha + e.$$

$x \in \mathbb{R}^N$ . In this example  $N = 2048$ .  $\Psi$  is a complete dictionary of size  $N \times N$ . Thus we have  $D = N$  and  $\alpha \in \mathbb{R}^N$ .  $\Phi \in \mathbb{R}^{M \times N}$ . In this example, the number of measurements  $M = 600$ . The measurement vector  $y \in \mathbb{R}^M$ . For this problem we chose  $e = 0$ .

Sparse signal recovery problem is denoted as

$$\hat{\alpha} = \text{recovery}(\Phi \Psi, y, K).$$

where  $\hat{\alpha}$  is a  $K$ -sparse approximation of  $\alpha$ .

Closely examining the coefficients in  $\alpha$  we can note that  $\max(\alpha_i) = 78.0546$ . Further if we put different thresholds over magnitudes of entries in  $\alpha$  we can find the number of coefficients higher than the threshold as listed in table 2. A choice of  $M = 600$  looks quite reasonable given the decay of entries in  $\alpha$ .

TABLE 2. Entries in wavelet representation of piecewise cubic polynomial signal higher than a threshold

Threshold	Entries higher than threshold
1	129
1E-1	173
1E-2	186
1E-4	197
1E-8	199
1E-12	200

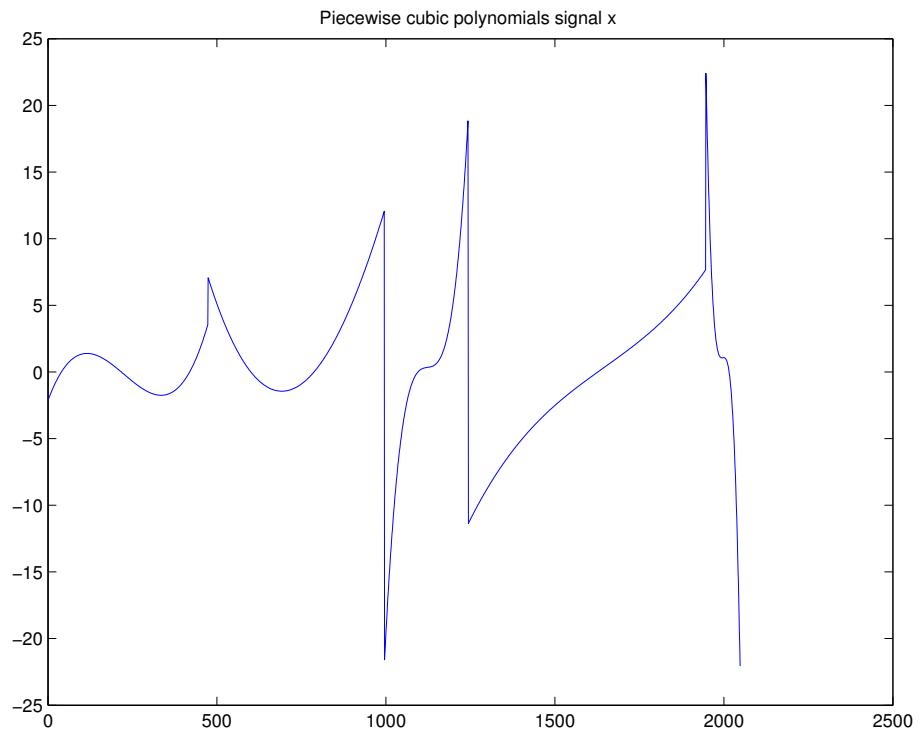


FIGURE 2.6. A piecewise cubic polynomials signal

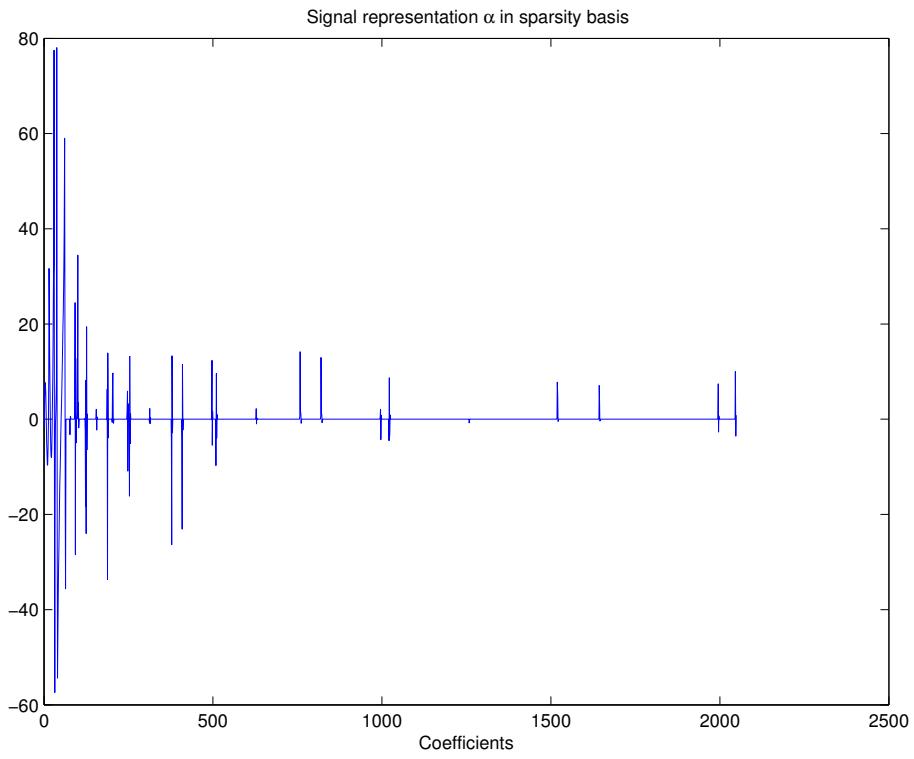


FIGURE 2.7. Sparse representation of signal in wavelet basis

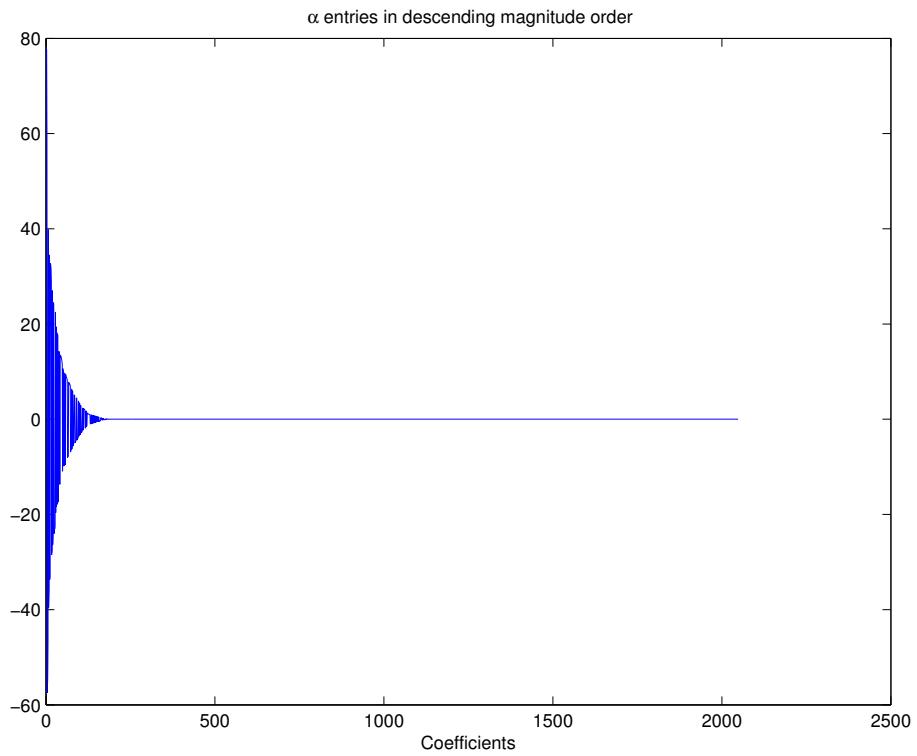


FIGURE 2.8. Wavelet coefficients sorted by magnitude

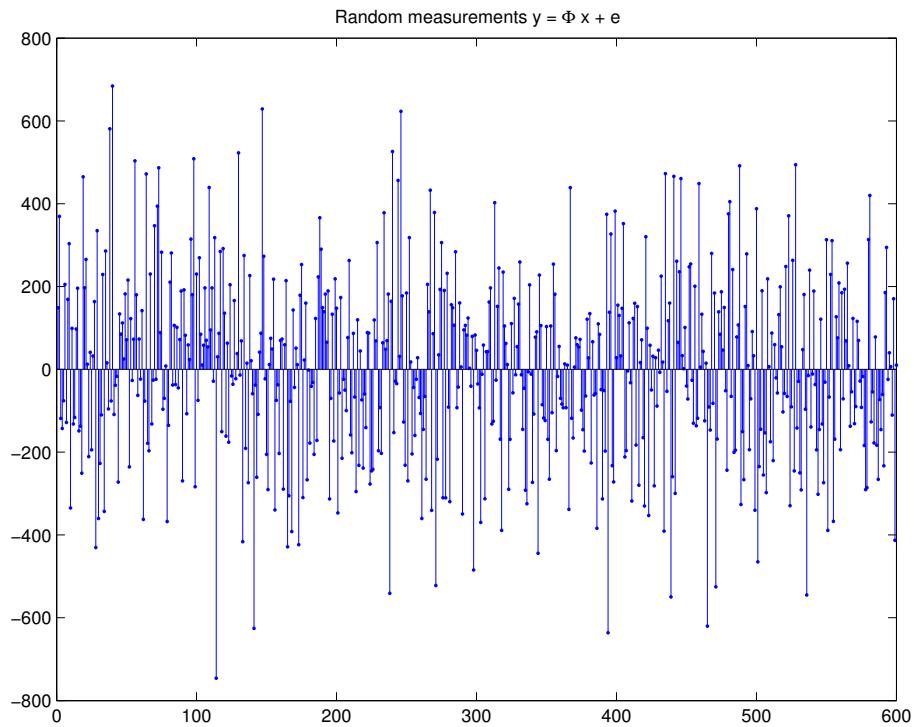


FIGURE 2.9. Measurement vector  $y = \Phi x + e$

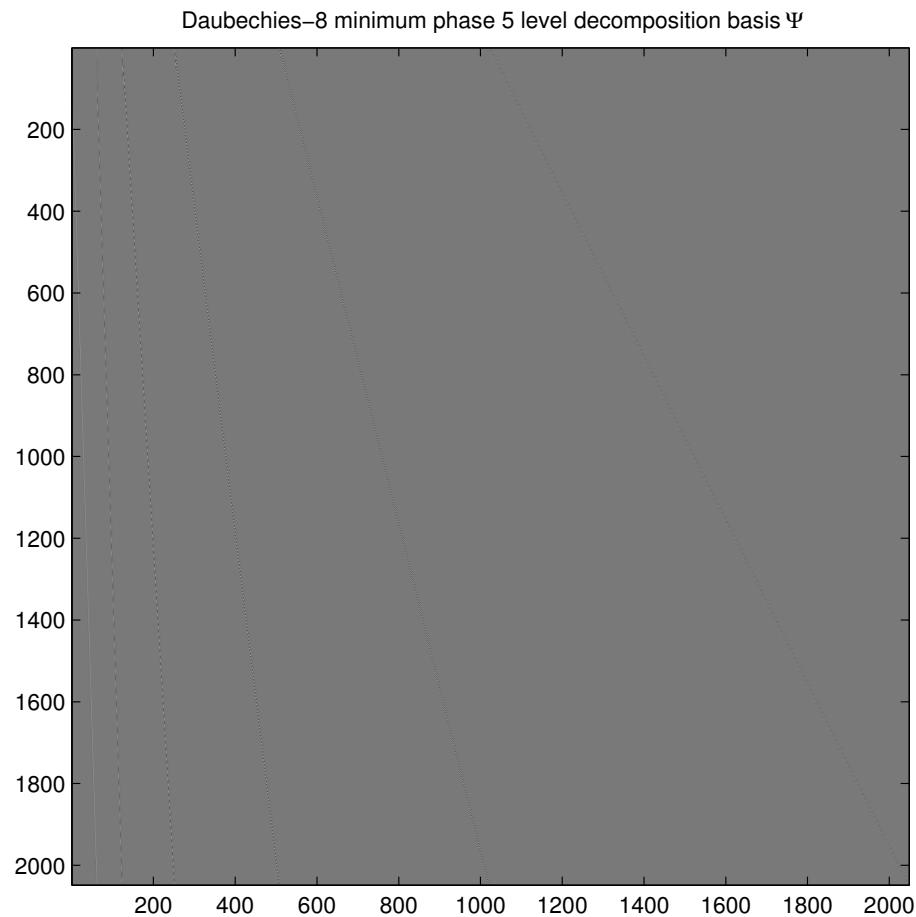


FIGURE 2.10. Daubechies-8 wavelet basis

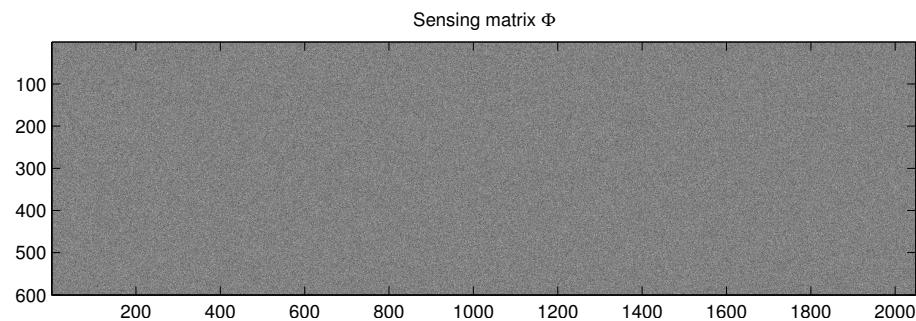


FIGURE 2.11. Gaussian sensing matrix  $\Phi$

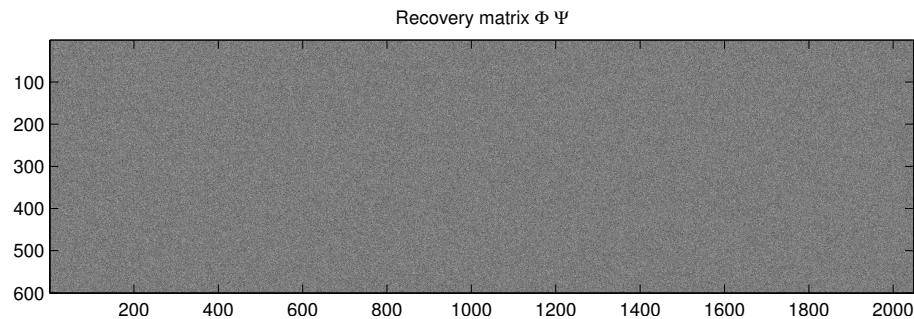


FIGURE 2.12. Recovery matrix  $\Phi\Psi$

## 2.10. Digest

This section summarizes results in this chapter.

$l_2$  regularization

$$\begin{aligned} & \underset{x}{\text{minimize}} \quad \|x\|_2 \\ & \text{subject to} \quad y = \Phi x. \end{aligned} \tag{P}_2$$

Lagrangian:

$$\mathcal{L}(x) = \|x\|_2^2 + \lambda^H(\Phi x - y)$$

Optimal solution:

$$x^* = \Phi^H(\Phi\Phi^H)^{-1}y = \Phi^\dagger y.$$

$l_1$  regularization

$$\begin{aligned} & \underset{x}{\text{minimize}} \quad \|x\|_1 \\ & \text{subject to} \quad \Phi x = y. \end{aligned} \tag{P}_1$$

$\|x\|_1$  is not strictly convex.

At least one solution to  $l_1$  minimization has  $\|x\|_0 = M$ . Also for real case, it can be shown to be a linear programming problem.

**Two orthonormal bases  $\Psi$  and  $\mathcal{X}$ .**

Mutual coherence or proximity

$$\mu(\Psi, \mathcal{X}) = \max_{1 \leq i, j \leq N} |\langle \psi_i, \chi_j \rangle|.$$

Mutual coherence bounds

$$\frac{1}{\sqrt{N}} \leq \mu(\Psi, \mathcal{X}) \leq 1.$$

Mutual coherence of Dirac and Fourier bases

$$\mu(I, F) = \frac{1}{\sqrt{N}}.$$

Uncertainty principle for two ortho bases  $x = \Psi\alpha = \mathcal{X}\beta$ .

$$\|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu(\Psi, \mathcal{X})}.$$

And for unit norm vectors  $x$ :

$$\|\alpha\|_1 + \|\beta\|_1 \geq \frac{2}{\sqrt{\mu(\Psi, \mathcal{X})}}.$$

**Two ortho basis  $\mathcal{H} = [\Psi \quad \mathcal{X}]$**

Null space vectors and sparsity:

$$\|v\|_0 \geq \frac{2}{\mu(\mathcal{H})} \quad \forall v \in \mathcal{N}(\mathcal{H}).$$

Uncertainty principle for two ortho basis:  $x = \mathcal{H}\alpha = \mathcal{H}\beta$

$$\|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu(\mathcal{H})}.$$

Uniqueness principle for two ortho basis:

$$\|\alpha\|_0 < \frac{1}{\mu(\mathcal{H})}$$

implies that  $\alpha$  is sparsest (and unique) representation.

### Sparse and redundant representations

**Dictionary**

$$\mathcal{D} = \{d_\omega : \omega \in \Omega\}$$

with  $\|d_\omega\|_2 = 1$  and  $\mathcal{D}$  spans  $\mathbb{C}^N$ .  $D = |\Omega|$ . Thus the set of  $(\mathcal{D}, K)$ -sparse signals is given by

$$\Sigma_{(\mathcal{D}, K)} = \left\{ x \in \mathbb{C}^N : x = \sum_{\lambda \in \Lambda} b_\lambda d_\lambda \right\}.$$

for some index set  $\Lambda \subset \Omega$  with  $|\Lambda| = K$ .

**Sparse approximation problem:**

$$\min_{|\Lambda|=K} \min_b \left\| x - \sum_{\lambda \in \Lambda} b_\lambda d_\lambda \right\|_2.$$

**$(\mathcal{D}, K)$ -EXACT-SPARSE problem**

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} && \|\alpha\|_0 \\ & \text{subject to} && x = \Phi\alpha \\ & \text{and} && \|\alpha\|_0 \leq K \end{aligned}$$

**$(\mathcal{D}, K)$ -SPARSE approximation problem**

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} && \|x - \Phi\alpha\|_2 \\ & \text{subject to} && \|\alpha\|_0 \leq K. \end{aligned}$$

A **Sub-dictionary** is a linearly independent collection of atoms.  $\Lambda$  index set for subdictionary,  $\mathcal{D}_\Lambda$  represents the subdictionary. We say  $K = |\Lambda|$ .  $\mathcal{D}_\Lambda$  is full rank.  $G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda$  is its Gram matrix which is positive definite. The inverse Gram matrix lists the inner products between the dual vectors.

**norm relationships** Sign vector for a complex vector

$$\text{sgn}(v_i) = \begin{cases} \exp(j\angle v_i) & \text{if } |v_i| \neq 0; \\ 0 & \text{if } |v_i| = 0. \end{cases}$$

$l_1$  norm as inner product with sign vector:

$$\|v\|_1 = \text{sgn}(v)^H v = \langle v, \text{sgn}(v) \rangle.$$

Bounds on  $l_1$  norm w.r.t.  $l_2$  norm

$$\|v\|_2 \leq \|v\|_1 \leq \sqrt{N}\|v\|_2.$$

Bound on  $l_2$  norm w.r.t  $l_\infty$ -norm

$$\|v\|_2 \leq \sqrt{N}\|v\|_\infty.$$

Relationship between arbitrary norms

$$\|v\|_q \leq \|v\|_p \text{ whenever } p \leq q.$$

$l_1$  norm as inner product with  $\mathbf{1}$ :

$$\begin{aligned} \|v\|_1 &= \mathbf{1}^T |v| = \mathbf{1}^H |v|. \\ v^H \mathbf{1} v &= \|v\|_1^2. \end{aligned}$$

**The set of sparse signals**

$$\Sigma_K = \{x \in \mathbb{C}^N : \|x\|_0 \leq K\}.$$

Norm relationships for  $K$  sparse signals:

$$\frac{\|u\|_1}{\sqrt{K}} \leq \|u\|_2 \leq \sqrt{K}\|u\|_\infty.$$

**Compressible signals**  $K$ -term approximation  $x|_K$  is a best  $K$ -term approximation of  $x$ .

**Compressible signal:**  $x$  is some signal and  $\hat{x}$  is a permutation of  $x$  in the descending magnitude order. If

$$|\hat{x}_i| \leq R \cdot i^{-\frac{1}{p}} \quad \forall i = 1, 2, \dots, N.$$

then  $x$  is compressible.

**Tools for dictionary analysis**

**Spark** is minimum number of linearly dependent columns.

**Uniqueness-Spark sufficient condition:**

$$\|x^*\|_0 < \frac{\text{spark}(\mathcal{D})}{2}$$

**Coherence** of a dictionary  $\mathcal{D}$ :

$$\mu = \max_{j \neq k} |\langle d_{\omega_j}, d_{\omega_k} \rangle| = \max_{j \neq k} |(\mathcal{D}^H \mathcal{D})_{jk}|.$$

**Coherence for arbitrary matrix**

$$\mu(A) = \max_{j \neq k} \frac{|\langle a_j, a_k \rangle|}{\|a_j\|_2 \|a_k\|_2}.$$

**Coherence based lower bound for spark:**

$$\text{spark}(\Phi) \geq 1 + \frac{1}{\mu(\Phi)}.$$

**Lower bound for spark of two ortho basis**

$$\text{spark}(\mathcal{D}) \geq \frac{2}{\mu(\mathcal{D})}.$$

**Uniqueness-Coherence:**

$$\|x^*\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu(\mathcal{D})} \right)$$

sufficient condition for guaranteeing uniqueness.

**Bounds on eigen values of Gram matrices for sub-dictionaries:**  $K = |\Lambda|$ .  $G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda$ .

$$1 - (K - 1)\mu \leq \lambda(G) \leq 1 + (K - 1)\mu. \quad (2.10.1)$$

**Bounds on singular values of sub-dictionaries:**

$$\sqrt{1 - (K - 1)\mu} \leq \sigma(\mathcal{D}_\Lambda) \leq \sqrt{1 + (K - 1)\mu}. \quad (2.10.2)$$

**Coherence bounds for norms of embeddings of sparse vectors using sub-dictionaries**

$$(1 - (K - 1)\mu) \|v\|_2^2 \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq (1 + (K - 1)\mu) \|v\|_2^2.$$

**Babel function**

$$\mu_1(p) \triangleq \max_{|\Lambda|=p} \max_{\psi} \sum_{\Lambda} |\langle \psi, d_\lambda \rangle|,$$

where the vector  $\psi$  ranges over the atoms indexed by  $\Omega \setminus \Lambda$ .

**Babel function upper bound from coherence:**

$$\mu_1(p) \leq p \mu(\mathcal{D}).$$

**Lower bound on spark from Babel function:**

$$\text{spark}(\mathcal{D}) \geq \min_{1 \leq p \leq N} \{p : \mu_1(p - 1) \geq 1\}.$$

Babel function and singular values of subdictionaries:  $|\Lambda| = K$ .

$$1 - \mu_1(K - 1) \leq \sigma^2(\mathcal{D}_\Lambda) \leq 1 + \mu_1(K - 1).$$

Babel function and singular value lower bound:

$$\mu_1(K - 1) < 1 \implies \sigma^2(\mathcal{D}_\Lambda) \geq 1 - \mu_1(K - 1).$$

Babel function and uncertainty principle: If  $\mu_1(K - 1) < 1$  and a signal has a sparse representation with  $k$  non-zero entries, then any other representation will have at least  $K - k + 1$  non-zero entries.

Norms of Gram matrix of subdictionary:

$$\|G\|_\infty = \|G\|_1 \leq 1 + \mu_1(K - 1). \quad (2.10.3)$$

Norms of inverse Gram matrix of subdictionary:

$$\|G^{-1}\|_\infty = \|G^{-1}\|_1 \leq \frac{1}{1 - \mu_1(K - 1)}$$

given that  $\mu_1(K - 1) < 1$ .

## CHAPTER 3

# More Tools for Dictionary and Random Matrix Analysis

In this chapter we cover the results for a wide variety of tools which are used for analysis of dictionaries, random matrices (like sensing matrices). The tools include restricted isometry property, coherence, spark, etc..

We also spend time in deeper understanding of concepts related to dimensionality reduction from high dimensional spaces to lower dimensional spaces using random projections. We delve upon the notion of stable embeddings of points from high dimensional spaces to lower dimensional spaces.

### 3.1. Restricted isometry property

**Definition 3.1** A matrix  $\Phi \in \mathbb{C}^{M \times N}$  is said to satisfy the **RIP (restricted isometry property)** of order  $K$  with a constant  $\delta \in (0, 1)$  if the following holds:

$$(1 - \delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta)\|x\|_2^2 \quad (3.1.1)$$

for every  $x \in \Sigma_K$  where

$$\Sigma_K = \{x \in \mathbb{C}^N : \|x\|_0 \leq K\}$$

is the set of all  $K$ -sparse vectors in  $\mathbb{C}^N$ .

**Definition 3.2** [Restricted isometry constant] If a matrix  $\Phi \in \mathbb{C}^{M \times N}$  satisfies RIP of order  $K$  then the smallest value of  $\delta$  (denoted as  $\delta_K$ ) for which the following holds

$$(1 - \delta) \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta) \|x\|_2^2 \quad \forall x \in \Sigma_K \quad (3.1.2)$$

is known as the  **$K$ -th restricted isometry constant** for  $\Phi$ . It is also written in short as  **$K$ -th RIP constant**.

We write the bounds as in terms of  $\delta_K$  as

$$(1 - \delta_K) \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2 \quad \forall x \in \Sigma_K. \quad (3.1.3)$$

Some remarks are in order.

- $\Phi$  maps a vector  $x \in \Sigma_K \subseteq \mathbb{C}^N$  into  $\mathbb{C}^M$  as a vector  $\Phi x$  (usually  $M < N$ ).
- We will call  $\Phi x \in \mathbb{C}^M$  as an **embedding** of  $x \in \mathbb{C}^N$  into  $\mathbb{C}^M$ .
- RIP quantifies the idea as to how much the squared length of a sparse signal changes during this embedding process.
- We can compare matrices satisfying RIP with orthonormal bases. An orthonormal basis or the corresponding unitary matrix preserves the length of a vector exactly (see ??). A matrix  $\Phi$  satisfying RIP of order  $K$  is able to preserve the length of  $K$  sparse signals approximately (the approximation range given by  $\delta_K$ ). In this sense we can say that  $\Phi$  implements a **restricted almost orthonormal system** [10]. By restricted we mean that orthonormality is limited to  $K$ -sparse signals. By almost we mean that the squared length is not preserved exactly. Rather it is preserved approximately.
- An arbitrary matrix  $\Phi$  need not satisfy RIP of any order at all.
- If  $\Phi$  satisfies RIP of order  $K$  then it is easy to see that  $\Phi$  satisfies RIP of any order  $L < K$  (since  $\Sigma_L \subset \Sigma_K$  whenever  $L < K$ ).
- If  $\Phi$  satisfies RIP of order  $K$  then it may or many not satisfy RIP of order  $L > K$ .

- Restricted isometry constant is a function of sparsity level  $K$  of the signal  $x \in \mathbb{C}^N$ .

**Example 3.1: Restricted isometry constant** As a running example in this section we will use following matrix

$$\Phi = \frac{1}{2} \begin{bmatrix} 1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 \\ -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix} \in \mathbb{R}^{4 \times 8}.$$

Consider

$$x = \begin{pmatrix} -2 & 0 & 0 & 0 & 0 & -3 & -1 & 0 \end{pmatrix}$$

which is a 3-sparse vector in  $\mathbb{R}^8$ .

We have

$$y = \Phi x = \begin{pmatrix} 0 & -1 & 3 & 3 \end{pmatrix}$$

Now

$$\begin{aligned} \|x\|_2^2 &= 14, & \|x\|_2 &= 3.7417 \\ \|y\|_2^2 &= 19, & \|y\|_2 &= 4.3589 \end{aligned}$$

We note that

$$\frac{\|y\|_2^2}{\|x\|_2^2} = 1.3571.$$

With this much information, all we can say that  $\delta_3 \geq .3571$  for this matrix  $\Phi$  since we haven't examined all possible 3-sparse vectors.

Still what is comforting to note is that for this particular example, the distance hasn't increased by a large factor.  $\square$

For a given  $K$ -sparse vector  $x$ , let  $J$  denote the support of  $x$ , i.e.

$$J = \{1 \leq i \leq N : x_i \neq 0\}.$$

In the running example

$$J = \{1, 6, 7\}$$

1143. MORE TOOLS FOR DICTIONARY AND RANDOM MATRIX ANALYSIS

We define  $x_J \in \mathbb{C}^K$  to be the vector formed by keeping the elements in  $x$  indexed by  $J$  and dropping of other elements (the zero elements). Note that the order of elements is preserved.

In the running example,

$$x_J = \begin{pmatrix} -2 & -3 & -1 \end{pmatrix}$$

Let  $\Phi_J$  be the corresponding sub-matrix by choosing columns from  $\Phi$  indexed by the set  $J$ . Note that the order of elements is preserved.

In the running example

$$\Phi_J = \frac{1}{2} \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{bmatrix} \in \mathbb{R}^{4 \times 3}.$$

It is easy to see that

$$y = \Phi x = \Phi_J x_J. \quad (3.1.4)$$

There are  $\binom{N}{K}$  ways of choosing a  $K$ -sparse support for  $x$ . Thus we have to consider  $\binom{N}{K}$  corresponding sub-matrices  $\Phi_J$ .

For each such sub-matrix  $\Phi_J$ , the RIP bounds can be rewritten as

$$(1 - \delta_K) \|x\|_2^2 \leq \|\Phi_J x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2 \quad (3.1.5)$$

for every  $x \in \mathbb{C}^N$ .

Note that

$$\|\Phi_J x\|_2^2 = (\Phi_J x)^H (\Phi_J x) = x^H \Phi_J^H \Phi_J x. \quad (3.1.6)$$

**Theorem 3.1** *An  $M \times N$  matrix  $\Phi$  cannot satisfy RIP of order  $K > M$ .*

PROOF. Since every  $\phi_j \in \mathbb{C}^M$  hence any set of  $M + 1$  columns in  $\Phi$  is linearly dependent. Thus there exists a non-zero  $M + 1$  sparse signal  $x \in \mathbb{C}^N$  such that  $\Phi x = 0$  (it belongs to the null space of the

chosen  $M + 1$  columns). RIP (3.1.1) requires that a non-zero vector be embedded as a non-zero vector. Thus  $\Phi$  cannot satisfy RIP of order  $M + 1$ . The argument can be easily extended for any  $K > M$ .  $\square$

**Theorem 3.2** *If  $\Phi$  satisfies RIP of order  $l$  then it satisfies RIP of order  $k$  where  $k < l$ .*

PROOF. Every  $k$  sparse signal is also  $l$  sparse signal. Thus if  $\Phi$  satisfies RIP of order  $l$  then it automatically satisfies RIP of order  $k < l$ .  $\square$

**Theorem 3.3** *Let  $\Phi$  satisfy RIP of order  $k$  and  $l$  where  $k < l$ . Then  $\delta_k \leq \delta_l$ . In other words, restricted isometry constants are non-decreasing.*

PROOF. Since every  $k$  sparse signal is also  $l$  sparse signal, hence for every  $x \in \Sigma_k$  following must be satisfied

$$(1 - \delta_k)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_k)\|x\|_2^2$$

and

$$(1 - \delta_l)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_l)\|x\|_2^2.$$

Since  $\delta_k$  is smallest such value for which these inequalities are satisfied hence  $\delta_l$  cannot be smaller than  $\delta_k$ .  $\square$

### 3.1.1. The first restricted isometry constant

We consider the simplest case where  $K = 1$ . We can write  $\Phi$  in terms of its column vectors

$$\Phi = [\phi_1 \ \dots \ \phi_N]$$

Now a 1-sparse vector  $x$  consists of only one non-zero value. Say that  $x$  is non-zero at index  $j$ . Then  $\Phi x$  is nothing but  $x_j \phi_j$ . With this the restricted isometry inequality can be written as

$$(1 - \delta_1)|x_j|^2 \leq \|x_j \phi_j\|_2^2 \leq (1 + \delta_1)|x_j|^2.$$

Dividing by  $|x_j|^2$  we get

$$(1 - \delta_1) \leq \|\phi_j\|_2^2 \leq (1 + \delta_1).$$

Let us formalize this in the following theorem.

**Theorem 3.4** *If a matrix  $\Phi$  satisfies RIP of order  $K \geq 1$  then the squared lengths of columns of  $\Phi$  satisfy the following bounds*

$$1 - \delta_1 \leq \|\phi_j\|_2^2 \leq 1 + \delta_1 \quad \forall 1 \leq j \leq N. \quad (3.1.7)$$

When  $\delta_1 = 0$  then all columns of  $\Phi$  are unit norm. Now if columns of  $\Phi$  span  $\mathbb{C}^M$  then  $\Phi$  can also be considered as a dictionary for  $\mathbb{C}^M$  (see definition 2.3).

REMARK. A dictionary (definition 2.3) satisfies RIP of order 1 with  $\delta_1 = 0$ .

### 3.1.2. Sums and differences of sparse vectors

**Theorem 3.5** *Let  $x, y \in \mathbb{C}^N$  with  $x \in \Sigma_k$  and  $y \in \Sigma_l$ . i.e.  $\|x\|_0 \leq k$  and  $\|y\|_0 \leq l$ . Then*

$$(1 - \delta_{k+l})\|x \pm y\|_2^2 \leq \|\Phi x \pm \Phi y\|_2^2 \leq (1 + \delta_{k+l})\|x \pm y\|_2^2 \quad (3.1.8)$$

*as long as  $\Phi$  satisfies RIP of order  $k + l$ .*

PROOF. We know that

$$\|x \pm y\|_0 \leq \|x\|_0 + \|y\|_0 = k + l.$$

Thus  $x \pm y \in \Sigma_{k+l}$ . The result follows.  $\square$

### 3.1.3. Distance between sparse vectors

Let  $x, y \in \Sigma_K$ . Then clearly  $x - y \in \Sigma_{2K}$ .

The  $l_2$  distance between vectors is given by

$$d(x, y) = \|x - y\|_2 = \sqrt{(x - y)^H(x - y)}$$

Now if  $\Phi$  satisfies RIP of order  $2K$  then we can see that it approximately preserves  $l_2$  distances between  $K$ -sparse vectors.

**Theorem 3.6** *Let  $x, y \in \Sigma_K \subset \mathbb{C}^N$ . Let  $\Phi x, \Phi y \in \mathbb{C}^M$  be corresponding embeddings. If  $\Phi$  satisfies RIP of order  $2K$ , then*

$$(1 - \delta_{2K})d^2(x, y) \leq d^2(\Phi x, \Phi y) \leq (1 + \delta_{2K})d^2(x, y). \quad (3.1.9)$$

PROOF. Since  $\Phi$  satisfies RIP of order  $2K$  hence for every vector  $v \in \Sigma_{2K}$  we have

$$(1 - \delta_{2K})\|v\|_2^2 \leq \|\Phi v\|_2^2 \leq (1 + \delta_{2K})\|v\|_2^2.$$

But then  $x - y \in \Sigma_{2K}$  for every  $x, y \in \Sigma_K$  and

$$d^2(x, y) = \|x - y\|_2$$

and

$$d^2(\Phi x, \Phi y) = \|\Phi x - \Phi y\|_2^2 = \|\Phi(x - y)\|_2^2.$$

Thus we have the result.  $\square$

### 3.1.4. RIP with unit length sparse vectors

Sometimes it is convenient to state RIP in terms of unit length sparse vectors.

**Theorem 3.7** *Let  $x$  be some arbitrary unit length (i.e.  $\|x\|_2 = 1$ ) vector belonging to  $\Sigma_K$ . A matrix  $\Phi$  is said to satisfy RIP of order  $K$  if and only if the following holds*

$$(1 - \delta_K) \leq \|\Phi x\|_2^2 \leq (1 + \delta_K) \quad (3.1.10)$$

for every  $x \in \Sigma_K$  with  $\|x\|_2 = 1$ .

PROOF. If  $\Phi$  satisfies RIP of order  $K$  then by putting  $\|x\|_2 = 1$  in (3.1.1) we get (3.1.10).

Now the converse. We assume (3.1.10) holds for all unit norm vectors  $x \in \Sigma_K$ . We need to show that (3.1.1) holds for all  $x \in \Sigma_K$ .

For  $x = 0$  the bounds in (3.1.1) are trivially satisfied.

Let  $x \in \Sigma_K$  be some non-zero vector. Let  $\hat{x} = \frac{x}{\|x\|_2}$ . Clearly  $\hat{x}$  is unit length. Hence

$$\begin{aligned} (1 - \delta_K) &\leq \|\Phi\hat{x}\|_2^2 \leq (1 + \delta_K) \\ \implies (1 - \delta_K) &\leq \left\| \Phi \frac{x}{\|x\|_2} \right\|_2^2 \leq (1 + \delta_K) \\ \implies (1 - \delta_K) \|x\|_2^2 &\leq \|\Phi x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2 \end{aligned} \tag{3.1.11}$$

Thus  $\Phi$  satisfies RIP of order  $K$ . □

### 3.1.5. Singular and eigen values of $K$ -sub matrices

Consider any index set  $J \subset \{1, \dots, N\}$  with  $|J| = K$ . Let  $\Phi_J$  be a sub matrix of  $\Phi$  consisting of columns indexed by  $J$ . Assume  $K \leq M$ . We define

$$G \triangleq \Phi_J^H \Phi_J \in \mathbb{C}^{K \times K} \tag{3.1.12}$$

as the Gram matrix for columns of  $\Phi_J$  (see ??).

We consider the eigen values of  $G$  given by

$$Gx = \lambda x$$

for some  $x \in \mathbb{C}^K$  and  $x \neq 0$ . We will show that eigen values of  $G$  are bounded by RIP constant.

In the running example

$$G = \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

Eigen values of  $G$  are  $(0.2929, 1, 1.7071)$ .

**Theorem 3.8** Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ . Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $K$  columns. Then the eigen values of  $G = \Phi_J^H \Phi_J$  lie in the range  $[1 - \delta_K, 1 + \delta_K]$ .

PROOF. We note that  $G \in \mathbb{C}^{K \times K}$ .

Let  $\lambda$  be some eigen value of  $G$  and  $x \in \mathbb{C}^K$  be a corresponding eigen vector.

$$\begin{aligned} Gx &= \lambda x \\ \implies x^H G x &= x^H \lambda x \\ \implies x^H \Phi_J^H \Phi_J x &= \lambda \|x\|_2^2 \\ \implies \|\Phi_J x\|_2^2 &= \lambda \|x\|_2^2 \end{aligned}$$

From (3.1.5) we recall that  $\delta_K$  RIP bounds apply for each vector in  $x \in \mathbb{C}^K$  for a  $K$ -column sub-matrix  $\Phi_J$  given by

$$(1 - \delta_K) \|x\|_2^2 \leq \|\Phi_J x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2.$$

Thus

$$\begin{aligned} (1 - \delta_K) \|x\|_2^2 &\leq \lambda \|x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2 \\ \implies (1 - \delta_K) &\leq \lambda \leq (1 + \delta_K) \quad \text{since } x \neq 0. \end{aligned}$$

□

**Corollary 3.9.** Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ . Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $K$  columns. Then the Gram matrix  $G = \Phi_J^H \Phi_J$  is full rank and invertible.

Moreover  $G$  is positive definite.

PROOF. Clearly from theorem 3.8 all eigen values of  $G$  are non-zero. Hence their product is non-zero. Thus  $\det(G)$  is non-zero. Hence  $G$  is invertible.

Since from theorem 3.8 all eigen values are positive, hence,  $G$  is positive definite.  $\square$

**Theorem 3.10** *Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ . Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $K$  columns. Then the singular values of  $\Phi_J$  are non-zero and they are in the range given by*

$$\sqrt{1 - \delta_K} \leq \sigma \leq \sqrt{1 + \delta_K} \quad (3.1.13)$$

where  $\sigma$  is a singular value of  $\Phi_J$ .

PROOF. This is straight forward application of ?? and theorem 3.8. Eigen values of  $\Phi_J^H \Phi_J$  are nothing but squares of the singular values of  $\Phi_J$ .  $\square$

**Corollary 3.11.** *Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ . Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $k$  columns where  $k \leq K$ . Then the singular values of  $\Phi_J$  are non-zero and they are in the range given by*

$$\sqrt{1 - \delta_k} \leq \sigma \leq \sqrt{1 + \delta_k} \quad (3.1.14)$$

where  $\sigma$  is a singular value of  $\Phi_J$ .

PROOF. Let  $\sigma$  be a singular value of  $\Phi_J$ . From theorem 3.10 we have

$$\sqrt{1 - \delta_k} \leq \sigma \leq \sqrt{1 + \delta_k}.$$

From theorem 3.3 we have  $\delta_k \leq \delta_K$ . Thus

$$1 - \delta_K \leq 1 - \delta_k, \quad 1 + \delta_k \leq 1 + \delta_K.$$

Thus

$$\sqrt{1 - \delta_K} \leq \sigma \leq \sqrt{1 + \delta_K}.$$

$\square$

**Theorem 3.12** *Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ . Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $k$  columns where  $k \leq K$ . Then*

the eigen values of  $\Phi_J^H \Phi_J + rI$  lie in the range

$$[1 - \delta_K + r, 1 + \delta_K + r]$$

Moreover consider  $\Delta = \Phi_J^H \Phi_J - I$ . Then

$$\|\Delta\|_2 \leq \delta_K.$$

**PROOF.** From theorem 3.8 eigen values of  $\Phi_J^H \Phi_J$  lie in the range  $[1 - \delta_K, 1 + \delta_K]$ . From ??  $\lambda$  is an eigen value of  $\Phi_J^H \Phi_J$  if and only if  $\lambda + r$  is an eigen value of  $\Phi_J^H \Phi_J + rI$ . Hence the result.

Now for  $\Delta = \Phi_J^H \Phi_J - I$  the eigen values lie in the range  $[-\delta_K, \delta_K]$ . Thus for every eigen value of  $\Delta$  we have  $\lambda \leq \delta_K$ . Since  $\Delta$  is Hermitian, its spectral norm is nothing but its largest eigen value. Hence

$$\|\Delta\|_2 \leq \delta_K.$$

□

From previous few results we see that bound over eigen values of  $\Phi_J^H \Phi_J$  given by  $(1 - \delta_K) \leq \lambda \leq (1 + \delta_K)$  is a necessary condition for  $\Phi$  to satisfy RIP of order  $K$ . We now show that this is also a sufficient condition.

**Theorem 3.13** *Let  $\Phi$  be an  $M \times N$  matrix with  $M \leq N$ . Let  $J \subset \{1, \dots, N\}$  be an index set with  $|J| = K \leq M$ . Let  $\Phi_J$  be the  $K$ -column sub-matrix of  $\Phi$  indexed by  $J$ . Let  $G = \Phi_J^H \Phi_J$  be the Gram matrix of columns of  $\Phi_J$ . Let the eigen values of  $G$  be  $\lambda$ . If there exists a number  $\delta \in (0, 1)$  such that*

$$1 - \delta \leq \lambda \leq 1 + \delta$$

*for every eigen value of  $G$  for every  $K$  column sub-matrix of  $\Phi$ , then  $\Phi$  satisfies RIP of order  $K$ .*

*Alternatively, let  $\Delta = G - I$ . If*

$$\|\Delta\|_2 \leq \delta < 1$$

then  $\Phi$  satisfies RIP of order  $K$ .

Alternatively if singular values of  $\Phi_J$  satisfy

$$\sqrt{1-\delta} \leq \sigma \leq \sqrt{1+\delta}$$

for every  $\Phi_J$  then  $\Phi$  satisfies RIP of order  $K$ .

**PROOF.** We note that eigen values of  $G$  are related to eigen values of  $\Delta$  by the relation (see ??)

$$\lambda_G - 1 = \lambda_\Delta \iff \Lambda_G = 1 + \lambda_\Delta.$$

So

$$\|\Delta\|_2 \leq \delta \iff -\delta \leq \lambda_\Delta \leq \delta \iff 1 - \delta \leq \lambda_G \leq 1 + \delta.$$

Thus the first two sufficient conditions are equivalent. Lastly the eigen values of  $G$  are squares of singular values of  $\Phi_J$ , thus all sufficient conditions are equivalent.

Now let  $x \in \Sigma_K$  be an arbitrary vector and let  $J = \text{supp}(x)$ . Clearly  $|J| \leq K$ . If  $|J| < K$  then augment  $J$  by adding some indices arbitrarily till we get  $|J| = K$ . Clearly  $x_J$  is an arbitrary vector in  $\mathbb{C}^K$  and  $\Phi x = \Phi_J x_J$ . Now let  $\lambda_1$  be the largest and  $\lambda_k$  be the smallest eigen value of  $G = \Phi_J^H \Phi_J$ .  $G$  is Hermitian and all its eigen values are positive, hence it is positive definite. From ?? we get

$$\lambda_k \|x\|_2^2 \leq x^H G x \leq \lambda_1 \|x\|_2^2 \quad \forall x \in \mathbb{C}^K.$$

Applying the limits on the eigen values and using  $x^H G x = \|\Phi_J x\|_2^2$ , we get

$$(1 - \delta) \|x\|_2^2 \leq \|\Phi_J x\|_2^2 \leq (1 + \delta) \|x\|_2^2 \quad \forall x \in \mathbb{C}^K.$$

Since this holds for every index set  $J$  with  $|J| = K$  hence an equivalent statement is

$$(1 - \delta) \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta) \|x\|_2^2 \quad \forall x \in \Sigma_K \subset \mathbb{C}^N.$$

Thus  $\Phi$  indeed satisfies RIP of order  $K$  with some  $\delta_K$  not larger than  $\delta$ .  $\square$

**Theorem 3.14** Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ .

Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $k$  columns where  $k \leq K$ . Let  $\Phi_J^\dagger$  be its Moore-Penrose pseudo-inverse. Then the singular values of  $\Phi_J^\dagger$  are non-zero and they are in the range given by

$$\frac{1}{\sqrt{1 + \delta_K}} \leq \sigma \leq \frac{1}{\sqrt{1 - \delta_K}} \quad (3.1.15)$$

where  $\sigma$  is a singular value of  $\Phi_J^\dagger$ .

PROOF. Construction of pseudo inverse of a matrix through its singular value decomposition is discussed in ???. ?? shows that if  $\sigma$  is a non-zero singular value of  $\Phi_J^\dagger$  then  $\frac{1}{\sigma}$  is a non-zero singular value of  $\Phi_J$ . From corollary 3.11 we have that if  $\frac{1}{\sigma}$  is a singular value of  $\Phi_J$  then,

$$\sqrt{1 - \delta_K} \leq \frac{1}{\sigma} \leq \sqrt{1 + \delta_K}$$

Inverting the terms in the inequalities we get our result.  $\square$

**Theorem 3.15** Eigen values of  $G = \Phi_J^H \Phi_J$  provide a lower bound on  $\delta_K$  given by

$$\delta_K \geq \max(1 - \lambda_{min}, \lambda_{max} - 1)$$

where  $J$  is some index set choosing  $K$  columns of  $\Phi$  and  $\delta_K$  is the  $K$ -th restricted isometry constant for  $\Phi$ .

In other words, singular values of  $\Phi_J$  provide a lower bound on  $\delta_K$  given by

$$\delta_K \geq \max(1 - \sigma_{min}^2, \sigma_{max}^2 - 1)$$

PROOF. Obvious.  $\square$

In the running example, the bounds tell us that

$$\delta_3 \geq 0.7071.$$

Certainly we have to consider all possible  $\binom{N}{K}$  sub-matrices  $\Phi_J$  to come up with an overall lower bound on  $\delta_K$ .

This result doesn't provide us any upper bound on  $\delta_K$ .

**Theorem 3.16** *Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ .*

*Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $k$  columns where  $k \leq K$ . Then*

$$\|\Phi_J x\|_2 \leq \sqrt{1 + \delta_K} \|x\|_2 \quad \forall x \in \mathbb{C}^k. \quad (3.1.16)$$

*Moreover*

$$\|\Phi_J^H y\|_2 \leq \sqrt{1 + \delta_K} \|y\|_2 \quad \forall y \in \mathbb{C}^M. \quad (3.1.17)$$

PROOF. We note that  $\Phi_J$  is an  $M \times k$  matrix. Let  $\sigma_1$  be the largest singular value of  $\Phi_J$ . Then by ?? we have

$$\|\Phi_J x\|_2 \leq \sigma_1 \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

and

$$\|\Phi_J^H y\|_2 \leq \sigma_1 \|y\|_2 \quad \forall y \in \mathbb{C}^M.$$

From theorem 3.10 and corollary 3.11 we get

$$\sigma_1 \leq \sqrt{1 + \delta_K}.$$

This completes the proof. □

First inequality is essentially a restatement of restricted isometry property in eq. (3.1.5). Second inequality is interesting. In compressed sensing terms,  $y$  is a measurement vector and we are using  $\Phi_J^H$  to project  $y$  back into  $\mathbb{C}^N$  over a  $k$  sparse support identified by  $J$ . The inequality provides an upper bound on how much the length can increase during this operation.

**Theorem 3.17** *Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ .*

*Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $k$  columns where  $k \leq K$ . Let*

$\Phi_J^\dagger$  be its Moore-Penrose pseudo-inverse. Then

$$\|\Phi_J^\dagger y\|_2 \leq \frac{1}{\sqrt{1 - \delta_K}} \|y\|_2 \quad \forall y \in \mathbb{C}^M. \quad (3.1.18)$$

PROOF. We note that  $\Phi_J^\dagger$  is an  $k \times M$  matrix. Let  $\sigma_1$  be the largest singular value of  $\Phi_J^\dagger$ . Then by ?? we have

$$\|\Phi_J^\dagger y\|_2 \leq \sigma_1 \|y\|_2 \quad \forall y \in \mathbb{C}^M.$$

From theorem 3.14 we see that singular values of  $\Phi_J^\dagger$  satisfy the inequalities

$$\frac{1}{\sqrt{1 + \delta_K}} \leq \sigma \leq \frac{1}{\sqrt{1 - \delta_K}}.$$

Thus

$$\sigma_1 \leq \frac{1}{\sqrt{1 - \delta_K}}.$$

Plugging it in we get

$$\|\Phi_J^\dagger y\|_2 \leq \frac{1}{\sqrt{1 - \delta_K}} \|y\|_2 \quad \forall y \in \mathbb{C}^M.$$

□

In previous theorem we saw that back-projection using  $\Phi_J^H$  had an upper bound on how much the length of measurement vector could increase. In this theorem we see another upper bound on how much the length of measurement vector can increase when back projected using the pseudo inverse of  $\Phi_J$ .

**Theorem 3.18** *Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ .*

*Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $k$  columns where  $k \leq K$ . Then*

$$(1 - \delta_K) \|x\|_2 \leq \|\Phi_J^H \Phi_J x\|_2 \leq (1 + \delta_K) \|x\|_2 \quad \forall x \in \mathbb{C}^k. \quad (3.1.19)$$

*Moreover*

$$\frac{1}{1 + \delta_K} \|x\|_2 \leq \|\Phi_J^H \Phi_J^{-1} x\|_2 \leq \frac{1}{1 - \delta_K} \|x\|_2 \quad \forall x \in \mathbb{C}^k. \quad (3.1.20)$$

PROOF. We note that  $\Phi_J$  is a full column rank tall matrix. We recall that all singular values of  $\Phi_J$  are positive and are bounded by (corollary 3.11):

$$\sqrt{1 - \delta_K} \leq \sigma_k \leq \dots \leq \sigma_1 \leq \sqrt{1 + \delta_K}$$

where  $\sigma_1, \dots, \sigma_k$  are the singular values of  $\Phi_J$  (in descending order).

We note that  $\Phi_J^H \Phi_J$  is an  $k \times k$  matrix which is invertible (corollary 3.9).

Now from ?? we get

$$\sigma_k^2 \|x\|_2 \leq \|\Phi_J^H \Phi_J x\|_2 \leq \sigma_1^2 \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

Applying the bounds on  $\sigma_i$  we get the result

$$(1 - \delta_K) \|x\|_2 \leq \|\Phi_J^H \Phi_J x\|_2 \leq (1 + \delta_K) \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

Now from ?? we have the bounds for  $(\Phi_J^H \Phi_J)^{-1}$  given by

$$\frac{1}{\sigma_1^2} \|x\|_2 \leq \|(\Phi_J^H \Phi_J)^{-1} x\|_2 \leq \frac{1}{\sigma_k^2} \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

Applying the bounds on  $\sigma_i$  we get the result

$$\frac{1}{1 + \delta_K} \|x\|_2 \leq \|(\Phi_J^H \Phi_J)^{-1} x\|_2 \leq \frac{1}{1 - \delta_K} \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

□

In the sequel we will discuss that  $\Phi^H \Phi x$  can work as a very good proxy for the signal  $x$ . The results in this theorem are very comforting in this regard.

**Theorem 3.19** *Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ . Let  $\Phi_J$  be any sub matrix of  $\Phi$  with  $k$  columns where  $k \leq K$ . Then*

$$\|(\Phi_J^H \Phi_J - I)x\|_2 \leq \delta_K \|x\|_2 \quad \forall x \in \mathbb{C}^k. \quad (3.1.21)$$

PROOF. From theorem 3.12 we get

$$\|\Phi_J^H \Phi_J - I\|_2 \leq \delta_k \leq \delta_K.$$

Thus since spectral norm is subordinate

$$\|(\Phi_J^H \Phi_J - I)x\|_2 \leq \|\Phi_J^H \Phi_J - I\|_2 \|x\|_2 \leq \delta_K \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

□

### 3.1.6. Approximate orthogonality

We are going to show that disjoint sets of columns from  $\Phi$  span nearly orthogonal subspaces. This property is proved in [29].

**Theorem 3.20** *Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ . Let  $S$  and  $T$  denote index sets over the columns of  $\Phi$  with  $|S| + |T| \leq K$  and  $S \cap T = \emptyset$ . i.e.  $S$  and  $T$  are disjoint index sets. Let  $\Phi_S$  and  $\Phi_T$  denote corresponding sub-matrices consisting of columns indexed by  $S$  and  $T$  respectively. Then*

$$\|\Phi_S^H \Phi_T\|_2 \leq \delta_K \quad (3.1.22)$$

where  $\|\cdot\|_2$  denotes the 2-norm or spectral norm of a matrix.

**PROOF.** Consider  $R = S \cup T$ . Consider the sub-matrix  $\Phi_R$ . Construct another matrix  $\Psi = \Phi_R^H \Phi_R - I$ . The off-diagonal entries of  $\Psi$  are nothing but inner products of columns of  $\Phi_R$ . We note that every entry in the matrix  $\Phi_S^H \Phi_T$  is an entry in  $\Psi$ . Moreover,  $\Phi_S^H \Phi_T$  is a sub-matrix of  $\Psi$ .

The spectral norm of a sub-matrix is never greater than the spectral norm of the matrix containing it. Thus

$$\|\Phi_S^H \Phi_T\|_2 \leq \|\Phi_R^H \Phi_R - I\|_2.$$

From theorem 3.12 the eigen values of  $\Phi_R^H \Phi_R - I$  satisfy

$$1 - \delta_K - 1 \leq \lambda \leq 1 + \delta_K - 1.$$

Thus the spectral norm of  $\Phi_R^H \Phi_R - I$  which is its largest eigen value (see ??) satisfies

$$\|\Phi_R^H \Phi_R - I\|_2 \leq \delta_K.$$

Plugging back we get

$$\|\Phi_S^H \Phi_T\|_2 \leq \delta_K.$$

□

This result has a useful corollary. It establishes the approximate orthogonality between a set of columns in  $\Phi$  and portion of a sparse vector not covered by those columns.

**Corollary 3.21.** *Let  $\Phi$  satisfy the RIP of order  $K$  where  $K \leq M$ . Let  $T \subset \{1, \dots, N\}$  be an index set and let  $x \in \mathbb{C}^N$  be some vector. Let  $S = \text{supp}(x)$ . Further let us assume that  $K \geq |T \cup S|$ . Define  $R = S \setminus T$ .*

*Then the following holds*

$$\|\Phi_T^H \Phi x_R\|_2 \leq \delta_K \|x_R\|_2 \quad (3.1.23)$$

*where  $x_R$  is obtained by keeping entries in  $x$  indexed by  $R$  while setting others to 0 (see definition 2.10).*

PROOF. Since

$$\Phi x = \sum_i \phi_i x_i$$

and  $x_R$  is zero on entries not indexed by  $R$ , hence

$$\Phi x_R = \Phi_R x_R$$

where on the R.H.S.  $x_R \in \mathbb{C}^{|R|}$  by dropping the 0 entries from it not indexed by  $R$  (see definition 2.10). Thus we have

$$\|\Phi_T^H \Phi x_R\|_2 = \|\Phi_T^H \Phi_R x_R\|_2.$$

From ?? we know that any operator norm is subordinate. Thus

$$\|\Phi_T^H \Phi_R x_R\|_2 \leq \|\Phi_T^H \Phi_R\|_2 \|x_R\|_2.$$

Since  $K \geq |T \cup S|$  hence we have

$$|R| = |S \setminus T| \leq K.$$

Further  $T$  and  $R$  are disjoint. Applying theorem 3.20 we get

$$\|\Phi_T^H \Phi_R\|_2 \leq \delta_K.$$

Putting back, we get our desired result

$$\|\Phi_T^H \Phi x_R\|_2 \leq \delta_K \|x_R\|_2.$$

□

### 3.1.7. Signal proxy

We can use the results so far to formalize the idea of signal proxy.

**Theorem 3.22** *Let  $x$  be a  $k$ -sparse signal. Let  $\Phi$  satisfy the RIP of order  $k + l$  or higher. Let  $p$  be defined as*

$$p = (\Phi^H \Phi x)|_l \quad (3.1.24)$$

*i.e.  $p$  is obtained by keeping the  $l$  largest entries in  $b = \Phi^H \Phi x$ . Then the following holds*

$$\|p\|_2 \leq (1 + \delta_l + \delta_{k+l})\|x\|_2. \quad (3.1.25)$$

PROOF. Let  $A = \text{supp}(x)$ , then  $|A| \leq k$ . Let  $B = \text{supp}(p)$ . Then  $|B| \leq l$ . Clearly

$$p = (\Phi^H \Phi x)|_l = (\Phi^H \Phi x)_B.$$

From lemma 2.20 we get

$$p = \Phi_B^H \Phi x.$$

Let  $C = A \setminus B$ . Since  $x$  is supported on  $A$  only, hence we can write

$$x = x_B + x_C.$$

Thus from corollary 2.19 we get ( $B$  and  $C$  are disjoint)

$$\Phi x = \Phi_B x_B + \Phi_C x_C.$$

Thus we have

$$p = \Phi_B^H \Phi_B x_B + \Phi_B^H \Phi_C x_C.$$

Using triangle inequality we can write

$$\|p\|_2 \leq \|\Phi_B^H \Phi_B x_B\|_2 + \|\Phi_B^H \Phi_C x_C\|_2.$$

Theorem 3.18 gives us

$$\|\Phi_B^H \Phi_B x_B\|_2 \leq (1 + \delta_l)\|x_B\|_2.$$

Since  $B$  and  $C$  are disjoint, hence Theorem 3.20 gives us

$$\|\Phi_B^H \Phi_C x_C\|_2 \leq \delta_{k+l}\|x_C\|_2.$$

Finally

$$\|p\|_2 \leq (1 + \delta_l + \delta_{k+l})\|x\|_2.$$

□

### 3.1.8. RIP and inner product

Let  $x$  and  $x'$  be two different vectors in  $\mathbb{C}^N$  such that their support is disjoint. i.e. if

$$T = \text{supp}(x) \subseteq \{1, \dots, N\}$$

and

$$T' = \text{supp}(x') \subseteq \{1, \dots, N\}$$

then  $T \cap T' = \emptyset$ .

Clearly

$$\|x\|_0 = |T|$$

and

$$\|x'\|_0 = |T'|.$$

Since the support of  $x$  and  $x'$  are disjoint hence it is straightforward that

$$\langle x, x' \rangle = 0.$$

What can we say about the inner product of their corresponding embedded vectors  $\Phi x$  and  $\Phi x'$ ?

Following theorem provides an upper bound on the magnitude of the inner product when the signal vectors  $x, x'$  belong to the Euclidean space  $\mathbb{R}^N$ . This result is adapted from [8].

**Theorem 3.23** *For all  $x, x' \in \mathbb{R}^N$  supported on disjoint subsets  $T, T' \subseteq \{1, \dots, N\}$  with  $|T| < k$  and  $|T'| < k'$  we have*

$$|\langle \Phi x, \Phi x' \rangle| \leq \delta_{k+k'} \|x\|_2 \|x'\|_2 \quad (3.1.26)$$

*where  $\delta_{k+k'}$  is the restricted isometry constant for the sparsity level  $k + k'$ .*

PROOF. Let  $\hat{x} = \frac{x}{\|x\|_2}$  and  $\hat{x}' = \frac{x'}{\|x'\|_2}$  be the corresponding unit norm vectors.

Then

$$\langle \Phi x, \Phi x' \rangle = \langle \Phi \hat{x}, \Phi \hat{x}' \rangle \|x\|_2 \|x'\|_2.$$

So if we prove the bound for unit norm vectors, then it will be straightforward to prove the bound for arbitrary vectors.

Let us assume now that  $x, x'$  are unit norm. We need to show that

$$|\langle \Phi x, \Phi x' \rangle| \leq \delta_{k+k'}.$$

With the help of parallelogram identity, we have

$$\langle \Phi x, \Phi x' \rangle = \frac{1}{4} (\|\Phi x + \Phi x'\|_2^2 - \|\Phi x - \Phi x'\|_2^2)$$

thus

$$|\langle \Phi x, \Phi x' \rangle| = \frac{1}{4} |\|\Phi x + \Phi x'\|_2^2 - \|\Phi x - \Phi x'\|_2^2|.$$

Now

$$\|x \pm x'\|_2^2 = \|x\|_2^2 + \|x'\|_2^2 \pm 2\langle x, x' \rangle = \|x\|_2^2 + \|x'\|_2^2 = 2$$

since  $x, x'$  are orthogonal and unit norm.

Thus from theorem 3.5 we have

$$(1 - \delta_{k+k'})\|x \pm x'\|_2^2 \leq \|\Phi x \pm \Phi x'\|_2^2 \leq (1 + \delta_{k+k'})\|x \pm x'\|_2^2 \quad (3.1.27)$$

$$\implies 2(1 - \delta_{k+k'}) \leq \|\Phi x \pm \Phi x'\|_2^2 \leq 2(1 + \delta_{k+k'}) \quad (3.1.28)$$

Clearly the maximum value of  $\|\Phi x + \Phi x'\|_2^2$  can be  $2(1 + \delta_{k+k'})$  while the minimum value of  $\|\Phi x - \Phi x'\|_2^2$  can be  $2(1 - \delta_{k+k'})$ .

This gives us the upper bound

$$|\langle \Phi x, \Phi x' \rangle| \leq \frac{1}{4} (2(1 + \delta_{k+k'}) - 2(1 - \delta_{k+k'})) = \delta_{k+k'}.$$

Finally when  $x, x'$  are not unit norm, the bound generalizes to

$$|\langle \Phi x, \Phi x' \rangle| \leq \delta_{k+k'} \|x\|_2 \|x'\|_2.$$

□

A variation of this result is presented below:

**Theorem 3.24** [17] Let  $u, v \in \mathbb{R}^N$  be given and let

$$K = \max(\|u + v\|_0, \|u - v\|_0).$$

Let  $\Phi$  satisfy RIP of order  $K$  with the constant  $\delta_K$ . Then

$$|\langle \Phi u, \Phi v \rangle - \langle u, v \rangle| \leq \delta_K \|u\|_2 \|v\|_2. \quad (3.1.29)$$

This result is more general as it doesn't require  $u, v$  to be supported on disjoint index sets. All it requires is them to be sufficiently sparse.

PROOF. As, in the previous result, it is sufficient to prove it for the case where  $\|u\|_2 = \|v\|_2 = 1$ . The simplified inequality becomes

$$|\langle \Phi u, \Phi v \rangle - \langle u, v \rangle| \leq \delta_K.$$

Clearly

$$\|u \pm v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 \pm 2\langle u, v \rangle = 2 \pm 2\langle u, v \rangle.$$

Due to RIP, we have

$$(1 - \delta_K)(2 \pm 2\langle u, v \rangle) \leq \|\Phi(u \pm v)\|_2^2 \leq (1 + \delta_K)(2 \pm 2\langle u, v \rangle).$$

From the parallelogram identity, we have

$$\langle \Phi u, \Phi v \rangle = \frac{1}{4} (\|\Phi(u + v)\|_2^2 - \|\Phi(u - v)\|_2^2). \quad (3.1.30)$$

Taking the upper bound on  $\|\Phi(u + v)\|_2^2$  and the lower bound on  $\|\Phi(u - v)\|_2^2$  in (3.1.30), we obtain

$$\langle \Phi u, \Phi v \rangle \leq \frac{1}{2} ((1 + \delta_K)(1 + \langle u, v \rangle) - (1 - \delta_K)(1 - \langle u, v \rangle)).$$

Simplifying, we get

$$\langle \Phi u, \Phi v \rangle \leq \langle u, v \rangle + \delta_K.$$

At the same time, taking the lower bound on  $\|\Phi(u+v)\|_2^2$  and the upper bound on  $\|\Phi(u-v)\|_2^2$  in (3.1.30), we obtain

$$\langle \Phi u, \Phi v \rangle \geq \frac{1}{2} ((1 - \delta_K)(1 + \langle u, v \rangle) - (1 + \delta_K)(1 - \langle u, v \rangle)).$$

Simplifying, we get

$$\langle \Phi u, \Phi v \rangle \geq \langle u, v \rangle - \delta_K.$$

Combining the two results, we obtain

$$|\langle \Phi u, \Phi v \rangle - \langle u, v \rangle| \leq \delta_K.$$

□

For the complex case, the result can be generalized if we choose a bilinear inner product rather than the usual sesquilinear inner product.

**Theorem 3.25** *Let  $u, v \in \mathbb{C}^N$  be given and let*

$$K = \max(\|u + v\|_0, \|u - v\|_0).$$

. Let the complex space  $\mathbb{C}^N$  be equipped with the bilinear inner product

$$\langle u, v \rangle_B \triangleq \operatorname{Re}(\langle u, v \rangle)$$

i.e. the real part of the usual inner product.

Let  $\Phi$  satisfy RIP of order  $K$  with the constant  $\delta_K$ . Then

$$|\langle \Phi u, \Phi v \rangle_B - \langle u, v \rangle_B| \leq \delta_K \|u\|_2 \|v\|_2. \quad (3.1.31)$$

**PROOF.** Recall that the norm induced by the bilinear inner product  $\langle u, v \rangle_B$  is the usual  $l_2$  norm since

$$\langle u, u \rangle_B = \operatorname{Re}(\langle u, u \rangle) = \operatorname{Re}(\|u\|_2^2) = \|u\|_2^2.$$

Let us just work out the parallelogram identity for the complex case

$$\begin{aligned} \|x \pm y\|_2^2 &= \langle x \pm y, x \pm y \rangle_B \\ &= \langle x, x \rangle_B + \langle y, y \rangle_B \pm \langle x, y \rangle_B \pm \langle y, x \rangle_B \\ &= \langle x, x \rangle_B + \langle y, y \rangle_B \pm 2\langle x, y \rangle_B \end{aligned}$$

due to the bilinearity of the inner product.

We can see that the rest of the proof is identical to the proof of (3.24).  $\square$

### 3.1.9. RIP and orthogonal projection

The first result in this section is presented for real matrices. The generalization for complex matrices will be done later.

Let  $\Lambda \subset \{1, \dots, N\}$  be an index set. Let  $\Phi \in \mathbb{R}^{M \times N}$  satisfy RIP of order  $K$  with the restricted isometry constant  $\delta_K$ .

Assume that the columns of  $\Phi_\Lambda$  are linearly independent.

We can define the pseudo inverse as

$$\Phi_\Lambda^\dagger = (\Phi_\Lambda^H \Phi_\Lambda)^{-1} \Phi_\Lambda^H. \quad (3.1.32)$$

The orthogonal projection operator to the column space for  $\Phi_\Lambda$  is given by

$$P_\Lambda = \Phi_\Lambda \Phi_\Lambda^\dagger. \quad (3.1.33)$$

The orthogonal projection operator onto the orthogonal complement of  $\mathcal{C}(\Phi_\Lambda)$  (column space of  $\Phi_\Lambda$ ) is given by

$$P_\Lambda^\perp = I - P_\Lambda. \quad (3.1.34)$$

Both  $P_\Lambda$  and  $P_\Lambda^\perp$  satisfy the usual properties like  $P = P^H$  and  $P^2 = P$ .

We further define

$$\Psi_\Lambda = P_\Lambda^\perp \Phi. \quad (3.1.35)$$

We are orthogonalizing the columns in  $\Phi$  against  $\mathcal{C}(\Phi_\Lambda)$ , i.e. taking the component of the column which is orthogonal to the column space of  $\Phi_\Lambda$ . Obviously the columns in  $\Psi_\Lambda$  corresponding to the index set  $\Lambda$  would be 0.

We now present a result which shows that the matrix  $\Psi_\Lambda$  satisfies a modified version of RIP.

**Theorem 3.26** [17] If  $\Phi$  satisfies the RIP of order  $K$  with isometry constant  $\delta_K$ , and  $\Lambda \subset \{1, \dots, N\}$  with  $|\Lambda| < K$ , then the matrix  $\Psi_\Lambda$  satisfies the modified version of RIP as

$$\left(1 - \frac{\delta_K}{1 - \delta_K}\right) \|x\|_2^2 \leq \|\Psi_\Lambda x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2 \quad (3.1.36)$$

for all  $x \in \mathbb{R}^N$  such that  $\|x\|_0 \leq K - |\Lambda|$  and  $\text{supp}(x) \cap \Lambda = \emptyset$ .

In words, if  $\Phi$  satisfies RIP of order  $K$ , then  $\Psi_\Lambda$  acts as an approximate isometry on every  $(K - |\Lambda|)$ -sparse vector supported on  $\Lambda^c$ .

PROOF. From the definition of  $\Psi_\Lambda$ , we have

$$\Psi_\Lambda x = (I - P_\Lambda)\Phi x = \Phi x - P_\Lambda \Phi x.$$

Alternatively

$$\Phi x = \Psi_\Lambda x + P_\Lambda \Phi x.$$

Since  $P_\Lambda$  is an orthogonal projection, hence the vectors  $P_\Lambda \Phi x$  and  $\Psi_\Lambda x = P_\Lambda^\perp \Phi x$  are orthogonal. Thus, we can write

$$\|\Phi x\|_2^2 = \|P_\Lambda \Phi x\|_2^2 + \|\Psi_\Lambda x\|_2^2. \quad (3.1.37)$$

We need to show that  $\|\Phi x\|_2 \approx \|\Psi_\Lambda x\|_2$  or alternatively that  $\|P_\Lambda \Phi x\|_2$  is small under the conditions of the theorem.

Since  $P_\Lambda \Phi x$  is orthogonal to  $\Psi_\Lambda x$ , hence

$$\begin{aligned} \langle P_\Lambda \Phi x, \Phi x \rangle &= \langle P_\Lambda \Phi x, \Psi_\Lambda x + P_\Lambda \Phi x \rangle \\ &= \langle P_\Lambda \Phi x, P_\Lambda \Phi x \rangle + \langle P_\Lambda \Phi x, \Psi_\Lambda x \rangle \\ &= \langle P_\Lambda \Phi x, P_\Lambda \Phi x \rangle \\ &= \|P_\Lambda \Phi x\|_2^2. \end{aligned} \quad (3.1.38)$$

Since  $P_\Lambda$  is a projection onto the  $\mathcal{C}(\Phi_\Lambda)$  (column space of  $\Phi_\Lambda$ ), there exists a vector  $z \in \mathbb{C}^N$ , such that  $P_\Lambda \Phi x = \Phi z$  and  $\text{supp}(z) \subseteq \Lambda$ .

Since  $\text{supp}(x) \cap \Lambda = \emptyset$ , hence  $\langle x, z \rangle = 0$ .

We also note that  $\|x + z\|_0 = \|x - z\|_0 \leq K$ .

Invoking theorem 3.24, we have

$$|\langle \Phi z, \Phi x \rangle| \leq \delta_K \|z\|_2 \|x\|_2.$$

Alternatively

$$|\langle P_\Lambda \Phi x, \Phi x \rangle| \leq \delta_K \|z\|_2 \|x\|_2.$$

From RIP, we have

$$\sqrt{1 - \delta_K} \|z\|_2 \leq \|\Phi z\|_2$$

and

$$\sqrt{1 - \delta_K} \|x\|_2 \leq \|\Phi x\|_2.$$

Thus

$$(1 - \delta_K) \|z\|_2 \|x\|_2 \leq \|\Phi z\|_2 \|\Phi x\|_2.$$

This gives us

$$|\langle P_\Lambda \Phi x, \Phi x \rangle| \leq \frac{\delta_K}{1 - \delta_K} \|P_\Lambda \Phi x\|_2 \|\Phi x\|_2.$$

Applying (3.1.38), we get

$$\|P_\Lambda \Phi x\|_2^2 \leq \frac{\delta_K}{1 - \delta_K} \|P_\Lambda \Phi x\|_2 \|\Phi x\|_2.$$

Cancelling the common term, we get

$$\|P_\Lambda \Phi x\|_2 \leq \frac{\delta_K}{1 - \delta_K} \|\Phi x\|_2.$$

Trivially, we have  $\|P_\Lambda \Phi x\|_2 \geq 0$ .

Applying these bounds on (3.1.37), we obtain

$$\left(1 - \left(\frac{\delta_K}{1 - \delta_K}\right)^2\right) \|\Phi x\|_2^2 \leq \|\Psi_\Lambda x\|_2^2 \leq \|\Phi x\|_2^2.$$

Finally, using the RIP again with

$$(1 - \delta_K) \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2$$

we obtain

$$\left(1 - \left(\frac{\delta_K}{1 - \delta_K}\right)^2\right) (1 - \delta_K) \|x\|_2^2 \leq \|\Psi_\Lambda x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2.$$

Simplifying

$$\begin{aligned} \left(1 - \left(\frac{\delta_K}{1 - \delta_K}\right)^2\right) (1 - \delta_K) &= \frac{1 + \delta_K^2 - 2\delta_K - \delta_K^2}{1 - \delta_K} \\ &= \frac{1 - 2\delta_K}{1 - \delta_K} \\ &= 1 - \frac{\delta_K}{1 - \delta_K}. \end{aligned}$$

Thus, we get the intended result in (3.1.36).  $\square$

### 3.1.10. RIP for higher orders

If  $\Phi$  satisfies RIP of order  $K$ , does it satisfy RIP of some other order  $K' > K$ ? There are some results available to answer this question.

**Theorem 3.27** *Let  $c$  and  $k$  be integers and let  $\Phi$  satisfy RIP of order  $2k$ .  $\Phi$  satisfies RIP of order  $ck$  with a restricted isometry constant*

$$\delta_{ck} \leq c\delta_{2k} \quad (3.1.39)$$

*if  $c\delta_{2k} < 1$ .*

Note that this is only a sufficient condition. Thus if  $c\delta_{2k} \geq 1$  we are not claiming whether  $\Phi$  satisfies RIP of order  $ck$  or not.

**PROOF.** For  $c = 1$ ,  $\delta_k \leq \delta_{2k}$ . For  $c = 2$ ,  $\delta_{2k} \leq 2\delta_{2k}$ . These two cases are trivial. We now consider the case for  $c \geq 3$ .

Let  $S$  be an arbitrary index set of size  $ck$ . Let

$$\Delta = \Phi_S^H \Phi_S - I.$$

From theorem 3.13, a sufficient condition for  $\Phi$  to satisfy RIP of order  $ck$  is that

$$\|\Delta\|_2 < 1$$

for all index sets  $S$  with  $|S| = ck$ .

Thus if we can show that

$$\|\Delta\|_2 \leq c\delta_{2k}$$

we would have shown that  $\Phi$  satisfies RIP of order  $ck$ .

We note that  $\Phi_S$  is of size  $M \times ck$  thus  $\Delta$  is of size  $ck \times ck$ . We partition  $\Delta$  into a block matrix of size  $c \times c$

$$\Delta = \begin{bmatrix} \Delta_{11} & \Delta_{12} & \dots & \Delta_{1c} \\ \Delta_{21} & \Delta_{22} & \dots & \Delta_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{c1} & \Delta_{c2} & \dots & \Delta_{cc} \end{bmatrix} \quad (3.1.40)$$

where each entry  $\Delta_{ij}$  is a square matrix of size  $k \times k$ .

Each diagonal matrix  $\Delta_{ii}$  corresponds to some  $\Phi_T^H \Phi_T - I$  where  $|T| = k$ . Thus we have (see theorem 3.12)

$$\|\Delta_{ii}\|_2 \leq \delta_k.$$

The off-diagonal matrices  $\Delta_{ij}$  are

$$\Delta_{ij} = \Phi_P^H \Phi_Q$$

where  $P$  and  $Q$  are disjoint index sets with  $|P| = |Q| = k$  with  $|P \cup Q| = 2k$ . Thus from the approximate orthogonality condition (theorem 3.20) we have

$$\|\Delta_{ij}\|_2 \leq \delta_{2k}.$$

Finally we apply Gershgorin circle theorem for block matrices (see ??). This gives us

$$|\|\Delta\|_2 - \|\Delta_{ii}\|_2| \leq \sum_{j \neq i} \|\Delta_{ij}\| \text{ for some } i \in \{1, 2, \dots, n\}.$$

Thus we have

$$\begin{aligned}
|\|\Delta\|_2 - \delta_k| &\leq \sum_{j \neq i} \delta_{2k} \\
\implies |\|\Delta\|_2 - \delta_k| &\leq (c-1)\delta_{2k} \\
\implies \|\Delta\|_2 &\leq \delta_k + (c-1)\delta_{2k} \\
\implies \|\Delta\|_2 &\leq \delta_{2k} + (c-1)\delta_{2k} \\
\implies \|\Delta\|_2 &\leq c\delta_{2k}.
\end{aligned}$$

We have shown that  $\|\Delta\|_2 \leq c\delta_{2k} < 1$  thus  $\delta_{ck} \leq \|\Delta\|_2$ , hence  $\Phi$  indeed satisfies RIP of order  $ck$ .  $\square$

This theorem helps us extend RIP from an order  $K$  to higher orders. Naturally if  $\delta_{2k}$  isn't sufficiently small, the bound isn't useful.

### 3.1.11. Bounds on norms of embeddings of arbitrary signals

So far we have considered only sparse signals while analyzing the embedding properties of a RIP satisfying matrix  $\Phi$ . In this subsection we wish to explore bounds on the  $l_2$  norm of an arbitrary signal when embedded by  $\Phi$ . This result is adapted from [29].

**Theorem 3.28** *Let  $\Phi$  be an  $M \times N$  matrix satisfying*

$$\|\Phi x\|_2 \leq \sqrt{1 + \delta_K} \|x\|_2 \quad \forall x \in \Sigma_K. \quad (3.1.41)$$

*Then for every signal  $x \in \mathbb{C}^N$ , the following holds:*

$$\|\Phi x\|_2 \leq \sqrt{1 + \delta_K} \left[ \|x\|_2 + \frac{1}{\sqrt{K}} \|x\|_1 \right]. \quad (3.1.42)$$

We note that the theorem requires  $\Phi$  to satisfy only the upper bound of RIP property (3.1.1). The proof is slightly involved.

**PROOF.** We note that the bound is trivially true for  $x = 0$ . Hence in the following we will consider only for  $x \neq 0$ .

1403. MORE TOOLS FOR DICTIONARY AND RANDOM MATRIX ANALYSIS

Consider an arbitrary index set  $\Lambda \subset \{1, 2, \dots, N\}$  such that  $|\Lambda| \leq K$ .

Consider the unit ball in the Banach space  $l_2(\Lambda)$  given by

$$B_2^\Lambda = \{x \in \mathbb{C}^N \mid \text{supp}(x) = \Lambda \text{ and } \|x\|_2 \leq 1\} \quad (3.1.43)$$

i.e. the set of all signals whose support is  $\Lambda$  and whose  $l_2$  norm is less than or equal to 1.

Now define a convex body

$$S = \text{conv} \left\{ \bigcup_{|\Lambda| \leq K} B_2^\Lambda \right\} \subset \mathbb{C}^N. \quad (3.1.44)$$

We recall from ?? that if  $x$  and  $y$  belong to  $S$  then their convex combination  $\theta x + (1 - \theta)y$  with  $\theta \in [0, 1]$  must lie in  $S$ . Further it can be verified that  $S$  is a compact convex set with non-empty interior. Hence its a convex body.

Consider any  $x \in B_2^{\Lambda_1}$  and  $y \in B_2^{\Lambda_2}$ .

From (3.1.41) and (3.1.43) we have

$$\|\Phi x\|_2 \leq \sqrt{1 + \delta_K} \|x\|_2 \leq \sqrt{1 + \delta_K}$$

and

$$\|\Phi y\|_2 \leq \sqrt{1 + \delta_K} \|y\|_2 \leq \sqrt{1 + \delta_K}$$

Now let

$$z = \theta x + (1 - \theta)y \text{ where } \theta \in [0, 1].$$

Then

$$\|z\|_2 = \|\theta x + (1 - \theta)y\|_2 \leq \theta \|x\|_2 + (1 - \theta) \|y\|_2 \leq \theta + (1 - \theta) = 1.$$

Further

$$\begin{aligned}
\|\Phi z\|_2 &= \|\Phi(\theta x + (1 - \theta)y)\|_2 \\
&\leq \|\Phi\theta x\|_2 + \|\Phi(1 - \theta)y\|_2 \\
&= \theta\|\Phi x\|_2 + (1 - \theta)\|\Phi y\|_2 \\
&\leq \theta\sqrt{1 + \delta_K} + (1 - \theta)\sqrt{1 + \delta_K} \\
&\leq \sqrt{1 + \delta_K}.
\end{aligned}$$

It can be shown that for every vector  $x \in S$  we have  $\|x\|_2 \leq 1$  and  $\|\Phi x\|_2 \leq \sqrt{1 + \delta_K}$ .

We now define another convex body

$$\Gamma = \left\{ x : \|x\|_2 + \frac{1}{\sqrt{K}}\|x\|_1 \leq 1 \right\} \subset \mathbb{C}^N. \quad (3.1.45)$$

We quickly verify the convexity property Let  $x, y \in \Gamma$ . Let

$$z = \theta x + (1 - \theta)y \quad \text{where } \theta \in [0, 1].$$

Then

$$\begin{aligned}
&\|z\| + \frac{1}{\sqrt{K}}\|z\|_1 \\
&= \|\theta x + (1 - \theta)y\|_2 + \frac{1}{\sqrt{K}}\|\theta x + (1 - \theta)y\|_1 \\
&\leq \theta\|x\|_2 + (1 - \theta)\|y\|_2 + \frac{\theta}{\sqrt{K}}\|x\|_1 + \frac{(1 - \theta)}{\sqrt{K}}\|y\|_1 \\
&= \theta \left[ \|x\|_2 + \frac{1}{\sqrt{K}}\|x\|_1 \right] + (1 - \theta) \left[ \|y\|_2 + \frac{1}{\sqrt{K}}\|y\|_1 \right] \\
&\leq \theta + (1 - \theta) = 1.
\end{aligned}$$

Thus  $z \in \Gamma$ . This analysis shows that all convex combinations of elements in  $\Gamma$  belong to  $\Gamma$ . Thus  $\Gamma$  is convex. Further it can be verified that  $\Gamma$  is a compact convex set with non-empty interior. Hence its a convex body. For any  $x \in \mathbb{C}^N$  one can find a  $y \in \Gamma$  by simply applying an appropriate non-zero scale  $y = cx$  where the scale factor  $c$  depends on  $x$ .

For a moment suppose that  $\Gamma \subset S$ . Then if  $y \in \Gamma$  the following are true:

$$\|y\|_2 + \frac{1}{\sqrt{K}}\|y\|_1 \leq 1$$

and

$$\|\Phi y\|_2 \leq \sqrt{1 + \delta_K}.$$

Now consider an arbitrary non-zero  $x \in \mathbb{C}^N$ . Let

$$\alpha = \|x\|_2 + \frac{1}{\sqrt{K}}\|x\|_1.$$

Define

$$y = \frac{1}{\alpha}x.$$

Then

$$\|y\|_2 + \frac{1}{\sqrt{K}}\|y\|_1 = \frac{1}{\alpha} \left( \|x\|_2 + \frac{1}{\sqrt{K}}\|x\|_1 \right) = 1.$$

Thus  $y \in \Gamma$  and

$$\begin{aligned} & \|\Phi y\|_2 \leq \sqrt{1 + \delta_K} \\ \implies & \left\| \Phi \frac{1}{\alpha}x \right\|_2 \leq \sqrt{1 + \delta_K} \\ \implies & \|\Phi x\|_2 \leq \sqrt{1 + \delta_K}\alpha \\ \implies & \|\Phi x\|_2 \leq \sqrt{1 + \delta_K} \left( \|x\|_2 + \frac{1}{\sqrt{K}}\|x\|_1 \right) \quad \forall x \in \mathbb{C}^N \end{aligned} \tag{3.1.46}$$

which is our intended result. Hence if we show that  $\Gamma \subset S$  holds, we would have proven our theorem. We will achieve this by showing that every vector  $x \in \Gamma$  can be shown to be a convex combination of vectors in  $S$ .

We start with an arbitrary  $x \in \Gamma$ . Let  $I = \text{supp}(x)$ . We partition  $I$  into disjoint sets of size  $K$ . Let there be  $J + 1$  such sets given by

$$I = \bigcup_{j=0}^J I_j.$$

Let  $I_0$  index the  $K$  largest entries in  $x$  (magnitude wise). Let  $I_1$  be next  $K$  largest entries and so on. Since  $|I|$  may not be a multiple of  $K$ , hence the last index set  $I_J$  may not have  $K$  indices. We define

$$x_{I_j}(i) = \begin{cases} x(i) & \text{if } i \in I_j; \\ 0 & \text{otherwise.} \end{cases}$$

Thus we can write

$$x = \sum_{j=0}^J x_{I_j}.$$

Now let

$$\theta_j = \|x_{I_j}\|_2 \quad \text{and} \quad y_j = \frac{1}{\theta_j} x_{I_j}.$$

We can write

$$x = \sum_{j=0}^J \theta_j y_j.$$

In this construction of  $x$  we can see that  $1 \geq \theta_0 \geq \theta_1 \geq \dots \geq \theta_J \geq 0$ .

Also  $y_j \in S$  since  $y_j$  is a unit norm  $K$  sparse vector eq. (3.1.44).

We will now show that  $\sum_j \theta_j \leq 1$ . This will imply that  $x$  is a convex combination of vectors from  $S$ . But since  $S$  is convex hence  $x \in S$ . This will imply that  $K \subset S$ . The proof will be complete.

Pick any  $j \in \{1, \dots, J\}$ . Since  $x_{I_j}$  is  $K$ -sparse hence due to lemma 2.16 we have

$$\theta_j = \|x_{I_j}\|_2 \leq \sqrt{K} \|x_{I_j}\|_\infty.$$

It is easy to see that  $I_{j-1}$  identifies exactly  $K$  non-zero entries in  $x$  and each of non-zero entries in  $x_{I_{j-1}}$  is larger than the largest entry in  $x_{I_j}$  (magnitude wise). Thus we have

$$\|x_{I_{j-1}}\|_1 = \sum_{i \in I_{j-1}} |x_i| \geq \sum_{i \in I_{j-1}} \|x_{I_j}\|_\infty = K \|x_{I_j}\|_\infty.$$

Thus

$$\|x_{I_j}\|_\infty \leq \frac{1}{K} \|x_{I_{j-1}}\|_1.$$

Combining the two inequalities we get

$$\theta_j \leq \frac{1}{\sqrt{K}} \|x_{I_{j-1}}\|_1.$$

This lets us write

$$\sum_{j=1}^J \theta_j \leq \sum_{j=1}^J \frac{1}{\sqrt{K}} \|x_{I_{j-1}}\|_1 \leq \frac{1}{\sqrt{K}} \|x\|_1$$

since

$$\|x\|_1 = \sum_{j=0}^J \|x_{I_j}\|_1 \geq \sum_{j=1}^J \|x_{I_{j-1}}\|_1.$$

Finally

$$\theta_0 = \|x_{I_0}\|_2 \leq \|x\|_2.$$

This gives us the inequality

$$\sum_{j=0}^J \theta_j \leq \|x\|_2 + \frac{1}{\sqrt{K}} \|x\|_1 \leq 1$$

since  $x \in \Gamma$ . Recalling our steps we can express  $x$  as

$$x = \theta_j y_j$$

where  $y_j \in S$  and  $\sum \theta_j \leq 1$  implies that  $x \in S$  since  $S$  is convex. Thus  $\Gamma \subset S$ . This completes the proof.  $\square$

### 3.1.12. A general form of RIP

A more general restricted isometry bound can be for an arbitrary matrix  $\Phi$  can be as follows

$$\alpha \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq \beta \|x\|_2^2 \tag{3.1.47}$$

where  $0 < \alpha \leq \beta < \infty$ .

Its straightforward to scale  $\Phi$  to match the bounds in (3.1.1).

Let  $\delta_K = \frac{\beta-\alpha}{\alpha+\beta}$ . Then  $1 - \delta_K = \frac{2\alpha}{\alpha+\beta}$  and  $1 + \delta_K = \frac{2\beta}{\alpha+\beta}$ .

Putting in (3.1.1) we get

$$\begin{aligned} \frac{2\alpha}{\alpha + \beta} \|x\|_2^2 &\leq \|\Phi x\|_2^2 \leq \frac{2\beta}{\alpha + \beta} \|x\|_2^2 \\ \implies \alpha \|x\|_2^2 &\leq \|\sqrt{\frac{\alpha + \beta}{2}} \Phi x\|_2^2 \leq \beta \|x\|_2^2 \end{aligned}$$

Thus by multiplying  $\Phi$  with  $\sqrt{2/(\alpha + \beta)}$  we can transform the more general bound to the form of (3.1.1).

### 3.1.13. Finding out RIP constants

**Definition 3.3** The optimal value of RIP constant of  $K$ -th order  $\delta_K$  denoted as  $\delta_K^*$  can be obtained by solving the following optimization problem.

$$\begin{aligned} \underset{0 < \delta_K < 1}{\text{minimize}} \quad & \delta_K \\ \text{subject to} \quad & (1 - \delta_K) \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2 \quad \forall x \in \Sigma_K. \end{aligned} \tag{3.1.48}$$

Of course this problem isn't easy to solve. In fact it has been shown in [3] that this problem is NP-hard.

### 3.1.14. RIP and coherence

Here we establish a relationship between the RIP constants and coherence of a dictionary.

Rather than a general matrix  $\Phi$ , we restrict our attention to a dictionary  $\mathcal{D} \in \mathbb{C}^{N \times D}$ . We assume that the dictionary is overcomplete ( $D > N$ ) and full rank  $\text{rank}(\mathcal{D}) = N$ . Dictionary is assumed to satisfy RIP of some order.

**Theorem 3.29** *Let  $\mathcal{D}$  satisfy RIP of order  $K$ . Then*

$$\delta_K \leq (K - 1)\mu(\mathcal{D}). \tag{3.1.49}$$

PROOF. We recall that  $\delta_K$  is the smallest constant  $\delta$  satisfying

$$(1 - \delta)\|x\|_2^2 \leq \|\mathcal{D}x\|_2^2 \leq (1 + \delta)\|x\|_2^2 \quad \forall x \in \Sigma_K.$$

Let  $\Lambda$  be any index set with  $|\Lambda| = K$ . Then

$$\|\mathcal{D}x\|_2^2 = \|\mathcal{D}_\Lambda x_\Lambda\|_2^2 \quad \forall x \in \mathbb{C}^\Lambda.$$

Since  $\mathcal{D}$  satisfies RIP of order  $K$ , hence  $\mathcal{D}_\Lambda$  is a sub dictionary.

We **recall** that

$$(1 - (K - 1)\mu)\|v\|_2^2 \leq \|\mathcal{D}_\Lambda v\|_2^2 \leq (1 + (K - 1)\mu)\|v\|_2^2.$$

Since  $\delta_K$  is smallest possible constant, hence

$$1 + \delta_K \leq 1 + (K - 1)\mu \implies \delta_K \leq (K - 1)\mu(\mathcal{D}).$$

□

### 3.2. Johnson Lindenstrauss theorem

Consider a high dimensional Euclidean space  $\mathbb{R}^N$ . We are talking about thousands to millions of dimensions. For example, an HD image has  $1080 \times 1920 = 2073600$  pixels. Consider a finite set of points  $S$  in this space. Now consider another Euclidean space  $\mathbb{R}^M$  with much lower dimensionality (i.e.  $M \ll N$ ). Map the points in  $S$  into this subspace through some mapping  $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ . Is it possible to approximately preserve the distances between the points in  $S$  while mapping to the subspace?

More specifically, the objective is to find out a mapping  $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$  such that for any  $x, y \in S$ , the following holds

$$(1 - \delta)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \delta)\|x - y\|_2^2 \quad (3.2.1)$$

where  $\delta \in (0, 1)$ .

For example if  $\delta = 0.1$  And let  $\|x - y\|_2^2 = d^2$ . Then we want

$$0.9d^2 \leq \|f(x) - f(y)\|_2^2 \leq 1.1d^2$$

for a suitably chosen  $M$  and  $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ .

We call such a map  $f$  a **low distortion embedding** of the data set  $S$  from the ambient space  $\mathbb{R}^N$  to a lower dimensional space  $\mathbb{R}^M$ .

It may not be feasible to enforce this bound for all values of  $M < N$  but if we can find a suitable  $M \ll N$ , then we can reduce our high-dimensional data set  $S$  to a much lower dimensional data set  $f(S)$ .

Since such a mapping guarantees that distances between points in our data set are approximately preserved, hence many of inferencing tasks can still be computed with the embedded data set. Carrying out such tasks with the original high-dimensional data set might be computationally infeasible, yet doing the same with the lower dimensional dataset might be much more computationally tractable in specific applications.

The Johnson-Lindenstrauss theorem (JL theorem in short) provides us specific guarantees that if  $M = \mathcal{O}\left(\frac{\ln K}{\delta^2}\right)$ , then such a map  $f$  can indeed be found. In this section we will gradually develop the JL theorem.

The development in this section closely follows [16].

We first state the theorem.

**Theorem 3.30** *For any  $0 < \delta < 1$  and any integer  $K$ , let  $M$  be a positive integer such that*

$$M \geq 4 \left( \frac{\delta^2}{2} - \frac{\delta^3}{3} \right)^{-1} \ln K. \quad (3.2.2)$$

*Then for any set  $S$  of  $K$  points in  $\mathbb{R}^N$ , there is a map  $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$  such that for all  $x, y \in S$ , the following holds:*

$$(1 - \delta)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \delta)\|x - y\|_2^2. \quad (3.2.3)$$

*Furthermore, such a map can be found in randomized polynomial time.*

The theorem doesn't specifically require  $M \ll N$ . Rather, it doesn't even care much about the dimension of ambient space  $N$ . The bound

on  $M$  depends on the required  $\delta$  and number of points in the set  $S$  given by  $K = |S|$ .

fig. 3.1 shows how the required dimension of subspace  $\mathbb{R}^M$  varies w.r.t. the number of points in  $S$  given by  $K = |S|$ .

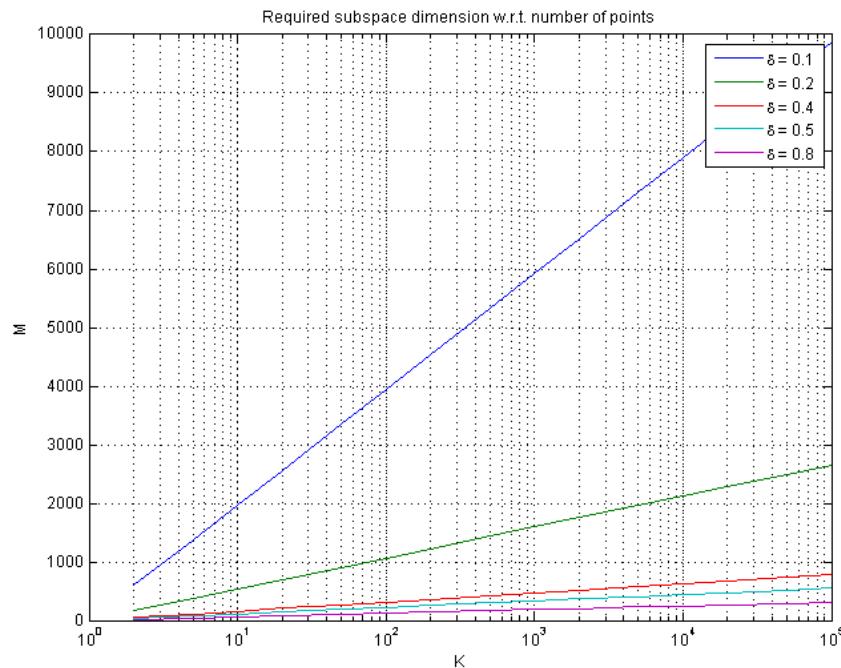


FIGURE 3.1. Required subspace dimension  $M$  for  $K$  points for restricted isometry constant  $\delta$

Let us more closely look at specific set of numbers for  $\delta = 0.1$ .

```

K =      2, M=595
K =      5, M=1380
K =     10, M=1974
K =    20, M=2568
K =   50, M=3354
K =  100, M=3948
K = 200, M=4542
K = 500, M=5327
  
```

```

K = 1000, M=5921
K = 2000, M=6516
K = 5000, M=7301
K = 10000, M=7895
K = 20000, M=8489
K = 50000, M=9275
K = 100000, M=9869

```

As we can see for  $\delta = 0.1$ , when the number of points is small, required  $M$  tends to be quite high. But as  $K$  increases,  $M$  increases only logarithmically. So if we have 100 thousand images in the ambient space of 2 million pixels, we can map it to a Euclidean subspace of dimension  $M = 10000$  while still preserving distances (squared) between those images within 10% variation. Such a dimensionality reduction could be a huge time and space saver in applications like face recognition.

In the sequel we will develop the proof of this theorem through fairly standard line of reasoning based on concentration of measures.

Let  $X_1, \dots, X_N$  be  $N$  independent Gaussian  $\mathcal{N}(0, 1)$  random variables constituting a random vector

$$X = (X_1, \dots, X_N). \quad (3.2.4)$$

Let

$$Y = \frac{X}{\|X\|_2} = (Y_1, \dots, Y_N). \quad (3.2.5)$$

Clearly  $\|Y\|_2 = 1$ . Thus  $Y$  is a point on the surface of the unit hypersphere in  $\mathbb{R}^N$ . It can be shown that  $Y$  follows a uniform distribution over the surface of unit hypersphere.

We note that by symmetry of design  $Y_i$  are identically distributed. Thus we have

$$\mathbb{E}(Y_1^2) = \mathbb{E}(Y_2^2) = \dots = \mathbb{E}(Y_N^2).$$

But  $\sum Y_i^2 = 1$ . Hence we have

$$\mathbb{E}(Y_i^2) = \frac{1}{N}.$$

Let  $Z \in \mathbb{R}^M$  be the projection of  $Y$  onto its first  $M$  coordinates. i.e.

$$Z = (Y_1, \dots, Y_M). \quad (3.2.6)$$

Let

$$L = \|Z\|_2^2 = Y_1^2 + \dots + Y_M^2. \quad (3.2.7)$$

Clearly

$$\mu = \mathbb{E}(L) = \frac{M}{N}. \quad (3.2.8)$$

The following lemma shows that  $L$  is also fairly tightly concentrated around  $\mu$ .

**Lemma 3.31** *Let  $M < N$ . Then*

(1) *If  $\beta < 1$ , then*

$$\mathbb{P}\left[L \leq \beta \frac{M}{N}\right] \leq \beta^{\frac{M}{2}} \left(1 + \frac{(1-\beta)M}{N-M}\right)^{(N-M)/2} \leq \exp\left(\frac{M}{2}(1-\beta + \ln \beta)\right). \quad (3.2.9)$$

(2) *If  $\beta > 1$ , then*

$$\mathbb{P}\left[L \geq \beta \frac{M}{N}\right] \leq \beta^{\frac{M}{2}} \left(1 + \frac{(1-\beta)M}{N-M}\right)^{(N-M)/2} \leq \exp\left(\frac{M}{2}(1-\beta + \ln \beta)\right). \quad (3.2.10)$$

Above we have shown that when a random unit norm vector has been projected to a fixed  $M$  dimensional subspace, its length concentrates strongly around its mean  $\frac{M}{N}$ .

Now consider the problem of projecting a given unit norm vector  $x$  to a random  $M$ -dimensional subspace  $V$ . Choose an orthonormal basis for  $\mathbb{R}^N$  such that each point in the subspace  $V$  can be expressed as  $(v_1, \dots, v_M, 0, \dots, 0)$ . Since the subspace  $V$  is randomly chosen, hence the corresponding orthonormal basis is also randomly chosen, thus in this basis the original vector  $x$  can be expressed as a random unit norm vector and the projection into  $V$  is same as picking up the first  $M$  elements of representation of  $x$  in this basis. Thus the two problems above are equivalent.

Hence, if a unit norm vector  $x \in \mathbb{R}^N$  is projected to an arbitrary  $M$ -dimensional subspace then the squared length of its projection has a mean  $\mu = \frac{M}{N}$  and it strongly concentrates around its mean.

We are now ready to complete the proof of JL theorem.

PROOF. We are given that

- $\delta \in (0, 1)$
- $M$  satisfies the following inequality

$$M \geq 4 \left( \frac{\delta^2}{2} - \frac{\delta^3}{3} \right)^{-1} \ln K.$$

We need to find a map  $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$  and show that with this map

$$(1 - \delta) \|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \delta) \|x - y\|_2^2. \quad (3.2.11)$$

holds for every  $x, y \in S$ .

We note that if  $x = y$ , then this bound holds trivially. Thus we need to consider only those cases where  $x \neq y$ .

We will construct  $f$  as a linear mapping.

If  $M \geq N$ , then the theorem is trivial. For any point  $x \in S$ , let  $x$  be given by

$$x = (x_1, \dots, x_N).$$

We define  $v = f(x)$  by the rule

$$v = (x_1, \dots, x_N, 0, \dots, 0).$$

This mapping trivially preserves all distances with  $\delta = 0$ .

For  $M < N$ , we consider a randomly chosen  $M$ -dimensional subspace  $V$  of  $\mathbb{R}^N$ .

Let elements of  $S$  be given as

$$S = \{s_1, \dots, s_K\}.$$

The required bounds on length in this theorem can be restated as

$$(1 - \delta) \leq \left\| f \frac{(s_i - s_j)}{\|s_i - s_j\|_2} \right\|_2^2 \leq (1 + \delta)$$

which should hold for every pair  $(s_i \in S, s_j \in S)$  with  $i \neq j$ .

Let

$$x = s_i - s_j.$$

and

$$y = \frac{x}{\|x\|_2} = \frac{(s_i - s_j)}{\|s_i - s_j\|_2}.$$

We can rewrite the required bounds as

$$(1 - \delta) \leq \|fy\|_2^2 \leq (1 + \delta)$$

for some fixed unit norm vector  $y$  as defined above.

Thus we are considering the problem of distribution of the length squared of projection of a fixed unit norm vector on to a randomly chosen  $M$ -dimensional subspace.

Let  $P_V$  be the orthonormal projection matrix for the subspace  $V$ .

Let  $v_i \in V$  be the projection of  $s_i \in S$ . i.e.

$$v_i = P_V s_i.$$

We define

$$L = \|v_i - v_j\|_2^2 = \|P_V s_i - P_V s_j\|_2^2 = \|P_V x\|_2^2$$

and

$$\mu = \frac{M}{N} \|s_i - s_j\|_2^2 = \frac{M}{N} \|x\|_2^2.$$

Let

$$Z = P_V y$$

be a random vector.

Let

$$\beta = (1 - \delta) < 1.$$

Now

$$\begin{aligned}\mathbb{P} \left( \|Z\|_2^2 \leq \beta \frac{M}{N} \right) &= \mathbb{P} \left( \|\|x\|_2 Z\|_2^2 \leq \beta \frac{M}{N} \|x\|_2^2 \right) \\ &= \mathbb{P} \left( \|P_V x\|_2^2 \leq \beta \mu \right) \\ &= \mathbb{P} (L \leq (1 - \delta)\mu).\end{aligned}$$

Applying previous lemma we have:

$$\begin{aligned}\mathbb{P}[L \leq (1 - \delta)\mu] &\leq \exp \left( \frac{M}{2} (1 - (1 - \delta) + \ln(1 - \delta)) \right) \\ &\leq \exp \left( \frac{M}{2} \left( \delta - \left( \delta + \frac{\delta^2}{2} \right) \right) \right) = \exp \left( -\frac{M\delta^2}{4} \right) \\ &\leq \exp(-2 \ln K) = \frac{1}{K^2}\end{aligned}$$

In 2nd line we use the property that

$$\ln(1 - x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \dots \leq -x - \frac{x^2}{2} \quad \forall x \in [0, 1).$$

In third line we go as follows

$$\frac{M\delta^2}{4} \geq \left( \frac{1}{2} - \frac{\delta}{3} \right)^{-1} \ln K$$

where

$$0 < \delta < 1 \implies \frac{1}{2} - \frac{\delta}{3} < \frac{1}{2} \implies \left( \frac{1}{2} - \frac{\delta}{3} \right)^{-1} > 2.$$

Thus

$$\frac{M\delta^2}{4} \geq 2 \ln K$$

hence

$$\exp \left( -\frac{M\delta^2}{4} \right) \leq \exp(-2 \ln K).$$

Again

$$\begin{aligned}
\mathbb{P}\left(\|Z\|_2^2 \leq \beta \frac{M}{N}\right) &= \mathbb{P}\left(\frac{N}{M}\|Z\|_2^2 \leq \beta\right) \\
&= \mathbb{P}\left(\left\|\sqrt{\frac{N}{M}}Z\right\|_2^2 \leq \beta\right) \\
&= \mathbb{P}\left(\left\|\sqrt{\frac{N}{M}}Z\right\|_2^2 \leq \beta\right) \\
&= \mathbb{P}\left(\left\|\sqrt{\frac{N}{M}}P_Vy\right\|_2^2 \leq \beta\right) \\
&= \mathbb{P}\left(\left\|\sqrt{\frac{N}{M}}P_Vx\right\|_2^2 \leq \beta\|x\|_2^2\right) \\
&= \mathbb{P}\left(\left\|\sqrt{\frac{N}{M}}P_V(s_i - s_j)\right\|_2^2 \leq (1 - \delta)\|s_i - s_j\|_2^2\right)
\end{aligned}$$

So we choose

$$f = \sqrt{\frac{N}{M}}P_V \quad (3.2.12)$$

and establish that

$$\mathbb{P}(\|f(s_i) - f(s_j)\|_2^2 \leq (1 - \delta)\|s_i - s_j\|_2^2) \leq \frac{1}{K^2}.$$

Similarly we apply 2nd part of previous lemma for  $\beta = 1 + \delta$  and use the inequality

$$\ln(1 + x) \leq x - \frac{x^2}{2} + \frac{x^3}{3} \quad \forall 0 \leq x < 1.$$

$$\begin{aligned}
\mathbb{P}(L \geq (1 + \delta)\mu) &\leq \exp\left(\frac{M}{2}(1 - (1 + \delta) + \ln(1 + \delta))\right) \\
&\leq \exp\left(\frac{M}{2}\left(-\delta + \left(\delta - \frac{\delta^2}{2} + \frac{\delta^3}{3}\right)\right)\right) \\
&= \exp\left(-\frac{M(\delta^2/2 - \delta^3/3)}{2}\right) \\
&\leq \exp(-2 \ln K) = \frac{1}{K^2}.
\end{aligned}$$

Thus the probability that the constraint

$$(1 - \delta)\|s_i - s_j\|_2^2 \leq \|f(s_i) - f(s_j)\|_2^2 \leq (1 + \delta)\|s_i - s_j\|_2^2$$

is not satisfied is at most  $\frac{2}{K^2}$ .

Number of possible pairs  $(s_i, s_j)$  is  $\binom{K}{2}$ .

Thus the probability that one or more pairs suffer a large distortion is at most

$$\binom{K}{2} \times \frac{2}{K^2} = 1 - \frac{1}{K}.$$

Hence the probability that  $f$  has the desired properties for approximate distance preservation is at least  $\frac{1}{K}$ .

Randomly choosing the  $M$ -dimensional subspace  $\mathcal{O}(K)$  times can boost the probability of finding the right  $f$  with the desired properties to the desired constant, giving the claimed randomized polynomial time algorithm for finding the map  $f$ .  $\square$

In the following we present a simple randomized algorithm for finding the desired mapping.

- (1) Compute  $M$  based on given  $\delta$  and  $K$ .
- (2) Construct a Gaussian random matrix  $G$  of size  $N \times M$ .
- (3) Using Gauss-Schmidt orthonormalization, orthonormalize the columns of  $G$  to construct  $P$ .
- (4) Define  $f = \sqrt{\frac{M}{N}}P'$ .

- (5) Compute  $f(s_i)$ .
- (6) Verify that approximate distance preservation constraints are met.
- (7) If yes, then return  $f$  as the desired mapping.
- (8) Otherwise go back to step 2.

### 3.3. Stable embeddings

Stable embeddings are a generalization of the idea of restricted isometry property. The discussion in this section is largely based on [18].

**Definition 3.4** Let  $\delta \in (0, 1)$  and  $U, V \subset \mathbb{C}^N$  be any two arbitrary sets. We say that a mapping  $\Phi : \mathbb{C}^N \rightarrow \mathbb{C}^M$  is a  **$\delta$ -stable embedding** of  $(U, V)$  if

$$(1 - \delta)\|u - v\|_2^2 \leq \|\Phi u - \Phi v\|_2^2 \leq (1 + \delta)\|u - v\|_2^2 \quad (3.3.1)$$

holds for all  $u \in U$  and for all  $v \in V$ .

A mapping satisfying this property is also called **bi-Lipschitz**.

**Example 3.2: RIP as stable embedding** A matrix satisfying RIP of order  $2K$  is equivalent to being a  $\delta_{2K}$ -stable embedding of  $(\Sigma_K, \Sigma_K)$  or of  $(\Sigma_{2K}, \{0\})$ .

Let  $\Phi$  satisfy RIP of order  $2K$ . Then we have

$$(1 - \delta_{2K})\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_{2K})\|x\|_2^2$$

for every  $x \in \Sigma_{2K}$  with  $\delta_{2K} \in (0, 1)$ .

Now let  $U = \Sigma_K$  and  $V = \Sigma_K$ . Clearly for every  $u \in U$  and every  $v \in V$ , we have  $(u - v) \in \Sigma_{2K}$ .

Thus we have

$$(1 - \delta_{2K})\|u - v\|_2^2 \leq \|\Phi u - \Phi v\|_2^2 \leq (1 + \delta_{2K})\|u - v\|_2^2$$

for all  $u \in U$  and for all  $v \in V$ .

Thus  $\Phi$  is a  $\delta_{2K}$ -stable embedding of  $(\Sigma_K, \Sigma_K)$ .

Similarly if  $U = \Sigma_{2K}$  and  $V = \{0\}$ , then for every  $u \in U$  and every  $v \in V$ , we have  $(u - v) = u \in \Sigma_{2K}$ . Following the logic above,  $\Phi$  is a  $\delta_{2K}$ -stable embedding of  $(\Sigma_{2K}, \{0\})$ .  $\square$

**Example 3.3: Stable embedding of sparse signals in an orthonormal basis** Let  $\Psi$  be some orthonormal basis in  $\mathbb{C}^N$ . We define the set of  $K$ -sparse signals in  $\Psi$  as

$$\Psi(\Sigma_K) = \{x : x = \Psi\alpha \text{ with } \|\alpha\|_0 \leq K\}. \quad (3.3.2)$$

The set of  $2K$ -sparse signals in  $\Psi$  will naturally be

$$\Psi(\Sigma_{2K}) = \{x : x = \Psi\alpha \text{ with } \|\alpha\|_0 \leq 2K\}. \quad (3.3.3)$$

Now let  $\mathcal{X} = \Phi\Psi$  be a matrix which satisfies RIP of order  $2K$ . Then  $\Phi$  is a  $\delta_{2K}$ -stable embedding of  $(\Psi(\Sigma_K), \Psi(\Sigma_K))$  or  $(\Psi(\Sigma_{2K}), \{0\})$ .

PROOF. Let  $U = \Psi(\Sigma_K)$  and  $V = \Psi(\Sigma_{2K})$ .

Let  $u \in U$  and  $v \in V$  be some arbitrary vectors. Then there exist  $\alpha, \beta \in \Sigma_K$  such that

$$u = \Psi\alpha, \quad v = \Psi\beta.$$

With this  $\alpha - \beta \in \Sigma_{2K}$ .

Since  $\mathcal{X}$  satisfies RIP of order  $2K$ , hence

$$(1 - \delta_{2K})\|\alpha - \beta\|_2^2 \leq \|\mathcal{X}(\alpha - \beta)\|_2^2 \leq (1 + \delta_{2K})\|\alpha - \beta\|_2^2$$

Since an orthonormal basis preserves  $l_2$  norms and distances hence

$$\|u - v\|_2^2 = \|\alpha - \beta\|_2^2.$$

Also

$$\mathcal{X}(\alpha - \beta) = \Phi\Psi(\alpha - \beta) = \Phi(\Psi\alpha - \Psi\beta) = \Phi(u - v).$$

Putting these back we get

$$(1 - \delta_{2K})\|u - v\|_2^2 \leq \|\Phi u - \Phi v\|_2^2 \leq (1 + \delta_{2K})\|u - v\|_2^2.$$

This establishes that  $\Phi$  is a  $\delta_{2K}$ -stable embedding of  $(\Psi(\Sigma_K), \Psi(\Sigma_K))$ .

Similar exercise establishes that  $\Phi$  is a  $\delta_{2K}$ -stable embedding of  $(\Psi(\Sigma_{2K}), \{0\})$ .  $\square$

We now provide a number of results related to stable embeddings.

### 3.3.1. Stable embeddings of finite sets of points

Consider the simple case where  $U$  and  $V$  are finite set of points in  $\mathbb{C}^N$ .

Let  $a = |U|$  and  $b = |V|$  be the number of elements in  $U$  and  $V$  respectively.

Thus

$$U = \{u_1, \dots, u_a\}$$

and

$$V = \{v_1, \dots, v_b\}.$$

**Lemma 3.32** *Let  $U$  and  $V$  be finite sets of points in  $\mathbb{C}^N$ . Fix  $\delta, \beta \in (0, 1)$ .*

*Let  $\Phi$  be an  $M \times N$  random matrix with i.i.d. entries chosen from a subgaussian distribution.*

*If*

$$M \geq \frac{\ln(|U||V|) + \ln\left(\frac{2}{\beta}\right)}{c\delta^2} \quad (3.3.4)$$

*then with probability exceeding  $1 - \beta$ ,  $\Phi$  is a  $\delta$ -stable embedding of  $(U, V)$ .*

### 3.3.2. Stable embeddings of K dimensional subspaces

We now consider the case where  $U = X$  is a  $K$ -dimensional subspace of  $\mathbb{C}^N$  and  $V = \{0\}$ .

Thus we wish to obtain a  $\Phi$  that nearly preserves the norm of any vector  $x \in X$ .

So rather than working with a finite set, here we have an uncountable set in hand.

Following lemma states the conditions under which this can be achieved.

**Lemma 3.33** Suppose that  $X$  is a  $K$ -dimensional subspace of  $\mathbb{C}^N$ .

Fix  $\delta, \beta \in (0, 1)$ .

Let  $\Phi$  be an  $M \times N$  random matrix with i.i.d. entries chosen from a subgaussian distribution.

If

$$M \geq 2 \frac{K \ln\left(\frac{42}{\delta}\right) + \ln\left(\frac{2}{\beta}\right)}{c\delta^2} \quad (3.3.5)$$

then with probability exceeding  $1 - \beta$ ,  $\Phi$  is a  $\delta$ -stable embedding of  $(X, \{0\})$ .

We can extend this result beyond a single  $K$ -dimensional subspace to all possible  $K$  dimensional subspaces that are defined w.r.t. an orthonormal basis  $\Psi$ .

**Lemma 3.34** Let  $\Psi$  be an orthonormal basis for  $\mathbb{C}^N$ . Fix  $\delta, \beta \in (0, 1)$ .

Let  $\Phi$  be an  $M \times N$  random matrix with i.i.d. entries chosen from a subgaussian distribution.

If

$$M > 2 \frac{K \ln\left(\frac{42eN}{\delta K}\right) + \ln\left(\frac{2}{\beta}\right)}{c\delta^2} \quad (3.3.6)$$

then with probability exceeding  $1 - \beta$ ,  $\Phi$  is a  $\delta$ -stable embedding of  $(\Psi(\Sigma_K), \{0\})$ .

### 3.4. Spark

We present some advanced results on spark of a dictionary.

#### 3.4.1. Upper bounds for spark

Whenever a set of atoms in a dictionary are linearly dependent, the dependence corresponds to some vector in its null space. Thus, identifying the spark of a dictionary essentially amounts of sifting through the vectors in its null space and finding one with smallest  $l_0$ -“norm”. This can be cast as an optimization problem:

$$\begin{aligned} & \underset{v}{\text{minimize}} && \|v\|_0 \\ & \text{subject to} && \mathcal{D}v = 0. \end{aligned} \tag{3.4.1}$$

Note that the solution  $v^*$  of this problem is not unique. If  $v^*$  is a solution that  $cv^*$  for any  $c \neq 0$  is also a solution. Spark is the optimum value of the objective function  $\|v\|_0$ .

We now define a sequence of optimization problems for  $k = 1, \dots, D$

$$\begin{aligned} & \underset{v}{\text{minimize}} && \|v\|_0 \\ & \text{subject to} && \mathcal{D}v = 0, v_k = 1. \end{aligned} \tag{R}_k$$

The  $k$ -th problem constrains the solution to choose atom  $d_k$  from the dictionary. Since the minimal set of linearly dependent atoms in  $\mathcal{D}$  will contain at least two vectors, hence  $\text{spark}(\mathcal{D})$  would correspond to the optimal value of one (or more) of the problems  $(R_k)$ .

Formally, if we denote  $v_k^{0,*}$  as an optimal vector for the problem  $(R_k)$ , then

$$\text{spark}(\mathcal{D}) = \underset{1 \leq k \leq D}{\text{minimize}} \|v_k^{0,*}\|_0. \tag{3.4.2}$$

Thus, solving (3.4.1) is equivalent to solving all  $D$  problems specified by  $(R_k)$  and then finding the minimum  $l_0$ -“norm” amongst them. The problems  $(R_k)$  are still computationally intractable.

We now change each of the  $l_0$ -“norm” ( $R_k$ ) minimization problems to  $l_1$ -“norm” minimization problems.

$$\begin{aligned} & \underset{v}{\text{minimize}} \quad \|v\|_1 \\ & \text{subject to } \mathcal{D}v = 0, v_k = 1. \end{aligned} \tag{Q}_k$$

Let us indicate an optimal solution of ( $Q_k$ ) as  $v_k^{1,*}$ . Since  $\mathcal{D}v_k^{1,*} = 0$ , hence  $v_k^{1,*}$  is feasible for ( $R_k$ ). Thus,

$$\|v_k^{0,*}\|_0 \leq \|v_k^{1,*}\|_0.$$

This gives us the relationship

$$\text{spark}(\mathcal{D}) \leq \underset{1 \leq k \leq D}{\text{minimize}} \|v_k^{1,*}\|_0. \tag{3.4.3}$$

We formally state the upper bound on  $\text{spark}(\mathcal{D})$  in the following theorem [20].

**Theorem 3.35** *Let  $\mathcal{D}$  be a dictionary. Then*

$$\text{spark}(\mathcal{D}) \leq \underset{1 \leq k \leq D}{\text{minimize}} \|v_k^{1,*}\|_0 \tag{3.4.4}$$

*where  $v_k^{1,*}$  is a solution of the problem ( $Q_k$ ).*

### 3.5. Coherence

In this section we develop some advanced bounds using coherence of a dictionary.

As usual, we will be considering an overcomplete dictionary  $\mathcal{D} \in \mathbb{C}^{N \times D}$  consisting of  $D$  atoms. The coherence of  $\mathcal{D}$  is denoted by  $\mu(\mathcal{D})$ . In short we will simply write it as  $\mu$ . A sub-dictionary will be indexed by an index set  $\Lambda$  consisting of linearly independent atoms.

**Theorem 3.36** *Suppose that  $(K - 1)\mu < 1$  and assume that  $|\Lambda| \leq K$ . Then*

$$\|\mathcal{D}_{\Lambda}^{\dagger}\|_{2 \rightarrow \infty} \leq \frac{1}{\sqrt{1 - (K - 1)\mu}}. \tag{3.5.1}$$

Equivalently, the rows of  $\mathcal{D}_\Lambda^\dagger$  have  $l_2$  norms no greater than  $\frac{1}{\sqrt{1-(K-1)\mu}}$ .

PROOF. We recall that the operator norm  $\|\mathcal{D}_\Lambda^\dagger\|_{2 \rightarrow \infty}$  computes the maximum  $l_2$  norm among the rows of  $\mathcal{D}_\Lambda^\dagger$ . TODO COMPLETE ITS PROOF.  $\square$

**Definition 3.5** [20, 13] Let  $G = \mathcal{D}^H \mathcal{D}$  be the Gram matrix for dictionary  $\mathcal{D}$ . We define  $\mu_{1/2}(G)$  as the smallest number  $m$  such that the sum of magnitudes of a collection of  $m$  off-diagonal entries in a single row or column of the Gram matrix  $G$  is at least  $\frac{1}{2}$ .

This quantity was introduced in [20] for developing more accurate bounds compared to bounds based on coherence. At that time the idea of Babel function was not available. A careful examination reveals that  $\mu_{1/2}(G)$  can be related to Babel function.

**Theorem 3.37** [20]

$$\mu_{1/2}(G) \geq \frac{1}{2\mu}. \quad (3.5.2)$$

PROOF. Since  $\mu$  is the maximum absolute value of any off diagonal term in  $G = \mathcal{D}^H \mathcal{D}$ , hence sum of any  $m$  terms, say  $T$ , is bounded by

$$T \leq m\mu.$$

Thus

$$T \geq \frac{1}{2} \implies m\mu \geq \frac{1}{2} \implies m \geq \frac{1}{2\mu}.$$

Since  $\mu_{1/2}(G)$  is the minimum number of off diagonal terms whose sum exceeds  $1/2$ , hence

$$\mu_{1/2}(G) \geq \frac{1}{2\mu}.$$

$\square$

**Theorem 3.38** [20]

$$\text{spark}(\mathcal{D}) \geq 2\mu_{1/2}(G) + 1. \quad (3.5.3)$$

PROOF. Let  $h \in \mathcal{N}(\mathcal{D})$ . Then

$$\mathcal{D}h = 0 \implies Gh = \mathcal{D}^H \mathcal{D}h = 0.$$

Subtracting both sides with  $h$  we get

$$Gh - h = (G - I)h = -h. \quad (3.5.4)$$

Let  $\Lambda = \text{supp}(h)$ . By taking columns indexed by  $\Lambda$  from  $G - I$  and corresponding entries in  $h$ , we can write:

$$(G - I)_\Lambda h_\lambda = -h.$$

Taking  $l_\infty$  norm on both sides we get

$$\|h\|_\infty = \|(G - I)_\Lambda h_\lambda\|_\infty.$$

We know that

$$\|(G - I)_\Lambda h_\lambda\|_\infty \leq \|(G - I)_\Lambda\|_\infty \|h_\lambda\|_\infty$$

and it is easy to see that:

$$\|h_\lambda\|_\infty = \|h\|_\infty.$$

Thus

$$\|h\|_\infty \leq \|(G - I)_\Lambda\|_\infty \|h\|_\infty.$$

This gives us

$$\|(G - I)_\Lambda\|_\infty \geq 1.$$

But  $\|(G - I)_\Lambda\|_\infty$  is nothing but the maximum sum of magnitudes of off diagonal entries in  $G$  along a row in  $G_\Lambda$ .

Consider any row in  $(G - I)_\Lambda$ . One of the entries in the row (on the main diagonal of  $G - I$ ) is 0. Thus, there are a maximum of  $|\Lambda| - 1$  non zero entries in the row.

$\Lambda$  is smallest when  $|\Lambda| = \text{spark}(\mathcal{D})$ . For such a  $\Lambda$ , there exists a row in  $G$  such that the sum of  $\text{spark}(\mathcal{D}) - 1$  off diagonal entries in the row exceeds 1.

Let  $n$  denote the minimum number of off diagonal elements on a row or a column of  $G$  such that the sum of their magnitudes exceeds one. Clearly

$$\text{spark}(\mathcal{D}) - 1 \geq n.$$

It is easy to see that

$$n \geq 2\mu_{1/2}(G)$$

i.e. minimum number of off diagonal elements summing up to 1 or more is at least twice the minimum number of off diagonal elements summing up to  $\frac{1}{2}$  or more on any row (or column due to Hermitian property). Thus

$$\text{spark}(\mathcal{D}) - 1 \geq 2\mu_{1/2}(G).$$

Rewriting, we get

$$\text{spark}(\mathcal{D}) \geq 2\mu_{1/2}(G) + 1.$$

□

### 3.6. Babel function

In this section we develop further results on **Babel function**.

We start with a more general development of Babel function for a pair of dictionaries.

When we consider a single dictionary, we will use  $\mathcal{D}$  as the dictionary.

When considering a pair of dictionaries of equal size, we would typically label them as  $\Phi$  and  $\Psi$  with both  $\Phi, \Psi \in \mathbb{C}^{N \times D}$ . We will assume that the dictionaries are full rank as they span the signal space  $\mathbb{C}^N$ .

Why a pair of dictionaries? We consider  $\Phi$  as a modeling dictionary from which the sparse signals

$$x \approx \Phi\alpha$$

are built.

$\Psi$  on the other hand is the sensing dictionary which will be used to compute correlations with the signal  $x$  and try to estimate the approximation  $\alpha$ .

Ideally,  $\Phi$  and  $\Psi$  should be same. But in real life, we may not know  $\Phi$  correctly. Hence,  $\Psi$  would be a dictionary slightly different from  $\Phi$ .

### 3.6.1. p-Babel functions

See [25] for reference.

**Definition 3.6** Consider an index set  $\Lambda \subset \{1, \dots, D\}$  indexing a subset of atoms in  $\Phi$  and  $\Psi$ . The  **$p$ -Babel function** over  $\Lambda$  is defined as

$$\mu_p(\Phi, \Psi, \Lambda) \triangleq \sup_{l \notin \Lambda} \left( \sum_{j \in \Lambda} |\langle \phi_j, \psi_l \rangle|^p \right)^{\frac{1}{p}}. \quad (3.6.1)$$

What is going on here?

Consider the row vector

$$v^l = \psi_l^H \Phi_\Lambda.$$

This vector contains inner products of modeling atoms in  $\Phi$  indexed by  $\Lambda$  with the sensing atom  $\psi_l$ .

Now

$$\|v^l\|_p = \left( \sum_i |v_i^l|^p \right)^{\frac{1}{p}} = \left( \sum_{j \in \Lambda} |\langle \phi_j, \psi_l \rangle|^p \right)^{\frac{1}{p}}$$

This is the term in (3.6.1). Thus

$$\mu_p(\Phi, \Psi, \Lambda) = \sup_{l \notin \Lambda} \|v^l\|_p.$$

$\|v^l\|_p$  is a measure of correlation of the sensing atom  $\psi_l$  with a group of modeling atoms in  $\Phi$  indexed by  $\Lambda$ .

$\mu_p(\Phi, \Psi, \Lambda)$  attempts to find out a sensing atom from  $\Psi$  outside the index set  $\Lambda$  which is most correlated to the group of modeling atoms in  $\Phi$  indexed by  $\Lambda$  and returns the maximum correlation value.

Different choices of  $p$ -norm lead to different correlation values.

We can also measure a correlation of sensing and modeling atoms inside the index set  $\Lambda$ .

**Definition 3.7** A complement to the  $p$ -Babel function measures the amount of correlation between atoms **inside** the support  $\Lambda$ :

$$\mu_p^{\text{in}}(\Phi, \Psi, \Lambda) \triangleq \sup_{i \in \Lambda} \mu_p(\Phi_\Lambda, \Psi_\Lambda, \Lambda \setminus \{i\}). \quad (3.6.2)$$

$\mu_p(\Phi_\Lambda, \Psi_\Lambda, \Lambda \setminus \{i\})$  computes the correlation of  $i$ -th sensing atom in  $\Psi$  with the modeling atoms in  $\Phi$  indexed by  $\Lambda \setminus \{i\}$  i.e. all modeling atoms in  $\Lambda$  except the  $i$ -th modeling atom.

Finally  $\mu_p^{\text{in}}(\Phi, \Psi, \Lambda)$  finds the maximum correlation of any sensing atom inside  $\Lambda$  with modeling atoms inside  $\Lambda$  (leaving the corresponding modeling atom).

So far, we have focused our attention to a specific index set  $\Lambda$ . We now consider all index sets with  $|\Lambda| \leq K$ .

**Definition 3.8** The Babel function for a pair of dictionaries  $\Phi$  and  $\Psi$  as a function of the sparsity level  $K$  is defined as

$$\mu_p(\Phi, \Psi, K) \triangleq \sup_{|\Lambda| \leq K} \mu_p(\Phi, \Psi, \Lambda). \quad (3.6.3)$$

Correspondingly, the complement of Babel function is defined as

$$\mu_p^{\text{in}}(\Phi, \Psi, K) \triangleq \sup_{|\Lambda| \leq K} \mu_p^{\text{in}}(\Phi, \Psi, \Lambda). \quad (3.6.4)$$

**REMARK.** It is straightforward to see that

$$\mu_p^{\text{in}}(\Phi, \Psi, K) \leq \mu_p(\Phi, \Psi, K - 1). \quad (3.6.5)$$

Now consider the special case where  $\mathcal{D} = \Phi = \Psi$ , i.e. the sensing and modeling dictionaries are same.

We obtain

$$\mu_p(\mathcal{D}, \Lambda) = \sup_{l \notin \Lambda} \left( \sum_{j \in \Lambda} |\langle d_j, d_l \rangle|^p \right)^{\frac{1}{p}}. \quad (3.6.6)$$

$$\mu_p^{\text{in}}(\mathcal{D}, \Lambda) = \sup_{i \in \Lambda} \mu_p(\mathcal{D}_\Lambda, \Lambda \setminus \{i\}). \quad (3.6.7)$$

$$\mu_p(\mathcal{D}, K) = \sup_{|\Lambda| \leq K} \mu_p(\mathcal{D}, \Lambda). \quad (3.6.8)$$

$$\mu_p^{\text{in}}(\mathcal{D}, K) = \sup_{|\Lambda| \leq K} \mu_p^{\text{in}}(\mathcal{D}, \Lambda). \quad (3.6.9)$$

Further by choosing  $p = 1$ , we get

$$\mu_1(\mathcal{D}, \Lambda) = \sup_{l \notin \Lambda} \left( \sum_{j \in \Lambda} |\langle d_j, d_l \rangle| \right). \quad (3.6.10)$$

$$\mu_1^{\text{in}}(\mathcal{D}, \Lambda) = \sup_{i \in \Lambda} \mu_1(\mathcal{D}_\Lambda, \Lambda \setminus \{i\}). \quad (3.6.11)$$

$$\mu_1(\mathcal{D}, K) = \sup_{|\Lambda| \leq K} \mu_1(\mathcal{D}, \Lambda). \quad (3.6.12)$$

$$\mu_1^{\text{in}}(\mathcal{D}, K) = \sup_{|\Lambda| \leq K} \mu_1^{\text{in}}(\mathcal{D}, \Lambda). \quad (3.6.13)$$

Finally compare this definition of  $\mu_1(\mathcal{D}, K)$  with the standard definition of **Babel function** as

$$\mu_1(K) = \max_{|\Lambda|=K} \max_{\psi} \sum_{\Lambda} |\langle \psi, d_\lambda \rangle|, \quad (3.6.14)$$

where the vector  $\psi$  ranges over the atoms indexed by  $\Omega \setminus \Lambda$ .

We also know that  $\mu_1(K)$  is an increasing function of  $K$ . Thus, replacing  $|\Lambda| = K$  with  $|\Lambda| \leq K$  doesn't make any difference to the value of  $\mu_1(K)$ .

Careful observation shows that the definitions of  $\mu_1(K)$  in (3.6.14) and  $\mu_1(\mathcal{D}, K)$  in (3.6.12) are exactly the same.

### 3.7. Exact recovery coefficient

In this section we will develop a measure of similarity between a sub-dictionary and the remaining atoms from the dictionary.

As usual,  $\mathcal{D}$  is our dictionary  $\Lambda$  indexes a linearly independent set of atoms giving a subdictionary  $\mathcal{D}_\Lambda$ .

**Definition 3.9** The **Exact Recovery Coefficient** [34, 35, 37] for a subdictionary  $\mathcal{D}_\Lambda$  is defined as

$$\text{ERC}(\mathcal{D}_\Lambda) = 1 - \max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^\dagger d_\omega\|_1. \quad (3.7.1)$$

We will also use the notation  $\text{ERC}(\Lambda)$  when the dictionary is clear from context.

The quantity is called exact recovery coefficient since for a number of algorithms the criteria  $\text{ERC}(\Lambda) > 0$  is a sufficient condition for exact recovery of sparse representations.

#### 3.7.1. ERC and Babel function

We present a lower bound on  $\text{ERC}(\Lambda)$  in terms of Babel function.

**Theorem 3.39** Suppose that  $|\Lambda| = k \leq K$ . A lower bound on Exact Recovery Coefficient is

$$\text{ERC}(\Lambda) \geq \frac{1 - \mu_1(K-1) - \mu_1(K)}{1 - \mu_1(K-1)} \quad (3.7.2)$$

*It follows that  $\text{ERC}(\Lambda) > 0$  whenever*

$$\mu_1(K-1) + \mu_1(K) < 1. \quad (3.7.3)$$

PROOF. Let us expand the pseudo-inverse  $\mathcal{D}_\Lambda^\dagger$ .

$$\begin{aligned} \max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^\dagger d_\omega\|_1 &= \max_{\omega \notin \Lambda} \left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \mathcal{D}_\Lambda^H d_\omega \right\|_1 \\ &\leq \|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1}\|_{1 \rightarrow 1} \max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^H d_\omega\|_1. \end{aligned}$$

For the Gram matrix  $G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda$  we recall that:

$$\|G^{-1}\|_1 \leq \frac{1}{1 - \mu_1(k-1)} \leq \frac{1}{1 - \mu_1(K-1)}.$$

For the other term we have

$$\max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^H d_\omega\|_1 = \max_{\omega \notin \Lambda} \sum_{\lambda \in \Lambda} |\langle d_\omega, d_\lambda \rangle| \leq \mu_1(k) \leq \mu_1(K).$$

Thus, we get

$$\max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^\dagger d_\omega\|_1 \leq \frac{\mu_1(K)}{1 - \mu_1(K-1)}.$$

Putting back in the definition of Exact Recovery Coefficient:

$$\text{ERC}(\Lambda) = 1 - \max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^\dagger d_\omega\|_1 \geq 1 - \frac{\mu_1(K)}{1 - \mu_1(K-1)}.$$

This completes the bound on ERC. Now, we verify the condition for  $\text{ERC}(\Lambda) > 0$ .

$$\begin{aligned} \mu_1(K) + \mu_1(K-1) < 1 &\iff \mu_1(K) < 1 - \mu_1(K-1) \\ &\iff \frac{\mu_1(K)}{1 - \mu_1(K-1)} < 1 \\ &\iff 1 - \frac{\mu_1(K)}{1 - \mu_1(K-1)} > 0. \end{aligned}$$

Thus, if  $\mu_1(K) + \mu_1(K-1) < 1$ , then the lower bound on  $\text{ERC}(\Lambda)$  is positive leading to  $\text{ERC}(\Lambda) > 0$ .  $\square$

### 3.7.2. ERC and coherence

On the same lines we develop a coherence bound for ERC.

**Theorem 3.40** Suppose that  $|\Lambda| = k \leq K$ . A lower bound on Exact Recovery Coefficient is

$$\text{ERC}(\Lambda) \geq \frac{1 - (2K - 1)\mu}{1 - (K - 1)\mu}. \quad (3.7.4)$$

It follows that  $\text{ERC}(\Lambda) > 0$  whenever

$$K\mu \leq \frac{1}{2}. \quad (3.7.5)$$

PROOF. Following the proof of theorem 3.39 for the Gram matrix  $G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda$  have:

$$\|G^{-1}\|_1 \leq \frac{1}{1 - \mu_1(K - 1)} \leq \frac{1}{1 - (K - 1)\mu}.$$

For the other term we have

$$\max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^H d_\omega\|_1 \leq \mu_1(K) \leq K\mu.$$

Thus, we get

$$\max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^\dagger d_\omega\|_1 \leq \frac{K\mu}{1 - (K - 1)\mu}.$$

Putting back in the definition of Exact Recovery Coefficient:

$$\text{ERC}(\Lambda) \geq 1 - \frac{K\mu}{1 - (K - 1)\mu} = \frac{1 - (2K - 1)\mu}{1 - (K - 1)\mu}.$$

This completes the bound on ERC. Now, we verify the condition for  $\text{ERC}(\Lambda) > 0$ .

$$K\mu \leq \frac{1}{2} \implies 2K\mu \leq 1 \implies 1 - 2K\mu \geq 0 \implies 1 - 2K\mu + \mu > 0.$$

And

$$K\mu \leq \frac{1}{2} \implies 1 - K\mu \geq \frac{1}{2} \implies 1 - K\mu + \mu \geq \frac{1}{2} + \mu.$$

Thus  $K\mu \leq \frac{1}{2}$  ensures that both numerator and denominator for the coherence lower bound on  $\text{ERC}(\Lambda)$  are positive leading to  $\text{ERC}(\Lambda) > 0$ .  $\square$

A more accurate bound on  $K$  is presented in the next theorem.

**Theorem 3.41**  $\text{ERC}(\Lambda) > 0$  holds whenever

$$K < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right) \quad (3.7.6)$$

where  $K = |\Lambda|$ .

PROOF. Assuming  $1 - (K - 1)\mu > 0$ , we have

$$\begin{aligned} \frac{1 - (2K - 1)\mu}{1 - (K - 1)\mu} &> 0 \\ \iff 1 - (2K - 1)\mu &> 0 \\ \iff 1 &> (2K - 1)\mu \\ \iff 2K - 1 &< \frac{1}{\mu} \\ \iff K &< \frac{1}{2} \left( 1 + \frac{1}{\mu} \right). \end{aligned}$$

From theorem 3.40, we have

$$\text{ERC}(\Lambda) \geq \frac{1 - (2K - 1)\mu}{1 - (K - 1)\mu}.$$

Thus under the given conditions, we have

$$\text{ERC}(\Lambda) > 0.$$

We also need to show that under these conditions

$$1 - (K - 1)\mu > 0.$$

$$\begin{aligned}
2K - 1 &< \frac{1}{\mu} \\
\implies 2K - 2 &< \frac{1}{\mu} - 1 \\
\implies 2(K - 1)\mu &< 1 - \mu \\
\implies -(K - 1)\mu &> \frac{\mu}{2} - \frac{1}{2} \\
\implies 1 - (K - 1)\mu &> \frac{1}{2} + \frac{\mu}{2} \\
\implies 1 - (K - 1)\mu &> 0.
\end{aligned}$$

□

### 3.7.3. Geometrical interpretation of ERC

**Definition 3.10** The **antipodal convex hull** [35] of a subdictionary  $\mathcal{D}_\Lambda$  is defined as the set of signals given by

$$\mathcal{A}_1(\Lambda) = \{\mathcal{D}_\Lambda x : x \in \mathbb{C}^\Lambda \text{ and } \|x\|_1 \leq 1\}. \quad (3.7.7)$$

It is the smallest convex set that contains every unit multiple of every atom.

We recall that  $P_\Lambda = \mathcal{D}_\Lambda \mathcal{D}_\Lambda^\dagger$  is the orthogonal projector on to the column space of  $\mathcal{D}_\Lambda$ . Therefore  $c_\omega = \mathcal{D}_\Lambda^\dagger d_\omega \in \mathbb{C}^\Lambda$  is a coefficient vector which can be used to synthesize this projection. In other words:

$$P_\Lambda d_\omega = \mathcal{D}_\Lambda \mathcal{D}_\Lambda^\dagger d_\omega = \mathcal{D}_\Lambda c_\omega.$$

Thus, the quantity  $1 - \|\mathcal{D}_\Lambda^\dagger d_\omega\|_1$  measures how far the projected atom  $P_\Lambda d_\omega$  lies from the boundary of  $\mathcal{A}_1(\Lambda)$ .

If every projected atom lies well within the antipodal convex hull, then it is possible to recover superpositions of atoms from  $\Lambda$ . This happens because coefficient associated with an atom outside  $\Lambda$  must be quite large to represent anything in the span of the subdictionary whenever  $\text{ERC}(\Lambda) > 0$ .

### 3.8. Digest

This section summarizes results in this chapter.

#### 3.8.1. Restricted Isometry Property

Restricted isometry property and RIP constants:

$$(1 - \delta_K) \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2 \quad \forall x \in \Sigma_K.$$

where  $\Phi \in \mathbb{C}^{M \times N}$ ,  $K \leq M \ll N$ .

RIP constants are non-decreasing:

$$\delta_k \leq \delta_l \text{ whenever } k < l.$$

First restricted isometry constant:

$$1 - \delta_1 \leq \|\phi_j\|_2^2 \leq 1 + \delta_1 \quad \forall 1 \leq j \leq N.$$

Sums and differences of signals:

$$(1 - \delta_{k+l}) \|x \pm y\|_2^2 \leq \|\Phi x \pm \Phi y\|_2^2 \leq (1 + \delta_{k+l}) \|x \pm y\|_2^2.$$

Approximate preservation of distances:

$$(1 - \delta_{2K}) d^2(x, y) \leq d^2(\Phi x, \Phi y) \leq (1 + \delta_{2K}) d^2(x, y).$$

RIP using unit length  $K$ -sparse vectors:

$$(1 - \delta_K) \leq \|\Phi x\|_2^2 \leq (1 + \delta_K)$$

We now consider sub-matrices of  $\Phi$  identified by an index set  $J \subset \{1, \dots, N\}$  with  $|J| = k \leq K$ .

**Eigen values of Gram matrices:** Let  $\lambda$  be an eigen value of  $\Phi_J^H \Phi_J$

$$1 - \delta_K \leq \lambda \leq 1 + \delta_K.$$

**Singular values:** Let  $\sigma$  be singular value of  $\Phi_J$ . Then

$$\sqrt{1 - \delta_K} \leq \sigma \leq \sqrt{1 + \delta_K}.$$

**Spectral norm:**

$$\|\Phi_J^H \Phi_J - I\|_2 = \|I - \Phi_J^H \Phi_J\|_2 \leq \delta_K.$$

**Sufficient condition for verifying RIP:** The eigen value bounds of Gram matrices are also sufficient conditions for verifying RIP of a matrix  $\Phi$ .

Singular values of  $\Phi_J^\dagger$ :

$$\frac{1}{\sqrt{1+\delta_K}} \leq \sigma \leq \frac{1}{\sqrt{1-\delta_K}}$$

Upper bound on norm of mappings for  $\Phi_J$  and  $\Phi_J^H$ :

$$\|\Phi_J x\|_2 \leq \sqrt{1+\delta_K} \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

$$\|\Phi_J^H y\|_2 \leq \sqrt{1+\delta_K} \|y\|_2 \quad \forall y \in \mathbb{C}^M.$$

Upper bound on norm of mappings for  $\Phi_J^\dagger$ :

$$\|\Phi_J^\dagger y\|_2 \leq \frac{1}{\sqrt{1-\delta_K}} \|y\|_2 \quad \forall y \in \mathbb{C}^M.$$

Bounds on mappings using Gram matrix:

$$(1-\delta_K) \|x\|_2 \leq \|\Phi_J^H \Phi_J x\|_2 \leq (1+\delta_K) \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

$$\frac{1}{1+\delta_K} \|x\|_2 \leq \|\Phi_J^H \Phi_J^{-1} x\|_2 \leq \frac{1}{1-\delta_K} \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

Upper bound for mapping using  $\Phi_J^H \Phi_J - I$ :

$$\|(\Phi_J^H \Phi_J - I)x\|_2 \leq \delta_K \|x\|_2 \quad \forall x \in \mathbb{C}^k.$$

**Approximate orthogonality:**  $S$  and  $T$  are disjoint index sets with  $|S \cup T| \leq K$ .

Spectral norm bound

$$\|\Phi_S^H \Phi_T\|_2 \leq \delta_K$$

Application:

$$\|\Phi_S^H \Phi_T x\|_2 \leq \delta_K \|x\|_2.$$

Columns of  $\Phi$  disjoint with support of  $x$ :  $R = \text{supp}(x) \setminus T$ .

$$\|\Phi_T^H \Phi x_R\|_2 \leq \delta_K \|x_R\|_2$$

RIP and inner product. **Inner product of signals with disjoint support:**

$$|\langle \Phi x, \Phi x' \rangle| \leq \delta_{k+k'} \|x\|_2 \|x'\|_2$$

Inner product of sparse real signals not necessarily disjoint support:

$$u, v \in \mathbb{R}^N, K = \max(\|u+v\|_0, \|u-v\|_0).$$

$$|\langle \Phi u, \Phi v \rangle - \langle u, v \rangle| \leq \delta_K \|u\|_2 \|v\|_2.$$

Complex space with bilinear inner product

$$|\langle \Phi u, \Phi v \rangle_B - \langle u, v \rangle_B| \leq \delta_K \|u\|_2 \|v\|_2.$$

RIP and orthogonal projection.  $P_\Lambda = \Phi_\Lambda \Phi_\Lambda^\dagger$ , and  $\Psi_\Lambda = (I - P_\Lambda)\Phi$  with  $|\Lambda| < K$ . **Modified RIP for  $\Psi_\Lambda$**

$$\left(1 - \frac{\delta_K}{1 - \delta_K}\right) \|x\|_2^2 \leq \|\Psi_\Lambda x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2$$

whenever  $\|x\|_0 \leq K - |\Lambda|$  and  $\text{supp}(x) \cap \Lambda = \emptyset$ .

Higher order RIP. **Upper bound on the RIP constant for higher order:**

$$\delta_{ck} \leq c\delta_{2k}.$$

**Bound on energy of embedding of arbitrary signal:**

$$\|\Phi x\|_2 \leq \sqrt{1 + \delta_K} \left[ \|x\|_2 + \frac{1}{\sqrt{K}} \|x\|_1 \right].$$

**Coherence bound for RIP constant of a dictionary**

$$\delta_K \leq (K - 1)\mu(\mathcal{D}).$$

### 3.8.2. Stable embeddings

#### 3.8.3. Spark

**Upper bound on spark**

$$\text{spark}(\mathcal{D}) \leq \underset{1 \leq k \leq D}{\text{minimize}} \|v_k^{1,*}\|_0$$

where  $v_k^{1,*}$  is a solution of the problem  $(Q_k)$ .

#### 3.8.4. Coherence

**Upper bound on the  $(2 \rightarrow \infty)$  norm of the pseudo-inverse of the sub-dictionary:** when  $(K - 1)\mu < 1$  and  $|\Lambda| \leq K$

$$\|\mathcal{D}_\Lambda^\dagger\|_{2 \rightarrow \infty} \leq \frac{1}{\sqrt{1 - (K - 1)\mu}}.$$

**$\mu_{1/2}(G)$**  Smallest number  $m$  such that the sum of magnitudes of a collection of  $m$  off-diagonal entries in a single row or column of the Gram matrix  $G$  is at least  $\frac{1}{2}$ .

**Coherence and  $\mu_{1/2}(G)$ :**

$$\mu_{1/2}(G) \geq \frac{1}{2\mu}.$$

Lower bound on spark in terms of  $\mu_{1/2}(G)$

$$\text{spark}(\mathcal{D}) \geq 2\mu_{1/2}(G) + 1.$$

### 3.8.5. Babel function

Modeling dictionary  $\Phi$ , sensing dictionary  $\Psi$ .

$p$ -Babel function for an index set

$$\mu_p(\Phi, \Psi, \Lambda) \triangleq \sup_{l \notin \Lambda} \left( \sum_{j \in \Lambda} |\langle \phi_j, \psi_l \rangle|^p \right)^{\frac{1}{p}}.$$

Complementary  $p$ -Babel function for an index set

$$\mu_p^{\text{in}}(\Phi, \Psi, \Lambda) \triangleq \sup_{i \in \Lambda} \mu_p(\Phi_\Lambda, \Psi_\Lambda, \Lambda \setminus \{i\}).$$

$p$  Babel function for a sparsity level

$$\mu_p(\Phi, \Psi, K) \triangleq \sup_{|\Lambda| \leq K} \mu_p(\Phi, \Psi, \Lambda).$$

$$\mu_p^{\text{in}}(\Phi, \Psi, K) \triangleq \sup_{|\Lambda| \leq K} \mu_p^{\text{in}}(\Phi, \Psi, \Lambda).$$

$$\mu_p^{\text{in}}(\Phi, \Psi, K) \leq \mu_p(\Phi, \Psi, K - 1).$$

For the case where  $\mathcal{D} = \Phi = \Psi$ :

$$\mu_p(\mathcal{D}, \Lambda) = \sup_{l \notin \Lambda} \left( \sum_{j \in \Lambda} |\langle d_j, d_l \rangle|^p \right)^{\frac{1}{p}}.$$

$$\mu_p^{\text{in}}(\mathcal{D}, \Lambda) = \sup_{i \in \Lambda} \mu_p(\mathcal{D}_\Lambda, \Lambda \setminus \{i\}).$$

$$\mu_p(\mathcal{D}, K) = \sup_{|\Lambda| \leq K} \mu_p(\mathcal{D}, \Lambda).$$

$$\mu_p^{\text{in}}(\mathcal{D}, K) = \sup_{|\Lambda| \leq K} \mu_p^{\text{in}}(\mathcal{D}, \Lambda).$$

For  $p = 1$

$$\begin{aligned}\mu_1(\mathcal{D}, \Lambda) &= \sup_{l \notin \Lambda} \left( \sum_{j \in \Lambda} |\langle d_j, d_l \rangle| \right). \\ \mu_1^{\text{in}}(\mathcal{D}, \Lambda) &= \sup_{i \in \Lambda} \mu_1(\mathcal{D}_{\Lambda \setminus \{i\}}). \\ \mu_1(\mathcal{D}, K) &= \sup_{|\Lambda| \leq K} \mu_1(\mathcal{D}, \Lambda). \\ \mu_1^{\text{in}}(\mathcal{D}, K) &= \sup_{|\Lambda| \leq K} \mu_1^{\text{in}}(\mathcal{D}, \Lambda).\end{aligned}$$

### 3.8.6. Exact recovery coefficient

**ERC**

$$\text{ERC}(\mathcal{D}_{\Lambda}) = 1 - \max_{\omega \notin \Lambda} \|\mathcal{D}_{\Lambda}^{\dagger} d_{\omega}\|_1.$$

**ERC and Babel function**

$$\text{ERC}(\Lambda) \geq \frac{1 - \mu_1(K-1) - \mu_1(K)}{1 - \mu_1(K-1)}.$$

**ERC and coherence**

$$\text{ERC}(\Lambda) \geq \frac{1 - (2K-1)\mu}{1 - (K-1)\mu}.$$

**Coherence based sufficient condition for ERC to be positive**

$$K < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right).$$

## CHAPTER 4

### Sensing Matrices

#### 4.1. Introduction

We will focus our attention to finite length signals.

Let  $x \in \mathbb{R}^N$  be our signal of interest where  $N$  is the number of signal components or *dimension* of the signal space ( $\mathbb{R}^N$ ).

Let us make  $M$  linear measurements of the signal. The measurements are given by

$$y = \Phi x \quad (4.1.1)$$

$y \in \mathbb{R}^M$  is our measurement vector in the measurement space ( $\mathbb{R}^M$ ) and  $M$  is the dimension of our measurement space.

$\Phi$  is an  $M \times N$  matrix known as the *sensing matrix*.

$M \ll N$ , hence  $\Phi$  achieves a *dimensionality reduction* over  $x$ .

We assume that measurements are *non-adaptive*. i.e. the matrix  $\Phi$  is predefined and doesn't depend on  $x$ .

The recovery process is denoted by

$$x' = \Delta y = \Delta(\Phi x) \quad (4.1.2)$$

where  $\Delta : \mathbb{R}^M \rightarrow \mathbb{R}^N$  is a (usually nonlinear) recovery algorithm.

Fundamental questions

- How should  $\Phi$  be designed so that information in  $x$  is preserved in  $y$ ?
- How do we recover  $x$  from  $y$ ?

We will look at three kinds of situations:

- (1) Signals are truly sparse. A signal has up to  $K(K \ll N)$  non-zero values only where  $K$  is known in advance. Measurement process is ideal and no noise is introduced during measurement. We will look for guarantees which can ensure exact recovery of signal from  $M(K < M \ll N)$  linear measurements.
- (2) Signals are not truly sparse but they have few  $K(K \ll N)$  values which dominate the signal. Thus if we approximate the signal by these  $K$  values, then approximation error is not noticeable. We again assume that there is no measurement noise being introduced. When we recover the signal, it will in general not be exact recovery. We expect the recovery error to be bounded (by approximation error). Also in special cases where the signal turns out to be  $K$ -sparse, we expect the recovery algorithm to recover the signal exactly. Such an algorithm with bounded recovery error will be called *robust*.
- (3) Signals are not sparse. Also there is measurement noise being introduced. We expect recovery algorithm to minimize error and thus perform *stable* recovery in the presence of measurement noise.

## 4.2. Recovery of exactly sparse signals

The null space of a matrix  $\Phi$  is denoted as

$$\mathcal{N}(\Phi) = \{v \in \mathbb{R}^N : \Phi v = 0\}. \quad (4.2.1)$$

The set of  $K$ -sparse signals is defined as

$$\Sigma_K = \{x \in \mathbb{R}^N : \|x\|_0 \leq K\}. \quad (4.2.2)$$

**Example 4.1: K sparse signals** Let  $N = 10$ .

- $x = (1, 2, 1, -1, 2, -3, 4, -2, 2, -2) \in \mathbb{R}^{10}$  is not a sparse signal.
- $x = (0, 0, 0, 0, 1, 0, 0, -1, 0, 0) \in \mathbb{R}^{10}$  is a 2-sparse signal. Its also a 4 sparse signal.

□

**Lemma 4.1** If  $a$  and  $b$  are two  $K$  sparse signals then  $a - b$  is a  $2K$  sparse signal.

PROOF.  $(a - b)_i$  is non zero only if at least one of  $a_i$  and  $b_i$  is non-zero. Hence number of non-zero components of  $a - b$  cannot exceed  $2K$ . Hence  $a - b$  is a  $2K$ -sparse signal. □

**Example 4.2: Difference of K sparse signals** Let  $N = 5$ .

- Let  $a = (0, 1, -1, 0, 0)$  and  $b = (0, 2, 0, -1, 0)$ . Then  $a - b = (0, -1, -1, 1, 0)$  is a 3 sparse hence 4 sparse signal.
- Let  $a = (0, 1, -1, 0, 0)$  and  $b = (0, 2, -1, 0, 0)$ . Then  $a - b = (0, -1, -2, 0, 0)$  is a 2 sparse hence 4 sparse signal.
- Let  $a = (0, 1, -1, 0, 0)$  and  $b = (0, 0, 0, 1, -1)$ . Then  $a - b = (0, 1, -1, -1, 1)$  is a 4 sparse signal.

□

**Lemma 4.2** A sensing matrix  $\Phi$  uniquely represents all  $x \in \Sigma_K$  if and only if  $\mathcal{N}(\Phi) \cap \Sigma_{2K} = \emptyset$ . i.e.  $\mathcal{N}(\Phi)$  contains no vectors in  $\Sigma_{2K}$ .

PROOF. Let  $a$  and  $b$  be two  $K$  sparse signals. Then  $\Phi a$  and  $\Phi b$  are corresponding measurements. Now if  $\Phi$  allows recovery of all  $K$  sparse signals, then  $\Phi a \neq \Phi b$ . Thus  $\Phi(a - b) \neq 0$ . Thus  $a - b \notin \mathcal{N}(\Phi)$ .

Let  $x \in \mathcal{N}(\Phi) \cap \Sigma_{2K}$ . Thus  $\Phi x = 0$  and  $\#x \leq 2K$ . Then we can find  $y, z \in \Sigma_K$  such that  $x = z - y$ . Thus  $m = \Phi z = \Phi y$ . But then,  $\Phi$  doesn't uniquely represent  $y, z \in \Sigma_K$ .  $\square$

There are many equivalent ways of characterizing above condition.

#### 4.2.1. The spark

We recall from definition 2.16, that spark of a matrix  $\Phi$  is defined as the minimum number of columns which are linearly dependent.

**Definition 4.1** A signal  $x \in \mathbb{R}^N$  is called an **explanation** of a measurement  $y \in \mathbb{R}^M$  w.r.t. sensing matrix  $\Phi$  if  $y = \Phi x$ .

**Theorem 4.3** For any measurement  $y \in \mathbb{R}^M$ , there exists at most one signal  $x \in \Sigma_K$  such that  $y = \Phi x$  if and only if  $\text{spark}(\Phi) > 2K$ .

PROOF. We need to show

- If for every measurement, there is only one  $K$ -sparse explanation, then  $\text{spark}(\Phi) > 2K$ .
- If  $\text{spark}(\Phi) > 2K$  then for every measurement, there is only one  $K$ -sparse explanation.

Assume that for every  $y \in \mathbb{R}^M$  there exists at most one signal  $x \in \mathbb{R}^N$  such that  $y = \Phi x$ .

Now assume that  $\text{spark}(\Phi) \leq 2K$ . Thus there exists a set of at most  $2K$  columns which are linearly dependent.

Thus there exists  $v \in \Sigma_{2K}$  such that  $\Phi v = 0$ . Thus  $v \in \mathcal{N}(\Phi)$ .

Thus  $\Sigma_{2K} \cap \mathcal{N}(\Phi) \neq \emptyset$ .

Hence  $\Phi$  doesn't uniquely represent each signal  $x \in \Sigma_K$ . A contradiction.

Hence  $\text{spark}(\Phi) > 2K$ .

Now suppose that  $\text{spark}(\Phi) > 2K$ .

Assume that for some  $y$  there exist two different  $K$ -sparse explanations  $x, x'$  such that  $y = \Phi x = \Phi x'$ .

Thus  $\Phi(x - x') = 0$ . Thus  $x - x' \in \mathcal{N}(\Phi)$  and  $x - x' \in \Sigma_{2K}$ .

Thus  $\text{spark}(\Phi) \leq 2K$ . A contradiction.

□

Since  $\text{spark}(\Phi) \in [2, M + 1]$  and we require that  $\text{spark}(\Phi) > 2K$  hence we require that  $M \geq 2K$ .

### 4.3. Recovery of approximately sparse signals

Spark is a useful criteria for characterization of sensing matrices for truly sparse signals. But this doesn't work well for *approximately* sparse signals. We need to have more restrictive criteria on  $\Phi$  for ensuring recovery of approximately sparse signals from compressed measurements.

In this context we will deal with two types of errors:

**Approximation error:** Let us approximate a signal  $x$  using only  $K$  coefficients. Let us call the approximation as  $\hat{x}$ . Thus  $e_a = (x - \hat{x})$  is approximation error.

**Recovery error:** Let  $\Phi$  be a sensing matrix. Let  $\Delta$  be a recovery algorithm. Then  $x' = \Delta(\Phi x)$  is the recovered signal vector. The error  $e_r = (x - x')$  is recovery error.

In this section we will

- Formalize the notion of null space property (NSP) of a matrix  $\Phi$ .
- Describe a measure for performance of an arbitrary recovery algorithm  $\Delta$ .

- Establish the connection between NSP and performance guarantee for recovery algorithms.

Suppose we approximate  $x$  by a  $K$ -sparse signal  $\hat{x} \in \Sigma_K$ , then the minimum error for  $l_p$  norm is given by

$$\sigma_K(x)_p = \min_{\hat{x} \in \Sigma_K} \|x - \hat{x}\|_p. \quad (4.3.1)$$

Specific  $\hat{x} \in \Sigma_K$  for which this minimum is achieved is the best  $K$ -term approximation.

In the following, we will need some new notation.

Let  $I = \{1, 2, \dots, N\}$  be the set of indices for signal  $x \in \mathbb{R}^N$ .

Let  $\Lambda \subset I$  be a subset of indices.

Let  $\Lambda^c = I \setminus \Lambda$ .

$x_\Lambda$  will denote a signal vector obtained by setting the entries of  $x$  indexed by  $\Lambda^c$  to zero.

**Example 4.3:**  $x_\Lambda$

Let  $N = 4$ . Then  $I = \{1, 2, 3, 4\}$ . Let  $\Lambda = \{1, 3\}$ . Then  $\Lambda^c = \{2, 4\}$ .

Now let  $x = (-1, 1, 2, -4)$ . Then  $x_\Lambda = (-1, 0, 2, 0)$ .

□

$\Phi_\Lambda$  will denote a  $M \times N$  matrix obtained by setting the columns of  $\Phi$  indexed by  $\Lambda^c$  to zero.

**Example 4.4:**  $\Phi_\Lambda$

Let  $N = 4$ . Then  $I = \{1, 2, 3, 4\}$ . Let  $\Lambda = \{1, 3\}$ . Then  $\Lambda^c = \{2, 4\}$ .

Now let  $x = (-1, 1, 2, -4)$ . Then  $x_\Lambda = (-1, 0, 2, -4)$ .

Now let

$$\Phi = \begin{pmatrix} 1 & 0 & -1 & 1 \\ -1 & -2 & 2 & 3 \end{pmatrix}$$

Then

$$\Phi_\Lambda = \begin{pmatrix} 1 & 0 & -1 & 0 \\ -1 & 0 & 2 & 0 \end{pmatrix}$$

□

**Definition 4.2** A matrix  $\Phi$  satisfies the **null space property (NSP)** of order  $K$  if there exists a constant  $C > 0$  such that,

$$\|h_\Lambda\|_2 \leq C \frac{\|h_{\Lambda^c}\|_1}{\sqrt{K}} \quad (4.3.2)$$

holds  $\forall h \in \mathcal{N}(\Phi)$  and  $\forall \Lambda$  such that  $|\Lambda| \leq K$ .

- Let  $h$  be  $K$  sparse. Thus choosing the indices on which  $h$  is non-zero, I can construct a  $\Lambda$  such that  $|\Lambda| \leq K$  and  $h_{\Lambda^c} = 0$ . Thus  $\|h_{\Lambda^c}\|_1 = 0$ . Hence above condition is not satisfied. Thus such a vector  $h$  should not belong to  $\mathcal{N}(\Phi)$  if  $\Phi$  satisfies NSP.
- Essentially vectors in  $\mathcal{N}(\Phi)$  shouldn't be concentrated in a small subset of indices.
- If  $\Phi$  satisfies NSP then the only  $K$ -sparse vector in  $\mathcal{N}(\Phi)$  is  $h = 0$ .

**4.3.0.1. Measuring the performance of a recovery algorithm.** Let  $\Delta : \mathbb{R}^M \rightarrow \mathbb{R}^N$  represent a recovery method to recover approximately sparse  $x$  from  $y$ .

$l_2$  recovery error is given by

$$\|\Delta(\Phi x) - x\|_2.$$

$l_1$  error for  $K$ -term approximation is given by  $\sigma_K(x)_1$ .

We will be interested in guarantees of the form

$$\|\Delta(\Phi x) - x\|_2 \leq C \frac{\sigma_K(x)_1}{\sqrt{K}} \quad (4.3.3)$$

Why, this recovery guarantee formulation?

- Exact recovery of  $K$ -sparse signals.  $\sigma_K(x)_1 = 0$  if  $x \in \Sigma_K$ .
- Robust recovery of non-sparse signals
- Recovery dependent on how well the signals are approximated by  $K$ -sparse vectors.
- Such guarantees are known as **instance optimal** guarantees.
- Also known as **uniform** guarantees.

Why the specific choice of norms?

- Different choices of  $l_p$  norms lead to different guarantees.
- $l_2$  norm on the LHS is a typical least squares error.
- $l_2$  norm on the RHS will require prohibitively large number of measurements.
- $l_1$  norm on the RHS helps us keep the number of measurements less.

If an algorithm  $\Delta$  provides instance optimal guarantees as defined above, what kind of requirements does it place on the sensing matrix  $\Phi$ ?

We show that NSP of order  $2K$  is a necessary condition for providing uniform guarantees.

**Theorem 4.4** Let  $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}^M$  denote a sensing matrix and  $\Delta : \mathbb{R}^M \rightarrow \mathbb{R}^N$  denote an arbitrary recovery algorithm. If the pair  $(\Phi, \Delta)$  satisfies instance optimal guarantee (4.3.3), then  $\Phi$  satisfies NSP of the order  $2K$ .

PROOF. We are given that

- $(\Phi, \Delta)$  form an encoder-decoder pair.

- Together, they satisfy instance optimal guarantee (4.3.3).
- Thus they are able to recover all sparse signals exactly.
- For non-sparse signals, they are able to recover their  $K$ -sparse approximation with bounded recovery error.

We need to show that if  $h \in \mathcal{N}(\Phi)$ , then  $h$  satisfies

$$\|h_\Lambda\|_2 \leq C \frac{\|h_{\Lambda^c}\|_1}{\sqrt{2K}}$$

where  $\Lambda$  corresponds to  $2K$  largest magnitude entries in  $h$ .

Note that we have used  $2K$  in this expression, since we need to show that  $\Phi$  satisfies NSP of order  $2K$ .

Let  $h \in \mathcal{N}(\Phi)$ .

Let  $\Lambda$  be the indices corresponding to the  $2K$  largest entries of  $h$ .

Thus

$$h = h_\Lambda + h_{\Lambda^c}.$$

Split  $\Lambda$  into  $\Lambda_0$  and  $\Lambda_1$  such that  $|\Lambda_0| = |\Lambda_1| = K$ .

Now

$$h_\Lambda = h_{\Lambda_0} + h_{\Lambda_1}.$$

Let

$$x = h_{\Lambda_0} + h_{\Lambda^c}.$$

Let

$$x' = -h_{\Lambda_1}.$$

Then

$$h = x - x'.$$

By assumption  $h \in \mathcal{N}(\Phi)$

Thus

$$\Phi h = \Phi(x - x') = 0 \implies \Phi x = \Phi x'.$$

But since  $x' \in \Sigma_K$  (recall that  $\Lambda_1$  indexes only  $K$  entries) and  $\Delta$  is able to recover all  $K$ -sparse signals exactly, hence

$$x' = \Delta(\Phi x')$$

Thus

$$\Delta(\Phi x) = \Delta(\Phi x') = x'.$$

i.e. the recovery algorithm  $\Delta$  recovers  $x'$  for the signal  $x$ . Certainly  $x'$  is not  $K$ -sparse.

Finally we also have (since  $h$  contains some additional non-zero entries)

$$\|h_\Lambda\|_2 \leq \|h\|_2 = \|x - x'\|_2 = \|x - \Delta(\Phi x)\|_2.$$

But as per instance optimal recovery guarantee (4.3.3) for  $(\Phi, \Delta)$  pair, we have

$$\|\Delta(\Phi x) - x\|_2 \leq C \frac{\sigma_K(x)_1}{\sqrt{K}}.$$

Thus

$$\|h_\Lambda\|_2 \leq C \frac{\sigma_K(x)_1}{\sqrt{K}}.$$

But

$$\sigma_K(x)_1 = \min_{\hat{x} \in \Sigma_K} \|x - \hat{x}\|_1.$$

Recall that  $x = h_{\Lambda_0} + h_{\Lambda^c}$  where  $\Lambda_0$  indexes  $K$  entries of  $h$  which are (magnitude wise) larger than all entries indexed by  $\Lambda^c$ . Thus the best  $l_1$ -norm  $K$  term approximation of  $x$  is given by  $h_{\Lambda_0}$ .

Hence

$$\sigma_K(x)_1 = \|h_{\Lambda^c}\|_1.$$

Thus we finally have

$$\|h_\Lambda\|_2 \leq C \frac{\|h_{\Lambda^c}\|_1}{\sqrt{K}} = \sqrt{2}C \frac{\|h_{\Lambda^c}\|_1}{\sqrt{2K}} \quad \forall h \in \mathcal{N}(\Phi).$$

Thus  $\Phi$  satisfies the NSP of order  $2K$ .  $\square$

It turns out that NSP of order  $2K$  is also sufficient to establish a guarantee of the form above for a practical recovery algorithm.

#### 4.4. Recovery in presence of measurement noise

Measurement vector in the presence of noise is given by

$$y = \Phi x + e \quad (4.4.1)$$

where  $e$  is the measurement noise or error.  $\|e\|_2$  is the  $l_2$  size of measurement error.

Recovery error as usual is given by

$$\|\Delta(y) - x\|_2 = \|\Delta(\Phi x + e) - x\|_2 \quad (4.4.2)$$

**Stability** of a recovery algorithm is characterized by comparing variation of recovery error w.r.t. measurement error.

NSP is both necessary and sufficient for establishing guarantees of the form:

$$\|\Delta(\Phi x) - x\|_2 \leq C \frac{\sigma_K(x)_1}{\sqrt{K}}$$

These guarantees do not account for presence of noise during measurement.

We need stronger conditions for handling noise.

The restricted isometry property for sensing matrices comes to our rescue.

##### 4.4.1. Restricted isometry property

We recall from definition 3.1 that a matrix  $\Phi$  satisfies the **restricted isometry property** (RIP) of order  $K$  if there exists  $\delta_K \in (0, 1)$  such that

$$(1 - \delta_K)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K)\|x\|_2^2 \quad (4.4.3)$$

holds for all  $x \in \Sigma_K = \{x : \|x\|_0 \leq K\}$ .

- If a matrix satisfies RIP of order  $K$ , then we can see that it *approximately* preserves the size of a  $K$ -sparse vector.
- If a matrix satisfies RIP of order  $2K$ , then we can see that it *approximately* preserves the distance between any two  $K$ -sparse vectors since difference vectors would be  $2K$  sparse (see theorem 3.6).
- We say that the matrix is *nearly orthonormal* for sparse vectors.
- If a matrix satisfies RIP of order  $K$  with a constant  $\delta_K$ , it automatically satisfies RIP of any order  $K' < K$  with a constant  $\delta_{K'} \leq \delta_K$ .

#### 4.4.2. Stability

Informally a recovery algorithm is stable if recovery error is small in the presence of small measurement error.

Is RIP necessary and sufficient for sparse signal recovery from noisy measurements?

Let us look at the necessary part.

We will define a notion of stability of the recovery algorithm.

**Definition 4.3** Let  $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}^M$  be a sensing matrix and  $\Delta : \mathbb{R}^M \rightarrow \mathbb{R}^N$  be a recovery algorithm. We say that the pair  $(\Phi, \Delta)$  is  **$C$ -stable** if for any  $x \in \Sigma_K$  and any  $e \in \mathbb{R}^M$  we have that

$$\|\Delta(\Phi x + e) - x\|_2 \leq C\|e\|_2. \quad (4.4.4)$$

- Error is added to the measurements.
- LHS is  $l_2$  norm of recovery error.
- RHS consists of scaling of the  $l_2$  norm of measurement error.
- The definition says that recovery error is bounded by a multiple of the measurement error.

- Thus adding a small amount of measurement noise shouldn't be causing arbitrarily large recovery error.

It turns out that  $C$ -stability requires  $\Phi$  to satisfy RIP.

**Theorem 4.5** *If a pair  $(\Phi, \Delta)$  is  $C$ -stable then*

$$\frac{1}{C} \|x\|_2 \leq \|\Phi x\|_2 \quad (4.4.5)$$

for all  $x \in \Sigma_{2K}$ .

PROOF. Remember that any  $x \in \Sigma_{2K}$  can be written in the form of  $x = y - z$  where  $y, z \in \Sigma_K$ .

So let  $x \in \Sigma_{2K}$ . Split it in the form of  $x = y - z$  with  $y, z \in \Sigma_K$ .

Define

$$e_y = \frac{\Phi(z - y)}{2} \quad \text{and} \quad e_z = \frac{\Phi(y - z)}{2}$$

Thus

$$e_y - e_z = \Phi(z - y) \implies \Phi y + e_y = \Phi z + e_z$$

We have

$$\Phi y + e_y = \Phi z + e_z = \frac{\Phi(y + z)}{2}.$$

Also we have

$$\|e_y\|_2 = \|e_z\|_2 = \frac{\|\Phi(y - z)\|_2}{2} = \frac{\|\Phi x\|_2}{2}$$

Let

$$y' = \Delta(\Phi y + e_y) = \Delta(\Phi z + e_z)$$

Since  $(\Phi, \Delta)$  is  $C$ -stable, hence we have

$$\|y' - y\|_2 \leq C\|e_y\|_2.$$

also

$$\|y' - z\|_2 \leq C\|e_z\|_2.$$

Using the triangle inequality

$$\begin{aligned}\|x\|_2 &= \|y - z\|_2 = \|y - y' + y' - z\|_2 \\ &\leq \|y - y'\|_2 + \|y' - z\|_2 \\ &\leq C\|e_y\|_2 + C\|e_z\|_2 = C(\|e_y\|_2 + \|e_z\|_2) = C\|\Phi x\|_2\end{aligned}$$

Thus we have  $\forall x \in \Sigma_{2K}$

$$\frac{1}{C}\|x\|_2 \leq \|\Phi x\|_2$$

□

This theorem gives us the lower bound for RIP property of order  $2K$  in (4.4.3) with  $\delta_{2K} = 1 - \frac{1}{C^2}$  as a necessary condition for  $C$ -stable recovery algorithms.

Note that smaller the constant  $C$ , lower is the bound on recovery error (w.r.t. measurement error). But as  $C \rightarrow 1$ ,  $\delta_{2K} \rightarrow 0$ , thus reducing the impact of measurement noise requires sensing matrix  $\Phi$  to be designed with tighter RIP constraints.

This result doesn't require an upper bound on the RIP property in (4.4.3).

It turns out that If  $\Phi$  satisfies RIP, then this is also sufficient for a variety of algorithms to be able to successfully recover a sparse signal from noisy measurements. We will discuss this later.

#### 4.4.3. Measurement bounds

As stated in previous section, for a  $(\Phi, \Delta)$  pair to be  $C$ -stable we require that  $\Phi$  satisfies RIP of order  $2K$  with a constant  $\delta_{2K}$ .

Let us ignore  $\delta_{2K}$  for the time being and look at relationship between  $M$ ,  $N$  and  $K$ .

We have a sensing matrix  $\Phi$  of size  $M \times N$  and expect it to provide RIP of order  $2K$ .

How many measurements  $M$  are necessary?

We will assume that  $K < N/2$ . This assumption is valid for approximately sparse signals.

Before we start figuring out the bounds, let us develop a special subset of  $\Sigma_K$  sets.

Consider the set

$$U = \{x \in \{0, +1, -1\}^N : \|x\|_0 = K\} \quad (4.4.6)$$

Some explanation: By  $A^N$  we mean  $A \times A \times \cdots \times A$  i.e.  $N$  times Cartesian product of  $A$ .

When we say  $\|x\|_0 = K$ , we mean that only  $K$  terms in each member of  $U$  can be non-zero (i.e.  $-1$  or  $+1$ ).

So  $U$  is a set of signal vectors  $x$  of length  $N$  where each sample takes values from  $\{0, +1, -1\}$  and number of allowed non-zero samples is fixed at  $K$ .

An example below explains it further.

**Example 4.5:**  $U$  for  $N = 6$  and  $K = 2$

Each vector in  $U$  will have 6 elements out of which 2 can be non zero. There are  $\binom{6}{2}$  ways of choosing the non-zero elements. Some of those sets are listed below as examples:

$$\begin{aligned} &(+1, +1, 0, 0, 0, 0) \\ &(+1, -1, 0, 0, 0, 0) \\ &(0, -1, 0, +1, 0, 0) \\ &(0, -1, 0, +1, 0, 0) \\ &(0, 0, 0, 0, -1, +1) \\ &(0, 0, -1, -1, 0, 0) \end{aligned}$$

□

Revisiting

$$U = \{x \in \{0, +1, -1\}^N : \|x\|_0 = K\}$$

It's now obvious that

$$\|x\|_2^2 = K \quad \forall x \in U. \quad (4.4.7)$$

Since there are  $\binom{N}{K}$  ways of choosing  $K$  non-zero elements and each non zero element can take either of the two values  $+1$  or  $-1$ , hence the cardinality of set  $U$  is given by:

$$|U| = \binom{N}{K} 2^K \quad (4.4.8)$$

By definition

$$U \subset \Sigma_K. \quad (4.4.9)$$

Further Let  $x, y \in U$ .

Then  $x - y$  will have a maximum of  $2K$  non-zero elements. The non-zero elements would have values  $\in \{-2, -1, 1, 2\}$ .

Thus  $\|x - y\|_0 = R \leq 2K$ .

Further  $\|x - y\|_2^2 \geq R$ . Explain!

Hence

$$\|x - y\|_0 \leq \|x - y\|_2^2 \quad \forall x, y \in U. \quad (4.4.10)$$

We now state a lemma which will help us in getting to the bounds.

**Lemma 4.6** *Let  $K$  and  $N$  satisfying  $K < \frac{N}{2}$  be given. There exists a set  $X \subset \Sigma_K$  such that for any  $x \in X$  we have  $\|x\|_2 \leq \sqrt{K}$  and for any  $x, y \in X$  with  $x \neq y$ ,*

$$\|x - y\|_2 \geq \sqrt{\frac{K}{2}}. \quad (4.4.11)$$

and

$$\ln |X| \geq \frac{K}{2} \ln \left( \frac{N}{K} \right). \quad (4.4.12)$$

PROOF. We just need to find one set  $X$  which satisfies the requirements of this lemma. We have to construct a set  $X$  such that

- $\|x\|_2 \leq \sqrt{K} \quad \forall x \in X.$
- $\|x - y\|_2 \geq \sqrt{\frac{K}{2}} \quad \forall x, y \in X.$
- $\ln |X| \geq \frac{K}{2} \ln \left( \frac{N}{K} \right)$  or equivalently  $|X| \geq \left( \frac{N}{K} \right)^{\frac{K}{2}}.$

We will construct  $X$  by picking vectors from  $U$ . Thus  $X \subset U$ .

Since  $x \in X \subset U$  hence  $\|x\|_2 = \sqrt{K} \leq \sqrt{K} \quad \forall x \in X.$

Consider any fixed  $x \in U$ .

How many elements  $y$  are there in  $U$  such that  $\|x - y\|_2^2 < \frac{K}{2}$  ?

Define

$$U_x^2 = \left\{ y \in U : \|x - y\|_2^2 < \frac{K}{2} \right\} \quad (4.4.13)$$

Clearly by requirements in the lemma, if  $x \in X$  then  $U_x^2 \cap X = \emptyset$ . i.e. no vector in  $U_x^2$  belongs to  $X$ .

How many elements are there in  $U_x^2$ ? Let us find an upper bound.

$\forall x, y \in U$  we have  $\|x - y\|_0 \leq \|x - y\|_2$ .

If  $x$  and  $y$  differ in  $\frac{K}{2}$  or more places, then naturally  $\|x - y\|_2^2 \geq \frac{K}{2}$ .

Hence if  $\|x - y\|_2^2 < \frac{K}{2}$  then  $\|x - y\|_0 < \frac{K}{2}$  hence  $\|x - y\|_0 \leq \frac{K}{2}$  for any  $x, y \in U_x^2$ .

So define

$$U_x^0 = \left\{ y \in U : \|x - y\|_0 \leq \frac{K}{2} \right\} \quad (4.4.14)$$

We have

$$U_x^2 \subseteq U_x^0 \quad (4.4.15)$$

Thus we have an upper bound given by

$$|U_x^2| \leq |U_x^0|. \quad (4.4.16)$$

Let us look at  $U_x^0$  carefully.

We can choose  $\frac{K}{2}$  indices where  $x$  and  $y$  *may* differ in  $\binom{N}{\frac{K}{2}}$  ways.

At each of these  $\frac{K}{2}$  indices,  $y_i$  can take value as one of  $(0, +1, -1)$ .

Thus We have an upper bound

$$|U_x^2| \leq |U_x^0| \leq \binom{N}{\frac{K}{2}} 3^{\frac{K}{2}}. \quad (4.4.17)$$

We now describe an iterative process for building  $X$  from vectors in  $U$ .

Say we have added  $j$  vectors to  $X$  namely  $x_1, x_2, \dots, x_j$ .

Then

$$(U_{x_1}^2 \cup U_{x_2}^2 \cup \dots \cup U_{x_j}^2) \cap X = \emptyset$$

Number of vectors in  $U_{x_1}^2 \cup U_{x_2}^2 \cup \dots \cup U_{x_j}^2$  is bounded by  $j \binom{N}{\frac{K}{2}} 3^{\frac{K}{2}}$ .

Thus we have at least

$$\binom{N}{K} 2^K - j \binom{N}{\frac{K}{2}} 3^{\frac{K}{2}} \quad (4.4.18)$$

vectors left in  $U$  to choose from for adding in  $X$ .

We can keep adding vectors to  $X$  till there are no more suitable vectors left.

So we can construct a set of size  $|X|$  provided

$$|X| \binom{N}{\frac{K}{2}} 3^{\frac{K}{2}} \leq \binom{N}{K} 2^K \quad (4.4.19)$$

Now

$$\frac{\binom{N}{K}}{\binom{N}{\frac{K}{2}}} = \frac{\left(\frac{K}{2}\right)! (N - \frac{K}{2})!}{K!(N - K)!} = \prod_{i=1}^{\frac{K}{2}} \frac{N - K + i}{K/2 + i}$$

Note that  $\frac{N-K+i}{K/2+i}$  is a decreasing function of  $i$ .

Its minimum value is achieved for  $i = \frac{K}{2}$  as  $(\frac{N}{K} - \frac{1}{2})$ .

So we have

$$\begin{aligned} \frac{N - K + i}{K/2 + i} &\geq \frac{N}{K} - \frac{1}{2} \\ \implies \prod_{i=1}^{\frac{K}{2}} \frac{N - K + i}{K/2 + i} &\geq \left(\frac{N}{K} - \frac{1}{2}\right)^{\frac{K}{2}} \\ \implies \frac{\binom{N}{K}}{\binom{N}{\frac{K}{2}}} &\geq \left(\frac{N}{K} - \frac{1}{2}\right)^{\frac{K}{2}} \end{aligned}$$

Rephrasing (4.4.19) we have

$$|X| \left(\frac{3}{4}\right)^{\frac{K}{2}} \leq \frac{\binom{N}{K}}{\binom{N}{\frac{K}{2}}} \quad (4.4.20)$$

So if

$$|X| \left(\frac{3}{4}\right)^{\frac{K}{2}} \leq \left(\frac{N}{K} - \frac{1}{2}\right)^{\frac{K}{2}}$$

then (4.4.19) will be satisfied.

Now it is given that  $K < \frac{N}{2}$ . So we have:

$$\begin{aligned}
K &< \frac{N}{2} \\
\implies \frac{N}{K} &> 2 \\
\implies \frac{N}{4K} &> \frac{1}{2} \\
\implies \frac{N}{K} - \frac{N}{4K} &< \frac{N}{K} - \frac{1}{2} \\
\implies \frac{3N}{4K} &< \frac{N}{K} - \frac{1}{2} \\
\implies \left(\frac{3N}{4K}\right)^{\frac{K}{2}} &< \left(\frac{N}{K} - \frac{1}{2}\right)^{\frac{K}{2}}
\end{aligned}$$

Thus we have

$$\left(\frac{N}{K}\right)^{\frac{K}{2}} \left(\frac{3}{4}\right)^{\frac{K}{2}} < \frac{\binom{N}{K}}{\binom{N}{\frac{K}{2}}} \quad (4.4.21)$$

Choose

$$|X| = \left(\frac{N}{K}\right)^{\frac{K}{2}} \quad (4.4.22)$$

Clearly this value of  $|X|$  satisfies (4.4.19). Hence  $X$  can have at least these many elements. Thus

$$\begin{aligned}
|X| &\geq \left(\frac{N}{K}\right)^{\frac{K}{2}} \\
\implies \ln |X| &\geq \frac{K}{2} \ln \left(\frac{N}{K}\right)
\end{aligned}$$

which completes the proof. □

We can now establish following bound on the required number of measurements to satisfy RIP.

At this moment, we won't worry about exact value of  $\delta_{2K}$ . We will just assume that  $\delta_{2K}$  is small in range  $(0, \frac{1}{2}]$ .

**Theorem 4.7** *Let  $\Phi$  be an  $M \times N$  matrix that satisfies RIP of order  $2K$  with constant  $\delta_{2K} \in (0, \frac{1}{2}]$ . Then*

$$M \geq CK \ln \left( \frac{N}{K} \right) \quad (4.4.23)$$

where  $C = \frac{1}{2 \ln(\sqrt{24}+1)} \approx 0.28173$ .

PROOF. Since  $\Phi$  satisfies RIP of order  $2K$  we have

$$\begin{aligned} (1 - \delta_{2K})\|x\|_2^2 &\leq \|\Phi x\|_2^2 \leq (1 + \delta_{2K})\|x\|_2^2 \quad \forall x \in \Sigma_{2K}. \\ \implies (1 - \delta_{2K})\|x - y\|_2^2 &\leq \|\Phi x - \Phi y\|_2^2 \leq (1 + \delta_{2K})\|x - y\|_2^2 \quad \forall x, y \in \Sigma_K. \end{aligned}$$

Also

$$\delta_{2K} \leq \frac{1}{2} \implies 1 - \delta_{2K} > \frac{1}{2} \text{ and } 1 + \delta_{2K} \leq \frac{3}{2}$$

Consider the set  $X \subset U \subset \Sigma_K$  developed in lemma 4.6.

We have

$$\begin{aligned} \|x - y\|_2^2 &\geq \frac{K}{2} \quad \forall x, y \in X \\ \implies (1 - \delta_{2K})\|x - y\|_2^2 &\geq \frac{K}{4} \\ \implies \|\Phi x - \Phi y\|_2^2 &\geq \frac{K}{4} \\ \implies \|\Phi x - \Phi y\|_2 &\geq \sqrt{\frac{K}{4}} \quad \forall x, y \in X \end{aligned}$$

Also

$$\begin{aligned} \|\Phi x\|_2^2 &\leq (1 + \delta_{2K})\|x\|_2^2 \leq \frac{3}{2}\|x\|_2^2 \quad \forall x \in X \subset \Sigma_K \subset \Sigma_{2K} \\ \implies \|\Phi x\|_2 &\leq \sqrt{\frac{3}{2}}\|x\|_2 \leq \sqrt{\frac{3K}{2}} \quad \forall x \in X. \end{aligned}$$

since  $\|x\|_2 \leq \sqrt{K} \quad \forall x \in X$ .

So we have a lower bound:

$$\|\Phi x - \Phi y\|_2 \geq \sqrt{\frac{K}{4}} \quad \forall x, y \in X. \quad (4.4.24)$$

and an upper bound:

$$\|\Phi x\|_2 \leq \sqrt{\frac{3K}{2}} \quad \forall x \in X. \quad (4.4.25)$$

What do these bounds mean? Let us start with the lower bound.

$\Phi x$  and  $\Phi y$  are projections of  $x$  and  $y$  in  $\mathbb{R}^M$  (measurement space).

Construct  $l_2$  balls of radius  $\sqrt{\frac{K}{4}}/2 = \sqrt{\frac{K}{16}}$  in  $\mathbb{R}^M$  around  $\Phi x$  and  $\Phi y$ .

Lower bound says that these balls are disjoint. Since  $x, y$  are arbitrary, this applies to every  $x \in X$ .

Upper bound tells us that all vectors  $\Phi x$  lie in a ball of radius  $\sqrt{\frac{3K}{2}}$  around origin in  $\mathbb{R}^M$ .

Thus the set of all balls lies within a larger ball of radius  $\sqrt{\frac{3K}{2}} + \sqrt{\frac{K}{16}}$  around origin in  $\mathbb{R}^M$ .

So we require that the volume of the larger ball MUST be greater than the sum of volumes of  $|X|$  individual balls.

Since volume of an  $l_2$  ball of radius  $r$  is proportional to  $r^M$ , we have:

$$\begin{aligned} \left( \sqrt{\frac{3K}{2}} + \sqrt{\frac{K}{16}} \right)^M &\geq |X| \cdot \left( \sqrt{\frac{K}{16}} \right)^M \\ \cdot \implies (\sqrt{24} + 1)^M &\geq |X| \\ \implies M &\geq \frac{\ln |X|}{\ln(\sqrt{24} + 1)} \end{aligned}$$

Again from lemma 4.6 we have

$$\ln |X| \geq \frac{K}{2} \ln \left( \frac{N}{K} \right).$$

Putting back we get

$$M \geq \frac{\frac{K}{2} \ln\left(\frac{N}{K}\right)}{\ln(\sqrt{24} + 1)}$$

which establishes a lower bound on the number of measurements  $M$ .

□

#### **Example 4.6: Lower bounds on $M$ for RIP of order $2K$**

- (1)  $N = 1000, K = 100 \implies M \geq 65$ .
- (2)  $N = 1000, K = 200 \implies M \geq 91$ .
- (3)  $N = 1000, K = 400 \implies M \geq 104$ .

□

Some remarks are in order:

- The theorem only establishes a necessary lower bound on  $M$ . It doesn't mean that if we choose an  $M$  larger than the lower bound then  $\Phi$  will have RIP of order  $2K$  with any constant  $\delta_{2K} \in (0, \frac{1}{2}]$ .
- The restriction  $\delta_{2K} \leq \frac{1}{2}$  is arbitrary and is made for convenience. In general, we can work with  $0 < \delta_{2K} \leq \delta_{\max} < 1$  and develop the bounds accordingly.
- This result fails to capture dependence of  $M$  on the RIP constant  $\delta_{2K}$  directly. *Johnson-Lindenstrauss lemma* helps us resolve this which concerns embeddings of finite sets of points in low-dimensional spaces.
- We haven't made significant efforts to optimize the constants. Still they are quite reasonable.

#### **4.5. The RIP and the NSP**

RIP and NSP are connected. If a matrix  $\Phi$  satisfies RIP then it also satisfies NSP (under certain conditions).

Thus RIP is strictly stronger than NSP (under certain conditions).

We will need following lemma which applies to any arbitrary  $h \in \mathbb{R}^N$ . The lemma will be proved later.

**Lemma 4.8** Suppose that  $\Phi$  satisfies RIP of order  $2K$ , and let  $h \in \mathbb{R}^N, h \neq 0$  be arbitrary. Let  $\Lambda_0$  be any subset of  $\{1, 2, \dots, N\}$  such that  $|\Lambda_0| \leq K$ .

Define  $\Lambda_1$  as the index set corresponding to the  $K$  entries of  $h_{\Lambda_0^c}$  with largest magnitude, and set  $\Lambda = \Lambda_0 \cup \Lambda_1$ . Then

$$\|h_\Lambda\|_2 \leq \alpha \frac{\|h_{\Lambda_0^c}\|_1}{\sqrt{K}} + \beta \frac{|\langle \Phi h_\Lambda, \Phi h \rangle|}{\|h_\Lambda\|_2}, \quad (4.5.1)$$

where

$$\alpha = \frac{\sqrt{2}\delta_{2K}}{1 - \delta_{2K}}, \beta = \frac{1}{1 - \delta_{2K}}. \quad (4.5.2)$$

Let us understand this lemma a bit. If  $h \in \mathcal{N}(\Phi)$ , then the lemma simplifies to

$$\|h_\Lambda\|_2 \leq \alpha \frac{\|h_{\Lambda_0^c}\|_1}{\sqrt{K}} \quad (4.5.3)$$

- $\Lambda_0$  maps to the initial few ( $K$  or less) elements we chose.
- $\Lambda_0^c$  maps to all other elements.
- $\Lambda_1$  maps to largest (in magnitude)  $K$  elements of  $\Lambda_0^c$ .
- $h_\Lambda$  contains a maximum of  $2K$  non-zero elements.
- $\Phi$  satisfies RIP of order  $2K$ .
- Thus  $(1 - \delta_{2K})\|h_\Lambda\|_2 \leq \|\Phi h_\Lambda\|_2 \leq (1 + \delta_{2K})\|h_\Lambda\|_2$ .

We now state the connection between RIP and NSP.

**Theorem 4.9** Suppose that  $\Phi$  satisfies RIP of order  $2K$  with  $\delta_{2K} < \sqrt{2} - 1$ . Then  $\Phi$  satisfies the NSP of order  $2K$  with constant

$$C = \frac{\sqrt{2}\delta_{2K}}{1 - (1 + \sqrt{2})\delta_{2K}} \quad (4.5.4)$$

PROOF. We are given

$$(1 - \delta_{2K})\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_{2K})\|x\|_2^2$$

holds for all  $x \in \Sigma_{2K}$  where  $\delta_{2K} < \sqrt{2} - 1$ .

We have to show that:

$$\|h_\Lambda\|_2 \leq C \frac{\|h_{\Lambda^c}\|_1}{\sqrt{K}}$$

holds  $\forall h \in \mathcal{N}(\Phi)$  and  $\forall \Lambda$  such that  $|\Lambda| \leq 2K$ .

Let  $h \in \mathcal{N}(\Phi)$ . Then  $\Phi h = 0$ .

Let  $\Lambda_m$  denote the  $2K$  largest entries of  $h$ . Then

$$\|h_\Lambda\|_2 \leq \|h_{\Lambda_m}\|_2 \quad \forall \Lambda : |\Lambda| \leq 2K.$$

Similarly

$$\|h_{\Lambda^c}\|_1 \geq \|h_{\Lambda_m^c}\|_1 \quad \forall \Lambda : |\Lambda| \leq 2K.$$

Thus if we show that  $\Phi$  satisfies NSP of order  $2K$  for  $\Lambda_m$ , i.e.

$$\|h_{\Lambda_m}\|_2 \leq C \frac{\|h_{\Lambda_m^c}\|_1}{\sqrt{K}}$$

then we would have shown it for all  $\Lambda$  such that  $|\Lambda| \leq 2K$ . So let  $\Lambda = \Lambda_m$ .

We can divide  $\Lambda$  into two components  $\Lambda_0$  and  $\Lambda_1$  of size  $K$  each.

Since  $\Lambda$  maps to the largest  $2K$  entries in  $h$  hence whatever entries we choose in  $\Lambda_0$ , the largest  $K$  entries in  $\Lambda_0^c$  will be  $\Lambda_1$ .

Hence as per lemma 4.8 above, we have

$$\|h_\Lambda\|_2 \leq \alpha \frac{\|h_{\Lambda_0^c}\|_1}{\sqrt{K}} \tag{4.5.5}$$

Also

$$\Lambda = \Lambda_0 \cup \Lambda_1 \implies \Lambda_0 = \Lambda \setminus \Lambda_1 = \Lambda \cap \Lambda_1^c \implies \Lambda_0^c = \Lambda_1 \cup \Lambda^c$$

Thus we have

$$\|h_{\Lambda_0^c}\|_1 = \|h_{\Lambda_1}\|_1 + \|h_{\Lambda^c}\|_1 \quad (4.5.6)$$

We have to get rid of  $\Lambda_1$ .

Since  $h_{\Lambda_1} \in \Sigma_K$ , by applying lemma 2.16 we get

$$\|h_{\Lambda_1}\|_1 \leq \sqrt{K} \|h_{\Lambda_1}\|_2$$

Hence

$$\|h_{\Lambda}\|_2 \leq \alpha \left( \|h_{\Lambda_1}\|_2 + \frac{\|h_{\Lambda^c}\|_1}{\sqrt{K}} \right) \quad (4.5.7)$$

But since  $\Lambda_1 \subset \Lambda$ , hence  $\|h_{\Lambda_1}\|_2 \leq \|h_{\Lambda}\|_2$ , hence

$$\|h_{\Lambda}\|_2 \leq \alpha \left( \|h_{\Lambda}\|_2 + \frac{\|h_{\Lambda^c}\|_1}{\sqrt{K}} \right) \quad (4.5.8)$$

$$\implies (1 - \alpha) \|h_{\Lambda}\|_2 \leq \alpha \frac{\|h_{\Lambda^c}\|_1}{\sqrt{K}} \quad (4.5.9)$$

$$\implies \|h_{\Lambda}\|_2 \leq \frac{\alpha}{1 - \alpha} \frac{\|h_{\Lambda^c}\|_1}{\sqrt{K}} \quad \text{if } \alpha \leq 1. \quad (4.5.10)$$

Note that the inequality is also satisfied for  $\alpha = 1$  in which case, we don't need to bring  $1 - \alpha$  to denominator.

Now

$$\begin{aligned} \alpha &\leq 1 \\ \implies \frac{\sqrt{2}\delta_{2K}}{1 - \delta_{2K}} &\leq 1 \\ \implies \sqrt{2}\delta_{2K} &\leq 1 - \delta_{2K} \\ \implies (\sqrt{2} + 1)\delta_{2K} &\leq 1 \\ \implies \delta_{2K} &\leq \sqrt{2} - 1 \end{aligned}$$

Putting

$$C = \frac{\alpha}{1 - \alpha} = \frac{\sqrt{2}\delta_{2K}}{1 - (1 + \sqrt{2})\delta_{2K}} \quad (4.5.11)$$

we see that  $\Phi$  satisfies NSP of order  $2K$  whenever  $\Phi$  satisfies RIP of order  $2K$  with  $\delta_{2K} \leq \sqrt{2} - 1$ .

□

Note that for  $\delta_{2K} = \sqrt{2} - 1$ ,  $C = \infty$ .

#### 4.6. Matrices satisfying RIP

The natural question at this moment is how to construct matrices which satisfy RIP.

There are two different approaches

- Deterministic approach
- Randomized approach

Known deterministic approaches so far tend to require  $M$  to be very large ( $O(K^2 \ln N)$  or  $O(KN^\alpha)$ ).

We can overcome this limitation by randomizing matrix construction.

Construction process:

- Input  $M$  and  $N$ .
- Generate  $\Phi$  by choosing  $\Phi_{ij}$  as independent realizations from some probability distribution.

Suppose that  $\Phi$  is drawn from normal distribution.

It can be shown that the rank of  $\Phi$  is  $M$  with probability 1.

**Example 4.7: Random matrices are full rank.** We can verify this fact by doing a small computer simulation.

```

1 M = 6;
2 N = 20;
3 trials = 10000;
4 numFullRankMatrices = 0;
5 for i=1:trials
    % Create a random matrix of size M x N

```

```

7 A = rand(M,N);
9 % Obtain its rank
10 R = rank(A);
11 % Check whether the rank equals M or not
12 if R == M
13     numFullRankMatrices = numFullRankMatrices + 1;
14 end
15 end
16 fprintf('Number of trials: %d\n', trials);
17 fprintf('Number of full rank matrices: %d\n', numFullRankMatrices);
18 percentage = numFullRankMatrices*100/trials;
19 fprintf('Percentage of full rank matrices: %.2f %%\n', percentage);

```

LISTING 4.1. demoRandomMatrixRank.m

Above program generates a number of random matrices and measures their ranks. It verifies whether they are full rank or not.

Here is a sample output:

```

>> demoRandomMatrixRank
Number of trials: 10000
Number of full rank matrices: 10000
Percentage of full rank matrices: 100.00 %

```

□

Thus if we choose  $M = 2K$ , any subset of  $2K$  columns will be linearly independent.

Thus the matrix with satisfy RIP with some  $\delta_{2K} > 0$ .

But this construction doesn't tell us exact value of  $\delta_{2K}$ .

In order to find out  $\delta_{2K}$ , we must consider all possible  $K$ - dimensional subspaces of  $\mathbb{R}^N$ .

This is computationally impossible for reasonably large  $N$  and  $K$ .

What is the alternative?

We can start with a chosen value of  $\delta_{2K}$  and try to construct a matrix which matches it.

Before we proceed further, we should take a detour and review sub-Gaussian distributions in ??.

We now state the main theorem of this section.

**Theorem 4.10** Suppose that  $X = [X_1, X_2, \dots, X_M]$  where each  $X_i$  is i.i.d. with  $X_i \sim \text{Sub}(c^2)$  and  $\mathbb{E}(X_i^2) = \sigma^2$ . Then

$$\mathbb{E}(\|X\|_2^2) = M\sigma^2 \quad (4.6.1)$$

Moreover, for any  $\alpha \in (0, 1)$  and for any  $\beta \in [c^2/\sigma^2, \beta_{max}]$ , there exists a constant  $\kappa^* \geq 4$  depending only on  $\beta_{max}$  and the ratio  $\sigma^2/c^2$  such that

$$\mathbb{P}(\|X\|_2^2 \leq \alpha M\sigma^2) \leq \exp\left(-\frac{M(1-\alpha)^2}{\kappa^*}\right) \quad (4.6.2)$$

and

$$\mathbb{P}(\|X\|_2^2 \geq \beta M\sigma^2) \leq \exp\left(-\frac{M(\beta-1)^2}{\kappa^*}\right) \quad (4.6.3)$$

PROOF. □

#### 4.6.1. Conditions on random distribution for RIP

Let us get back to our business of constructing a matrix  $\Phi$  using random distributions which satisfies RIP with a given  $\delta$ .

We will impose some conditions on the random distribution.

- (1) We require that the distribution will yield a matrix that is norm-preserving. This requires that

$$\mathbb{E}(\Phi_{ij}^2) = \frac{1}{M} \quad (4.6.4)$$

Hence variance of distribution should be  $\frac{1}{M}$ .

- (2) We require that distribution is a sub-Gaussian distribution i.e. there exists a constant  $c > 0$  such that

$$\mathbb{E}(\exp(\Phi_{ij}t)) \leq \exp\left(\frac{c^2 t^2}{2}\right) \quad (4.6.5)$$

This says that the moment generating function of the distribution is dominated by a Gaussian distribution.

In other words, tails of the distribution decay at least as fast as the tails of a Gaussian distribution.

We will further assume that entries of  $\Phi$  are strictly sub-Gaussian. i.e. they must satisfy (4.6.5) with

$$c^2 = \mathbb{E}(\Phi_{ij}^2) = \frac{1}{M}$$

Under these conditions we have the following result. This is proven later.

**Corollary 4.11.** *Suppose that  $\Phi$  is an  $M \times N$  matrix whose entries  $\Phi_{ij}$  are i.i.d. with  $\Phi_{ij}$  drawn according to a strictly sub-Gaussian distribution with  $c^2 = \frac{1}{M^2}$ .*

*Let  $Y = \Phi x$  for  $x \in \mathbb{R}^N$ . Then for any  $\epsilon > 0$  and any  $x \in \mathbb{R}^N$ ,*

$$\mathbb{E}(\|Y\|_2^2) = \|x\|_2^2 \quad (4.6.6)$$

and

$$\mathbb{P}(\|Y\|_2^2 - \|x\|_2^2 \geq \epsilon \|x\|_2^2) \leq 2 \exp\left(-\frac{M\epsilon^2}{\kappa^*}\right) \quad (4.6.7)$$

where  $\kappa^* = \frac{2}{1-\ln(2)} \approx 6.5178$ .

This means that the norm of a sub-Gaussian random vector strongly concentrates about its mean.

#### 4.6.2. Sub Gaussian random matrices satisfy the RIP

Using this result we now state that sub-Gaussian matrices satisfy the RIP.

**Theorem 4.12** *Fix  $\delta \in (0, 1)$ . Let  $\Phi$  be an  $M \times N$  random matrix whose entries  $\Phi_{ij}$  are i.i.d. with  $\Phi_{ij}$  drawn according to a strictly*

sub-Gaussian distribution with  $c^2 = \frac{1}{M}$ . If

$$M \geq \kappa_1 K \ln \left( \frac{N}{K} \right), \quad (4.6.8)$$

then  $\Phi$  satisfies the RIP of order  $K$  with the prescribed  $\delta$  with probability exceeding  $1 - 2e^{-\kappa_2 M}$ , where  $\kappa_1$  is arbitrary and

$$\kappa_2 = \frac{\delta^2}{2\kappa^*} - \frac{1}{\kappa_1} \ln \left( \frac{42e}{\delta} \right) \quad (4.6.9)$$

We note that this theorem achieves  $M$  of the same order as the lower bound obtained in theorem 4.7 up to a constant.

This is much better than deterministic approaches.

#### 4.6.3. Advantages of random construction

There are a number of advantages of the random sensing matrix construction approach:

- One can show that for random construction, the measurements are *democratic*. This means that all measurements are equal in importance and it is possible to recover the signal from any sufficiently large subset of the measurements.

Thus by using random  $\Phi$  one can be robust to the loss of loss or corruption of a small fraction of measurements.

- In general we are more interested in  $x$  which is sparse in some basis  $\Psi$ . In this setting, we require that  $\Phi\Psi$  satisfy the RIP. Deterministic construction would explicitly require taking  $\Psi$  into account.

But if  $\Phi$  is random, we can avoid this issue.

If  $\Phi$  is Gaussian and  $\Psi$  is an orthonormal basis, then one can easily show that  $\Phi\Psi$  will also have a Gaussian distribution.

Thus if  $M$  is high,  $\Phi\Psi$  will also satisfy RIP with very high probability.

Similar results hold for other sub-Gaussian distributions as well.

## CHAPTER 5

# Dictionaries and Sensing Matrices

This chapter covers various results for popularly used dictionaries and sensing matrices.

### 5.1. Dirac-DCT dictionary

**Definition 5.1** The Dirac-DCT dictionary is a two-ortho dictionary consisting of the union of the Dirac and the DCT bases.

This dictionary is suitable for real signals since both Dirac and DCT are totally real bases  $\in \mathbb{R}^{N \times N}$ .

The dictionary is obtained by combining the  $N \times N$  identity matrix (Dirac basis) with the  $N \times N$  DCT matrix for signals in  $\mathbb{R}^N$ .

Let  $\Psi_{\text{DCT},N}$  denote the DCT matrix for  $\mathbb{R}^N$ . Let  $I_N$  denote the identity matrix for  $\mathbb{R}^N$ . Then

$$\mathcal{D}_{\text{DCT}} = [I_N \quad \Psi_{\text{DCT},N}]. \quad (5.1.1)$$

Let

$$\Psi_{\text{DCT},N} = [\psi_1 \quad \psi_2 \quad \dots \quad \psi_N]$$

The  $k$ -th column of  $\Psi_{\text{DCT},N}$  is given by

$$\psi_k(n) = \sqrt{\frac{2}{N}} \Omega_k \cos\left(\frac{\pi}{2N}(2n-1)(k-1)\right), n = 1, \dots, N, \quad (5.1.2)$$

with  $\Omega_k = \frac{1}{\sqrt{2}}$  for  $k = 1$  and  $\Omega_k = 1$  for  $2 \leq k \leq N$ .

Note that for  $k = 1$ , the entries become

$$\sqrt{\frac{2}{N}} \frac{1}{\sqrt{2}} \cos 0 = \sqrt{\frac{1}{N}}.$$

Thus, the  $l_2$  norm of  $\psi_1$  is 1. We can similarly verify the  $l_2$  norm of other columns also. They are all one.

**Theorem 5.1** [25] *The Dirac-DCT dictionary has coherence  $\sqrt{\frac{2}{N}}$ .*

PROOF. The coherence of a two ortho basis where one basis is Dirac basis is given by the magnitude of the largest entry in the other basis. For  $\Psi_{DCT,N}$ , the largest value is obtained when  $\Omega_k = 1$  and the cos term evaluates to 1. Clearly,

$$\mu(\mathcal{D}_{DCT}) = \sqrt{\frac{2}{N}}.$$

□

**Theorem 5.2** [25] *The  $p$  Babel function for Dirac-DCT dictionary is given by*

$$\mu_p(k) = k^{\frac{1}{p}}\mu \quad \forall 1 \leq k \leq N. \quad (5.1.3)$$

*In particular, the standard Babel function is given by*

$$\mu_1(k) = k\mu \quad (5.1.4)$$

PROOF. TODO prove it. □

## 5.2. Grassmannian frames

### 5.3. Rademacher sensing matrices

In this section we collect several results related to Rademacher sensing matrices.

**Definition 5.2** A Rademacher sensing matrix  $\Phi \in \mathbb{R}^{M \times N}$  with  $M < N$  is constructed by drawing each entry  $\phi_{ij}$  independently

from a Radamacher random distribution given by

$$\mathbb{P}_X(x) = \frac{1}{2}\delta\left(x - \frac{1}{\sqrt{M}}\right) + \frac{1}{2}\delta\left(x + \frac{1}{\sqrt{M}}\right). \quad (5.3.1)$$

Thus  $\phi_{ij}$  takes a value  $\pm\frac{1}{\sqrt{M}}$  with equal probability.

We can remove the scale factor  $\frac{1}{\sqrt{M}}$  out of the matrix  $\Phi$  writing

$$\Phi = \frac{1}{\sqrt{M}}\mathcal{X}$$

With that we can draw individual entries of  $\mathcal{X}$  from a simpler Rademacher distribution given by

$$\mathbb{P}_X(x) = \frac{1}{2}\delta(x - 1) + \frac{1}{2}\delta(x + 1). \quad (5.3.2)$$

Thus entries in  $\mathcal{X}$  take values of  $\pm 1$  with equal probability.

This construction is useful since it allows us to implement the multiplication with  $\Phi$  in terms of just additions and subtractions. The scaling can be implemented towards the end in the signal processing chain.

We note that

$$\mathbb{E}(\phi_{ij}) = 0. \quad (5.3.3)$$

$$\mathbb{E}(\phi_{ij}^2) = \frac{1}{M}. \quad (5.3.4)$$

Actually we have a better result with

$$\phi_{ij}^2 = \frac{1}{M}. \quad (5.3.5)$$

We can write

$$\Phi = \begin{bmatrix} \phi_1 & \dots & \phi_N \end{bmatrix}$$

where  $\phi_j \in \mathbb{R}^M$  is a Rademacher random vector with independent entries.

We note that

$$\mathbb{E}(\|\phi_j\|_2^2) = \mathbb{E}\left(\sum_{i=1}^M \phi_{ij}^2\right) = \sum_{i=1}^M (\mathbb{E}(\phi_{ij}^2)) = M \frac{1}{M} = 1. \quad (5.3.6)$$

Actually in this case we also have

$$\|\phi_j\|_2^2 = 1. \quad (5.3.7)$$

Thus the squared length of each of the columns in  $\Phi$  is 1.

**Lemma 5.3** [A tail bound for Rademacher random vectors] *Let  $z \in \mathbb{R}^M$  be a Rademacher random vector with i.i.d entries  $z_i$  that take a value  $\pm \frac{1}{\sqrt{M}}$  with equal probability. Let  $u \in \mathbb{R}^M$  be an arbitrary unit norm vector. Then*

$$\mathbb{P}(|\langle z, u \rangle| > \epsilon) \leq 2 \exp\left(-\epsilon^2 \frac{M}{2}\right). \quad (5.3.8)$$

PROOF. This can be proven using Hoeffding inequality. To be elaborated later.  $\square$

A particular application of this lemma is when  $u$  itself is another (independently chosen) unit norm Rademacher random vector.

The lemma establishes that the probability of inner product of two independent unit norm Rademacher random vectors being large is very very small. In other words, independently chosen unit norm Rademacher random vectors are incoherent with high probability. This is a very useful result as we will see later in measurement of coherence of Rademacher sensing matrices.

### 5.3.1. Joint correlation

Columns of  $\Phi$  satisfy a joint correlation property ([38]) which is described in following lemma.

**Lemma 5.4** Let  $\{u_k\}$  be a sequence of  $K$  vectors (where  $u_k \in \mathbb{R}^M$ ) whose  $l_2$  norms do not exceed one. Independently choose  $z \in \mathbb{R}^M$  to be a random vector with i.i.d. entries  $z_i$  that take a value  $\pm \frac{1}{\sqrt{M}}$  with equal probability. Then

$$\mathbb{P}\left(\max_k |\langle z, u_k \rangle| \leq \epsilon\right) \geq 1 - 2K \exp\left(-\epsilon^2 \frac{M}{2}\right). \quad (5.3.9)$$

PROOF. Let us call  $\gamma = \max_k |\langle z, u_k \rangle|$ .

We note that if for any  $u_k$ ,  $\|u_k\|_2 < 1$  and we increase the length of  $u_k$  by scaling it, then  $\gamma$  will not decrease and hence  $\mathbb{P}(\gamma \leq \epsilon)$  will not increase. Thus if we prove the bound for vectors  $u_k$  with  $\|u_k\|_2 = 1 \forall 1 \leq k \leq K$ , it will be applicable for all  $u_k$  whose  $l_2$  norms do not exceed one. Hence we will assume that  $\|u_k\|_2 = 1$ .

From lemma 5.3 we have

$$\mathbb{P}(|\langle z, u_k \rangle| > \epsilon) \leq 2 \exp\left(-\epsilon^2 \frac{M}{2}\right).$$

Now the event

$$\left\{\max_k |\langle z, u_k \rangle| > \epsilon\right\} = \bigcup_{k=1}^K \{|\langle z, u_k \rangle| > \epsilon\}$$

i.e. if any of the inner products (absolute value) is greater than  $\epsilon$  then the maximum is greater.

We recall Boole's inequality which states that

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i).$$

Thus

$$\mathbb{P}\left(\max_k |\langle z, u_k \rangle| > \epsilon\right) \leq 2K \exp\left(-\epsilon^2 \frac{M}{2}\right).$$

This gives us

$$\begin{aligned}\mathbb{P} \left( \max_k |\langle z, u_k \rangle| \leq \epsilon \right) &= 1 - \mathbb{P} \left( \max_k |\langle z, u_k \rangle| > \epsilon \right) \\ &\geq 1 - 2K \exp \left( -\epsilon^2 \frac{M}{2} \right).\end{aligned}\tag{5.3.10}$$

□

### 5.3.2. Coherence of Rademacher sensing matrix

We show that coherence of Rademacher sensing matrix is fairly small with high probability (adapted from [38]).

**Lemma 5.5** [Coherence of Rademacher sensing matrix] *Fix  $\delta \in (0, 1)$ . For an  $M \times N$  Rademacher sensing matrix  $\Phi$  as defined in definition 5.2, the coherence statistic*

$$\mu \leq \sqrt{\frac{4}{M} \ln \left( \frac{N}{\delta} \right)}\tag{5.3.11}$$

*with probability exceeding  $1 - \delta$ .*

PROOF. We recall the definition of coherence as

$$\mu = \max_{j \neq k} |\langle \phi_j, \phi_k \rangle| = \max_{j < k} |\langle \phi_j, \phi_k \rangle|.$$

Since  $\Phi$  is a Rademacher sensing matrix hence each column of  $\Phi$  is unit norm column. Consider some  $1 \leq j < k \leq N$  identifying columns  $\phi_j$  and  $\phi_k$ . We note that they are independent of each other. Thus from lemma 5.3 we have

$$\mathbb{P} (|\langle \phi_j, \phi_k \rangle| > \epsilon) \leq 2 \exp \left( -\epsilon^2 \frac{M}{2} \right).$$

Now there are  $\frac{N(N-1)}{2}$  such pairs of  $(j, k)$ . Hence by applying Boole's inequality

$$\mathbb{P} \left( \max_{j < k} |\langle \phi_j, \phi_k \rangle| > \epsilon \right) \leq 2 \frac{N(N-1)}{2} \exp \left( -\epsilon^2 \frac{M}{2} \right) \leq N^2 \exp \left( -\epsilon^2 \frac{M}{2} \right).$$

Thus we have

$$\mathbb{P}(\mu > \epsilon) \leq N^2 \exp\left(-\epsilon^2 \frac{M}{2}\right).$$

What we need to do now is to choose a suitable value of  $\epsilon$  so that the R.H.S. of this inequality is simplified.

We choose

$$\epsilon^2 = \frac{4}{M} \ln\left(\frac{N}{\delta}\right).$$

This gives us

$$\epsilon^2 \frac{M}{2} = 2 \ln\left(\frac{N}{\delta}\right) \implies \exp\left(-\epsilon^2 \frac{M}{2}\right) = \left(\frac{\delta}{N}\right)^2.$$

Putting back we get

$$\mathbb{P}(\mu > \epsilon) \leq N^2 \left(\frac{\delta}{N}\right)^2 \leq \delta^2.$$

This justifies why we need  $\delta \in (0, 1)$ .

Finally

$$\mathbb{P}\left(\mu \leq \sqrt{\frac{4}{M} \ln\left(\frac{N}{\delta}\right)}\right) = \mathbb{P}(\mu \leq \epsilon) = 1 - \mathbb{P}(\mu > \epsilon) > 1 - \delta^2$$

and

$$1 - \delta^2 > 1 - \delta$$

which completes the proof.  $\square$

## 5.4. Gaussian sensing matrices

In this section we collect several results related to Gaussian sensing matrices.

**Definition 5.3** A Gaussian sensing matrix  $\Phi \in \mathbb{R}^{M \times N}$  with  $M < N$  is constructed by drawing each entry  $\phi_{ij}$  independently from a Gaussian random distribution  $\mathcal{N}(0, \frac{1}{M})$ .

We note that

$$\mathbb{E}(\phi_{ij}) = 0. \quad (5.4.1)$$

$$\mathbb{E}(\phi_{ij}^2) = \frac{1}{M}. \quad (5.4.2)$$

We can write

$$\Phi = \begin{bmatrix} \phi_1 & \dots & \phi_N \end{bmatrix}$$

where  $\phi_j \in \mathbb{R}^M$  is a Gaussian random vector with independent entries.

We note that

$$\mathbb{E}(\|\phi_j\|_2^2) = \mathbb{E}\left(\sum_{i=1}^M \phi_{ij}^2\right) = \sum_{i=1}^M (\mathbb{E}(\phi_{ij}^2)) = M \frac{1}{M} = 1. \quad (5.4.3)$$

Thus the expected value of squared length of each of the columns in  $\Phi$  is 1.

#### 5.4.1. Joint correlation

Columns of  $\Phi$  satisfy a joint correlation property ([38]) which is described in following lemma.

**Lemma 5.6** *Let  $\{u_k\}$  be a sequence of  $K$  vectors (where  $u_k \in \mathbb{R}^M$ ) whose  $l_2$  norms do not exceed one. Independently choose  $z \in \mathbb{R}^M$  to be a random vector with i.i.d.  $\mathcal{N}(0, \frac{1}{M})$  entries. Then*

$$\mathbb{P}\left(\max_k |\langle z, u_k \rangle| \leq \epsilon\right) \geq 1 - K \exp\left(-\epsilon^2 \frac{M}{2}\right). \quad (5.4.4)$$

PROOF. Let us call  $\gamma = \max_k |\langle z, u_k \rangle|$ .

We note that if for any  $u_k$ ,  $\|u_k\|_2 < 1$  and we increase the length of  $u_k$  by scaling it, then  $\gamma$  will not decrease and hence  $\mathbb{P}(\gamma \leq \epsilon)$  will not increase. Thus if we prove the bound for vectors  $u_k$  with  $\|u_k\|_2 = 1 \forall 1 \leq k \leq K$ , it will be applicable for all  $u_k$  whose  $l_2$  norms do not exceed one. Hence we will assume that  $\|u_k\|_2 = 1$ .

Now consider  $\langle z, u_k \rangle$ . Since  $z$  is a Gaussian random vector, hence  $\langle z, u_k \rangle$  is also a Gaussian random vector. Since  $\|u_k\| = 1$  hence

$$\langle z, u_k \rangle \sim \mathcal{N} \left( 0, \frac{1}{M} \right).$$

We recall a well known tail bound for Gaussian random variables which states that

$$\mathbb{P}_X(|x| > \epsilon) = \sqrt{\frac{2}{\pi}} \int_{\epsilon\sqrt{N}}^{\infty} \exp \left( -\frac{x^2}{2} \right) dx \leq \exp \left( -\epsilon^2 \frac{M}{2} \right).$$

Now the event

$$\left\{ \max_k |\langle z, u_k \rangle| > \epsilon \right\} = \bigcup_{k=1}^K \{ |\langle z, u_k \rangle| > \epsilon \}$$

i.e. if any of the inner products (absolute value) is greater than  $\epsilon$  then the maximum is greater.

We recall Boole's inequality which states that

$$\mathbb{P} \left( \bigcup_i A_i \right) \leq \sum_i \mathbb{P}(A_i).$$

Thus

$$\mathbb{P} \left( \max_k |\langle z, u_k \rangle| > \epsilon \right) \leq K \exp \left( -\epsilon^2 \frac{M}{2} \right).$$

This gives us

$$\begin{aligned} \mathbb{P} \left( \max_k |\langle z, u_k \rangle| \leq \epsilon \right) &= 1 - \mathbb{P} \left( \max_k |\langle z, u_k \rangle| > \epsilon \right) \\ &\geq 1 - K \exp \left( -\epsilon^2 \frac{M}{2} \right). \end{aligned} \tag{5.4.5}$$

□

## 5.5. Partial Fourier sensing matrices

In this section we collect several results related to partial Fourier sensing matrices.

**5.6. Digest**

## CHAPTER 6

# Basis Pursuit for Sparse Recovery

### 6.1. Introduction

We recall following sparse approximation and CS recovery problems.

#### 6.1.1. Sparse representation formulations

Given a signal  $x \in \mathbb{C}^N$  which is known to have a sparse representation in a dictionary  $\mathcal{D}$ , the exact-sparse recovery problem is:

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_0 \text{ subject to } x = \mathcal{D}\alpha. \quad (\text{P}_0)$$

When  $x \in \mathbb{C}^N$  doesn't have a sparse representation in  $\mathcal{D}$ , a  $K$ -sparse approximation of  $x$  in  $\mathcal{D}$  can be obtained by solving the following problem:

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|x - \mathcal{D}\alpha\|_2 \text{ subject to } \|\alpha\|_0 \leq K. \quad (\text{P}_0^K)$$

Here  $x$  is modeled as  $x = \mathcal{D}\alpha + e$  where  $\alpha$  denotes a sparse representation of  $x$  and  $e$  denotes the approximation error.

A different way to formulate the approximation problem is to provide an upper bound to the acceptable approximation error  $\|e\|_2 \leq \epsilon$  and try to find sparsest possible representation within this approximation error bound as

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_0 \text{ subject to } \|x - \mathcal{D}\alpha\|_2 \leq \epsilon. \quad (\text{P}_0^\epsilon)$$

#### 6.1.2. CS formulations

In the context of compressed sensing, for simplicity, we assume the sparsifying dictionary to be the Dirac basis (i.e.  $\mathcal{D} = I$  and  $N =$

$D$ ). Further, we assume signal  $x$  to be  $K$ -sparse in  $\mathbb{C}^N$ . With the sensing matrix  $\Phi$  and the measurement vector  $y$ , the CS sparse recovery problem in the absence of measurement noise (i.e.  $y = \Phi x$ ) is stated as:

$$\hat{x} = \arg \min_{x \in \mathbb{C}^N} \|x\|_0 \text{ subject to } y = \Phi x. \quad (\text{CS}_0)$$

In the presence of measurement noise (i.e.  $y = \Phi x + e$ ), the recovery problem takes the form of

$$\hat{x} = \arg \min_{x \in \mathbb{C}^N} \|y - \Phi x\|_2 \text{ subject to } \|x\|_0 \leq K. \quad (\text{CS}_0^K)$$

when a bound on sparsity is provided, or alternatively:

$$\hat{x} = \arg \min_{x \in \mathbb{C}^N} \|x\|_0 \text{ subject to } \|y - \Phi x\|_2 \leq \epsilon. \quad (\text{CS}_0^\epsilon)$$

when a bound on the measurement noise is provided.

### 6.1.3. Basis pursuit formulations

In this chapter, we look at the basis pursuit based methods to solve these problems.

**Basis Pursuit** (BP) [14] suggests the convex relaxation of  $(P_0)$  by replacing  $l_0$ -“norm” with  $l_1$ -norm.

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_1 \text{ subject to } x = \mathcal{D}\alpha. \quad (P_1)$$

For real signals, it can be implemented as a linear program. For complex signals, it can be implemented as a second order cone program.

In the presence of approximation error  $(P_0^\epsilon)$ , where  $x = \mathcal{D}\alpha + e$  with  $\alpha$  being a  $K$ -sparse approximate representation of  $x$  in  $\mathcal{D}$  we can formulate corresponding  $l_1$ -minimization problem as:

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_1 \text{ subject to } \|x - \mathcal{D}\alpha\|_2 \leq \epsilon \quad (P_1^\epsilon)$$

where  $\epsilon \geq \|e\|_2$  provides an upper bound on the approximation error. This version is known as **basis pursuit with inequality constraints** (BPIC). The dual problem constructed using Lagrange multipliers is

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_1 + \lambda \|x - \mathcal{D}\alpha\|_2^2. \quad (P_1^\lambda)$$

This is known as **basis pursuit denoising**(BPDN). With appropriate choice of  $\lambda$ , the two problems BPIC and BPDN are equivalent. This formulation attempts to minimize the  $l_1$ -norm subject to a penalty term over the approximation error. The Lagrangian constant  $\lambda$  controls how large the penalty due to approximation error will be.

Note that the constraint  $\|x - \mathcal{D}\alpha\|_2 \leq \epsilon$  is equivalent to  $\|x - \mathcal{D}\alpha\|_2^2 \leq \epsilon^2$ . We have used the squared version to construct the dual BPDN problem since the term  $\|x - \mathcal{D}\alpha\|_2^2$  is easier to differentiate and work with.

Efficient solvers are available to solve BP, BPIC, BPDN problems using convex optimization techniques. They are usually polynomial time and involve sophisticated algorithms for implementation. The good part is a guarantee that a globally unique solution can be found (since the problem is convex). The not so good part is that convex optimization methods are still quite computationally intensive.

An alternative formulation of BPDN is as follows.

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \frac{1}{2} \|x - \mathcal{D}\alpha\|_2^2 + \gamma \|\alpha\|_1. \quad (\text{P}_1^\gamma)$$

The difference in the two formulations is essentially with which term the Lagrangian constant ( $\lambda$  or  $\gamma$ ) is placed. By choosing  $\lambda = 1/(2\gamma)$ , the two formulations are essentially the same (with a scale factor in the objective function). This formulation attempts to minimize the approximation error subject to an  $l_1$ -norm penalty. Thus, the two formulations differentiate w.r.t. which term is minimized and which term is considered as penalty.

Basis pursuit is not an algorithm but a principle which says that for most real life problems, the solution of  $l_0$ -minimization problem is same as the solution of  $l_1$ -minimization problem. Actual algorithms for solving the basis pursuit formulation of sparse recovery problem come from convex optimization literature.

## 6.2. Basis Pursuit

We start our discussion with the analysis of exact-sparse case.

As part of our theoretical analysis, we would like to explore conditions under which the problems  $(P_0)$  and  $(P_1)$  are equivalent i.e. there exists a unique solution to both of them and the solution is identical. Under such conditions, the NP-hard problem  $(P_0)$  can be easily replaced with a tractable  $(P_1)$  problem which is convex and solvable in polynomial time.

### 6.2.1. Two-Ortho-Case

Further simplifying, we consider the case where the dictionary  $\mathcal{D}$  is a two-ortho-basis

$$\mathcal{D} = [\Psi \quad \Phi]$$

with  $\Psi$  and  $\Phi$  both being orthonormal bases for  $\mathbb{C}^N$ . Clearly,  $\mathcal{D} \in \mathbb{C}^{N \times 2N}$  and  $D = 2N$ . We denote

$$\Omega = \{1, 2, \dots, 2N\}$$

as the index set for the representation vectors  $\alpha$ .

The representation  $\alpha$  of a signal  $x$  in  $\mathcal{D}$  can be written as

$$x = \mathcal{D}\alpha = [\Psi \quad \Phi] \begin{bmatrix} \alpha^p \\ \alpha^q \end{bmatrix} = \Psi\alpha^p + \Phi\alpha^q.$$

We can assign

$$k_p = \|\alpha^p\|_0 \quad \text{and} \quad k_q = \|\alpha^q\|_0.$$

Total sparsity of  $\alpha$  is given by

$$K = \|\alpha\|_0 = k_p + k_q.$$

Whenever  $K \ll N$ , we have a sparse representation. Further, let  $S_p \subseteq \{1, \dots, N\}$  be the support corresponding to  $\alpha^p$  part of  $\alpha$  (i.e.  $S_p = \text{supp}(\alpha^p)$ ) and  $S_q \subseteq \{1, \dots, N\}$  be the support corresponding to  $\alpha^q$  part of  $\alpha$  (i.e.  $S_q = \text{supp}(\alpha^q)$ ). Clearly,  $|S_p| = k_p$  and  $|S_q| = k_q$ . Note that  $S_p$  and  $S_q$  need not be disjoint. But,  $S_p$  and  $S_q + N$  are disjoint.

In fact,  $\text{supp}(\alpha) = S_p \cup (S_q + N)$ .  $\mathbf{1}_p \in \mathbb{C}^N$  will denote the indicator vector for  $S_p$  i.e.  $\mathbf{1}_p(i) = 0 \forall i \notin S_p$  and  $\mathbf{1}_p(i) = 1 \forall i \in S_p$ . Similarly,  $\mathbf{1}_q \in \mathbb{C}^N$  will denote the indicator vector for  $S_q$ .  $\mathbf{1} \in \mathbb{C}^N$  will denote the vector  $\{1, \dots, 1\}$ . Also,  $\mathbf{1} \in \mathbb{C}^{N \times N}$  will denote a square matrix of all ones. Note that  $\mathbf{1} = \mathbf{1} \cdot \mathbf{1}^T$ .

We now state our main result for equivalence of solutions of  $(P_0)$  and  $(P_1)$  for the two ortho-case. Going forward, we will simply use  $\mu$  to refer to the coherence of  $\mathcal{D}$  (i.e.  $\mu(\mathcal{D})$ ).

**Theorem 6.1** [Equivalence-Basis Pursuit-Two-Ortho Case] *Let  $\mathcal{D}$  be a two-ortho-basis dictionary  $\mathcal{D} = [\Psi \Phi]$ . Let  $x = \mathcal{D}\alpha$ , where  $x$  is known. If a  $K$ -sparse representation  $\alpha$  exists with  $k_p \geq k_q$  such that  $(k_p, k_q)$  obey*

$$2\mu^2 k_p k_q + \mu k_p - 1 < 0 \quad (6.2.1)$$

*, then  $\alpha$  is the unique solution of both problems  $(P_0)$  and  $(P_1)$ .*

*A weaker condition is: if*

$$\|\alpha\|_0 = K = k_p + k_q < \frac{\sqrt{2} - 0.5}{\mu} \quad (6.2.2)$$

*, then  $\alpha$  is a unique ( $K$ -sparse) solution to both  $(P_0)$  and  $(P_1)$ .*

**PROOF.** Let  $\alpha$  be solution of  $(P_0)$ . Clearly,  $\alpha$  is a feasible vector to  $(P_1)$  though it need not be an optimal solution. We have to find criteria under which  $\alpha$  is optimal and no other solution  $\beta$  is optimal.

Towards this, we consider the set of alternative solutions to  $(P_1)$  given by

$$C = \{\beta | \beta \neq \alpha, \|\beta\|_1 \leq \|\alpha\|_1, \|\beta\|_0 > \|\alpha\|_0 \text{ and } \mathcal{D}(\alpha - \beta) = 0\}.$$

This set contains all solutions to  $(P_1)$  which are different from  $\alpha$ , have larger support, satisfy the linear system of equations  $x = \mathcal{D}\alpha$  and have  $l_1$  norm less than or equal to  $\alpha$ . If this set is non-empty, then there

exists a solution to basis pursuit which is not same as  $\alpha$ . If this set is empty, then the solutions of  $(P_0)$  and  $(P_1)$  coincide.

Towards the end of this proof, we show that  $(6.2.1) \implies (6.2.2)$ . Due to  $(6.2.2)$ ,

$$\|\alpha\|_0 = K = k_p + k_q < \frac{\sqrt{2} - 0.5}{\mu} = \frac{0.414}{\mu} < \frac{1}{\mu}.$$

Thus, if  $\alpha$  satisfies  $(6.2.1)$ , then it is necessarily the sparsest possible representation. All other representations are denser (i.e. have more non-zero entries). Thus,  $\|\beta\|_0 > \|\alpha\|_0$  for every  $\beta \in C$ .

Writing  $e = \beta - \alpha \iff \beta = e + \alpha$ , we have

$$\|\beta\|_1 \leq \|\alpha\|_1 \iff \|e + \alpha\|_1 - \|\alpha\|_1 \leq 0.$$

Thus, we can rewrite  $C$  as

$$C_s = \{e | e \neq 0, \|e + \alpha\|_1 - \|\alpha\|_1 \leq 0 \text{ and } \mathcal{D}e = 0\}.$$

In order to show that  $C$  is empty, we will show that a larger set containing  $C_s$  is also empty. Essentially, we wish to consider a larger set whose volume can be assessed. If that set is empty due to  $(6.2.1)$ , then  $C$  would also be empty and we would have completed the proof.

We start by the requirement  $\|e + \alpha\|_1 - \|\alpha\|_1 \leq 0$ . Let  $\alpha = \begin{bmatrix} \alpha^p \\ \alpha^q \end{bmatrix}$  and  $e = \begin{bmatrix} e^p \\ e^q \end{bmatrix}$ , where  $p$  and  $q$  refer to parts corresponding to the orthonormal bases  $\Psi$  and  $\Phi$  respectively (as described at the beginning of this section). Note that even if  $\alpha^p$  and  $\alpha^q$  are sparse,  $e^p$  and  $e^q$  need not be. In fact, support of  $e^p$  and  $e^q$  could be very different from  $S_p$  and  $S_q$ .

We can now write

$$\begin{aligned} 0 \geq \|e + \alpha\|_1 - \|\alpha\|_1 &= \left( \sum_{i=1}^N |e_i^p + \alpha_i^p| - |\alpha_i^p| \right) + \left( \sum_{i=1}^N |e_i^q + \alpha_i^q| - |\alpha_i^q| \right) \\ &= \sum_{i \notin S_p} |e_i^p| + \sum_{i \notin S_q} |e_i^q| \\ &\quad + \left( \sum_{i \in S_p} |e_i^p + \alpha_i^p| - |\alpha_i^p| \right) + \left( \sum_{i \in S_q} |e_i^q + \alpha_i^q| - |\alpha_i^q| \right) \end{aligned}$$

What is going on here? We are splitting the sums to the sums over indices in the supports  $S_p$  and  $S_q$  and over indices outside the two supports. i.e. the indices  $i \in S_p$  and  $i \notin S_p$ , similarly  $i \in S_q$  and  $i \notin S_q$ .

For  $i \notin S_p$ ,  $\alpha_i^p = 0$  leading to  $|e_i^p + \alpha_i^p| - |\alpha_i^p| = |e_i^p|$ . Ditto for  $i \notin S_q$ .

We recall from triangle inequality that  $|a + b| \geq |b| - |a| \forall a, b \in \mathbb{C}^N$  which implies  $|a + b| - |b| \geq -|a|$ . Thus,

$$|e_i^p + \alpha_i^p| - |\alpha_i^p| \geq -|e_i^p| \forall i \in S_p$$

and

$$|e_i^q + \alpha_i^q| - |\alpha_i^q| \geq -|e_i^q| \forall i \in S_q.$$

With this, the above condition can be relaxed as <sup>1</sup>

$$0 \geq \|e + \alpha\|_1 - \|\alpha\|_1 \geq \sum_{i \notin S_p} |e_i^p| + \sum_{i \notin S_q} |e_i^q| - \sum_{i \in S_p} |e_i^p| - \sum_{i \in S_q} |e_i^q|.$$

Every  $e$  satisfying this inequality will also satisfy the requirements of  $C$ . To simplify notation we can write

$$\sum_{i \in S_p} |e_i^p| = \mathbf{1}_p^T |e^p| \text{ and } \sum_{i \in S_q} |e_i^q| = \mathbf{1}_q^T |e^q|.$$

Then we have

$$\|e^p\|_1 = \sum_{i \in S_p} |e_i^p| + \sum_{i \notin S_p} |e_i^p| \iff \sum_{i \notin S_p} |e_i^p| = \|e^p\|_1 - \sum_{i \in S_p} |e_i^p| = \|e^p\|_1 - \mathbf{1}_p^T |e^p|.$$

---

<sup>1</sup>Note that the triangle inequality goes for the worst case condition

Similarly,

$$\sum_{i \notin S_q} |e_i^q| = \|e^q\|_1 - \mathbf{1}_q^T |e^q|.$$

Thus,

$$\sum_{i \notin S_p} |e_i^p| + \sum_{i \notin S_q} |e_i^q| - \sum_{i \in S_p} |e_i^p| - \sum_{i \in S_q} |e_i^q| = \|e^p\|_1 - 2\mathbf{1}_p^T |e^p| + \|e^q\|_1 - 2\mathbf{1}_q^T |e^q|.$$

We can now define the set

$$C_s^1 = \{e \mid \|e^p\|_1 + \|e^q\|_1 - 2\mathbf{1}_p^T |e^p| - 2\mathbf{1}_q^T |e^q| \leq 0 \text{ and } \mathcal{D}e = 0\}. \quad (6.2.3)$$

Clearly,  $C_s \subseteq C_s^1$  and if  $C_s^1$  is empty, then  $C_s$  will also be empty. Note that this formulation of  $C_s^1$  is dependent only on the support of  $\alpha$  and not on values in  $\alpha$ .

We now turn back to the requirement  $\mathcal{D}e = 0$  and relax it further. We note that,

$$\mathcal{D}e = \begin{bmatrix} \Psi & \Phi \end{bmatrix} \begin{bmatrix} e^p \\ e^q \end{bmatrix} = \Psi e^p + \Phi e^q = 0.$$

Multiplying by  $\Psi^H$  we get

$$e^p + \Psi^H \Phi e^q = 0 \iff e^p = -\Psi^H \Phi e^q$$

since  $\Psi^H \Psi = I$  (unitary matrix). Similarly multiplying with  $\Phi^H$ , we obtain

$$\Phi^H \Psi e^p + e^q = 0 \iff e^q = -\Phi^H \Psi e^p.$$

Note that entries in  $\Psi^H \Phi$  and  $\Phi^H \Psi$  are inner products between columns of  $\mathcal{D}$ , hence their magnitudes are upper bounded by  $\mu$  (coherence). Denote  $B = \Psi^H \Phi$  and consider the product  $v = \Psi^H \Phi e^q = Be^q$ . Then

$$v_i = \sum_{j=1}^N B_{ij} e_j^q.$$

Thus,

$$|v_i| = \left| \sum_{j=1}^N B_{ij} e_j^q \right| \leq \sum_{j=1}^N |B_{ij} e_j^q| \leq \mu \sum_{j=1}^N |e_j^q| = \mu \mathbf{1}^T |e^q|.$$

Applying this result on  $e^p$  we get,

$$|e^p| = |\Psi^H \Phi e^q| \preceq \mu \mathbf{1} |e^q|.$$

Similarly,

$$|e^q| = |\Phi^H \Psi e^p| \preceq \mu \mathbf{1} |e^p|.$$

Note that since  $\mathbf{1} = \mathbf{1} \cdot \mathbf{1}^T$ , it is a rank-1 matrix.

We now construct a set  $C_s^2$  as

$$C_s^2 = \left\{ e \left| \begin{array}{l} e \neq 0 \\ \|e^p\|_1 + \|e^q\|_1 - 2\mathbf{1}_p^T |e^p| - 2\mathbf{1}_q^T |e^q| \leq 0 \\ |e^p| \preceq \mu \mathbf{1} |e^q| \\ \text{and } |e^q| \preceq \mu \mathbf{1} |e^p| \end{array} \right. \right\}. \quad (6.2.4)$$

Clearly,  $C_s^1 \subseteq C_s^2$  since for every  $e \in C_s^1$ ,  $\mathcal{D}e = 0 \implies e \in C_s^2$ .

We now define  $f^p = |e^p|$  and  $f^q = |e^q|$  as the absolute value vectors.

Correspondingly, let us define  $f = |e| = \begin{bmatrix} f^p \\ f^q \end{bmatrix}$ . Clearly,  $\|e^p\|_1 = \mathbf{1}^T f^p$  and  $\|e^q\|_1 = \mathbf{1}^T f^q$ . Further  $f^p \succeq 0$  i.e. every entry in  $f^p$  is non-negative. Similarly,  $f^q \succeq 0$ . We can then rewrite  $C_s^2$  as

$$C_f = \left\{ f \left| \begin{array}{l} f \neq 0 \\ \mathbf{1}^T f^p + \mathbf{1}^T f^q - 2\mathbf{1}_p^T f^p - 2\mathbf{1}_q^T f^q \leq 0 \\ f^p \preceq \mu \mathbf{1} f^q \\ f^q \preceq \mu \mathbf{1} f^p \\ \text{and } f^p \succeq 0, f^q \succeq 0 \end{array} \right. \right\}. \quad (6.2.5)$$

We note that if  $f \in C_f$ , then for all  $c \geq 0$ ,  $cf \in C_f$ . Thus, in order to study (the emptiness of)  $C_f$ , it is sufficient to study unit  $l_1$ -norm vectors  $f \in C_f$ . Now

$$\|f\|_1 = \mathbf{1}^T f = \mathbf{1}^T f^p + \mathbf{1}^T f^q$$

since  $f \succeq 0$ . This leads to:

$$\|f\|_1 = 1 \iff \mathbf{1}^T f^p + \mathbf{1}^T f^q = 1.$$

We construct the new set of unit  $l_1$ -norm vectors

$$C_r = \left\{ f \left| \begin{array}{l} f \neq 0 \\ 1 - 2\mathbb{1}_p^T f^p - 2\mathbb{1}_q^T f^q \leq 0 \\ f^p \preceq \mu \mathbf{1} f^q \\ f^q \preceq \mu \mathbf{1} f^p \\ \mathbb{1}^T f^p + \mathbb{1}^T f^q = 1 \\ \text{and } f^p \succeq 0, f^q \succeq 0 \end{array} \right. \right\}. \quad (6.2.6)$$

Clearly  $C_r \neq \emptyset \iff C_f \neq \emptyset$ . Note that the constraint  $1 - 2\mathbb{1}_p^T f^p - 2\mathbb{1}_q^T f^q \leq 0$  can be rewritten as

$$\mathbb{1}_p^T f^p + \mathbb{1}_q^T f^q \geq \frac{1}{2}$$

The set  $C_r$  is much easier to analyze since

- If has no explicit dependency on  $\mathcal{D}$ .  $\mathcal{D}$  is represented only by a single parameter, its coherence  $\mu$ .
- All constraints are simple linear constraints. Thus finding the elements of  $C_f$  can be formulated as a linear programming problem.
- The order of non-zero entries inside  $f^p$  and  $f^q$  doesn't have any influence on the requirements for  $f$  to belong to  $C_r$ . Thus, without loss of generality, we can focus on vectors for which the first  $k_p$  entries are non-zero in  $f^p$  and first  $k_q$  entries are non-zero in  $f^q$  respectively.

In order to find vectors in  $C_r$ , we can solve following linear program.

$$\begin{aligned}
& \underset{f^p, f^q}{\text{maximize}} && \mathbf{1}_p^T f^p + \mathbf{1}_q^T f^q \\
& \text{subject to} && f^p \preceq \mu \mathbf{1} f^q \\
& && f^q \preceq \mu \mathbf{1} f^p \\
& && \mathbf{1}^T (f^p + f^q) = 1 \\
& && f^p \succeq 0, f^q \succeq 0
\end{aligned} \tag{6.2.7}$$

$f = 0$  is a feasible vector for this linear program. Hence a solution does exist for this program. What is interesting is the value of the objective function for the optimal solution. Let  $f^{p*}, f^{q*}$  be (an) optimal solution for this linear program. If  $\mathbf{1}_p^T f^{p*} + \mathbf{1}_q^T f^{q*} \geq \frac{1}{2}$ , then  $f^*$  satisfies all the requirements of  $C_r$  and  $C_r$  is indeed not empty. This doesn't guarantee that  $C$  will also be non-empty though. On the contrary, if  $\mathbf{1}_p^T f^{p*} + \mathbf{1}_q^T f^{q*} < \frac{1}{2}$ , then  $C_r$  is indeed empty (as one of the requirements cannot be met), hence  $C_f$  is also empty leading to  $C \subset C_f$  being empty too. Thus, a condition which leads to  $\mathbf{1}_p^T f^{p*} + \mathbf{1}_q^T f^{q*} < \frac{1}{2}$  is a sufficient condition for equivalence of  $(P_0)$  and  $(P_1)$ .

Consider a solution  $f$  to (6.2.7). Let  $\|f^p\|_1 = \mathbf{1}^T f^p = c$ . Since  $\mathbf{1}^T (f^p + f^q) = 1$ , hence  $\|f^q\|_1 = \mathbf{1}^T f^q = 1 - c$ .

We note that

$$\mathbf{1} f^p = \mathbf{1} \cdot \mathbf{1}^T f^p = \|f^p\|_1 \mathbf{1} = c \mathbf{1}.$$

Similarly,

$$\mathbf{1} f^q = (1 - c) \mathbf{1}.$$

Thus, first two constraints change into

$$\begin{aligned}
f^p &\preceq (1 - c)\mu \mathbf{1} \\
f^q &\preceq c\mu \mathbf{1}.
\end{aligned} \tag{6.2.8}$$

Since the objective is to maximize  $\mathbf{1}_p^T f^p + \mathbf{1}_q^T f^q$ , it is natural to maximize non-zero entries in  $f^p$  and  $f^q$  corresponding to  $S_p$  and  $S_q$ . A straight-forward option is to choose first  $k_p$  entries in  $f^p$  to be  $(1 - c)\mu$  and first  $k_q$  entries in  $f^q$  to be  $c\mu$ . Other entries can be chosen arbitrarily to meet the requirement that  $\mathbf{1}^T (f^p + f^q) = 1$ . With this choice,

we have

$$\mathbb{1}_p^T f^p + \mathbb{1}_q^T f^q = k_p(1 - c)\mu + k_q c\mu = \mu(k_p - c(k_p - k_q)).$$

We recall that we have chosen  $k_p \geq k_q$ . Thus, the expression is maximized if  $c$  is chosen to be as small as possible.

The choice of  $c$  must meet following conditions on  $l_1$ -norms. (Basically the sum of first  $k_p$  terms of  $f_p$  must not be more than the  $l_1$  norm of  $f^p$ . Ditto for  $f^q$ ).

$$\begin{aligned} \|f^p\|_1 &= \mathbb{1}^T f^p = c \geq k_p(1 - c)\mu \\ \|f^q\|_1 &= \mathbb{1}^T f^q = 1 - c \geq k_q c\mu. \end{aligned} \tag{6.2.9}$$

Simplifying these inequalities we get

$$\begin{aligned} c \geq k_p(1 - c)\mu &\implies c \geq \frac{k_p\mu}{1 + k_p\mu} \\ 1 - c \geq k_q c\mu &\implies c \leq \frac{1}{1 + k_q\mu}. \end{aligned} \tag{6.2.10}$$

Since these two conditions must be satisfied, hence we require  $k_p, k_q$  to meet

$$\frac{k_p\mu}{1 + k_p\mu} \leq \frac{1}{1 + k_q\mu} \implies k_p k_q \leq \frac{1}{\mu^2}. \tag{6.2.11}$$

We will verify later that this condition is met if (6.2.1) holds. Assuming the condition is met, obviously the smallest possible value of  $c$  is given by  $\frac{k_p\mu}{1 + k_p\mu}$ . The maximum value of objective function then becomes

$$\begin{aligned} \mathbb{1}_p^T f^p + \mathbb{1}_q^T f^q &= \mu(k_p - c(k_p - k_q)) \\ &= \mu \left( k_p - \frac{k_p\mu}{1 + k_p\mu}(k_p - k_q) \right) \\ &= \frac{k_p\mu + k_p k_q \mu^2}{1 + k_p\mu}. \end{aligned} \tag{6.2.12}$$

Finally, for BP to succeed, we require this expression to be strictly less than half. This gives us

$$\frac{k_p\mu + k_p k_q \mu^2}{1 + k_p\mu} < \frac{1}{2} \implies 2k_p k_q \mu^2 + k_p \mu - 1 < 0 \tag{6.2.13}$$

which is the sufficient condition for BP to succeed in the theorem.

We now show that (6.2.1)  $\implies$  the weaker condition (6.2.2). From (6.2.1) we can write  $k_q$  as

$$2k_p k_q \mu^2 + k_p \mu - 1 < 0 \implies 2k_p k_q \mu^2 < 1 - k_p \mu \implies k_q < \frac{1 - k_p \mu}{2k_p \mu^2}.$$

Thus,

$$\begin{aligned} \|\alpha\|_0 = k_p + k_q &< k_p + \frac{1 - k_p \mu}{2k_p \mu^2} \\ &= \frac{2\mu^2 k_p^2 + 1 - \mu k_p}{2\mu^2 k_p} \\ &= \frac{1}{\mu} \cdot \frac{2\mu^2 k_p^2 + 1 - \mu k_p}{2\mu k_p}. \end{aligned} \tag{6.2.14}$$

We define  $u = \mu k_p$  and rewrite above as

$$\|\alpha\|_0 < \frac{1}{\mu} \frac{2u^2 - u + 1}{2u}.$$

The weaker condition can now be obtained by minimizing the upper bound on R.H.S. of this equation. We define

$$f(u) = \frac{2u^2 - u + 1}{2u}.$$

Differentiating and equating with 0, we get

$$f'(u) = \frac{2u^2 - 1}{2u^2} = 0.$$

The optimal value is obtained when  $u = \pm\sqrt{0.5}$ . Since both  $\mu$  and  $k_p$  are positive quantities, hence the negative value for  $u$  is rejected and we get  $u = \sqrt{0.5}$ . This gives us

$$\|\alpha\|_0 < \frac{1}{\mu} \frac{2 - \sqrt{0.5}}{2\sqrt{0.5}} = \frac{\sqrt{2} - 0.5}{\mu}.$$

Lastly, the property that arithmetic mean is greater than or equal to geometric mean gives us

$$k_p k_q \leq \frac{(k_p + k_q)^2}{4} < \frac{(\sqrt{2} - 0.5)^2}{4\mu^2} < \frac{1}{\mu^2}.$$

□

### 6.2.2. General Case

We now consider the case where  $\mathcal{D} \in \mathbb{C}^{N \times D}$  is an arbitrary (redundant) dictionary. We will require that  $\mathcal{D}$  is full row rank. If  $\mathcal{D}$  is not a full row rank matrix then some of its columns (atoms) can be removed to make it so.

We develop sufficient conditions under which solutions of  $(P_0)$  and  $(P_1)$  match for the general case [20, 21].

**Theorem 6.2** [Equivalence-Basis Pursuit] *Let  $\mathcal{D}$  be an arbitrary full rank redundant dictionary. Let  $x = \mathcal{D}\alpha$ , where  $x$  is known. If a sparse representation  $\alpha$  exists obeying*

$$\|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right), \quad (6.2.15)$$

*then  $\alpha$  is the unique solution of both  $(P_0)$  and  $(P_1)$ .*

PROOF. We start with defining the set of alternative feasible vectors to  $(P_1)$ :

$$C = \left\{ \beta \left| \begin{array}{l} \beta \neq \alpha \\ \|\beta\|_1 \leq \|\alpha\|_1 \\ \|\beta\|_0 > \|\alpha\|_0 \\ \text{and } \mathcal{D}(\beta - \alpha) = 0 \end{array} \right. \right\}. \quad (6.2.16)$$

This set contains all possible representations that are different from  $\alpha$ , have larger support, satisfy  $\mathcal{D}\beta = x$  and have a better (or at least as good)  $l_1$ -norm. We need to show that if (6.2.15) holds, then the set  $C$  will be empty. Otherwise, BP would choose a solution different than  $\alpha$ .

Since

$$\|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right)$$

it is necessarily the unique and sparsest solution. Thus any other  $\beta$  which satisfies  $\mathcal{D}\beta = x$ , is denser (i.e.  $\|\beta\|_0 > \|\alpha\|_0$ ). So this condition in  $C$  is redundant.

Following the proof of theorem 6.1, we define

$$e = \beta - \alpha.$$

We can then rewrite  $C$  as

$$C_s = \{e | e \neq 0, \|e + \alpha\|_1 - \|\alpha\|_1 \leq 0, \text{ and } \mathcal{D}e = 0\}. \quad (6.2.17)$$

Again, we will enlarge the set  $C_s$  and show that even the larger set is empty when (6.2.15) holds.

We start with the requirement  $\|e + \alpha\|_1 - \|\alpha\|_1 \leq 0$ . A simple permutation of columns of  $\mathcal{D}$  can bring the non-zero entries in  $\alpha$  to the beginning. Thus, without loss of generality, we assume that first  $K$  entries in  $\alpha$  are non-zero and the rest are zero. We can now rewrite the requirement as

$$\|e + \alpha\|_1 - \|\alpha\|_1 = \sum_{j=1}^K (|e_j + \alpha_j| - |\alpha_j|) + \sum_{j>K} |e_j| \leq 0. \quad (6.2.18)$$

Using the inequality  $|a + b| - |b| \geq -|a|$ , we can relax above condition as

$$-\sum_{j=1}^K |e_j| + \sum_{j>K} |e_j| \leq 0. \quad (6.2.19)$$

Let  $\mathbb{1}_K$  denote a vector with  $K$  ones at the beginning and rest 0s. Then,

$$\sum_{j=1}^K |e_j| = \mathbb{1}_K^T |e|.$$

Further,

$$\sum_{j>K} |e_j| = \|e\|_1 - \sum_{j=1}^K |e_j| = \mathbb{1}^T |e| - \mathbb{1}_K^T |e|.$$

Thus, we can rewrite above inequality as

$$\mathbb{1}^T |e| - 2\mathbb{1}_K^T |e| \leq 0. \quad (6.2.20)$$

We can now define

$$C_s^1 = \{e | e \neq 0, \mathbb{1}^T |e| - 2\mathbb{1}_K^T |e| \leq 0, \text{ and } \mathcal{D}e = 0\}. \quad (6.2.21)$$

Clearly  $C_s \subseteq C_s^1$ . We will now relax the requirement of  $\mathcal{D}e = 0$ . Multiplying by  $\mathcal{D}^H$ , we get

$$\mathcal{D}^H \mathcal{D}e = 0. \quad (6.2.22)$$

If  $e \in C_s^1$ , it will also satisfy above equation. Moreover, if  $e$  satisfies above, then  $e$  belongs to the null space of  $\mathcal{D}^H \mathcal{D}$ . Since  $\mathcal{D}$  is full rank, hence  $e$  has to be in the null space of  $\mathcal{D}$  also. Thus the two conditions  $\mathcal{D}e = 0$  and  $\mathcal{D}^H \mathcal{D}e = 0$  are equivalent. We note that off-diagonal entries in  $\mathcal{D}^H \mathcal{D}$  are bounded by  $\mu$  while the main diagonal consists of all ones. So, we can write

$$\mathcal{D}^H \mathcal{D}e = 0 \iff (\mathcal{D}^H \mathcal{D} - I + I)e = 0 \iff -e = (\mathcal{D}^H \mathcal{D} - I)e. \quad (6.2.23)$$

Suppose  $v = Gu$ . Then  $v_i = \sum_j G_{ij}u_j$ . Thus

$$|v_i| = \left| \sum_j G_{ij}u_j \right| \leq \sum_j |G_{ij}u_j| = \sum_j |G_{ij}| |u_j|.$$

This gives us  $|v| \preceq |G||v|$  where  $\preceq$  indicates component wise inequality.

Taking an entry-wise absolute value on both sides, we get

$$|e| = |(\mathcal{D}^H \mathcal{D} - I)e| \preceq |\mathcal{D}^H \mathcal{D} - I||e| \preceq \mu(\mathbf{1} - I)|e|. \quad (6.2.24)$$

The last part is due to the fact that all entries in the vector  $|e|$  and the matrix  $|\mathcal{D}^H \mathcal{D} - I|$  are non-negative and the entries in  $|\mathcal{D}^H \mathcal{D} - I|$  are dominated by  $\mu$ . Further,

$$\begin{aligned} |e| \preceq \mu(\mathbf{1} - I)|e| &\iff (1 + \mu)|e| \preceq \mu\mathbf{1}|e| = \mu\|e\|_1\mathbf{1} \\ &\iff |e| \preceq \frac{\mu\|e\|_1}{1 + \mu}\mathbf{1}. \end{aligned} \quad (6.2.25)$$

In the above we used the fact that  $\mathbf{1}|e| = \mathbf{1}\mathbf{1}^T|e| = \mathbf{1}\|e\|_1$ . We can now define a new set

$$C_s^2 = \left\{ e \left| \begin{array}{l} e \neq 0, \mathbf{1}^T|e| - 2\mathbf{1}_K^T|e| \leq 0 \\ \text{and } |e| \preceq \frac{\mu\|e\|_1}{1 + \mu}\mathbf{1} \end{array} \right. \right\}. \quad (6.2.26)$$

Clearly,  $C_s^1 \subseteq C_s^2$ . We note that  $C_s^2$  is unbounded since if  $e \in C_s^2$ , then  $ce \in C_s^2 \forall c \neq 0$ . Thus, in order to study its behavior, it is sufficient

to consider the set of vectors with unit norm vectors  $\|e\|_1 = 1$ . We construct the new set as

$$C_r = \left\{ e \left| \|e\|_1 = 1, 1 - 2\mathbb{1}_K^T |e| \leq 0 \text{ and } |e| \preceq \frac{\mu}{1+\mu} \mathbb{1} \right. \right\}. \quad (6.2.27)$$

Note that we replaced  $\mathbb{1}^T |e| = \|e\|_1 = 1$  in formulating the description of  $C_r$  and the condition  $e \neq 0$  is automatically enforced since  $\|e\|_1 = 1$ . Clearly  $C_s^2 = \emptyset \iff C_r = \emptyset$ .

In order to satisfy the requirement  $1 - 2\mathbb{1}_K^T |e| \leq 0$ , we need to have  $\mathbb{1}_K^T |e|$  as large as possible. Since this quantity only considers first  $K$  entries in  $e$ , hence the energy in  $e$  should be concentrated inside the first  $K$  entries to maximize this quantity. However, entries in  $e$  are restricted by the third requirement in  $C_r$ . We can maximize it by choosing

$$|e_j| = \frac{\mu}{1+\mu}$$

for first  $K$  entries in  $e$ . We then get

$$1 - 2\mathbb{1}_K^T |e| = 1 - 2K \frac{\mu}{1+\mu} \leq 0. \quad (6.2.28)$$

This gives us

$$\begin{aligned} 1 - 2K \frac{\mu}{1+\mu} \leq 0 &\iff 1 + \mu \leq 2K\mu \\ &\iff 2K \geq \frac{1+\mu}{\mu} \\ &\iff K \geq \frac{1}{2} \left( 1 + \frac{1}{\mu} \right). \end{aligned} \quad (6.2.29)$$

This is a necessary condition for  $C_r$  to be non-empty. Thus, if

$$K < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right)$$

then, the requirement  $1 - 2\mathbb{1}_K^T |e| \leq 0$  is not satisfied and  $C_r$  is empty. Consequently,  $C$  is empty and the theorem is proved.

□

We present another result which is based on  $\mu_{1/2}(G)$  measure of the Gram matrix of the dictionary.

**Theorem 6.3** [20] *Let  $x = \mathcal{D}\alpha$  and  $\|\alpha\|_0 < \mu_{1/2}(G)$ , then the unique solution of both  $(P_0)$  and  $(P_1)$ .*

PROOF. Let  $\Lambda = \text{supp}(\alpha)$  and  $K = |\Lambda|$ . As per the theorem, let  $K < \mu_{1/2}(G)$ . We show that any vector in the null space of  $\mathcal{D}$  exhibits less than 50% concentration on  $\Lambda$ , i.e. for every  $h \in \mathcal{N}(\mathcal{D})$

$$\sum_{k \in \Lambda} |h_k| < \frac{1}{2} \|h\|_1. \quad (6.2.30)$$

Now

$$\mathcal{D}h = 0 \implies Gh = \mathcal{D}^H \mathcal{D}h = 0.$$

Subtracting both sides with  $h$  we get

$$Gh - h = (G - I)h = -h. \quad (6.2.31)$$

Let  $F$  denote an  $K \times D$  matrix formed from the rows of  $G - I$  corresponding to the indices in  $\Lambda$ . Then

$$(G - I)h = -h \implies \|Fh\|_1 = \sum_{k \in \Lambda} |h_k|.$$

Basically  $h_k$  for some  $k \in \Lambda$  is the inner product of some row in  $F$  with  $h$ .

We know that

$$\|Fh\|_1 \leq \|F\|_1 \|h\|_1$$

where  $\|F\|_1$  is the max-column-sum norm of  $F$ . This gives us

$$\|F\|_1 \|h\|_1 \geq \sum_{k \in \Lambda} |h_k|.$$

In any column of  $F$  the number of entries is  $K$ . Actually one of them is 0 (corresponding to the diagonal entry in  $G$ ). Thus, leaving it the rest of the entries are  $K - 1$ . By assumption  $\mu_{1/2}(G) > K$ . Thus any set of

entries in a column which is less than  $K$  cannot have a sum exceeding  $\frac{1}{2}$ . This gives an upper bound on the max-column-sum of  $F$ . i.e.

$$\|F\|_1 < \frac{1}{2}.$$

Thus, we get

$$\sum_{k \in \Lambda} |h_k| \leq \|F\|_1 \|h\|_1 < \frac{1}{2} \|h\|_1$$

for every  $h \in \mathcal{N}(\mathcal{D})$ .

The rest follows from the fact that for any other  $\alpha'$  such that  $x = \mathcal{D}\alpha' = \mathcal{D}\alpha$ , we know that

$$\|\alpha'\|_1 > \|\alpha\|_1$$

whenever

$$\sum_{k \in \Lambda} |h_k| < \frac{1}{2} \|h\|_1$$

where  $h = \alpha - \alpha'$  (thus  $\mathcal{D}h = 0$ ).  $\square$

### 6.3. Stability of sparsest solution

In this section we discuss various results related to the stability of the sparsest solution for the  $(P_0^\epsilon)$  problem.

NOTE: some of this content may be moved to different chapters on further review.

For convenience, we restate the problem. We represent the signal  $x \in \mathbb{C}^N$  as  $x = \mathcal{D}\alpha + e$  where  $\alpha$  is a sparse approximation of  $x$  in  $\mathcal{D}$ .

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_0 \text{ subject to } \|x - \mathcal{D}\alpha\|_2 \leq \epsilon. \quad (P_0^\epsilon)$$

Since we only know  $x$  and  $\mathcal{D}$  (both  $\alpha$  and  $e$  are unknown to us), hence in general it is not possible to recover  $\alpha$  exactly. The term  $d = \alpha - \hat{\alpha}$  will represent the recovery error (note that it is different from the representation error  $e$ ). It is important for us to ensure that the solution of the recovery problem is stable i.e. in the presence of small approximation error  $\|e\|_2$ , the recovery error  $\|d\|_2$  should also be small. For the sparse

recovery problem  $(P_0^\epsilon)$ , we cannot provide a uniqueness guarantee as such. Still, we can identify criteria which ensure that the solution remains stable when the approximation error is bounded. Our analysis in this section will focus on identifying criteria which ensures this.

We start with generalizing the notion of spark for the noisy case. Suppose  $\alpha$  and  $\beta$  are two solutions of  $(P_0^\epsilon)$ . Then  $\|\mathcal{D}\alpha - x\|_2 \leq \epsilon$  as well as  $\|\mathcal{D}\beta - x\|_2 \leq \epsilon$ . Thus, both  $\mathcal{D}\alpha$  and  $\mathcal{D}\beta$  lie in a ball of radius  $\epsilon$  around  $x$ . Thus, the maximum distance between  $\mathcal{D}\alpha$  and  $\mathcal{D}\beta$  can be  $2\epsilon$ . Alternatively, using triangle inequality we have

$$\begin{aligned} \|\mathcal{D}(\alpha - \beta)\|_2 &= \|\mathcal{D}\alpha - x + x - \mathcal{D}\beta\|_2 \\ &\leq \|\mathcal{D}\alpha - x\|_2 + \|x - \mathcal{D}\beta\|_2 \leq 2\epsilon. \end{aligned} \tag{6.3.1}$$

If we define  $d = \alpha - \beta$ , then

$$\|\mathcal{D}d\|_2 \leq 2\epsilon. \tag{6.3.2}$$

### 6.3.1. $\text{spark}_\eta$

**Definition 6.1** Let  $A \in \mathbb{C}^{N \times D}$  be some matrix. Consider all possible sub-sets of  $K$  columns. Let each such set form sub-matrix  $A_\Lambda \in \mathbb{C}^{N \times K}$  where  $\Lambda$  denotes the index set of  $K$  indices chosen. We define  $\text{spark}_\eta(A)$  as the smallest possible  $K$  (number of columns) that guarantees

$$\min_{\Lambda} \sigma_K(A_\Lambda) \leq \eta \tag{6.3.3}$$

where  $\sigma_K$  denotes the smallest singular value (i.e.  $K$ -th singular value) of the sub-matrix  $A_\Lambda$ . Note that we are minimizing over all possible index sets  $\Lambda$  with  $|\Lambda| = K$ .

In words, this is the smallest number of columns (indexed by  $\Lambda$ ) that can be gathered from  $A$  such that the smallest singular value of  $A_\Lambda$  is no larger than  $\eta$ . i.e. there exists a sub-matrix of  $A$  consisting of  $\text{spark}_\eta(A)$  columns whose smallest singular value is  $\eta$  or below. At the same time, all submatrices of  $A$  with number of columns less than  $\text{spark}_\eta(A)$  have the smallest singular value larger than  $\eta$ .

When the smallest singular value is 0, then the columns are linearly dependent. Thus, by choosing  $\eta = 0$ , we get the smallest number of columns  $K$  which are linearly dependent. This matches with the definition of spark. Thus,

$$\text{spark}_0(A) = \text{spark}(A).$$

Since singular values are always non-negative, hence  $\eta \geq 0$ . When columns of  $A$  are unit-norm (the case of dictionaries), then any single column sub-matrix has a singular value of 1. Hence,

$$\text{spark}_1(A) = 1.$$

Choosing a value of  $\eta > 1$  doesn't make any difference since with a single column sub-matrix, we can show that

$$\text{spark}_\eta(A) = 1 \quad \forall \eta \geq 1.$$

Let  $\eta_1 > \eta_2$ . Let

$$K_2 = \text{spark}_{\eta_2}(A).$$

Thus, there exists a sub-matrix consisting of  $K_2$  columns of  $A$  whose smallest singular value is upper bounded by  $\eta_2$ . Since  $\eta_1 > \eta_2$ ,  $\eta_1$  also serves as an upper bound for the smallest singular value for this sub-matrix. Clearly then  $K_1 = \text{spark}_{\eta_1}(A) \leq K_2$ . Thus, we note that  $\text{spark}_\eta$  is a monotone decreasing function of  $\eta$ . i.e.

$$\text{spark}_{\eta_1}(A) \leq \text{spark}_{\eta_2}(A), \text{ whenever } \eta_1 > \eta_2.$$

Further, we recall that the spark of  $A$  is upper bounded by its rank plus one. Assuming  $A$  to be a full rank matrix, we get following inequality:

$$1 \leq \text{spark}_\eta(A) \leq \text{spark}_0(A) = \text{spark}(A) \leq N + 1 \quad \forall 0 \leq \eta \leq 1. \quad (6.3.4)$$

We recall that if  $Av = 0$  then  $\|v\|_0 \geq \text{spark}(A)$ . A similar property can be developed for  $\text{spark}_\eta(A)$  also.

**Theorem 6.4** *If  $\|Av\|_2 \leq \eta$  and  $\|v\|_2 = 1$ , then  $\|v\|_0 \geq \text{spark}_\eta(A)$ .*

PROOF. For contradiction, let us assume that  $K = \|v\|_0 < \text{spark}_\eta(A)$ . Let  $\Lambda = \text{supp}(v)$ . Then  $Av = A_\Lambda v_\Lambda$ . Also  $\|v_\Lambda\|_2 = \|v\|_2 = 1$ .

We recall that the smallest singular value of  $A_\Lambda$  is given by

$$\sigma_{\min}(A_\Lambda) = \inf_{\|x\|_2=1} \|A_\Lambda x\|_2.$$

Thus,

$$\|A_\Lambda x\|_2 \geq \sigma_{\min}(A_\Lambda) \text{ whenever } \|x\|_2 = 1.$$

Thus, in our particular case

$$\|A_\Lambda v_\Lambda\|_2 \geq \sigma_{\min}(A_\Lambda).$$

$A_\Lambda$  has  $K$  columns with  $K < \text{spark}_\eta(A)$ .

Thus, from the definition of  $\text{spark}_\eta(A)$

$$\sigma_{\min}(A_\Lambda) > \eta.$$

This gives us

$$\|Av\|_2 = \|A_\Lambda v_\Lambda\|_2 > \eta$$

which contradicts with the assumption that  $\|Av\|_2 \leq \eta$ .  $\square$

### 6.3.2. $\text{spark}_\eta$ and coherence

In the following, we will focus on the  $\text{spark}_\eta$  of a full rank dictionary  $\mathcal{D}$ . We now establish a connection between  $\text{spark}_\eta$  and coherence of a dictionary.

**Theorem 6.5** *Let  $\mathcal{D}$  be a full rank dictionary with coherence  $\mu$ .*

*Then*

$$\text{spark}_\eta(\mathcal{D}) \geq \frac{1 - \eta^2}{\mu} + 1. \quad (6.3.5)$$

PROOF. We recall from Gershgorin's theorem that for any square matrix  $A \in \mathbb{C}^{K \times K}$ , every eigen value  $\lambda$  of  $A$  satisfies

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \text{ for some } i \in \{1, \dots, K\}.$$

Now consider a matrix  $A$  with diagonal elements equal to 1 and off-diagonal elements bounded by a value  $\mu$ . Then

$$|\lambda - 1| \leq \sum_{j \neq i} |a_{ij}| \leq \sum_{j \neq i} \mu = (K - 1)\mu.$$

Thus,

$$-(K - 1)\mu \leq \lambda - 1 \leq (K - 1)\mu \iff 1 - (K - 1)\mu \leq \lambda \leq 1 + (K - 1)\mu$$

This gives us a lower bound on the smallest eigen value.

$$\lambda_{\min}(A) \geq 1 - (K - 1)\mu.$$

Now consider any index set  $\Lambda \subseteq \{1, \dots, D\}$  and consider the submatrix  $\mathcal{D}_\Lambda$  with  $|\Lambda| = \text{spark}_\eta(\mathcal{D}) = K$ . Define  $G = \mathcal{D}_\Lambda^H \mathcal{D}_\Lambda$ . The diagonal elements of  $G$  are one, while off-diagonal elements are bounded by  $\mu$ . Thus,

$$\begin{aligned} \lambda_{\min}(G) \geq 1 - (K - 1)\mu &\iff (K - 1)\mu \geq 1 - \lambda_{\min}(G) \\ &\iff K - 1 \geq \frac{1 - \lambda_{\min}(G)}{\mu} \\ &\iff K \geq \frac{1 - \lambda_{\min}(G)}{\mu} + 1. \end{aligned} \tag{6.3.6}$$

Since this applies to every sub-matrix  $\mathcal{D}_\Lambda$ , this in particular applies to the sub-matrix for which  $\sigma_{\min}(\mathcal{D}_\Lambda) \leq \eta$  holds. For this sub-matrix

$$\lambda_{\min}(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda) = \sigma_{\min}^2(\mathcal{D}_\Lambda) \leq \eta^2.$$

Thus

$$K = \text{spark}_\eta(\mathcal{D}) \geq \frac{1 - \lambda_{\min}(G)}{\mu} + 1 \geq \frac{1 - \eta^2}{\mu} + 1. \tag{6.3.7}$$

□

### 6.3.3. Uncertainty with $\text{spark}_\eta$

We now present an uncertainty result for the noisy case.

**Theorem 6.6** If  $\alpha_1$  and  $\alpha_2$  satisfy  $\|x - \mathcal{D}\alpha_i\|_2 \leq \epsilon, i = 1, 2$ , then

$$\|\alpha_1\|_0 + \|\alpha_2\|_0 \geq \text{spark}_\eta(\mathcal{D}), \text{ where } \eta = \frac{2\epsilon}{\|\alpha_1 - \alpha_2\|_2}. \quad (6.3.8)$$

PROOF. From triangle inequality we have

$$\|\mathcal{D}(\alpha_1 - \alpha_2)\|_2 \leq 2\epsilon.$$

We define  $\beta = \alpha_1 - \alpha_2$ . Then  $\|\mathcal{D}\beta\|_2 \leq 2\epsilon$ . Further define  $v = \beta/\|\beta\|_2$  as the normalized vector. Then

$$\|\mathcal{D}v\|_2 = \frac{\|\mathcal{D}\beta\|_2}{\|\beta\|_2} \leq \frac{2\epsilon}{\|\beta\|_2}.$$

Now define

$$\eta = \frac{2\epsilon}{\|\beta\|_2} = \frac{2\epsilon}{\|\alpha_1 - \alpha_2\|_2}.$$

Then from theorem 6.4 if  $\|\mathcal{D}v\|_2 \leq \eta$  with  $\|v\|_2 = 1$ , then  $\|v\|_0 \geq \text{spark}_\eta(\mathcal{D})$ . Finally,

$$\|\alpha_1\|_0 + \|\alpha_2\|_0 \geq \|\alpha_1 - \alpha_2\|_0 = \|\beta\|_0 = \|v\|_0 \geq \text{spark}_\eta(\mathcal{D}). \quad (6.3.9)$$

This concludes the proof.  $\square$

This result gives us a lower bound on the sum of sparsity levels of two different sparse representations of same vector  $x$  under the given bound approximation error.

#### 6.3.4. Localization of sparse representations

We can now develop a localization result for the sparse approximation up to a Euclidean ball. This is analogous to the uniqueness result in noiseless case.

**Theorem 6.7** Given a distance  $\delta \geq 0$  (bound on distance between two sparse representations) and  $\epsilon$  (bound on norm of approximation error), set  $\eta = 2\epsilon/\delta$ . Suppose there are two approximate

representations  $\alpha_i, i = 1, 2$  both obeying

$$\|x - \mathcal{D}\alpha_i\|_2 \leq \epsilon \text{ and } \|\alpha_i\|_0 \leq \frac{1}{2}\text{spark}_\eta(\mathcal{D}). \quad (6.3.10)$$

Then  $\|\alpha_1 - \alpha_2\|_2 \leq \delta$ .

PROOF. Since  $\|\alpha_i\|_0 \leq \frac{1}{2}\text{spark}_\eta(\mathcal{D})$ , hence

$$\|\alpha_1\|_0 + \|\alpha_2\|_0 \leq \text{spark}_\eta(\mathcal{D}).$$

From theorem 6.6, if we define

$$\nu = \frac{2\epsilon}{\|\alpha_1 - \alpha_2\|_2},$$

then

$$\|\alpha_1\|_0 + \|\alpha_2\|_0 \geq \text{spark}_\nu(\mathcal{D}).$$

Combining the two, we get

$$\text{spark}_\eta(\mathcal{D}) \geq \|\alpha_1\|_0 + \|\alpha_2\|_0 \geq \text{spark}_\nu(\mathcal{D}). \quad (6.3.11)$$

Because of the monotonicity of  $\text{spark}_\eta(\mathcal{D})$ , we have

$$\begin{aligned} \text{spark}_\eta(\mathcal{D}) \geq \text{spark}_\nu(\mathcal{D}) &\implies \eta \leq \nu \\ &\implies \frac{2\epsilon}{\delta} \leq \frac{2\epsilon}{\|\alpha_1 - \alpha_2\|_2} \\ &\implies \delta \geq \|\alpha_1 - \alpha_2\|_2 \end{aligned} \quad (6.3.12)$$

which completes our proof.  $\square$

This theorem says that if  $x$  has two different sufficiently sparse representations  $\alpha_i$  with small approximation errors, they fall within a small distance.

### 6.3.5. Stability of sparsest solution using coherence

We can now develop a stability result for the  $(P_0^e)$  problem in terms of coherence of the dictionary.

**Theorem 6.8** Consider an instance of the  $(P_0^\epsilon)$  problem defined by the triplet  $(\mathcal{D}, x, \epsilon)$ . Suppose that a sparse vector  $\alpha \in \mathbb{C}^D$  satisfies the sparsity constraint

$$\|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right)$$

and gives a representation of  $x$  to within error tolerance  $\epsilon$  (i.e.  $\|x - \mathcal{D}\alpha\|_2 \leq \epsilon$ ). Every solution  $\hat{\alpha}$  of  $(P_0^\epsilon)$  must obey

$$\|\hat{\alpha} - \alpha\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(2\|\alpha\|_0 - 1)}. \quad (6.3.13)$$

**PROOF.** Note that  $\alpha$  need not be sparsest possible representation of  $x$  within the approximation error  $\epsilon$ . But  $\alpha$  is a feasible solution of  $(P_0^\epsilon)$ . Now since  $\hat{\alpha}$  is an optimal solution of  $(P_0^\epsilon)$  (thus sparsest possible), hence it is at least as sparse as  $\alpha$  i.e.

$$\|\hat{\alpha}\|_0 \leq \|\alpha\|_0.$$

We recall that

$$\frac{1}{2} \text{spark}(\mathcal{D}) \geq \frac{1}{2} \left( 1 + \frac{1}{\mu} \right) > \|\alpha\|_0.$$

Thus, there exists a value  $\eta \geq 0$  such that

$$\frac{1}{2} \text{spark}(\mathcal{D}) \geq \frac{1}{2} \text{spark}_\eta(\mathcal{D}) \geq \|\alpha\|_0 \geq \|\hat{\alpha}\|_0.$$

From theorem 6.5 we recall that

$$\text{spark}_\eta(\mathcal{D}) \geq \frac{1 - \eta^2}{\mu} + 1. \quad (6.3.14)$$

Thus, we can find a suitable value of  $\eta \geq 0$  such that we can enforce a more stricter requirement:

$$\|\alpha\|_0 \leq \frac{1}{2} \left( \frac{1 - \eta^2}{\mu} + 1 \right) \leq \frac{1}{2} \text{spark}_\eta(\mathcal{D}). \quad (6.3.15)$$

From this we can develop an upper bound on  $\eta$  being

$$\begin{aligned} \|\alpha\|_0 \leq \frac{1}{2} \left( \frac{1 - \eta^2}{\mu} + 1 \right) &\iff 2\|\alpha\|_0\mu \leq 1 - \eta^2 + \mu \\ &\iff \eta^2 \leq 1 - \mu(2\|\alpha\|_0 - 1). \end{aligned} \quad (6.3.16)$$

If we choose  $\eta^2 = 1 - \mu(2\|\alpha\|_0 - 1)$ , then

$$\begin{aligned} \|\alpha\|_0 &= \frac{1}{2} \left( \frac{1-\eta^2}{\mu} + 1 \right) \implies \|\alpha\|_0 \leq \frac{1}{2} \text{spark}_\eta(\mathcal{D}) \\ &\implies \|\hat{\alpha}\|_0 \leq \|\alpha\|_0 \leq \frac{1}{2} \text{spark}_\eta(\mathcal{D}) \end{aligned} \quad (6.3.17)$$

continues to hold.

We have two solutions  $\alpha$  and  $\hat{\alpha}$  both of which satisfy

$$\|\alpha\|_0, \|\hat{\alpha}\|_0 \leq \frac{1}{2} \text{spark}_\eta(\mathcal{D})$$

and

$$\|x - \mathcal{D}\alpha\|_2, \|x - \mathcal{D}\hat{\alpha}\|_2 \leq \epsilon.$$

If we choose a  $\delta = \frac{2\epsilon}{\eta}$ , then applying theorem 6.7, we will get

$$\|\alpha - \hat{\alpha}\|_2^2 \leq \delta^2 = \frac{4\epsilon^2}{\eta^2} = \frac{4\epsilon^2}{1 - \mu(2\|\alpha\|_0 - 1)}. \quad (6.3.18)$$

□

### 6.3.6. Stability of sparsest solution using RIP

**Theorem 6.9** Consider an instance of the  $(P_0^\epsilon)$  problem defined by the triplet  $(\mathcal{D}, x, \epsilon)$ . Let  $\mathcal{D}$  satisfy RIP of order  $2K$ . Suppose that a sparse vector  $\alpha \in \mathbb{C}^D$  with  $\|\alpha\|_0 = K$  is a feasible solution of  $(P_0^\epsilon)$ . Then, every solution  $\hat{\alpha}$  of  $(P_0^\epsilon)$  must obey

$$\|\hat{\alpha} - \alpha\|_2^2 \leq \frac{4\epsilon^2}{1 - \delta_{2K}}. \quad (6.3.19)$$

Further, if

$$\|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right)$$

then the following also holds:

$$\|\hat{\alpha} - \alpha\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(2\|\alpha\|_0 - 1)}. \quad (6.3.20)$$

PROOF. Let  $\hat{\alpha}$  be an alternative solution to  $(P_0^\epsilon)$ . Defining  $\beta = \hat{\alpha} - \alpha$ , as usual

$$\|\mathcal{D}\beta\|_2 \leq 2\epsilon.$$

Further

$$\|\beta\|_0 = \|\alpha - \hat{\alpha}\|_0 \leq \|\alpha\|_0 + \|\hat{\alpha}\|_0 \leq 2K$$

since  $\|\hat{\alpha}\|_0 \leq \|\alpha\|_0 = K$ .

Since  $\mathcal{D}$  satisfies RIP of order  $2K$ , hence

$$(1 - \delta_{2K})\|\beta\|_2^2 \leq \|\mathcal{D}\beta\|_2^2 \leq (1 + \delta_{2K})\|\beta\|_2^2.$$

This gives us

$$(1 - \delta_{2K})\|\beta\|_2^2 \leq 4\epsilon^2.$$

Rewriting we get

$$\|\beta\|_2^2 \leq \frac{4\epsilon^2}{1 - \delta_{2K}}$$

which is the desired result.

We **recall** that

$$\delta_{2K} \leq (2K - 1)\mu.$$

Thus,

$$1 - \delta_{2K} \geq 1 - (2K - 1)\mu \implies \frac{4\epsilon^2}{1 - \delta_{2K}} \leq \frac{4\epsilon^2}{1 - (2K - 1)\mu}.$$

This is useful only if the denominator is positive, i.e.

$$1 - (2K - 1)\mu > 0 \implies \frac{1}{\mu} > 2K - 1 \implies K < \frac{1}{2} \left(1 + \frac{1}{\mu}\right).$$

Under this condition, we get the result

$$\|\beta\|_2^2 \leq \frac{4\epsilon^2}{1 - (2K - 1)\mu}.$$

□

## 6.4. BPIC

In the section, we present a stability guarantee result for BPIC.

**Theorem 6.10** Consider an instance of the  $(P_1^c)$  problem defined by the triplet  $(\mathcal{D}, x, \epsilon)$ . Suppose that a vector  $\alpha \in \mathbb{C}^D$  is a feasible

solution to  $(P_1^\epsilon)$  satisfying the sparsity constraint

$$\|\alpha\|_0 < \frac{1}{4} \left( 1 + \frac{1}{\mu(\mathcal{D})} \right).$$

The solution  $\hat{\alpha}$  of  $(P_1^\epsilon)$  must satisfy

$$\|\hat{\alpha} - \alpha\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(\mathcal{D})(4\|\alpha\|_0 - 1)}. \quad (6.4.1)$$

PROOF. As usual, we define  $\beta = \hat{\alpha} - \alpha$ . Then

$$\|\mathcal{D}\beta\|_2 = \|\mathcal{D}(\hat{\alpha} - \alpha)\|_2 = \|\mathcal{D}\hat{\alpha} - x + x - \mathcal{D}\alpha\|_2 \leq 2\epsilon. \quad (6.4.2)$$

We now rewrite the inequality in terms of the Gram matrix  $G = \mathcal{D}^H \mathcal{D}$ .

$$\begin{aligned} 4\epsilon^2 &\geq \|\mathcal{D}\beta\|_2^2 = \beta^H G \beta \\ &= \beta^H (G - I + I) \beta \\ &= \|\beta\|_2^2 + \beta^H (G - I) \beta. \end{aligned} \quad (6.4.3)$$

It is easy to show that:

$$-\|\beta\|^T |A| \|\beta\| \leq \beta^H A \beta \leq \|\beta\|^T |A| \|\beta\|$$

whenever  $A$  is Hermitian. To see this just notice that  $\beta^H A \beta$  is a real quantity. Hence  $\beta^H A \beta = \pm |\beta^H A \beta|$ . Now, using triangle inequality we can easily show that  $|\beta^H A \beta| \leq \|\beta\|^T |A| \|\beta\|$ .

Since  $G - I$  is Hermitian, hence

$$\beta^H (G - I) \beta \geq -\|\beta\|^T |G - I| \|\beta\|.$$

Now

$$|\beta\|^T |G - I| \|\beta\| = \sum_{i,j} |\beta_i| |d_i^H d_j - \delta_{ij}| |\beta_j| \leq \mu(\mathcal{D}) \sum_{i,j, i \neq j} |\beta_i| |\beta_j| = \mu(\mathcal{D}) |\beta|^T (\mathbf{1} - I) |\beta|.$$

Only the off-diagonal terms of  $G$  remain in the sum, which are all dominated by  $\mu(\mathcal{D})$ .

Thus we get

$$\begin{aligned} 4\epsilon^2 &\geq \|\beta\|_2^2 - |\beta|^T(\mathbf{1} - I)|\beta| \\ &= (1 + \mu(\mathcal{D}))\|\beta\|_2^2 - \mu(\mathcal{D})|\beta|^T\mathbf{1}|\beta| \\ &= (1 + \mu(\mathcal{D}))\|\beta\|_2^2 - \mu(\mathcal{D})\|\beta\|_1^2. \end{aligned} \quad (6.4.4)$$

This is valid since  $v^H\mathbf{1}v = \|v\|_1^2$ .

Since  $\hat{\alpha}$  is optimal solution of  $(P_1^\epsilon)$ , hence

$$\|\hat{\alpha}\|_1 = \|\beta + \alpha\|_1 \leq \|\alpha\|_1 \implies \|\beta + \alpha\|_1 - \|\alpha\|_1 \leq 0. \quad (6.4.5)$$

Let  $\Lambda = \text{supp}(\alpha)$  and  $K = |\Lambda|$ . By a simple permutation of columns of  $\mathcal{D}$ , we can bring the entries in  $\alpha$  to the first  $K$  entries making  $\Lambda = \{1, \dots, K\}$ . We will make this assumption going forward without loss of generality. Let  $\mathbf{1}_K$  be corresponding support vector (of ones in first  $K$  places and 0 in rest). From our previous analysis, we recall that

$$\|\beta + \alpha\|_1 - \|\alpha\|_1 \geq \|\beta\|_1 - 2\mathbf{1}_K^T|\beta|.$$

Thus

$$\|\beta\|_1 - 2\mathbf{1}_K^T|\beta| \leq 0 \implies \|\beta\|_1 \leq 2\mathbf{1}_K^T|\beta|. \quad (6.4.6)$$

$\mathbf{1}_K^T|\beta|$  is the sum of first  $K$  terms of  $|\beta|$ . Considering  $\beta_\Lambda$  as a vector  $\in \mathbb{C}^K$  and using the  $l_1$ - $l_2$  norm relation  $\|v\|_1 \leq \sqrt{K}\|v\|_2 \forall v \in \mathbb{C}^N$ , we get

$$\mathbf{1}_K^T|\beta| = \|\beta_\Lambda\|_1 \leq \sqrt{K}\|\beta_\Lambda\|_2 \leq \sqrt{K}\|\beta\|_2. \quad (6.4.7)$$

Thus,

$$\|\beta\|_1 \leq 2\mathbf{1}_K^T|\beta| \leq 2\sqrt{K}\|\beta\|_2. \quad (6.4.8)$$

Putting this back in the previous inequality

$$\begin{aligned} 4\epsilon^2 &\geq (1 + \mu(\mathcal{D}))\|\beta\|_2^2 - \mu(\mathcal{D})\|\beta\|_1^2 \\ &\geq (1 + \mu(\mathcal{D}))\|\beta\|_2^2 - \mu(\mathcal{D})4K\|\beta\|_2^2 \\ &= (1 - (4K - 1)\mu(\mathcal{D}))\|\beta\|_2^2. \end{aligned} \quad (6.4.9)$$

We note that this inequality is valid only if

$$1 - (4K - 1)\mu(\mathcal{D}) > 0.$$

This condition can be reformulated as

$$\|\alpha\|_0 = K < \frac{1}{4} \left( 1 + \frac{1}{\mu(\mathcal{D})} \right).$$

Rewriting the bound on  $\|\beta\|_2^2$  we get

$$\|\beta\|_2^2 \leq \frac{4\epsilon^2}{(1 - (4K - 1)\mu(\mathcal{D}))} \quad (6.4.10)$$

which is the desired result.  $\square$

## 6.5. $l_1$ penalty problem

In this section we will examine the  $l_1$  penalty problem more closely.

Let us recall the approximation error minimization with  $l_1$  penalty problem.

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \frac{1}{2} \|x - \mathcal{D}\alpha\|_2^2 + \gamma \|\alpha\|_1. \quad (\text{P}_1^\gamma)$$

We will focus on following issues in this section:

- Some results from convex analysis useful for our study
- Conditions for the minimization of  $(\text{P}_1^\gamma)$  over coefficients  $\alpha$  supported on a subdictionary  $\mathcal{D}_\Lambda$
- Conditions under which the unique minimizer for a subdictionary is also the global minimizer for  $(\text{P}_1^\gamma)$
- Application of  $(\text{P}_1^\gamma)$  for sparse signal recovery
- Application of  $(\text{P}_1^\gamma)$  for identification of sparse signals in presence of noise
- Application of  $(\text{P}_1^\gamma)$  for identification of sparse signals in presence of Gaussian noise

### 6.5.1. Convex analysis

We recall some definitions and results from convex analysis which will help us understand the minimizers for  $(\text{P}_1^\gamma)$  problem.

Convex analysis for real valued functions over a complex vector space is developed using the bilinear inner product defined as

$$\langle x, y \rangle_B = \operatorname{Re}(y^H x). \quad (6.5.1)$$

We can easily notice that  $\langle x, y \rangle_B = \langle y, x \rangle_B$ ,  $\langle x, x \rangle_B \geq 0$ ,  $\langle x, x \rangle_B = 0 \iff x = 0$  and this inner product is linear in its first argument. Thus, it is a valid inner product. Further, the norm induced by this inner product is  $\sqrt{\operatorname{Re}(x^H x)} = \|x\|_2$  which is the usual  $l_2$  norm.

The subscript  $B$  is there to distinguish it from the usual inner product for the complex vector space  $\langle x, y \rangle = y^H x$ . The two inner products are related as

$$\langle x, y \rangle_B = \operatorname{Re}(\langle x, y \rangle).$$

We consider real valued functions over the inner product space  $\mathbb{X} = (\mathbb{C}^D, \langle \cdot, \cdot \rangle_B)$ . Note that the dimension of  $\mathbb{X}$  is  $D$  which is equal to the number of atoms in the dictionary  $\mathcal{D}$ . A convex function  $f : \mathbb{X} \rightarrow \mathbb{R}$  satisfies usual definitions of

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \quad \forall 0 \leq \theta \leq 1.$$

The objective function for the problem  $(P_1^\gamma)$  is

$$L(\alpha) = \frac{1}{2} \|x - \mathcal{D}\alpha\|_2^2 + \gamma \|\alpha\|_1 \quad (6.5.2)$$

Clearly,  $L$  is a real valued function over  $\mathbb{X}$  and it is easy to see that it is a convex function. Moreover  $L(\alpha) \geq 0$  always.

For any function  $f : \mathbb{X} \rightarrow \mathbb{R}$ , its **subdifferential set** is defined as

$$\partial f(x) \triangleq \{g \in \mathbb{X} : f(y) \geq f(x) + \langle y - x, g \rangle_B \quad \forall y \in \mathbb{X}\}. \quad (6.5.3)$$

The elements of subdifferential set are called **subgradients**. If  $f$  possesses a gradient at  $x$ , then it is the unique subgradient at  $x$ . i.e.

$$\partial f(x) = \{\nabla f(x)\}$$

where  $\nabla f(x)$  is the gradient of  $f$  at  $x$ .

The subdifferential of a sum is the (Minkowski) sum of the subdifferentials. i.e.

$$\partial(f(x) + g(x)) = \partial f(x) + \partial g(x) = \{h_1 + h_2 \mid h_1 \in \partial f(x), h_2 \in \partial g(x)\}.$$

If  $f$  is a closed, proper convex function, then  $x$  is a global minimizer of  $f$  if and only if  $0 \in \partial f(x)$ .

We would be specifically interested in the subdifferential for the function  $\|\alpha\|_1$ .

**Theorem 6.11** *Let  $z \in \mathbb{X}$ . The vector  $g \in \mathbb{X}$  lies in the subdifferential  $\partial\|z\|_1$  if and only if*

- (a)  $|g_k| \leq 1$  whenever  $z_k = 0$ .
- (a)  $g_k = \text{sgn}(z_k)$  whenever  $z_k \neq 0$ .

We recall that the signum function for complex numbers is defined as

$$\text{sgn}(re^{j\theta}) = \begin{cases} e^{j\theta} & \text{if } r > 0; \\ 0 & \text{if } r = 0. \end{cases} \quad (6.5.4)$$

The proof is skipped.

REMARK.  $\|g\|_\infty = 1$  whenever  $z \neq 0$ .  $\|g\|_\infty \leq 1$  when  $z = 0$ .

### 6.5.2. Restricted minimizers

Suppose  $\Lambda$  index a sub-dictionary  $\mathcal{D}_\Lambda$ . Since  $\mathcal{D}_\Lambda$  is a linearly independent collection of atoms, hence a unique  $l_2$  best approximation  $\hat{x}_\Lambda$  of  $x$  using the atoms in  $\mathcal{D}_\Lambda$  can be obtained using the least square techniques. We define the orthogonal projection operator

$$P_\Lambda = \mathcal{D}_\Lambda \mathcal{D}_\Lambda^\dagger.$$

And we get

$$\hat{x}_\Lambda = P_\Lambda x.$$

Obviously the approximation is orthogonal to the residual, i.e.  $x - \hat{x}_\Lambda \perp \hat{x}_\Lambda$ . There is a unique coefficient vector  $c_\Lambda$  supported on  $\Lambda$  that

synthesizes the approximation  $\hat{x}_\Lambda$ .

$$c_\Lambda = \mathcal{D}_\Lambda^\dagger x = \mathcal{D}_\Lambda^\dagger \hat{x}_\Lambda.$$

We also have

$$\hat{x}_\Lambda = P_\Lambda x = \mathcal{D}_\Lambda c_\Lambda.$$

**Theorem 6.12** Let  $\Lambda$  index a linearly independent collection of atoms in  $\mathcal{D}$  and let  $\alpha_*$  minimize the objective function  $L(\alpha)$  over all coefficient vectors supported on  $\Lambda$  (i.e.  $\text{supp}(\alpha) \subseteq \Lambda$ ). A necessary and sufficient condition on such a minimizer is that

$$c_\Lambda - \alpha_* = \gamma(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} g \quad (6.5.5)$$

where the vector  $g$  is drawn from  $\partial\|\alpha_*\|_1$ . Moreover, the minimizer  $\alpha_*$  is unique.

PROOF. Since we are restricted  $\alpha$  to be supported on  $\Lambda$  (i.e.  $\alpha \in \mathbb{C}^\Lambda$ ), hence

$$\mathcal{D}\alpha = \mathcal{D}_\Lambda \alpha_\Lambda.$$

Now, both  $\hat{x}_\Lambda$  and  $\mathcal{D}_\Lambda \alpha_\Lambda$  belong to the column space of  $\mathcal{D}_\Lambda$  while  $x - \hat{x}_\Lambda$  is orthogonal to it, hence

$$x - \hat{x}_\Lambda \perp \hat{x}_\Lambda - \mathcal{D}\alpha.$$

Thus, using the Pythagorean theorem, we get

$$\|x - \mathcal{D}\alpha\|_2^2 = \|x - \hat{x}_\Lambda + \hat{x}_\Lambda - \mathcal{D}\alpha\|_2^2 = \|x - \hat{x}_\Lambda\|_2^2 + \|\hat{x}_\Lambda - \mathcal{D}\alpha\|_2^2.$$

We can rewrite  $L(\alpha)$  as

$$L(\alpha) = \frac{1}{2} \|x - \hat{x}_\Lambda\|_2^2 + \frac{1}{2} \|\hat{x}_\Lambda - \mathcal{D}\alpha\|_2^2 + \gamma \|\alpha\|_1.$$

Define

$$F(\alpha) = \frac{1}{2} \|\hat{x}_\Lambda - \mathcal{D}\alpha\|_2^2 + \gamma \|\alpha\|_1.$$

Then

$$L(\alpha) = \frac{1}{2} \|x - \hat{x}_\Lambda\|_2^2 + F(\alpha).$$

Note that the term  $\|x - \hat{x}_\Lambda\|_2^2$  is constant. Thus, minimizing  $L(\alpha)$  over the coefficient vectors supported on  $\Lambda$  is equivalent to minimizing  $F(\alpha)$  over the same support set. Note that

$$\mathcal{D}\alpha = \mathcal{D}_\Lambda\alpha_\Lambda \text{ and } \|\alpha\|_1 = \|\alpha_\Lambda\|_1.$$

We can write  $F(\alpha)$  as

$$F(\alpha) = \frac{1}{2}\|\hat{x}_\Lambda - \mathcal{D}_\Lambda\alpha_\Lambda\|_2^2 + \gamma\|\alpha_\Lambda\|_1.$$

Note that  $F(\alpha)$  depends only on entries in  $\alpha$  which are part of the support  $\Lambda$ . We can replace the variable  $\alpha_\Lambda$  with  $\alpha \in \mathbb{C}^\Lambda$  and rewrite  $F(\alpha)$  as

$$F(\alpha) = \frac{1}{2}\|\hat{x}_\Lambda - \mathcal{D}_\Lambda\alpha\|_2^2 + \gamma\|\alpha\|_1 \quad \forall \alpha \in \mathbb{C}^\Lambda.$$

Since atoms indexed by  $\Lambda$  are linearly independent, hence  $\mathcal{D}_\Lambda$  has full column rank. Thus, the quadratic term  $\|\hat{x}_\Lambda - \mathcal{D}_\Lambda\alpha\|_2^2$  is strictly convex. Since  $\|\alpha\|_1$  is also convex,  $F(\alpha)$  therefore is strictly convex and its minimizer is unique.

Since  $F$  is strictly convex and unconstrained, hence  $0 \in \partial F(\alpha_*)$  is a necessary and sufficient condition for the coefficient vector  $\alpha_*$  to minimize  $F(\alpha)$ .

The gradient of the first (quadratic) term is

$$(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda) \alpha - \mathcal{D}_\Lambda^H \hat{x}_\Lambda.$$

For the second term we have to consider its subdifferential  $\delta\|\alpha\|$ . Thus, at  $\alpha_*$  it follows that

$$(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda) \alpha_* - \mathcal{D}_\Lambda^H \hat{x}_\Lambda + \gamma g = 0$$

where  $g$  is some subgradient in  $\partial\|\alpha_*\|$ . Premultiplying with  $(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1}$  we get

$$\begin{aligned} \alpha_* - \mathcal{D}_\Lambda^\dagger \hat{x}_\Lambda + \gamma (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} g &= 0 \\ \implies \mathcal{D}_\Lambda^\dagger \hat{x}_\Lambda - \alpha_* &= \gamma (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} g. \end{aligned}$$

Finally, we recall that  $\mathcal{D}_\Lambda^\dagger \hat{x}_\Lambda = c_\Lambda$ . Thus, we get the desired result

$$c_\Lambda - \alpha_* = \gamma (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} g.$$

□

Some bounds follow as a result of this theorem.

**Theorem 6.13** Suppose that  $\Lambda$  index a subdictionary  $\mathcal{D}_\Lambda$  and let  $\alpha_*$  minimize the function  $(L)$  over all coefficient vectors supported on  $\Lambda$ . Then following bounds are in force:

$$\|c_\Lambda - \alpha_*\|_\infty \leq \gamma \|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1}\|_\infty. \quad (6.5.6)$$

$$\|\mathcal{D}_\Lambda(c_\Lambda - \alpha_*)\|_2 \leq \gamma \|\mathcal{D}_\Lambda^\dagger\|_{2 \rightarrow 1}. \quad (6.5.7)$$

PROOF. Starting with

$$c_\Lambda - \alpha_* = \gamma (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} g \quad (6.5.8)$$

we take the  $l_\infty$  norm on both sides and apply some norm bounds

$$\begin{aligned} \|c_\Lambda - \alpha_*\|_\infty &= \|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} g\|_\infty \\ &\leq \gamma \|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1}\|_\infty \|g\|_\infty \\ &\leq \gamma \|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1}\|_\infty. \end{aligned}$$

The last inequality is valid since from theorem 6.11 we have:  $\|g\|_\infty \leq 1$ .

Now let us multiply (6.5.8) with  $\mathcal{D}_\Lambda$  and apply  $l_2$  norm

$$\begin{aligned} \|\mathcal{D}_\Lambda(c_\Lambda - \alpha_*)\|_2 &= \|\gamma D \mathcal{D}_\Lambda (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} g\|_2 = \gamma \|(\mathcal{D}_\Lambda^\dagger)^H g\|_2 \\ &\leq \|(\mathcal{D}_\Lambda^\dagger)^H\|_{\infty \rightarrow 2} \|g\|_\infty \\ &= \|\mathcal{D}_\Lambda^\dagger\|_{2 \rightarrow 1} \|g\|_\infty \\ &\leq \|\mathcal{D}_\Lambda^\dagger\|_{2 \rightarrow 1}. \end{aligned}$$

In this derivation we used facts like  $\|A\|_{p \rightarrow q} = \|A^H\|_{q' \rightarrow p'}$ ,  $\|Ax\|_q \leq \|A\|_{p \rightarrow q} \|x\|_p$  and  $\|g\|_\infty \leq 1$ .

□

### 6.5.3. The correlation condition

So far we have established a condition which ensures that  $\alpha_*$  is a unique minimizer of  $L$  given that  $\alpha$  is supported on  $\Lambda$ . We now establish a sufficient condition under which  $\alpha_*$  is also a global minimizer for  $L(\alpha)$ .

**Theorem 6.14** [Correlation condition] *Assume that  $\Lambda$  indexes a subdictionary. Let  $\alpha_*$  minimize the function  $(L)$  over all coefficient vectors supported on  $\Lambda$ . Suppose that*

$$\|\mathcal{D}^H(x - \hat{x}_\Lambda)\|_\infty \leq \gamma \left[ 1 - \max_{\omega \notin \Lambda} |\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle| \right] \quad (6.5.9)$$

where  $g \in \partial \|\alpha_*\|_1$  is determined by (6.5.5). Then  $\alpha_*$  is the unique global minimizer of  $(L)$ .

Moreover, the condition

$$\|\mathcal{D}^H(x - \hat{x}_\Lambda)\|_\infty \leq \gamma \left[ 1 - \max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^\dagger d_\omega\|_1 \right] \quad (6.5.10)$$

guarantees that  $\alpha_*$  is the unique global minimizer of  $(L)$ .

**PROOF.** Let  $\alpha_*$  be the unique minimizer of  $(L)$  over coefficient vectors supported on  $\Lambda$ . Then, the value of the objective function  $L(\alpha)$  increases if we change any coordinate of  $\alpha_*$  indexed in  $\Lambda$ .

What we need is a condition which ensures that the value of objective function also increases if we change any other component of  $\alpha_*$  (not indexed by  $\Lambda$ ). If this happens, then  $\alpha_*$  will become a local minimizer of  $(L)$ . Further, since  $(L)$  is convex,  $\alpha_*$  will also be global minimizer.

Towards this, let  $\omega$  be some index not in  $\Lambda$  and  $e_\omega \in \mathbb{C}^D$  be corresponding unit vector. Let  $\delta e_\omega$  be a small perturbation introduced in  $\omega$ -th coordinate. ( $\delta \in \mathbb{C}$  is a small scalar, though need not be positive real) We need find a condition which ensures

$$L(\alpha_* + \delta e_\omega) - L(\alpha_*) > 0 \quad \forall \omega \notin \Lambda.$$

Let us expand the L.H.S. of this inequality:

$$\begin{aligned} L(\alpha_* + \delta e_\omega) - L(\alpha_*) &= \\ &\left[ \frac{1}{2} \|x - \mathcal{D}\alpha_* - \delta d_\omega\|_2^2 - \frac{1}{2} \|x - \mathcal{D}\alpha_*\|_2^2 \right] \\ &+ \gamma [\|\alpha_* + \delta e_\omega\|_1 - \|\alpha_*\|_1]. \end{aligned}$$

Here we used the fact that  $\mathcal{D}e_\omega = d_\omega$ .

Note that since  $\alpha_*$  is supported on  $\Lambda$  and  $\omega \notin \Lambda$ , hence

$$\|\alpha_* + \delta e_\omega\|_1 = \|\alpha_*\|_1 + \|\delta e_\omega\|_1.$$

Thus

$$\|\alpha_* + \delta e_\omega\|_1 - \|\alpha_*\|_1 = |\delta|.$$

We should also simplify the first bracket.

$$\begin{aligned} \|x - \mathcal{D}\alpha_*\|_2^2 &= (x - \mathcal{D}\alpha_*)^H (x - \mathcal{D}\alpha_*) \\ &= x^H x + \alpha_*^H \mathcal{D}^H \mathcal{D}\alpha_* - x^H \mathcal{D}\alpha_* - \alpha_*^H \mathcal{D}^H x. \end{aligned}$$

Similarly

$$\begin{aligned} \|x - \mathcal{D}\alpha_* - \delta d_\omega\|_2^2 &= (x - \mathcal{D}\alpha_* - \delta d_\omega)^H (x - \mathcal{D}\alpha_* - \delta d_\omega) \\ &= x^H x + \alpha_*^H \mathcal{D}^H \mathcal{D}\alpha_* - x^H \mathcal{D}\alpha_* - \alpha_*^H \mathcal{D}^H x \\ &\quad - (x - \mathcal{D}\alpha_*)^H \delta d_\omega - \delta d_\omega^H (x - \mathcal{D}\alpha_*) + \|\delta d_\omega\|_2^2. \end{aligned}$$

Cancelling the like terms we get

$$\|\delta d_\omega\|_2^2 - 2 \operatorname{Re}(\langle x - \mathcal{D}\alpha_*, \delta d_\omega \rangle).$$

Thus,

$$\begin{aligned} L(\alpha_* + \delta e_\omega) - L(\alpha_*) &= \\ &\frac{1}{2} \|\delta d_\omega\|_2^2 - \operatorname{Re}(\langle x - \mathcal{D}\alpha_*, \delta d_\omega \rangle) + \gamma |\delta|. \end{aligned}$$

Recall that since  $\alpha_*$  is supported on  $\Lambda$ , hence  $\mathcal{D}\alpha_* = \mathcal{D}_\Lambda \alpha_*$ .

We can further split the middle term by adding and subtracting  $\mathcal{D}_\Lambda c_\Lambda$ .

$$\begin{aligned} \operatorname{Re}(\langle x - \mathcal{D}_\Lambda \alpha_*, \delta d_\omega \rangle) &= \operatorname{Re}(\langle x - \mathcal{D}_\Lambda c_\Lambda + \mathcal{D}_\Lambda c_\Lambda - \mathcal{D}_\Lambda \alpha_*, \delta d_\omega \rangle) \\ &= \operatorname{Re}(\langle x - \mathcal{D}_\Lambda c_\Lambda, \delta d_\omega \rangle) + \operatorname{Re}(\langle \mathcal{D}_\Lambda(c_\Lambda - \alpha_*), \delta d_\omega \rangle) \end{aligned}$$

Thus, we can write

$$L(\alpha_* + \delta e_\omega) - L(\alpha_*) = \frac{1}{2} \|\delta d_\omega\|_2^2 - \operatorname{Re}(\langle x - \mathcal{D}_\Lambda c_\Lambda, \delta d_\omega \rangle) - \operatorname{Re}(\langle \mathcal{D}_\Lambda(c_\Lambda - \alpha_*), \delta d_\omega \rangle) + \gamma |\delta|.$$

The term  $\frac{1}{2} \|\delta d_\omega\|_2^2$  is strictly positive giving us

$$L(\alpha_* + \delta e_\omega) - L(\alpha_*) > -\operatorname{Re}(\langle x - \mathcal{D}_\Lambda c_\Lambda, \delta d_\omega \rangle) - \operatorname{Re}(\langle \mathcal{D}_\Lambda(c_\Lambda - \alpha_*), \delta d_\omega \rangle) + \gamma |\delta|.$$

Using lower triangle inequality we can write

$$L(\alpha_* + \delta e_\omega) - L(\alpha_*) > \gamma |\delta| - |\langle x - \mathcal{D}_\Lambda c_\Lambda, \delta d_\omega \rangle| - |\langle \mathcal{D}_\Lambda(c_\Lambda - \alpha_*), \delta d_\omega \rangle|.$$

Using linearity of inner product, we can take out  $|\delta|$ :

$$L(\alpha_* + \delta e_\omega) - L(\alpha_*) > |\delta| [\gamma - |\langle x - \mathcal{D}_\Lambda c_\Lambda, d_\omega \rangle| - |\langle \mathcal{D}_\Lambda(c_\Lambda - \alpha_*), d_\omega \rangle|]. \quad (6.5.11)$$

Let us simplify this expression. Since  $\alpha_*$  is a unique minimizer over coefficients in  $\mathbb{C}^\Lambda$ , hence using theorem 6.12

$$\begin{aligned} c_\Lambda - \alpha_* &= \gamma (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} g \\ \iff \mathcal{D}_\Lambda(c_\Lambda - \alpha_*) &= \gamma \mathcal{D}_\Lambda (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} g = (\mathcal{D}_\Lambda^\dagger)^H g. \end{aligned}$$

where  $g \in \partial \|\alpha_*\|_1$ . Thus

$$|\langle \mathcal{D}_\Lambda(c_\Lambda - \alpha_*), d_\omega \rangle| = \gamma |\langle (\mathcal{D}_\Lambda^\dagger)^H g, d_\omega \rangle| = \gamma |\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle|$$

Using the fact that  $\langle Ax, y \rangle = \langle x, A^H y \rangle$ . Also, we recall that  $\widehat{x}_\Lambda = \mathcal{D}_\Lambda c_\Lambda$ .

Putting the back in (6.5.11) we obtain:

$$L(\alpha_* + \delta e_\omega) - L(\alpha_*) > |\delta| \left[ \gamma - \gamma |\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle| - |\langle x - \widehat{x}_\Lambda, d_\omega \rangle| \right]. \quad (6.5.12)$$

In (6.5.12), the L.H.S. is positive (our real goal) whenever the term in the bracket on the R.H.S. is non-negative (since  $|\delta|$  is positive). Therefore we want that

$$\gamma - \gamma |\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle| - |\langle x - \widehat{x}_\Lambda, d_\omega \rangle| \geq 0.$$

This can be rewritten as

$$|\langle x - \widehat{x}_\Lambda, d_\omega \rangle| \leq \gamma \left[ 1 - |\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle| \right].$$

Since this condition should hold for every  $\omega \notin \Lambda$ , hence we maximize the L.H.S. and minimize the R.H.S. over  $\omega \notin \Lambda$ . We get

$$\max_{\omega \notin \Lambda} |\langle x - \hat{x}_\Lambda, d_\omega \rangle| \leq \min_{\omega \notin \Lambda} \gamma \left[ 1 - |\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle| \right] = \gamma \left[ 1 - \max_{\omega \notin \Lambda} |\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle| \right].$$

Recall that  $x - \hat{x}_\Lambda$  is orthogonal to the space spanned by atoms in  $\mathcal{D}_\Lambda$ . Hence

$$\max_{\omega \notin \Lambda} |\langle x - \hat{x}_\Lambda, d_\omega \rangle| = \max_{\omega} |\langle x - \hat{x}_\Lambda, d_\omega \rangle| = \|\mathcal{D}^H(x - \hat{x}_\Lambda)\|_\infty.$$

This gives us the desired sufficient condition

$$\|\mathcal{D}^H(x - \hat{x}_\Lambda)\|_\infty \leq \gamma \left[ 1 - \max_{\omega \notin \Lambda} |\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle| \right].$$

This condition still uses  $g$ . We know that  $\|g\|_\infty \leq 1$ . Let us simplify as follows:

$$\begin{aligned} |\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle| &= |(\mathcal{D}_\Lambda^\dagger d_\omega)^H g| \\ &= \|(\mathcal{D}_\Lambda^\dagger d_\omega)^H g\|_\infty \\ &\leq \|(\mathcal{D}_\Lambda^\dagger d_\omega)^H\|_\infty \|g\|_\infty \\ &= \|(\mathcal{D}_\Lambda^\dagger d_\omega)\|_1 \|g\|_\infty \\ &\leq \|(\mathcal{D}_\Lambda^\dagger d_\omega)\|_1. \end{aligned}$$

Another way to understand this is as follows. For any vector  $v \in \mathbb{C}^D$

$$\begin{aligned} |\langle v, g \rangle| &= \left| \sum_{i=1}^D \overline{g_i} v_i \right| \\ &\leq \sum_{i=1}^D |g_i| |v_i| \\ &\leq [\sum_{i=1}^D |v_i|] \|g\|_\infty \\ &\leq \|v\|_1. \end{aligned}$$

Thus

$$|\langle \mathcal{D}_\Lambda^\dagger d_\omega, g \rangle| \leq \|\mathcal{D}_\Lambda^\dagger d_\omega\|_1.$$

Thus, it is also sufficient that

$$\|\mathcal{D}^H(x - \hat{x}_\Lambda)\|_\infty \leq \gamma \left[ 1 - \max_{\omega \notin \Lambda} \|(\mathcal{D}_\Lambda^\dagger d_\omega)\|_1 \right].$$

□

#### 6.5.4. Exact recovery coefficient

We recall that **Exact Recovery Coefficient** for a subdictionary is defined as

$$\text{ERC}(\Lambda) = 1 - \max_{\omega \notin \Lambda} \|\mathcal{D}_\Lambda^\dagger d_\omega\|_1.$$

Thus, the sufficient condition can be rewritten as

$$\|\mathcal{D}^H(x - \hat{x}_\Lambda)\|_\infty \leq \gamma \text{ERC}(\Lambda).$$

Note that the L.H.S. in both sufficient conditions is always non-negative. Hence, if the R.H.S. is negative (i.e.  $\text{ERC}(\Lambda) < 0$ ), the sufficient condition is useless.

On the other hand if  $\text{ERC}(\Lambda) > 0$ , then a sufficiently high  $\gamma$  can always be chosen to satisfy the condition in (6.5.10). At the same time as  $\gamma \rightarrow \infty$ , the optimum minimizer is  $\alpha_* = 0$ .

How do we interpret the L.H.S.  $\|\mathcal{D}^H(x - \hat{x}_\Lambda)\|_\infty$ ?

**Definition 6.2** Given a non-zero signal  $v$  and a dictionary  $\mathcal{D}$ , define the function

$$\text{maxcor}(v) \triangleq \frac{\max_{\omega \in \Omega} |\langle v, d_\omega \rangle|}{\|v\|_2}. \quad (6.5.13)$$

If  $v = 0$ , then define  $\text{maxcor}(v) = 0$ . This is known as the **maximum correlation** [37] of a signal with a dictionary.

Essentially, for any signal we normalize it and then find out its maximum inner product (absolute value) with atoms in the dictionary  $\mathcal{D}$ . Obviously  $0 \leq \text{maxcor}(v) \leq 1$ .

REMARK.

$$\|\mathcal{D}^H v\|_\infty = \text{maxcor}(v) \|v\|_2. \quad (6.5.14)$$

We can now interpret

$$\|\mathcal{D}^H(x - \hat{x}_\Lambda)\|_\infty = \text{maxcor}(x - \hat{x}_\Lambda) \|x - \hat{x}_\Lambda\|_2.$$

Therefore, the sufficient condition in theorem 6.14 is strongest when the magnitude of the residual  $(x - \hat{x}_\Lambda)$  and its maximum correlation with the dictionary are both small.

Since the maximum correlation of the residual never exceeds one, hence we obtain following (much weaker result)

**Corollary 6.15.** *Let  $\Lambda$  index a subdictionary and let  $x$  be an input signal. Suppose that the residual vector  $x - \hat{x}_\Lambda$  satisfies*

$$\|x - \hat{x}_\Lambda\|_2 \leq \gamma \text{ERC}(\Lambda).$$

*Then any coefficient vector  $\alpha_*$  that minimizes the function  $(L)$  must be supported inside  $\Lambda$ .*

### 6.5.5. Applications of $l_1$ penalization

Having setup the basic results in place, we can now study the applications of  $(P_1^\gamma)$ .

**Theorem 6.16** *Let  $\Lambda$  index a subdictionary  $\mathcal{D}_\Lambda$  for which  $\text{ERC}(\Lambda) \geq 0$ . Suppose that  $x$  is an input signal whose  $l_2$  best approximation over  $\Lambda$  satisfies the correlation condition*

$$\|\mathcal{D}_\Lambda^H(x - \hat{x}_\Lambda)\|_\infty \leq \gamma \text{ERC}(\Lambda).$$

*Let  $\alpha_*$  solve the convex program  $(P_1^\gamma)$  with parameter  $\gamma$ . We may conclude that:*

- (1) *Support of  $\alpha_*$  is contained in  $\Lambda$  and*
- (2) *The distance between  $\alpha_*$  and the optimal coefficient vector  $c_\Lambda$  satisfies*

$$\|\alpha_* - c_\Lambda\|_\infty \leq \gamma \left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \right\|_\infty.$$

(3) In particular,  $\text{supp}(\alpha_*)$  contains every index  $\lambda$  in  $\Lambda$  for which

$$|c_\Lambda(\lambda)| > \gamma \left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \right\|_\infty.$$

(4) Moreover, the minimizer  $\alpha_*$  is unique.

PROOF. Since the sufficient condition for correlation condition theorem 6.14 are satisfied, hence  $\alpha_*$  which minimizes  $(L)$  over coefficient vectors in  $\mathbb{C}^\Lambda$  is also a global minimizer of  $(L)$ . Since  $\alpha_* \in \mathbb{C}^\Lambda$ , hence  $\text{supp}(\alpha_*) \subseteq \Lambda$ .

For claim 2, application of theorem 6.13 gives us

$$\|c_\Lambda - \alpha_*\|_\infty \leq \gamma \|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1}\|_\infty.$$

Since the convex function  $(L)$  is strictly convex, hence  $\alpha_*$  is unique global minimizer.

For claim 3, suppose  $\alpha_*(\lambda) = 0$  for some index  $\lambda \in \Lambda$  for which

$$|c_\Lambda(\lambda)| > \gamma \left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \right\|_\infty.$$

Then

$$|\alpha_*(\lambda) - c_\Lambda(\lambda)| = |c_\Lambda(\lambda)| > \gamma \left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \right\|_\infty.$$

But

$$\|\alpha_* - c_\Lambda\|_\infty \geq |\alpha_*(\lambda) - c_\Lambda(\lambda)|.$$

This violates the bound that

$$\|\alpha_* - c_\Lambda\|_\infty \leq \gamma \left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \right\|_\infty.$$

Thus,  $\text{supp}(\alpha_*)$  contains every index  $\lambda \in \Lambda$  for which

$$|c_\Lambda(\lambda)| > \gamma \left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \right\|_\infty.$$

□

We can formulate a simpler condition in terms of coherence of the dictionary.

**Theorem 6.17** Suppose that  $K\mu \leq \frac{1}{2}$ . Assume that  $|\Lambda| \leq K$  i.e.  $\Lambda$  contains at most  $K$  indices. Suppose that  $x$  is an input signal whose  $l_2$  best approximation over  $\Lambda$  denoted by  $\hat{x}_\Lambda$  satisfies the correlation condition

$$\|\mathcal{D}_\Lambda^H(x - \hat{x}_\Lambda)\|_\infty \leq \gamma \frac{1 - (2K - 1)\mu}{1 - (K - 1)\mu}.$$

Let  $\alpha_*$  solve the convex program  $(P_1^\gamma)$  with parameter  $\gamma$ . We may conclude that:

- (1) Support of  $\alpha_*$  is contained in  $\Lambda$  and
- (2) The distance between  $\alpha_*$  and the optimal coefficient vector  $c_\Lambda$  satisfies

$$\|\alpha_* - c_\Lambda\|_\infty \leq \gamma \frac{1}{1 - (K - 1)\mu}.$$

- (3) In particular,  $\text{supp}(\alpha_*)$  contains every index  $\lambda$  in  $\Lambda$  for which

$$|c_\Lambda(\lambda)| > \gamma \frac{1}{1 - (K - 1)\mu}.$$

- (4) Moreover, the minimizer  $\alpha_*$  is unique.

PROOF. We recall the coherence bounds on ERC as

$$\text{ERC}(\Lambda) \geq \frac{1 - (2K - 1)\mu}{1 - (K - 1)\mu}.$$

Thus,

$$\|\mathcal{D}_\Lambda^H(x - \hat{x}_\Lambda)\|_\infty \leq \gamma \frac{1 - (2K - 1)\mu}{1 - (K - 1)\mu} \leq \gamma \text{ERC}(\Lambda).$$

A direct application of theorem 6.16 validates claims 1 and 4.

We recall the upper bound on norm of inverse Gram matrix of a sub-dictionary as

$$\|G^{-1}\|_\infty = \|G^{-1}\|_1 \leq \frac{1}{1 - \mu_1(K - 1)} \leq \frac{1}{1 - (K - 1)\mu}.$$

Putting this in theorem 6.16 validates claims 2 and 3.

□

### 6.5.6. Application: Identifying sparse signals

We now show how one can recover an exactly sparse signal solving the convex program  $(P_1^\gamma)$ .

**Theorem 6.18** *Assume that  $\Lambda$  indexes a subdictionary for which  $\text{ERC}(\Lambda) \geq 0$ . Choose an arbitrary coefficient vector  $c_{\text{opt}}$  supported on  $\Lambda$ . Fix an input signal  $x = \mathcal{D}c_{\text{opt}}$ . Let  $\alpha_*(\gamma)$  denote the unique minimizer of  $(P_1^\gamma)$  with parameter  $\gamma$ . We may conclude that*

- i. *There is a positive number  $\gamma_0$  for which  $\gamma < \gamma_0$  implies that  $\text{supp}(\alpha_*(\gamma)) = \Lambda$ .*
- ii. *In the limit as  $\gamma \rightarrow 0$ , we have  $\alpha_*(\gamma) \rightarrow c_{\text{opt}}$ .*

PROOF. Since there is no noise, hence the best  $l_2$  approximation of  $x$  over  $\Lambda$

$$\hat{x}_\Lambda = x$$

itself and the corresponding coefficient vector is

$$c_\Lambda = c_{\text{opt}}.$$

Therefore

$$\|\mathcal{D}_\Lambda^H(x - \hat{x}_\Lambda)\|_\infty = 0 \leq \gamma \text{ERC}(\Lambda).$$

Thus, the correlation condition is in force for every positive value of  $\gamma$ . Thus, as per theorem 6.16, minimizer  $\alpha_*(\gamma)$  of the convex program  $(P_1^\gamma)$  must be supported inside  $\Lambda$ . Moreover, we have

$$\|\alpha_*(\gamma) - c_{\text{opt}}\|_\infty \leq \gamma \left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \right\|_\infty.$$

Clearly, as  $\gamma \rightarrow 0$ , we have  $\alpha_*(\gamma) \rightarrow c_{\text{opt}}$ .

Finally, recall that  $\text{supp}(\alpha_*(\gamma))$  contains very index  $\lambda$  in  $\Lambda$  for which

$$|c_{\text{opt}}(\lambda)| > \gamma \left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \right\|_\infty.$$

In order for every index in  $\Lambda$  to be part of  $\text{supp}(\alpha_*(\gamma))$ , we require

$$\frac{\min_{\gamma \in \Gamma} |c_{\text{opt}}(\lambda)|}{\left\| (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \right\|_\infty} > \gamma.$$

Choosing the L.H.S. to be  $\gamma_0$ , we get an explicit value of upper bound on  $\gamma$  such that  $\gamma < \gamma_0$  leads to complete discovery of support.  $\square$

### 6.6. BPIC for compressed sensing

In this section we study the issue of stable signal recovery from incomplete and inaccurate measurements using basis pursuit with inequality constraints approach.

The results are based on [12].

We recall the formulation of compressed sensing as

$$y = \Phi x + e$$

where  $x \in \mathbb{C}^N$  is our desired signal and  $x$  is  $K$ -sparse,  $\Phi \in \mathbb{C}^{M \times N}$  is our sensing matrix with appropriate RIP constants,  $e \in \mathbb{C}^M$  is the measurement noise introduced during sensing process and  $y \in \mathbb{C}^M$  is the vector of compressed measurements with  $K < M \ll N$ . The measurement noise  $l_2$  norm has an upper bound given by  $\|e\|_2 \leq \epsilon$ . We can think of  $e$  as a perturbation to the sensing system.

Let us **recall** the problem of sparse recovery from CS measurements with bound on measurement noise.

$$\hat{x} = \arg \min_{x \in \mathbb{C}^N} \|x\|_0 \text{ subject to } \|y - \Phi x\|_2 \leq \epsilon \quad (\text{CS}_0^\epsilon)$$

A convex relaxation of this problem is

$$\hat{x} = \arg \min_{x \in \mathbb{C}^N} \|x\|_1 \text{ subject to } \|y - \Phi x\|_2 \leq \epsilon. \quad (\text{CS}_1^\epsilon)$$

For stable recovery, we need to connect the measurement error norm bound  $\epsilon$  with the recovery error norm  $\|\hat{x} - x\|_2$  and identify the conditions under which the recovery error norm  $\|\hat{x} - x\|_2$  is bounded by some

small multiple of the measurement error upper bound  $\epsilon$ . Such a result establishes that if the measurement error is small, then the recovery error will also remain small.

In this section

- We will develop the guarantee for recovery of sparse signals in terms of restricted isometry property of  $\Phi$ .
- We will then generalize the guarantee for arbitrary signals.
- We will then specialize the general guarantee for compressible signals.

First of all, let us focus our attention to sparse signals.

We will denote **recovery error vector** as

$$h = \hat{x} - x. \quad (6.6.1)$$

Note that  $h$  need not belong to the null space of  $\Phi$ . This happens only when measurement noise is 0. Also, as a solution of  $(\text{CS}_1^\epsilon)$ ,  $\hat{x}$  need not be sparse.  $h$  need not be sparse either.

### 6.6.1. The recovery error vector

For most part of this section, we will study the behavior of the recovery error vector. As a result of this analysis we will obtain the recovery guarantees for the  $(\text{CS}_1^\epsilon)$  program.

**Theorem 6.19** *The recovery error vector of  $(\text{CS}_1^\epsilon)$   $h = \hat{x} - x$  satisfies*

$$\|\Phi h\|_2 \leq 2\epsilon. \quad (6.6.2)$$

*This constraint is also known as the **tube constraint**.*

PROOF. We are given that

$$\|y - \Phi \hat{x}\|_2 \leq \epsilon$$

and

$$\|y - \Phi x\|_2 \leq \epsilon.$$

This gives us

$$\|\Phi(x - \hat{x})\|_2 = \|y - \Phi\hat{x} + \Phi x - y\|_2 \leq \|y - \Phi\hat{x}\|_2 + \|y - \Phi x\|_2 \leq 2\epsilon.$$

□

**Theorem 6.20** *The recovery error vector of  $(\text{CS}_1^\epsilon)$   $h = \hat{x} - x$  satisfies*

$$\|h_{\Lambda^c}\|_1 \leq \|h_\Lambda\|_1 \quad (6.6.3)$$

where  $\Lambda = \text{supp}(x)$ . This constraint is also known as **cone constraint**.

PROOF. Since  $x$  is a feasible vector for  $(\text{CS}_1^\epsilon)$ , hence

$$\|\hat{x}\|_1 \leq \|x\|_1. \quad (6.6.4)$$

We can further split  $h$  as

$$h = h_\Lambda + h_{\Lambda^c}$$

i.e.  $h_\Lambda$  is the part of  $h$  with the same support as  $x$  and  $h_{\Lambda^c}$  is the rest of  $h$ .

Now, replacing  $\hat{x} = x + h$  in (6.6.4) we get

$$\begin{aligned} & \|x + h\|_1 \leq \|x\|_1 \\ \implies & \|x + h_\Lambda + h_{\Lambda^c}\|_1 \leq \|x\|_1 \\ \implies & \|x + h_\Lambda\|_1 + \|h_{\Lambda^c}\|_1 \leq \|x\|_1 \\ \implies & \|x\|_1 - \|h_\Lambda\|_1 + \|h_{\Lambda^c}\|_1 \leq \|x\|_1 \\ \implies & -\|h_\Lambda\|_1 + \|h_{\Lambda^c}\|_1 \leq 0 \\ \implies & \|h_{\Lambda^c}\|_1 \leq \|h_\Lambda\|_1 \end{aligned}$$

where we used the triangle inequality in  $\|x + h_\Lambda\|_1 \geq \|x\|_1 - \|h_\Lambda\|_1$ . □

We have established a relationship between the  $l_1$  norms of  $h_\Lambda$  and  $h_{\Lambda^c}$ .

Although, it is not possible to compare  $\|h_\Lambda\|_2$  and  $\|h_{\Lambda^c}\|_2$  at this stage, if we pick up some of the largest entries in  $h_{\Lambda^c}$  and include in  $\Lambda$ , then

we can say something concrete about the  $l_2$  norm of the aggregated vector.

Let us enumerate the indices in  $\Lambda^c$  as

$$n_1, n_2, \dots, n_{N-|\Lambda|}$$

in decreasing order of magnitude of entries in  $h_{\Lambda^c}$ .

We proceed by breaking  $\Lambda^c$  down into smaller subsets of size  $T$  (exact value would be specified later). The smaller sets will be denoted as  $\Lambda_j$ .

$\Lambda_1$  will contain indices of  $T$  largest (magnitude) entries in  $h_{\Lambda^c}$ .  $\Lambda_2$  will contain next  $T$  largest entries and so on, i.e. we set

$$\Lambda_j = \{n_l : (j-1)T + 1 \leq l \leq jT\}.$$

Total number of such sets would be  $J = \left\lceil \frac{N-|\Lambda|}{T} \right\rceil$ . Note that the last set may have fewer than  $T$  indices. One reason we break down  $\Lambda^c$  into smaller sets is because it is then possible to apply restricted isometry property on the vectors  $h_{\Lambda_j}$  individually.

$h$  need not be sparse.  $h_\Lambda$  is  $K$ -sparse.  $h_{\Lambda^c}$  need not be sparse either but  $h_{\Lambda_j}$  are  $T$ -sparse where  $T$  would be appropriately chosen.

Obviously

$$\|h_{\Lambda_j}\|_p \geq \|h_{\Lambda_k}\|_p \text{ whenever } j < k$$

for any  $p$ -norm.

We are now ready to include indices in  $\Lambda_1$  and develop a bound on  $\|h\|_2$  in terms of entries in  $h$  indexed by  $\Lambda$  and  $\Lambda_1$ .

**Theorem 6.21** Consider the index set  $\Gamma = \Lambda \cup \Lambda_1$ .

$$\|h\|_2^2 \leq (1 + \rho) \|h_\Gamma\|_2^2. \quad (6.6.5)$$

where  $\rho = \frac{|\Lambda|}{T}$

Before proving, we can easily see that if we choose  $T = 3|\Lambda|$ , then

$$\|h\|_2^2 \leq \frac{4}{3} \|h_\Gamma\|_2^2 \iff \|h_{\Gamma^c}\|_2^2 \leq \frac{1}{3} \|h_\Gamma\|_2^2.$$

This demonstrates the concentration of  $l_2$ -norm of  $h$  over the index set  $\Gamma = \Lambda \cup \Lambda_1$ .

PROOF. We recall that the  $k$ -th largest entry in  $h_{\Lambda^c}$  obeys

$$|h_{\Lambda^c}|_{(k)} \leq \frac{\|h_{\Lambda^c}\|_1}{k}.$$

$|h_{\Lambda^c}|$  denotes the vector of absolute values of entries in  $h_{\Lambda^c}$ . We are using the subscript  $(k)$  to indicate  $k$ -th largest entry in the vector.

Consider the vector  $h_{\Gamma^c} \in \mathbb{C}^N$ . The  $|\Lambda|$  entries of  $h$  corresponding to  $h_\Lambda$  are set to 0 in  $h_{\Gamma^c}$ .  $T$  largest entries of  $h_{\Lambda^c}$  are set to zero in  $h_{\Gamma^c}$ . Consequently, the largest entries in  $h_{\Gamma^c}$  start with  $T+1$ -th largest entries in  $h_{\Lambda^c}$ . Thus,

$$|h_{\Gamma^c}|_{(k-T)} \leq \frac{\|h_{\Lambda^c}\|_1}{k} \quad \forall T+1 \leq k \leq N - |\Lambda|.$$

This gives us

$$\|h_{\Gamma^c}\|_2^2 = \sum_{T+1}^{N-|\Lambda|} |h_{\Gamma^c}|_{(k-T)}^2 \leq \|h_{\Lambda^c}\|_1^2 \sum_{k=T+1}^{N-|\Lambda|} \frac{1}{k^2}.$$

Just to clarify, the entries over the index set  $\Gamma$  are known to be 0 in  $h_{\Lambda^c}$  by construction. Hence, they have been left out in the  $l_2$ -norm (squared) expansion in above.

It is possible to show that

$$\sum_{k=T+1}^{N-|\Lambda|} \frac{1}{k^2} \leq \frac{1}{T}.$$

This gives us

$$\|h_{\Gamma^c}\|_2^2 \leq \frac{\|h_{\Lambda^c}\|_1^2}{T}.$$

Applying the cone inequality (6.6.3) ( $\|h_{\Lambda^c}\|_1 \leq \|h_\Lambda\|_1$ ), we get

$$\|h_{\Gamma^c}\|_2^2 \leq \frac{\|h_\Lambda\|_1^2}{T}$$

We also recall the upper bound on  $l_1$ -norm in terms of  $l_2$ -norm for sparse signals as:

$$\|h_\Lambda\|_1 \leq \sqrt{|\Lambda|} \|h_\Lambda\|_2.$$

This gives us

$$\|h_{\Gamma^c}\|_2^2 \leq \frac{|\Lambda|}{T} \|h_\Lambda\|_2^2. \quad (6.6.6)$$

Since  $\|h_\Lambda\|_2^2 \leq \|h_\Gamma\|_2^2$ , we can also write

$$\|h_{\Gamma^c}\|_2^2 \leq \frac{|\Lambda|}{T} \|h_\Gamma\|_2^2.$$

Finally this gives us

$$\|h\|_2^2 = \|h_\Gamma\|_2^2 + \|h_{\Gamma^c}\|_2^2 \leq \left(1 + \frac{|\Lambda|}{T}\right) \|h_\Gamma\|_2^2 \quad (6.6.7)$$

Replacing the definition of  $\rho = \frac{|\Lambda|}{T}$ , we get

$$\|h\|_2^2 \leq (1 + \rho) \|h_\Gamma\|_2^2.$$

□

Looking back at (6.6.6), we see that  $T$  should be greater than  $|\Lambda|$  or  $\rho \leq 1$  in order to have a useful energy concentration over  $\Gamma$ .

In (6.6.6) we established  $\|h_{\Gamma^c}\|_2 \leq \sqrt{\rho} \|h_\Lambda\|_2$ . We can write

$$h_{\Gamma^c} = \sum_{j=2}^J h_{\Lambda_j}.$$

This leads us to

$$\|h_{\Gamma^c}\|_2 \leq \sum_{j=2}^J \|h_{\Lambda_j}\|_2.$$

At this moment, no relationship between the two upper bounds  $\sum_{j=2}^J \|h_{\Lambda_j}\|_2$  and  $\sqrt{\rho} \|h_\Lambda\|_2$  is obvious. Before proving main recovery guarantees of this section, we will need one more result which establishes the relationship between these two upper bounds.

**Theorem 6.22** *The split of  $h$  over the index sets  $\Lambda_j$  satisfies*

$$\sum_{j=2}^J \|h_{\Lambda_j}\|_2 \leq \sqrt{\rho} \|h_\Lambda\|_2 \quad (6.6.8)$$

where  $\rho = \frac{|\Lambda|}{T}$ .

PROOF. Consider the vectors  $h_{\Lambda_j}$  and  $h_{\Lambda_{j+1}}$ . By construction every non-zero entry in  $h_{\Lambda_{j+1}}$  is not larger (in magnitude) than every entry in  $h_{\Lambda_j}$ .

The average value of magnitude of entries in  $h_{\Lambda_j}$  is given by  $\frac{\|h_{\Lambda_j}\|_1}{T}$ . Thus, by construction

$$|h_{\Lambda_{j+1}}(i)| \leq \frac{\|h_{\Lambda_j}\|_1}{T} \quad \forall 1 \leq i \leq N.$$

Using this, we get a bound on the squared  $l_2$ -norm of  $h_{\Lambda_{j+1}}$  (only  $T$  entries in  $h_{\Lambda_{j+1}}$  can be non-zero) as

$$\|h_{\Lambda_{j+1}}\|_2^2 \leq \frac{\|h_{\Lambda_j}\|_1^2}{T}.$$

Now, summing over all the  $\Lambda_j$  with  $2 \leq j \leq J$ , we get

$$\begin{aligned} \sum_{j=2}^J \|h_{\Lambda_{j+1}}\|_2 &\leq \sum_{j=1}^{J-1} \frac{\|h_{\Lambda_j}\|_1}{\sqrt{T}} \\ &\leq \sum_{j=1}^J \frac{\|h_{\Lambda_j}\|_1}{\sqrt{T}} \\ &= \frac{\sum_{j=1}^J \|h_{\Lambda_j}\|_1}{\sqrt{T}} = \frac{\|h_{\Lambda^c}\|_1}{\sqrt{T}} \\ &\leq \frac{\|h_{\Lambda}\|_1}{\sqrt{T}} \\ &\leq \frac{\sqrt{|\Lambda|} \|h_{\Lambda}\|_2}{\sqrt{T}} \\ &= \sqrt{\frac{|\Lambda|}{T}} \|h_{\Lambda}\|_2 = \sqrt{\rho} \|h_{\Lambda}\|_2 \end{aligned}$$

which is the desired result. In between, we used the cone inequality and  $l_1$ - $l_2$  norm bounds for sparse vectors.  $\square$

### 6.6.2. Sparse case recovery guarantee

We have now developed enough understanding of the behavior of the recovery error vector. We get back to the job of establishing an upper

bound on the recovery error in terms of the measurement error upper bound.

Our main result is

**Theorem 6.23** *Let  $K$  be such that  $\delta_{3K} + 3\delta_{4K} < 2$ . Then, for any signal  $x$  supported on  $\Lambda$  with  $|\Lambda| \leq K$  and any perturbation  $e$  with  $\|e\|_2 \leq \epsilon$ , the solution  $\hat{x}$  to  $(\text{CS}_1^e)$  obeys*

$$\|\hat{x} - x\|_2 = \|h\|_2 \leq C_K \epsilon, \quad (6.6.9)$$

where the constant  $C_K$  may only depend on  $\delta_{4K}$ . For reasonable values of  $\delta_{4K}$ ,  $C_K$  is well behaved; e.g.  $C_K \approx 8.82$  for  $\delta_{4K} = 1/5$  and  $C_K = 10.47$  for  $\delta_{4K} = 1/4$ .

PROOF. We will work on  $\|\Phi h\|_2$ . We will expand it in terms of sparse components of  $h$  and then apply RIP to get a bound on  $\|h\|_2$ .

Recall that

$$h = h_\Lambda + h_{\Lambda_1} + h_{\Lambda_2} + \cdots + h_{\Lambda_J} = h_\Gamma + \sum_{j=2}^J h_{\Lambda_j}.$$

So, by multiplying with  $\Phi$  on both sides we get

$$\begin{aligned} \|\Phi h\|_2 &= \left\| \Phi h_\Gamma + \sum_{j=2}^J \Phi h_{\Lambda_j} \right\|_2 \\ &\geq \|\Phi h_\Gamma\|_2 - \left\| \sum_{j=2}^J \Phi h_{\Lambda_j} \right\|_2 \\ &\geq \|\Phi h_\Gamma\|_2 - \sum_{j \geq 2} \|\Phi h_{\Lambda_j}\|_2 \\ &\geq \sqrt{1 - \delta_{T+|\Lambda|}} \|h_\Gamma\|_2 - \sqrt{1 + \delta_T} \sum_{j=2}^J \|h_{\Lambda_j}\|_2. \end{aligned}$$

We used RIP in

$$\|\Phi h_\Gamma\|_2 \geq \sqrt{1 - \delta_{T+|\Lambda|}} \|h_\Gamma\|_2$$

since  $|\Gamma| = T + |\Lambda|$  and in

$$\|\Phi h_{\Lambda_j}\|_2 \leq \sqrt{1 + \delta_T} \|h_{\Lambda_j}\|_2$$

since  $|\Lambda_j| \leq T$ .

From (6.6.8) we have:

$$\sum_{j=2}^J \|h_{\Lambda_j}\|_2 \leq \sqrt{\rho} \|h_\Lambda\|_2.$$

This gives us

$$\begin{aligned} \|\Phi h\|_2 &\geq \sqrt{1 - \delta_{T+|\Lambda|}} \|h_\Gamma\|_2 - \sqrt{1 + \delta_T} \sqrt{\rho} \|h_\Lambda\|_2 \\ &\geq \left( \sqrt{1 - \delta_{T+|\Lambda|}} - \sqrt{\rho} \sqrt{1 + \delta_T} \right) \|h_\Gamma\|_2 \end{aligned} \quad (6.6.10)$$

using the fact that  $\|h_\Gamma\|_2 \geq \|h_\Lambda\|_2$ .

Now define a constant depending on  $|\Lambda|$  and  $T$  as

$$C_{|\Lambda|,T} \triangleq \sqrt{1 - \delta_{T+|\Lambda|}} - \sqrt{\rho} \sqrt{1 + \delta_T}. \quad (6.6.11)$$

Then, the inequality can be simplified as

$$\|\Phi h\|_2 \geq C_{|\Lambda|,T} \|h_\Gamma\|_2. \quad (6.6.12)$$

From tube constraint (6.6.2), we have  $\|\Phi h\|_2 \leq 2\epsilon$ . Thus,

$$\|h_\Gamma\|_2 \leq \frac{2\epsilon}{C_{|\Lambda|,T}}$$

provided the denominator  $C_{|\Lambda|,T}$  is positive.

Finally from (6.6.5)

$$\|h\|_2 \leq \sqrt{1 + \rho} \|h_\Gamma\|_2 \leq \frac{2\sqrt{1 + \rho}}{C_{|\Lambda|,T}} \epsilon.$$

Since  $|\Lambda| \leq K^2$ , hence  $\delta_{T+|\Lambda|} \leq \delta_{T+K}$ , leading to

$$C_{K,T} \geq C_{|\Lambda|,T}.$$

Thus,

$$\|h\|_2 \leq \frac{2\sqrt{1 + \rho}}{C_{K,T}} \epsilon.$$

---

<sup>2</sup>The RIP constants are non-decreasing function of sparsity level.

Choosing

$$C_K = \frac{2\sqrt{1+\rho}}{C_{K,T}}$$

we can rewrite this as

$$\|h\|_2 \leq C_K \epsilon$$

which indeed is the desired result.

It remains to show that  $C_{|\Lambda|,T}$  is indeed positive with reasonable values for specific choices of  $T$ . We take  $T = 3|\Lambda|$ . Then  $\rho = \frac{1}{3}$  and

$$C_{|\Lambda|,T} = \sqrt{1 - \delta_{4|\Lambda|}} - \sqrt{\frac{1}{3}\sqrt{1 + \delta_{3|\Lambda|}}}.$$

Now

$$\begin{aligned} & C_{|\Lambda|,T} > 0 \\ \iff & \sqrt{1 - \delta_{4|\Lambda|}} - 2\sqrt{1 + \delta_{3|\Lambda|}} > 0 \\ \iff & 1 - \delta_{4|\Lambda|} > \frac{1}{3}(1 + \delta_{3|\Lambda|}) \\ \iff & 3 - 3\delta_{4|\Lambda|} > 1 + \delta_{3|\Lambda|} \\ \iff & \delta_{3|\Lambda|} + 3\delta_{4|\Lambda|} < 2 \end{aligned}$$

Since  $|\Lambda| \leq K$ , hence  $\delta_{3|\Lambda|} \leq \delta_{3K}$  and  $\delta_{4|\Lambda|} \leq \delta_{4K}$ . Thus,

$$\delta_{3K} + 3\delta_{4K} < 2 \implies \delta_{3|\Lambda|} + 3\delta_{4|\Lambda|} < 2.$$

**REMARK.** We note that if  $\delta_{4K} < \frac{1}{2}$ , then  $\delta_{3K} \leq \delta_{4K}$  ensures that  $\delta_{3K} + 3\delta_{4K} < 2$ .

What remains is to show some explicit values of  $C_K$  with

$$C_K = \frac{2\sqrt{1+\rho}}{C_{K,T}} \approx \frac{2\sqrt{4/3}}{\sqrt{1 - \delta_{4K}} - \sqrt{\frac{1}{3}\sqrt{1 + \delta_{3K}}}}$$

for specific suitable choices of  $\delta_{4K}$  and  $\rho = \frac{1}{3}$ .

Choosing  $\delta_{4K} = \frac{1}{5}$  and assuming  $\delta_{3K} \approx \frac{1}{5}$ , we get

$$C_K \approx \frac{2\sqrt{4/3}}{\sqrt{.8} - \sqrt{1/3}\sqrt{1.2}} = 8.8155.$$

Similarly, for  $\delta_{4K} = .25$ , we get

$$C_K \approx \frac{2\sqrt{4/3}}{\sqrt{.75} - \sqrt{1/3}\sqrt{1.25}} = 10.4721.$$

□

### 6.6.3. General case recovery guarantee

We now relax the condition that  $x$  is sparse and consider the case for recovery of arbitrary  $x$  by (CS<sub>i</sub><sup>ε</sup>) program.

We recall that the best  $K$ -term approximation of  $x$  is given by  $x|_K$  which is a vector comprising of  $K$  largest entries (in absolute value) in  $x$  and having 0 everywhere else.

Let  $\Lambda = \text{supp}(x|_K)$ . We can consider  $x - x|_K$  as another noise introduced during the compressive sampling process. Define

$$g = x - x|_K$$

as the approximation error vector for the best  $K$ -sparse approximation of  $x$ . We note that  $x|_K = x_\Lambda$  and

$$x = x_\Lambda + x_{\Lambda^c} \implies g = x - x|_K = x - x_\Lambda = x_{\Lambda^c}.$$

A suitable recovery guarantee should be able to connect the recovery error  $\|h\|_2 = \|\hat{x} - x\|_2$  with the measurement error upper bound  $\epsilon$  and some measure of the approximation error  $g$ .

The tube constraint (6.6.2) continues to hold for the arbitrary signal  $x$ . But the cone constraint (6.6.3) doesn't hold anymore. Although, a variation does hold.

**Theorem 6.24** *The recovery error vector  $h = \hat{x} - x$  for the arbitrary vector  $x \in \mathbb{C}^N$  satisfies*

$$\|h_{\Lambda^c}\|_1 \leq \|h_\Lambda\|_1 + 2\|x_{\Lambda^c}\|_1 \quad (6.6.13)$$

where  $\Lambda = \text{supp}(x|_K)$ .

PROOF. We start with

$$\|x + h\|_1 \leq \|x\|_1.$$

Taking lower bounds (using triangle inequality) for the L.H.S.:

$$\begin{aligned} \|x + h\|_1 &= \|x_\Lambda + x_{\Lambda^c} + h_\Lambda + h_{\Lambda^c}\|_1 \\ &\geq \|x_\Lambda + h_\Lambda + h_{\Lambda^c}\|_1 - \|x_{\Lambda^c}\|_1 \\ &\geq \|x_\Lambda + h_{\Lambda^c}\|_1 - \|x_{\Lambda^c}\|_1 - \|h_\Lambda\|_1 \\ &= \|x_\Lambda\|_1 + \|h_{\Lambda^c}\|_1 - \|x_{\Lambda^c}\|_1 - \|h_\Lambda\|_1. \end{aligned}$$

At the same time on the R.H.S.:

$$\|x\|_1 = \|x_\Lambda\|_1 + \|x_{\Lambda^c}\|_1.$$

Putting them together, we get

$$\|x_\Lambda\|_1 + \|h_{\Lambda^c}\|_1 - \|x_{\Lambda^c}\|_1 - \|h_\Lambda\|_1 \leq \|x_\Lambda\|_1 + \|x_{\Lambda^c}\|_1.$$

Canceling the common term and rearranging we get

$$\|h_{\Lambda^c}\|_1 - \|h_\Lambda\|_1 \leq 2\|x_{\Lambda^c}\|_1.$$

Alternatively

$$\|h_{\Lambda^c}\|_1 \leq \|h_\Lambda\|_1 + 2\|x_{\Lambda^c}\|_1.$$

□

As before, we will now consider the vector  $h_{\Lambda^c}$  and divide  $\Lambda^c$  into index sets of size  $T$  with decreasing order of magnitudes of entries in  $h_{\Lambda^c}$ . There would be  $J$  such sets.

We will now develop a bound similar to theorem 6.21.

**Theorem 6.25** Consider the index set  $\Gamma = \Lambda \cup \Lambda_1$ .

$$\|h\|_2 \leq (1 + \sqrt{\rho})\|h_\Gamma\|_2 + 2\sqrt{\rho} \frac{\|x_{\Lambda^c}\|_1}{\sqrt{K}}. \quad (6.6.14)$$

where  $\rho = \frac{K}{T}$

PROOF. Proceeding as in the proof of theorem 6.21, we have

$$\|h_{\Gamma^c}\|_2 \leq \frac{\|h_{\Lambda^c}\|_1}{\sqrt{T}}.$$

We cannot apply the cone inequality here, but (6.6.13) is available.

This gives us

$$\|h_{\Gamma^c}\|_2 \leq \frac{\|h_{\Lambda}\|_1 + 2\|x_{\Lambda^c}\|_1}{\sqrt{T}}.$$

Using  $\|h_{\Lambda}\|_1 \leq \sqrt{K}\|h_{\Lambda}\|_2$  and  $\frac{1}{\sqrt{T}} = \frac{\sqrt{\rho}}{\sqrt{K}}$ , we obtain

$$\|h_{\Gamma^c}\|_2 \leq \sqrt{\rho} \cdot \left( \|h_{\Lambda}\|_2 + \frac{2\|x_{\Lambda^c}\|_1}{\sqrt{K}} \right).$$

Using  $\|h_{\Lambda}\|_2 \leq \|h_{\Gamma}\|_2$ , we get

$$\|h_{\Gamma^c}\|_2 \leq \sqrt{\rho} \cdot \left( \|h_{\Gamma}\|_2 + \frac{2\|x_{\Lambda^c}\|_1}{\sqrt{K}} \right).$$

Adding  $\|h_{\Gamma}\|_2$  on both sides and noting that

$$\|h\|_2 \leq \|h_{\Gamma}\|_2 + \|h_{\Gamma^c}\|_2$$

we get

$$\|h\|_2 \leq (1 + \sqrt{\rho})\|h_{\Gamma}\|_2 + 2\sqrt{\rho} \frac{\|x_{\Lambda^c}\|_1}{\sqrt{K}}.$$

□

Next we develop a result similar to theorem 6.22 for the general case

**Theorem 6.26** *The split of  $h$  over the index sets  $\Lambda_j$  satisfies*

$$\sum_{j=2}^J \|h_{\Lambda_{j+1}}\|_2 \leq \sqrt{\rho} \cdot \left( \|h_{\Lambda}\|_2 + \frac{2\|x_{\Lambda^c}\|_1}{\sqrt{K}} \right). \quad (6.6.15)$$

where  $\rho = \frac{|\Lambda|}{T}$ .

PROOF. Proceeding as in theorem 6.22 we obtain

$$\sum_{j=2}^J \|h_{\Lambda_{j+1}}\|_2 \leq \frac{\|h_{\Lambda^c}\|_1}{\sqrt{T}}.$$

Things change here as cone constraint doesn't apply anymore. Applying (6.6.13) we get

$$\sum_{j=2}^J \|h_{\Lambda_{j+1}}\|_2 \leq \frac{\|h_\Lambda\|_1 + 2\|x_{\Lambda^c}\|_1}{\sqrt{T}}.$$

Simplifying just like theorem 6.25, we get

$$\sum_{j=2}^J \|h_{\Lambda_{j+1}}\|_2 \leq \sqrt{\rho} \cdot \left( \|h_\Lambda\|_2 + \frac{2\|x_{\Lambda^c}\|_1}{\sqrt{K}} \right).$$

□

We are now in a position to develop a recovery guarantee for arbitrary signals in terms of  $\epsilon$  and  $\|x_{\Lambda^c}\|_1$ .

**Theorem 6.27** Suppose that  $x$  is an arbitrary vector in  $\mathbb{C}^N$  and let  $x|_K$  be the best  $K$ -term approximation of  $x$  in  $\mathbb{C}^N$ . Under the hypothesis of theorem 6.23, the solution  $\hat{x}$  to (CS $^\epsilon_1$ ) obeys

$$\|\hat{x} - x\|_2 \leq C_{1,K}\epsilon + C_{2,K} \cdot \frac{\|x - x|_K\|_1}{\sqrt{K}}. \quad (6.6.16)$$

For reasonable values of  $\delta_{4K}$  the constants in (6.6.16) are well behaved; e.g.  $C_{1,K} \approx 12.04$  and  $C_{2,K} \approx 8.77$  for  $\delta_{4K} = \frac{1}{5}$ .

We note that the bound in (6.6.16) is useful when  $\|x - x|_K\|_1$  is small.

In terms of  $g = x_{\Lambda^c}$  and  $h = \hat{x} - x$  we can write (6.6.16) as

$$\|h\|_2 \leq C_{1,K}\epsilon + C_{2,K} \cdot \frac{\|g\|_1}{\sqrt{K}}.$$

PROOF. Proceeding as in the proof of theorem 6.23

$$\|\Phi h\|_2 \geq \sqrt{1 - \delta_{T+K}} \|h_\Gamma\|_2 - \sqrt{1 + \delta_T} \sum_{j=2}^J \|h_{\Lambda_j}\|_2.$$

From (6.6.15), we have

$$\sum_{j=2}^J \|h_{\Lambda_{j+1}}\|_2 \leq \sqrt{\rho} \cdot \left( \|h_\Lambda\|_2 + \frac{2\|x_{\Lambda^c}\|_1}{\sqrt{K}} \right).$$

Thus,

$$\|\Phi h\|_2 \geq \sqrt{1 - \delta_{T+K}} \|h_\Gamma\|_2 - \sqrt{\rho} \sqrt{1 + \delta_T} \left( \|h_\Lambda\|_2 + \frac{2\|x_{\Lambda^c}\|_1}{\sqrt{K}} \right).$$

Using  $\|h_\Gamma\|_2 \geq \|h_\Lambda\|_2$ , we obtain

$$\|\Phi h\|_2 \geq (\sqrt{1 - \delta_{T+K}} - \sqrt{\rho} \sqrt{1 + \delta_T}) \|h_\Gamma\|_2 - \sqrt{\rho} \sqrt{1 + \delta_T} \frac{2\|x_{\Lambda^c}\|_1}{\sqrt{K}}.$$

Like in theorem 6.23, we introduce

$$C_{K,T} = \sqrt{1 - \delta_{T+K}} - \sqrt{\rho} \sqrt{1 + \delta_T}$$

and using the tube constraint (6.6.2) we get

$$C_{K,T} \|h_\Gamma\|_2 - \sqrt{\rho} \sqrt{1 + \delta_T} \frac{2\|x_{\Lambda^c}\|_1}{\sqrt{K}} \leq 2\epsilon$$

or

$$C_{K,T} \|h_\Gamma\|_2 \leq 2\epsilon + \sqrt{\rho} \sqrt{1 + \delta_T} \frac{2\|x_{\Lambda^c}\|_1}{\sqrt{K}}.$$

Under the hypothesis of theorem 6.23 we know that  $C_{K,T}$  is positive. This gives us

$$\|h_\Gamma\|_2 \leq \frac{2}{C_{K,T}} \cdot \left( \epsilon + \sqrt{\rho} \sqrt{1 + \delta_T} \frac{\|x_{\Lambda^c}\|_1}{\sqrt{K}} \right)$$

Finally, from (6.6.14), we have

$$\|h\|_2 \leq (1 + \sqrt{\rho}) \|h_\Gamma\|_2 + 2\sqrt{\rho} \frac{\|x_{\Lambda^c}\|_1}{\sqrt{K}}.$$

Define

$$C_{1,K} = \frac{2(1 + \sqrt{\rho})}{C_{K,T}}$$

and

$$C_{2,K} = \frac{2(1 + \sqrt{\rho})}{C_{K,T}} \sqrt{\rho} \sqrt{1 + \delta_T} + 2\sqrt{\rho}.$$

Then we get:

$$\|h\|_2 \leq C_{1,K} \epsilon + C_{2,K} \frac{\|x_{\Lambda^c}\|_1}{\sqrt{K}}$$

which is the desired result.

Choosing  $T = 3K$ , we have  $\rho = \frac{1}{3}$ . We get

$$C_{K,T} = \sqrt{1 - \delta_{4K}} - \sqrt{1/3} \sqrt{1 + \delta_{3K}}$$

Now choosing  $\delta_{4K} = 0.2$ , we have

$$C_{K,T} \approx 0.2620.$$

This gives us

$$C_{1,K} \approx 12.0421$$

and

$$C_{2,K} \approx 8.7708.$$

Note that we can also write

$$C_{2,K} = C_{1,K} \sqrt{\rho} \sqrt{1 + \delta_T} + 2\sqrt{\rho}.$$

□

#### 6.6.4. Compressible signal recovery guarantee

**Recall** the definition of a  $p$ -compressible signal  $x \in \mathbb{C}^N$  satisfying

$$|x_{(i)}| \leq R \cdot i^{-\frac{1}{p}} \quad \forall i = 1, 2, \dots, N.$$

where  $x_{(i)}$  represents the  $i$ -th largest (magnitude wise) entry in  $x$ .

We also **recall** that the  $l_1$  norm of approximation error for the  $K$ -sparse approximation  $x|_K$  satisfies

$$\|x - x|_K\|_1 \leq C_p \cdot R \cdot K^{1-\frac{1}{p}}$$

where

$$C_p = \left(\frac{1}{p} - 1\right)^{-1}.$$

Dividing by  $\sqrt{K}$  on both sides we get

$$\frac{\|x - x|_K\|_1}{\sqrt{K}} \leq C_p \cdot R \cdot K^{-\frac{1}{p} + \frac{1}{2}}.$$

Putting these bounds in theorem 6.27, we obtain a bound on the recovery error as

$$\|\hat{x} - x\|_2 \leq C_{1,K}\epsilon + C_{2,K} \cdot C_p \cdot R \cdot K^{-\frac{1}{p} + \frac{1}{2}}. \quad (6.6.17)$$

### 6.6.5. Concluding remarks

We note that the analysis in this section exclusively depends on the restricted isometry property of  $\Phi$ . Thus, the analysis is completely deterministic and applies for all sparse or compressible signals.

## 6.7. Digest

### Problem formulations

**Exact sparse problem:**

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_0 \text{ subject to } x = \mathcal{D}\alpha. \quad (\text{P}_0)$$

**Sparse recovery with sparsity bound:**

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|x - \mathcal{D}\alpha\|_2 \text{ subject to } \|\alpha\|_0 \leq K. \quad (\text{P}_0^K)$$

**Sparse recovery with approximation error bound**

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_0 \text{ subject to } \|x - \mathcal{D}\alpha\|_2 \leq \epsilon. \quad (\text{P}_0^\epsilon)$$

**Noiseless compressed sensing**

$$\hat{x} = \arg \min_{x \in \mathbb{C}^N} \|x\|_0 \text{ subject to } y = \Phi x. \quad (\text{CS}_0)$$

**CS recovery with sparsity bound**

$$\hat{x} = \arg \min_{x \in \mathbb{C}^N} \|y - \Phi x\|_2 \text{ subject to } \|x\|_0 \leq K. \quad (\text{CS}_0^K)$$

**CS recovery with measurement error bound**

$$\hat{x} = \arg \min_{x \in \mathbb{C}^N} \|x\|_0 \text{ subject to } \|y - \Phi x\|_2 \leq \epsilon. \quad (\text{CS}_0^\epsilon)$$

**Basis pursuit**

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_1 \text{ subject to } x = \mathcal{D}\alpha. \quad (\text{P}_1)$$

Basis pursuit with inequality constraints

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_1 \text{ subject to } \|x - \mathcal{D}\alpha\|_2 \leq \epsilon \quad (\text{P}_1^\epsilon)$$

Basis pursuit denoising with approximation error penalty

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_1 + \lambda \|x - \mathcal{D}\alpha\|_2^2. \quad (\text{P}_1^\lambda)$$

Basis pursuit denoising with  $l_1$  penalty

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \frac{1}{2} \|x - \mathcal{D}\alpha\|_2^2 + \gamma \|\alpha\|_1. \quad (\text{P}_1^\gamma)$$

**Basis pursuit** We examine conditions for equivalence of  $(\text{P}_0)$  and  $(\text{P}_1)$  problems.

**BP: Two ortho case**

$$\begin{aligned} \mathcal{D} &= \begin{bmatrix} \Psi & \Phi \end{bmatrix} \\ x = \mathcal{D}\alpha &= \begin{bmatrix} \Psi & \Phi \end{bmatrix} \begin{bmatrix} \alpha^p \\ \alpha^q \end{bmatrix} = \Psi\alpha^p + \Phi\alpha^q. \\ k_p &= \|\alpha^p\|_0 \quad \text{and} \quad k_q = \|\alpha^q\|_0. \end{aligned}$$

Basis pursuit equivalence two ortho case sufficient condition:

$$2\mu(\mathcal{D})^2 k_p k_q + \mu(\mathcal{D}) k_p - 1 < 0$$

Weaker sufficient condition:

$$\|\alpha\|_0 = K = k_p + k_q < \frac{\sqrt{2} - 0.5}{\mu(\mathcal{D})}$$

**BP: General case Equivalence-Basis Pursuit** sufficient condition:

$$\|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu(\mathcal{D})} \right).$$

**spark** <sub>$\eta$</sub>  Smallest possible  $K$  with  $K = |\Lambda|$  s.t.

$$\min_{\Lambda} \sigma_K(A_{\Lambda}) \leq \eta.$$

$$\text{spark}_0(\mathcal{D}) = \text{spark}(\mathcal{D}).$$

$$\text{spark}_1(\mathcal{D}) = 1.$$

$$\text{spark}_{\eta_1}(\mathcal{D}) \leq \text{spark}_{\eta_2}(\mathcal{D}), \text{ whenever } \eta_1 > \eta_2.$$

$$1 \leq \text{spark}_\eta(\mathcal{D}) \leq \text{spark}_0(\mathcal{D}) = \text{spark}(\mathcal{D}) \leq N + 1 \quad \forall 0 \leq \eta \leq 1.$$

**Null space vectors** If  $\|\mathcal{D}v\|_2 \leq \eta$  and  $\|v\|_2 = 1$ , then  $\|v\|_0 \geq \text{spark}_\eta(\mathcal{D})$ .

**spark<sub>η</sub>** and coherence

$$\text{spark}_\eta(\mathcal{D}) \geq \frac{1 - \eta^2}{\mu(\mathcal{D})} + 1.$$

**Uncertainty result for sparse representations**

$$\|\alpha_1\|_0 + \|\alpha_2\|_0 \geq \text{spark}_\eta(\mathcal{D}), \text{ where } \eta = \frac{2\epsilon}{\|\alpha_1 - \alpha_2\|_2}.$$

**Localization of sparse representations** sufficient condition:

$$\|\alpha_i\|_0 \leq \frac{1}{2} \text{spark}_\eta(\mathcal{D})$$

Representation error upper bound

$$\|\alpha_1 - \alpha_2\|_2 \leq \delta = \frac{2\epsilon}{\eta}.$$

**Stability coherence**

$$\|\hat{\alpha} - \alpha\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(\mathcal{D})(2\|\alpha\|_0 - 1)}.$$

whenever

$$\|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu(\mathcal{D})} \right).$$

**Stability RIP**

$$\|\hat{\alpha} - \alpha\|_2^2 \leq \frac{4\epsilon^2}{1 - \delta_{2K}}.$$

**BPIC stability guarantee**

$$\|\hat{\alpha} - \alpha\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(\mathcal{D})(4\|\alpha\|_0 - 1)}.$$

whenever

$$\|\alpha\|_0 < \frac{1}{4} \left( 1 + \frac{1}{\mu(\mathcal{D})} \right).$$

# CHAPTER 7

## Matching Pursuit Algorithms

### 7.1. Introduction

In this chapter we will review some matching pursuit algorithms which can help us solve the sparse approximation problem and the sparse recovery problem discussed in chapter 2.

The presentation in this chapter is based on a number of sources including [4, 7, 21, 29, 34, 38].

Let us recall the definitions of sparse approximation and recovery problems from previous chapters.

From definition 2.8 let  $\mathcal{D}$  be a signal dictionary with  $\Phi \in \mathbb{C}^{N \times D}$  being its synthesis matrix. The  $(\mathcal{D}, K)$ -SPARSE approximation can be written as

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} \quad \|x - \Phi\alpha\|_2 \\ & \text{subject to} \quad \|\alpha\|_0 \leq K. \end{aligned} \tag{P}_0^{\epsilon}$$

From definition 2.7 with the help of synthesis matrix  $\Phi$ , the  $(\mathcal{D}, K)$ -EXACT-SPARSE problem can be written as

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} \quad \|\alpha\|_0 \\ & \text{subject to} \quad x = \Phi\alpha \\ & \text{and} \quad \|\alpha\|_0 \leq K \end{aligned} \tag{P}_0$$

From definition 2.23 we recall the *sparse signal recovery from compressed measurements* problem as following. Let  $\Phi \in \mathbb{C}^{M \times N}$  be a sensing matrix. Let  $x \in \mathbb{C}^N$  be an unknown signal which is assumed to be sparse or compressible. Let  $y = \Phi x$  be a measurement vector in  $\mathbb{C}^M$ .

Then the signal recovery problem is to recover  $x$  from  $y$  subject to

$$y = \Phi x$$

assuming  $x$  to be  $K$  sparse or at least  $K$  compressible.

We note that sparse approximation problem and sparse recovery problems have pretty much same structure. They are in fact dual to each other. Thus we will see that the same set of algorithms can be adapted to solve both problems.

In the sequel we will see many variations of above problems.

### 7.1.1. Our first problem

We will start with attacking a very simple version of  $(\mathcal{D}, K)$ -EXACT-SPARSE problem.

Setting up notation

- $x \in \mathbb{C}^N$  is our signal of interest and it is known.
- $\mathcal{D}$  is the dictionary in which we are looking for a sparse representation of  $x$ .
- $\Phi \in \mathbb{C}^{N \times D}$  is the synthesis matrix for  $\mathcal{D}$ .
- The sparse representation of  $x$  in  $\mathcal{D}$  is given by

$$x = \Phi \alpha.$$

- It is assumed that  $\alpha \in \mathbb{C}^D$  is sparse with  $|\alpha|_0 \leq K$ .
- Also we assume that  $\alpha$  is the sparsest possible solution for  $x$  that we are looking.
- We know  $x$ , we know  $\Phi$ , we don't know  $\alpha$ . We are looking for it.

Thus we need to solve the optimization problem given by

$$\underset{\alpha}{\text{minimize}} \|\alpha\|_0 \text{ subject to } x = \Phi \alpha. \quad (7.1.1)$$

For the unknown vector  $\alpha$ , we need to find

- the sparsest support for the solution i.e.  $\{i|\alpha_i \neq 0\}$
- the non-zero values  $\alpha_i$  over this support.

If we are able to find the support for the solution  $\alpha$ , then we may assume that the non-zero values of  $\alpha$  can be easily computed by least squares methods.

Note that the support is discrete in nature (An index  $i$  either belongs to the support or it does not). Hence algorithms which will seek the support will also be discrete in nature.

We now build up a case for greedy algorithms before jumping into specific algorithms later.

Let us begin with a much simplified version of the problem.

Let the columns of the matrix  $\Phi$  be represented as

$$\Phi = [\phi_1 \ \phi_2 \ \dots \ \phi_N]. \quad (7.1.2)$$

Let  $\text{spark}(\Phi) > 2$ . Thus no two columns in  $\Phi$  are linearly dependent and as per theorem 2.23, for any  $x$ , there is at most only one 1-sparse explanation vector.

We now assume that such a representation exists and we would be looking for optimal solution vector  $\alpha^*$  that has only one non-zero value, i.e.  $\|\alpha^*\|_0 = 1$ .

Let  $i$  be the index at which  $\alpha_i^* \neq 0$ .

Thus  $x = \alpha_i^* \phi_i$ , i.e.  $x$  is a scalar multiple of  $\phi_i$  (the  $i$ -th column of  $\Phi$ ).

Off-course we don't know what is the value of index  $i$ .

We can find this by comparing  $x$  with each column of  $\Phi$  and find the column which best matches it.

Consider the least squares minimization problem:

$$\epsilon(j) = \underset{z_j}{\text{minimize}} \| \phi_j z_j - x \|_2. \quad (7.1.3)$$

where  $z_j \in \mathbb{C}$  is a scalar.

From linear algebra, it attempts to find the projection of  $x$  over  $\phi_j$  and  $\epsilon(j)$  represents the magnitude of error between  $x$  and the projection of  $x$  over  $\phi_j$ .

Optimal solution is given by

$$z_j^* = \frac{\phi_j^H x}{\|\phi_j\|_2^2} = \phi_j^H x \quad (7.1.4)$$

since columns of a dictionary are assumed to be unit norm.

Plugging it back into the expression of minimum squared error we get

$$\begin{aligned} \epsilon^2(j) &= \underset{z_j}{\text{minimize}} \| \phi_j z_j - x \|_2^2 \\ &= \| \phi_j \phi_j^H x - x \|_2^2 \\ &= \| x \|_2^2 - | \phi_j^H x |^2. \end{aligned}$$

Now since  $x$  is a scalar multiple of  $\phi_i$ , hence  $\epsilon(i) = 0$ , thus if we look at  $\epsilon(j)$  for  $j = 1, \dots, D$ , the minimum value 0 will be obtained for  $j = i$ .

And  $\epsilon(i) = 0$  means

$$\| x \|_2^2 - | \phi_i^H x |^2 = 0 \implies \| x \|_2^2 = | \phi_i^H x |^2. \quad (7.1.5)$$

This is a special case of Cauchy-Schwartz inequality when  $x$  and  $\phi_i$  are collinear.

The sparse representation is given by

$$\alpha = \begin{bmatrix} 0 \\ \vdots \\ z_i^* \\ \vdots \\ 0 \end{bmatrix}$$

Since  $x \in \mathbb{C}^N$  and  $\phi_j \in \mathbb{C}^N$ , hence computation of  $\epsilon(j)$  requires  $\mathcal{O}(N)$  time.

Since we may need to do it for all  $D$  columns, hence finding the index  $i$  takes  $\mathcal{O}(ND)$  time.

Now let us make our life more complex. We now suppose that  $\text{spark}(\Phi) > 2K$ . Thus a sparse representation  $\alpha$  of  $x$  with up to  $K$  non-zero values is unique if it exists(see again theorem 2.23). We assume it exists. Since  $x = \Phi\alpha$ ,  $x$  is a linear combination of up to  $K$  columns of  $\Phi$ .

One approach could be to check out all  $\binom{D}{K}$  possible subsets of  $K$  columns from  $\Phi$ .

But  $\binom{D}{K}$  is  $\mathcal{O}(D^K)$  and for each subset of  $K$  columns solving the least squares problem will take  $\mathcal{O}(NK^2)$  time. Hence overall complexity of the recovery process would be  $\mathcal{O}(D^K NK^2)$ . This is prohibitively expensive.

A way around is by adopting a greedy strategy in which we abandon the hopeless exhaustive search and attempt a series of single term updates in the solution vector  $\alpha$ .

Since this is an iterative procedure, let us call the approximation at each iteration as  $\alpha^k$  where  $k$  is the iteration index.

- We start with  $\alpha^0 = 0$ .
- At each iteration we choose one new column in  $\alpha^k$  and fill in a value.
- The column and value are chosen such that it maximally reduces the  $l_2$  error between  $x$  and the approximation. i.e.

$$\|x - \Phi\alpha^{k+1}\|_2 < \|x - \Phi\alpha^k\|_2 \quad (7.1.6)$$

and the error reduction is as high as possible.

- We stop when the  $l_2$  error reduces below a specific threshold.

We are now ready to explore different greedy algorithms.

## 7.2. Orthogonal Matching Pursuit for sparse approximation

### 7.2.1. The algorithm

The core **Orthogonal Matching Pursuit** algorithm is presented in [Figure 7.1](#). The algorithm is iterative.

- We start with the initial estimate of solution  $\alpha = 0$ .
- We also maintain the support of  $\alpha$  i.e. the set of indices for which  $\alpha$  is non-zero. We start with an empty support.
- In each ( $k$ -th) iteration we attempt to reduce the difference between the actual signal  $x$  and the approximate signal based on current solution  $\alpha^{k-1}$  given by  $r^{k-1} = x - \Phi\alpha^{k-1}$ .
- We do this by choosing a new index in  $\alpha$  given by  $j_0$  for the column  $\phi_{j_0}$  which most closely matches our current residual.
- We include this to our support for  $\alpha$ , estimate new solution vector  $\alpha$  and compute new residual.
- We stop when the residual magnitude is below a threshold  $\epsilon_0$  defined by us.

Each iteration of algorithm consists of following stages:

**Sweep:** For each column  $\phi_j$  in our synthesis matrix, we measure the projection of residual from previous iteration on the column and compute the magnitude of error between the projection and residual.

The square of minimum error for  $\phi_j$  is given by:

$$\epsilon^2(j) = \|r^{k-1}\|_2^2 - |\phi_j^H r^{k-1}|^2.$$

We can also note that minimizing over  $\epsilon(j)$  is equivalent to maximizing over the inner product of  $\phi_j$  with  $r^{k-1}$  though this just helps us reduce only  $N$  subtractions per iteration.

**Update support:** Ignoring the columns which have already been included in the support, we pick up the column which most closely resembles the residual of previous stage. i.e. the magnitude of error is minimum. We include the index of this column  $j_0$  in the support set  $S^k$ .

```

Input: Synthesis matrix  $\Phi \in \mathbb{C}^{N \times D}$  with  $\text{spark}(\Phi) > 2K \ll D$ 
Input: Threshold  $\epsilon_0$ 
Input: Signal  $x \in \mathbb{C}^N$ 
Output:  $K$ -sparse approximate representation  $\alpha \in \Sigma_K \subseteq \mathbb{C}^D$ 
          satisfying  $\|x - \Phi\alpha\|_2 \leq \epsilon_0$ 

// Initialization
 $k \leftarrow 0$ ; // Iteration counter
 $\alpha^0 \leftarrow 0$ ; // Solution vector  $\alpha \in \mathbb{C}^D$ 
 $r^0 \leftarrow x - \Phi\alpha^0 = x$ ; // Residual  $r \in \mathbb{C}^N$ 
 $S^0 \leftarrow \emptyset$ ; // Solution support  $S = \text{supp}(\alpha)$ 

while  $\|r^k\|_2 > \epsilon_0$  do
     $k \leftarrow k + 1$ ;
    // Sweep
    foreach  $j \leftarrow 1, \dots, D$  do
         $\epsilon^2(j) \leftarrow \underset{z_j}{\text{minimize}} \|z_j \phi_j - r^{k-1}\|_2^2 = \|r^{k-1}\|_2^2 - |\phi_j^H r^{k-1}|^2$ ;
        //  $z_j^* = \phi_j^H r^{k-1}$ 
    end
    // Update support
    Find  $j_0$  that minimizes  $\epsilon(j) \forall j \notin S^{k-1}$ ; // i.e.  $\epsilon(j_0) \leq \epsilon(j)$ 
     $S^k \leftarrow S^{k-1} \cup \{j_0\}$ ;
    // Update provisional solution
     $\alpha^k \leftarrow \underset{\alpha}{\text{minimize}} \|\Phi\alpha - x\|_2^2$  subject to  $\text{supp}(\alpha) = S^k$ ;
    // Update residual
     $r^k = x - \Phi\alpha^k$ ;
end

```

FIGURE 7.1. Orthogonal matching pursuit for sparse approximation

**Update provisional solution:** In this step we find the solution of minimizing  $\|\Phi\alpha - x\|^2$  over the support  $S^k$  as our next candidate solution vector.

By keeping  $\alpha_i = 0$  for  $i \notin S^k$  we are essentially leaving out corresponding columns  $\phi_i$  from our calculations.

Thus we pickup up only the columns specified by  $S^k$  from  $\Phi$ . Let us call this matrix as  $\Phi_{S^k}$ . The size of this matrix is  $N \times |S^k|$ . Let us call corresponding sub vector as  $\alpha_{S^k}$ .

E.g. suppose  $D = 4$ , then  $\Phi = [\phi_1 \ \phi_2 \ \phi_3 \ \phi_4]$ . Let  $S^k = \{1, 4\}$ . Then  $\Phi_{S^k} = [\phi_1 \ \phi_4]$  and  $\alpha_{S^k} = (\alpha_1, \alpha_4)$ .

Our minimization problem then reduces to minimizing  $\|\Phi_{S^k}\alpha_{S^k} - x\|_2$ .

We use standard least squares estimate for getting the coefficients for  $\alpha_{S^k}$  over these indices. We put back  $\alpha_{S^k}$  to obtain our new solution estimate  $\alpha^k$ .

In the running example after obtaining the values  $\alpha_1$  and  $\alpha_4$ , we will have  $\alpha^k = (\alpha_1, 0, 0, \alpha_4)$ .

The solution to this minimization problem is given by

$$\Phi_{S^k}^H(\Phi_{S^k}\alpha_{S^k} - x) = 0 \implies \alpha_{S^k} = (\Phi_{S^k}^H\Phi_{S^k})^{-1}\Phi_{S^k}^Hx$$

Interestingly we note that  $r^k = x - \Phi\alpha^k = x - \Phi_{S^k}\alpha_{S^k}$ , thus

$$\Phi_{S^k}^H r^k = 0$$

which means that columns in  $\Phi_{S^k}$  which are part of support  $S^k$  are necessarily orthogonal to the residual  $r^k$ . This implies that these columns will not be considered in the coming iterations for extending the support. This orthogonality is the reason behind the name of the algorithm as OMP.

**Update residual:** We finally update the residual vector to  $r^k$  based on new solution vector estimate.

### Example 7.1: $(D, K)$ -exact-sparse recovery with OMP

Let us consider a synthesis matrix of size  $10 \times 20$ . Thus  $N = 10$  and  $D = 20$ . In order to fit into the display, we will present the matrix in two 10 column parts.

$$\Phi_a = \frac{1}{\sqrt{10}} \begin{bmatrix} -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

$$\Phi_b = \frac{1}{\sqrt{10}} \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \\ -1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 \end{bmatrix}$$

with

$$\Phi = [\Phi_a \quad \Phi_b].$$

You may verify that each column is unit norm.

It is known that  $\text{rank}(\Phi) = 10$  and  $\text{spark}(\Phi) = 6$ . Thus if a signal  $x$  has a 2 sparse representation in  $\Phi$  then the representation is necessarily unique.

We now consider a signal  $x$  given by

$$x = \begin{pmatrix} 4.74342 & -4.74342 & 1.58114 & -4.74342 & -1.58114 \\ 1.58114 & -4.74342 & -1.58114 & -4.74342 & -4.74342 \end{pmatrix}.$$

For saving space, we have written it as an n-tuple over two rows. You should treat it as a column vector of size  $10 \times 1$ .

It is known that the vector has a two sparse representation in  $\Phi$ . Let us go through the steps of OMP and see how it works.

In step 0,  $r^0 = x$ ,  $\alpha^0 = 0$ , and  $S^0 = \emptyset$ .

We now compute absolute value of inner product of  $r^0$  with each of the columns. They are given by

$$\begin{pmatrix} 4 & 4 & 4 & 7 & 3 & 1 & 11 & 1 & 2 & 1 \\ 2 & 1 & 7 & 0 & 2 & 4 & 0 & 2 & 1 & 3 \end{pmatrix}$$

We quickly note that the maximum occurs at index 7 with value 11.

We modify our support to  $S^1 = \{7\}$ .

We now solve the least squares problem

$$\text{minimize } \|x - [\phi_7]\alpha_7\|_2^2.$$

The solution gives us  $\alpha_7 = 11.00$ . Thus we get

$$\alpha^1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 11 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Again note that to save space we have presented  $\alpha$  over two rows. You should consider it as a  $20 \times 1$  column vector.

This leaves us the residual as

$$r^1 = x - \Phi\alpha^1 = \begin{pmatrix} 1.26491 & -1.26491 & -1.89737 & -1.26491 & 1.89737 \\ -1.89737 & -1.26491 & 1.89737 & -1.26491 & -1.26491 \end{pmatrix}.$$

We can cross check that the residual is indeed orthogonal to the columns already selected, for

$$\langle r^1, \phi_7 \rangle = 0.$$

Next we compute inner product of  $r^1$  with all the columns in  $\Phi$  and take absolute values. They are given by

$$\begin{pmatrix} 0.4 & 0.4 & 0.4 & 0.4 & 0.8 & 1.2 & 0 & 1.2 & 2 & 1.2 \\ 2.4 & 3.2 & 4.8 & 0 & 2 & 0.4 & 0 & 2 & 1.2 & 0.8 \end{pmatrix}$$

We quickly note that the maximum occurs at index 13 with value 4.8.

We modify our support to  $S^1 = \{7, 13\}$ .

We now solve the least squares problem

$$\text{minimize} \|x - \begin{bmatrix} \phi_7 & \phi_{13} \end{bmatrix} \begin{bmatrix} \alpha_7 \\ \alpha_{13} \end{bmatrix}\|_2^2.$$

This gives us  $\alpha_7 = 10$  and  $\alpha_{13} = -5$ .

Thus we get

$$\alpha^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & -5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Finally the residual we get at step 2 is

$$r^2 = x - \Phi\alpha^2 = 10^{-14} \begin{pmatrix} 0 & 0 & -0.111022 & 0 & 0.111022 \\ -0.111022 & 0 & 0.111022 & 0 & 0 \end{pmatrix}$$

The magnitude of residual is very small. We conclude that our OMP algorithm has converged and we have been able to recover the exact 2 sparse representation of  $x$  in  $\Phi$ .  $\square$

### 7.2.2. Exact recovery conditions

In this section following Tropp in [34] we will closely look at some conditions under which OMP is guaranteed to recover the solution for  $(\mathcal{D}, K)$ -EXACT-SPARSE problem.

It is known that  $x = \Phi\alpha$  where  $\alpha$  contains at most  $K$  non-zero entries. Both the support and entries of  $\alpha$  are known.

Let  $\Lambda_{\text{opt}} = \text{supp}(\alpha)$  i.e. the set of indices at which optimal representation  $\alpha$  has non-zero entries. Then we can write

$$x = \sum_{i \in \Lambda} \alpha_i \phi_i.$$

From the synthesis matrix  $\Phi$  we can extract a  $N \times K$  matrix  $\Phi_{\text{opt}}$  whose columns are indexed by  $\Lambda_{\text{opt}}$ .

$$\Phi_{\text{opt}} \triangleq \begin{bmatrix} \phi_{\lambda_1} & \dots & \phi_{\lambda_K} \end{bmatrix}$$

where  $\lambda_i \in \Lambda_{\text{opt}}$ .

Thus we can also write

$$x = \Phi_{\text{opt}} \alpha_{\text{opt}}$$

where  $\alpha_{\text{opt}} \in \mathbb{C}^K$  is a vector of  $K$  complex entries.

Now the columns of optimum  $\Phi_{\text{opt}}$  are linearly independent. Hence  $\Phi_{\text{opt}}$  has full column rank.

We define another matrix  $\Psi_{\text{opt}}$  whose columns are the remaining  $D - K$  columns of  $\Phi$ . Thus  $\Psi_{\text{opt}}$  consists of atoms or columns which do not participate in the optimum representation of  $x$ .

OMP starts with an empty support. In every step, it picks up one column from  $\Phi$  and adds to the support of approximation. If we can ensure that it never selects any column from  $\Psi_{\text{opt}}$  we will be guaranteed that correct  $K$  sparse representation is recovered.

We will use mathematical induction and assume that OMP has succeeded in its first  $k$  steps and has chosen  $k$  columns from  $\Phi_{\text{opt}}$  so far. At this point it is left with the residual  $r^k$ .

In  $(k+1)$ -th iteration, we compute inner product of  $r^k$  with all columns in  $\Phi$  and choose the column which has highest inner product.

We note that maximum value of inner product of  $r^k$  with any of the columns in  $\Psi_{\text{opt}}$  is given by

$$\|\Psi_{\text{opt}}^H r^k\|_\infty.$$

Correspondingly maximum value of inner product of  $r^k$  with any of the columns in  $\Phi_{\text{opt}}$  is given by

$$\|\Phi_{\text{opt}}^H r^k\|_\infty.$$

Actually since we have already shown that  $r^k$  is orthogonal to the columns already chosen, hence they will not contribute to this equation.

In order to make sure that none of the columns in  $\Psi_{\text{opt}}$  is selected, we need

$$\|\Psi_{\text{opt}}^H r^k\|_\infty < \|\Phi_{\text{opt}}^H r^k\|_\infty.$$

**Definition 7.1** [Greedy selection ratio] We define a ratio

$$\rho(r^k) \triangleq \frac{\|\Psi_{\text{opt}}^H r^k\|_\infty}{\|\Phi_{\text{opt}}^H r^k\|_\infty}. \quad (7.2.1)$$

This ratio is known as **greedy selection ratio**.

We can see that as long as  $\rho(r^k) < 1$ , OMP will make a right decision at  $(k+1)$ -th stage. If  $\rho(r^k) = 1$  then there is no guarantee that OMP will make the right decision. We will assume pessimistically that OMP makes wrong decision in such situations.

We note that this definition of  $\rho(r^k)$  looks very similar to matrix  $p$ -norms defined in ???. It is suggested to review the properties of  $p$ -norms for matrices at this point.

We now present a condition which guarantees that  $\rho(r^k) < 1$  is always satisfied.

**Theorem 7.1** [Exact recovery for OMP] *A sufficient condition for Orthogonal Matching Pursuit to resolve  $x$  completely in  $K$  steps is that*

$$\max_{\psi} \|\Phi_{\text{opt}}^\dagger \psi\|_1 < 1, \quad (7.2.2)$$

where  $\psi$  ranges over columns in  $\Psi_{\text{opt}}$ .

Moreover, Orthogonal Matching Pursuit is a correct algorithm for  $(\mathcal{D}, K)$ -EXACT-SPARSE problem whenever the condition holds for every superposition of  $K$  atoms from  $\mathcal{D}$ .

PROOF. In (7.2.2)  $\Phi_{\text{opt}}^\dagger$  is the pseudo-inverse of  $\Phi$

$$\Phi_{\text{opt}}^\dagger = (\Phi_{\text{opt}}^H \Phi_{\text{opt}})^{-1} \Phi_{\text{opt}}^H.$$

What we need to show is if (7.2.2) holds true then  $\rho(r^k)$  will always be less than 1.

We note that the projection operator for the column span of  $\Phi_{\text{opt}}$  is given by

$$\Phi_{\text{opt}}(\Phi_{\text{opt}}^H \Phi_{\text{opt}})^{-1} \Phi_{\text{opt}}^H = (\Phi_{\text{opt}}^\dagger)^H \Phi_{\text{opt}}^H.$$

We also note that by assumption since  $x \in \mathcal{C}(\Phi_{\text{opt}})$  and the approximation at the  $k$ -th step,  $x^k = \Phi \alpha^k \in \mathcal{C}(\Phi_{\text{opt}})$ , hence  $r^k = x - x^k$  also belongs to  $\mathcal{C}(\Phi_{\text{opt}})$ .

Thus

$$r^k = (\Phi_{\text{opt}}^\dagger)^H \Phi_{\text{opt}}^H r^k$$

i.e. applying the projection operator for  $\Phi_{\text{opt}}$  on  $r^k$  doesn't change it.

Using this we can rewrite  $\rho(r^k)$  as

$$\rho(r^k) = \frac{\|\Psi_{\text{opt}}^H r^k\|_\infty}{\|\Phi_{\text{opt}}^H r^k\|_\infty} = \frac{\|\Psi_{\text{opt}}^H (\Phi_{\text{opt}}^\dagger)^H \Phi_{\text{opt}}^H r^k\|_\infty}{\|\Phi_{\text{opt}}^H r^k\|_\infty}.$$

We see  $\Phi_{\text{opt}}^H r^k$  appearing both in numerator and denominator.

Now consider the matrix  $\Psi_{\text{opt}}^H (\Phi_{\text{opt}}^\dagger)^H$  and recall the definition of matrix  $\infty$ -norm from ??

$$\|A\|_\infty = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} \geq \frac{\|Ax\|_\infty}{\|x\|_\infty} \quad \forall x \neq 0.$$

Thus

$$\|\Psi_{\text{opt}}^H (\Phi_{\text{opt}}^\dagger)^H\|_\infty \geq \frac{\|\Psi_{\text{opt}}^H (\Phi_{\text{opt}}^\dagger)^H \Phi_{\text{opt}}^H r^k\|_\infty}{\|\Phi_{\text{opt}}^H r^k\|_\infty}$$

which gives us

$$\rho(r^k) \leq \|\Psi_{\text{opt}}^H (\Phi_{\text{opt}}^\dagger)^H\|_\infty = \left\| \left( \Phi_{\text{opt}}^\dagger \Psi_{\text{opt}} \right)^H \right\|_\infty.$$

Finally we recall that  $\|A\|_\infty$  is max row sum norm while  $\|A\|_1$  is max column sum norm. Hence

$$\|A\|_\infty = \|A^T\|_1 = \|A^H\|_1$$

which means

$$\left\| \left( \Phi_{\text{opt}}^\dagger \Psi_{\text{opt}} \right)^H \right\|_\infty = \|\Phi_{\text{opt}}^\dagger \Psi_{\text{opt}}\|_1.$$

Thus

$$\rho(r^k) \leq \|\Phi_{\text{opt}}^\dagger \Psi_{\text{opt}}\|_1.$$

Now the columns of  $\Phi_{\text{opt}}^\dagger \Psi_{\text{opt}}$  are nothing but  $\Phi_{\text{opt}}^\dagger \psi$  where  $\psi$  ranges over columns of  $\Psi_{\text{opt}}$ .

Thus in terms of max column sum norm

$$\rho(r^k) \leq \max_\psi \|\Phi_{\text{opt}}^\dagger \psi\|_1$$

Thus assuming that OMP has made  $k$  correct decision and  $r^k$  lies in  $\mathcal{C}(\Phi_{\text{opt}})$ ,  $\rho(r^k) < 1$  whenever

$$\max_\psi \|\Phi_{\text{opt}}^\dagger \psi\|_1 < 1. \quad (7.2.3)$$

Finally the initial residual  $r^0 = 0$  which always lies in column space of  $\Phi_{\text{opt}}$ . By above logic, OMP will always select an optimal column in each step. Since the residual is always orthogonal to the columns already selected, hence it will never select the same column twice. Thus in  $K$  steps it will retrieve all  $K$  atoms which comprise  $x$ .  $\square$

### 7.2.3. Babel function estimates

There is a small problem with theorem 7.1. Since we don't know the support a-priori hence its not possible to verify that

$$\max_\psi \|\Phi_{\text{opt}}^\dagger \psi\|_1 < 1$$

holds. Off course verifying this for all  $K$  column sub-matrices is computationally prohibitive.

It turns out that Babel function (recall definition 2.21) is there to help. We show how Babel function guarantees that exact recovery condition for OMP holds.

**Theorem 7.2** Suppose that  $\mu_1$  is the Babel function for a dictionary  $\mathcal{D}$  with synthesis matrix  $\Phi$ . The exact recovery condition holds whenever

$$\mu_1(K-1) + \mu_1(K) < 1. \quad (7.2.4)$$

*Thus, Orthogonal Matching Pursuit is a correct algorithm for  $(\mathcal{D}, K)$ -EXACT-SPARSE problem whenever (7.2.4) holds.*

*In other words, for sufficiently small  $K$  for which (7.2.4) holds, OMP will recover any arbitrary superposition of  $K$  atoms from  $\mathcal{D}$ .*

PROOF. We can write

$$\max_{\psi} \|\Phi_{\text{opt}}^\dagger \psi\|_1 = \max_{\psi} \|(\Phi_{\text{opt}}^H \Phi_{\text{opt}})^{-1} \Phi_{\text{opt}}^H \psi\|_1$$

We recall from ?? that operator-norm is subordinate i.e.

$$\|Ax\|_1 \leq \|A\|_1 \|x\|_1.$$

Thus with  $A = (\Phi_{\text{opt}}^H \Phi_{\text{opt}})^{-1}$  we have

$$\|(\Phi_{\text{opt}}^H \Phi_{\text{opt}})^{-1} \Phi_{\text{opt}}^H \psi\|_1 \leq \|(\Phi_{\text{opt}}^H \Phi_{\text{opt}})^{-1}\|_1 \|\Phi_{\text{opt}}^H \psi\|_1.$$

With this we have

$$\max_{\psi} \|\Phi_{\text{opt}}^\dagger \psi\|_1 \leq \|(\Phi_{\text{opt}}^H \Phi_{\text{opt}})^{-1}\|_1 \max_{\psi} \|\Phi_{\text{opt}}^H \psi\|_1. \quad (7.2.5)$$

Now let us look at  $\|\Phi_{\text{opt}}^H \psi\|_1$  closely. There are  $K$  columns in  $\Phi_{\text{opt}}$ . For each column we compute its inner product with  $\psi$ . And then absolute sum of the inner product.

Also recall the definition of Babel function:

$$\mu_1(K) = \max_{|\Lambda|=K} \max_{\psi} \sum_{\Lambda} |\langle \psi, \phi_{\lambda} \rangle|.$$

Clearly

$$\max_{\psi} \|\Phi_{\text{opt}}^H \psi\|_1 = \max_{\psi} \sum_{\Lambda_{\text{opt}}} |\langle \psi, \phi_{\lambda_i} \rangle| \leq \mu_1(K). \quad (7.2.6)$$

We also need to provide a bound on  $\|(\Phi_{\text{opt}}^H \Phi_{\text{opt}})^{-1}\|_1$  which requires more work.

First note that since all columns in  $\Phi$  are unit norm, hence the diagonal of  $\Phi_{\text{opt}}^H \Phi_{\text{opt}}$  contains unit entries. Thus we can write

$$\Phi_{\text{opt}}^H \Phi_{\text{opt}} = I_K + A$$

where  $A$  contains the off diagonal terms in  $\Phi_{\text{opt}}^H \Phi_{\text{opt}}$ .

Looking carefully, each column of  $A$  lists the inner products between one atom of  $\Phi_{\text{opt}}$  and the remaining  $K - 1$  atoms. By definition of Babel function

$$\|A\|_1 = \max_k \sum_{j \neq k} |\langle \phi_{\lambda_k} \phi_{\lambda_j} \rangle| \leq \mu_1(K - 1).$$

Now whenever  $\|A\|_1 < 1$  then the Von Neumann series  $\sum(-A)^k$  converges to the inverse  $(I_K + A)^{-1}$ .

Thus we have

$$\begin{aligned} \|(\Phi_{\text{opt}}^H \Phi_{\text{opt}})^{-1}\|_1 &= \|(I_K + A)^{-1}\|_1 \\ &= \left\| \sum_{k=0}^{\infty} (-A)^k \right\|_1 \\ &\leq \sum_{k=0}^{\infty} \|A\|_1^k \\ &= \frac{1}{1 - \|A\|_1} \\ &\leq \frac{1}{1 - \mu_1(K - 1)}. \end{aligned} \tag{7.2.7}$$

Thus putting things together we get

$$\max_{\psi} \|\Phi_{\text{opt}}^\dagger \psi\|_1 \leq \frac{\mu_1(K)}{1 - \mu_1(K - 1)}.$$

Thus whenever

$$\mu_1(K - 1) + \mu_1(K) < 1.$$

we have

$$\frac{\mu_1(K)}{1 - \mu_1(K - 1)} < 1 \implies \max_{\psi} \|\Phi_{\text{opt}}^\dagger \psi\|_1 < 1.$$

□

### 7.2.4. Sparse approximation conditions

We now remove the assumption that  $x$  is  $K$ -sparse in  $\Phi$ . This is indeed true for all real life signals as they are not truly sparse.

In this section we will look at conditions under which OMP can successfully solve the  $(\mathcal{D}, K)$ -SPARSE approximation problem.

Let  $x$  be an arbitrary signal and suppose that  $\alpha_{\text{opt}}$  is an optimum  $K$ -term approximation representation of  $x$ . i.e.  $\alpha_{\text{opt}}$  is a solution to  $(P_0^e)$  and the optimal  $K$ -term approximation of  $x$  is given by

$$x_{\text{opt}} = \Phi \alpha_{\text{opt}}.$$

We note that  $\alpha_{\text{opt}}$  may not be unique.

Let  $\Lambda_{\text{opt}}$  be the support of  $\alpha_{\text{opt}}$  which identifies the atoms in  $\Phi$  that participate in the  $K$ -term approximation of  $x$ .

Once again let  $\Phi_{\text{opt}}$  be the sub-matrix consisting of columns of  $\Phi$  indexed by  $\Lambda_{\text{opt}}$ .

We assume that columns in  $\Phi_{\text{opt}}$  are linearly independent. This is easily established since if any atom in this set were linearly dependent on other atoms, we could always find a more optimal solution.

Again let  $\Psi_{\text{opt}}$  be the matrix of  $(D - K)$  columns which are not indexed by  $\Lambda_{\text{opt}}$ .

We note that if  $\Lambda_{\text{opt}}$  is identified then finding  $\alpha_{\text{opt}}$  is a straightforward least squares problem.

We now present a condition under which Orthogonal Matching Pursuit is able to recover the optimal atoms.

**Theorem 7.3** [General recovery for OMP] *Assume that  $\mu_1(K) < \frac{1}{2}$ , and suppose that at  $k$ -th iteration, the support  $S^k$  for  $\alpha^k$  consists only of atoms from an optimal  $k$ -term approximation of the signal  $x$ . At step  $(k+1)$ , Orthogonal Matching Pursuit will recover another*

atom indexed by  $\Lambda_{opt}$  whenever

$$\|x - \Phi\alpha^k\|_2 > \sqrt{1 + \frac{K(1 - \mu_1(K))}{(1 - 2\mu_1(K))^2}} \|x - \Phi\alpha_{opt}\|_2. \quad (7.2.8)$$

A few remarks are in order.

- $\|x - \Phi\alpha^k\|_2$  is the approximation error norm at  $k$ -th iteration.
- $\|x - \Phi\alpha_{opt}\|_2$  is the optimum approximation error after  $K$  iterations.
- The theorem says that OMP makes absolute progress whenever the current error is larger than optimum error by a factor.
- As a result of this theorem, we note that every optimal  $K$ -term approximation of  $x$  contains the same kernel of atoms. The optimum error is always independent of choice of atoms in  $K$  term approximation (since it is optimum). Initial error is also independent of choice of atoms (since initial support is empty). OMP always selects the same set of atoms by design.

**PROOF.** Let us assume that after  $k$  steps, OMP has recovered an approximation  $x^k$  given by

$$x^k = \Phi\alpha^k$$

where  $S^k = \text{supp}(\alpha^k)$  chooses  $k$  columns from  $\Phi$  all of which belong to  $\Phi_{opt}$ .

Let the residual at  $k$ -th stage be

$$r^k = x - x^k = x - \Phi\alpha^k.$$

Recalling from previous section, a sufficient condition for recovering another optimal atom is

$$\rho(r^k) = \frac{\|\Psi_{opt}^H r^k\|_\infty}{\|\Phi_{opt}^H r^k\|_\infty} < 1. \quad (7.2.9)$$

One difference from previous section is that  $r^k \notin \mathcal{C}(\Phi_{opt})$ .

We can write

$$r^k = x - x^k = (x - x_{\text{opt}}) + (x_{\text{opt}} - x^k).$$

Note that  $(x - x_{\text{opt}})$  is nothing but the residual left after  $K$  iterations.

We also note that since residual in OMP is always orthogonal to already selected columns, hence

$$\Phi_{\text{opt}}^H(x - x_{\text{opt}}) = 0.$$

We will now use these expressions to simplify  $\rho(r^k)$ .

$$\begin{aligned} \rho(r^k) &= \frac{\|\Psi_{\text{opt}}^H r^k\|_\infty}{\|\Phi_{\text{opt}}^H r^k\|_\infty} \\ &= \frac{\|\Psi_{\text{opt}}^H(x - x_{\text{opt}}) + \Psi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty}{\|\Phi_{\text{opt}}^H(x - x_{\text{opt}}) + \Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty} \\ &= \frac{\|\Psi_{\text{opt}}^H(x - x_{\text{opt}}) + \Psi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty}{\|\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty} \\ &\leq \frac{\|\Psi_{\text{opt}}^H(x - x_{\text{opt}})\|_\infty}{\|\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty} + \frac{\|\Psi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty}{\|\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty} \end{aligned} \tag{7.2.10}$$

We now define two new terms

$$\rho_{\text{err}}(r^k) \triangleq \frac{\|\Psi_{\text{opt}}^H(x - x_{\text{opt}})\|_\infty}{\|\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty} \tag{7.2.11}$$

and

$$\rho_{\text{opt}}(r^k) \triangleq \frac{\|\Psi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty}{\|\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty}. \tag{7.2.12}$$

With these we have

$$\rho(r^k) \leq \rho_{\text{opt}}(r^k) + \rho_{\text{err}}(r^k) \tag{7.2.13}$$

Now  $x_{\text{opt}}$  has an exact  $K$ -term representation in  $\Phi$  given by  $\alpha_{\text{opt}}$ . Hence  $\rho_{\text{opt}}(r^k)$  is nothing but  $\rho(r^k)$  for corresponding EXACT-SPARSE problem.

From the proof of theorem 7.2 we recall

$$\rho_{\text{opt}}(r^k) \leq \frac{\mu_1(K)}{1 - \mu_1(K-1)} \leq \frac{\mu_1(K)}{1 - \mu_1(K)} \quad (7.2.14)$$

since

$$\mu_1(K-1) \leq \mu_1(K) \implies 1 - \mu_1(K-1) \geq 1 - \mu_1(K).$$

The remaining problem is  $\rho_{\text{err}}(r^k)$ . Let us look at its numerator and denominator one by one.

$\|\Psi_{\text{opt}}^H(x - x_{\text{opt}})\|_\infty$  essentially is the maximum (absolute) inner product between any column in  $\Psi_{\text{opt}}$  with  $x - x_{\text{opt}}$ .

We can write

$$\|\Psi_{\text{opt}}^H(x - x_{\text{opt}})\|_\infty \leq \max_{\psi} |\psi^H(x - x_{\text{opt}})| \leq \max_{\psi} \|\psi\|_2 \|x - x_{\text{opt}}\|_2 = \|x - x_{\text{opt}}\|_2$$

since all columns in  $\Phi$  are unit norm. In between we used Cauchy-Schwartz inequality.

Now look at denominator  $\|\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty$  where  $(x_{\text{opt}} - x^k) \in \mathbb{C}^N$  and  $\Phi_{\text{opt}} \in \mathbb{C}^{N \times K}$ . Thus

$$\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k) \in \mathbb{C}^K.$$

Now for every  $v \in \mathbb{C}^K$  we have

$$\|v\|_2 \leq \sqrt{K} \|v\|_\infty.$$

Hence

$$\|\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_\infty \geq K^{-1/2} \|\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_2.$$

Since  $\Phi_{\text{opt}}$  has full column rank, hence its singular values are non-zero. Thus

$$\|\Phi_{\text{opt}}^H(x_{\text{opt}} - x^k)\|_2 \geq \sigma_{\min}(\Phi_{\text{opt}}) \|x_{\text{opt}} - x^k\|_2.$$

From corollary 2.35 we have

$$\sigma_{\min}(\Phi_{\text{opt}}) \geq \sqrt{1 - \mu_1(K-1)} \geq \sqrt{1 - \mu_1(K)}.$$

Combining these observations we get

$$\rho_{\text{err}}(r^k) \leq \frac{\sqrt{K}\|x - x_{\text{opt}}\|_2}{\sqrt{1 - \mu_1(K)}\|x_{\text{opt}} - x^k\|_2}. \quad (7.2.15)$$

Now from (7.2.13)  $\rho(r^k) < 1$  whenever  $\rho_{\text{opt}}(r^k) + \rho_{\text{err}}(r^k) < 1$ .

Thus a sufficient condition for  $\rho(r^k) < 1$  can be written as

$$\frac{\mu_1(K)}{1 - \mu_1(K)} + \frac{\sqrt{K}\|x - x_{\text{opt}}\|_2}{\sqrt{1 - \mu_1(K)}\|x_{\text{opt}} - x^k\|_2} < 1. \quad (7.2.16)$$

We need to simplify this expression a bit. Multiplying by  $(1 - \mu_1(K))$  on both sides we get

$$\begin{aligned} & \mu_1(K) + \frac{\sqrt{K}\sqrt{1 - \mu_1(K)}\|x - x_{\text{opt}}\|_2}{\|x_{\text{opt}} - x^k\|_2} < 1 - \mu_1(K) \\ \implies & \frac{\sqrt{K(1 - \mu_1(K))}\|x - x_{\text{opt}}\|_2}{\|x_{\text{opt}} - x^k\|_2} < 1 - 2\mu_1(K) \\ \implies & \|x_{\text{opt}} - x^k\|_2 > \frac{\sqrt{K(1 - \mu_1(K))}}{1 - 2\mu_1(K)}\|x - x_{\text{opt}}\|_2. \end{aligned} \quad (7.2.17)$$

We assumed  $\mu_1(K) < \frac{1}{2}$  thus  $1 - 2\mu_1(K) > 0$  which validates the steps above.

Finally we remember that  $(x - x_{\text{opt}}) \perp \mathcal{C}(\Phi_{\text{opt}})$  and  $(x_{\text{opt}} - x^k) \in \mathcal{C}(\Phi_{\text{opt}})$  thus  $(x - x_{\text{opt}})$  and  $(x_{\text{opt}} - x^k)$  are orthogonal to each other. Thus by applying Pythagorean theorem we have

$$\|x - x^k\|_2^2 = \|x - x_{\text{opt}}\|_2^2 + \|x_{\text{opt}} - x^k\|_2^2.$$

Thus we have

$$\|x - x^k\|_2^2 > \frac{K(1 - \mu_1(K))}{(1 - 2\mu_1(K))^2}\|x - x_{\text{opt}}\|_2^2 + \|x - x_{\text{opt}}\|_2^2.$$

This gives us a sufficient condition

$$\|x - x^k\|_2 > \sqrt{1 + \frac{K(1 - \mu_1(K))}{(1 - 2\mu_1(K))^2}}\|x - x_{\text{opt}}\|_2. \quad (7.2.18)$$

i.e. whenever (7.2.18) holds true, we have  $\rho(r^k) < 1$  which leads to OMP making a correct choice and choosing an atom from the optimal set.

Putting  $x^k = \Phi\alpha^k$  and  $x_{\text{opt}} = \Phi\alpha_{\text{opt}}$  we get back (7.2.8) which is the desired result.  $\square$

Theorem 7.3 establishes that as long as (7.2.8) holds for each of the steps from 1 to  $K$ , OMP will recover a  $K$  term optimum approximation  $x_{\text{opt}}$ . If  $x \in \mathbb{C}^N$  is completely arbitrary, then it may not be possible that (7.2.8) holds for all the  $K$  iterations. In this situation, a question arises as to what is the worst  $K$ -term approximation error that OMP will incur if (7.2.8) doesn't hold true all the way.

This is answered in following corollary of theorem 7.3.

**Corollary 7.4.** *Assume that  $\mu_1(K) < \frac{1}{2}$  and let  $x \in \mathbb{C}^N$  be a completely arbitrary signal. Orthogonal Matching Pursuit produces a  $K$ -term approximation  $x^K$  which satisfies*

$$\|x - x^K\|_2 \leq \sqrt{1 + C(\mathcal{D}, K)} \|x - x_{\text{opt}}\|_2 \quad (7.2.19)$$

where  $x_{\text{opt}}$  is the optimum  $K$ -term approximation of  $x$  in dictionary  $\mathcal{D}$  (i.e.  $x_{\text{opt}} = \Phi\alpha_{\text{opt}}$  where  $\alpha_{\text{opt}}$  is an optimal solution of  $(P_0^\epsilon)$ ).  $C(\mathcal{D}, K)$  is a constant depending upon the dictionary  $\mathcal{D}$  and the desired sparsity level  $K$ . An estimate of  $C(\mathcal{D}, K)$  is given by

$$C(\mathcal{D}, K) \leq \frac{K(1 - \mu_1(K))}{(1 - 2\mu_1(K))^2}. \quad (7.2.20)$$

**PROOF.** Suppose that OMP runs fine for first  $p$  steps where  $p < K$ . Thus (7.2.8) keeps holding for first  $p$  steps. We now assume that (7.2.8) breaks at step  $p + 1$  and OMP is no longer guaranteed to make an optimal choice of column from  $\Phi_{\text{opt}}$ . Thus at step  $p + 1$  we have

$$\|x - x^p\|_2 \leq \sqrt{1 + \frac{K(1 - \mu_1(K))}{(1 - 2\mu_1(K))^2}} \|x - x_{\text{opt}}\|_2. \quad (7.2.21)$$

Any further iterations of OMP will only reduce the error further (although not in an optimal way). This gives us

$$\|x - x^K\|_2 \leq \|x - x^p\|_2 \leq \sqrt{1 + \frac{K(1 - \mu_1(K))}{(1 - 2\mu_1(K))^2}} \|x - x_{\text{opt}}\|_2. \quad (7.2.22)$$

Choosing

$$C(\mathcal{D}, K) = \frac{K(1 - \mu_1(K))}{(1 - 2\mu_1(K))^2}$$

we can rewrite this as

$$\|x - x^K\|_2 \leq \sqrt{1 + C(\mathcal{D}, K)} \|x - x_{\text{opt}}\|_2.$$

□

This is a very useful result. It establishes that even if OMP is not able to recover the optimum  $K$ -term representation of  $x$ , it always constructs an approximation whose error lies within a constant factor of optimum approximation error where the constant factor is given by  $\sqrt{1 + C(\mathcal{D}, K)}$ .

If the optimum approximation error  $\|x - x_{\text{opt}}\|_2$  is small then  $\|x - x^K\|_2$  will also be not too large.

If  $\|x - x_{\text{opt}}\|_2$  is moderate, then the OMP may inflate the approximation error to a higher value. But in this case, probably sparse approximation is not the right tool for signal representation over the dictionary.

### 7.3. Orthogonal Matching Pursuit for Compressed Sensing

We now switch gears and consider adapting OMP for the problem of compressed sensing. This section is largely based on [38].

We start by fixing the notation recalling from definition 2.23. We will restrict our attention to the  $N$  dimensional Euclidean space as our signal space for the moment.  $x \in \mathbb{R}^N$  is a signal vector.  $\Phi \in \mathbb{R}^{M \times N}$  is the sensing matrix with  $M \ll N$ . The measurement vector  $y \in \mathbb{R}^M$  is given by

$$y = \Phi x$$

where  $\mathbb{R}^M$  is our measurement space.  $y$  is known,  $\Phi$  is known while  $x$  is unknown.  $x$  is assumed to be either  $K$ -sparse or  $K$ -compressible.

The sparse recovery problem can be written as

$$\begin{aligned} & \underset{x}{\text{minimize}} \quad \|y - \Phi x\|_2 \\ & \text{subject to} \quad \|x\|_0 \leq K. \end{aligned} \tag{7.3.1}$$

Though the problem looks similar to  $(\mathcal{D}, K)$ -SPARSE approximation problem, but there are differences since  $\Phi$  is not a dictionary (see section 2.8.1).

We will adapt OMP algorithm studied in previous section for the problem of sparse recovery in compressed sensing framework.

In the analysis of OMP for CS We will address following questions :

- How many measurements are required to recover  $x$  from  $y$  exactly if  $x$  is  $K$ -sparse?
- What kind of sensing matrices are admissible for OMP to work in CS framework?
- If  $x$  is not  $K$ -sparse, then how much maximum error is incurred?

### 7.3.1. The algorithm

OMP algorithm adapted to CS is presented in fig. 7.2.

Some remarks are in order

- The algorithm returns a  $K$ -term approximation of  $x$  given by  $\hat{x}$ .
- Each step of algorithm is identified by the iteration counter  $k$  which runs from 0 to  $K$ .
- At each step  $x^k$ ,  $y^k$  and  $r^k$  are estimated where  $x^k$  is the  $k$ -term estimate of  $x$ ,  $y^k$  is corresponding measurement vector and  $r^k$  is the residual between actual measurement vector  $y$  and the estimated measurement vector  $y^k$ .

- The support for  $\hat{x}$  is maintained in an index set  $\Lambda$ .
- At each iteration we add one more new index  $\lambda_k$  to  $\Lambda^{k-1}$  giving us  $\Lambda^k$ .
- We will use  $\Phi_{\Lambda^k} \in \mathbb{R}^{M \times k}$  to denote the sub-matrix constructed from the columns indexed by  $\Lambda^k$ . i.e. If  $\Lambda^k = \{\lambda_1, \dots, \lambda_k\}$ , then

$$\Phi_{\Lambda^k} = [\phi_{\lambda_1} \ \dots \ \phi_{\lambda_k}] .$$

- Similarly we will denote a vector  $x_{\Lambda^k}^k \in \mathbb{R}^k$  to denote a vector consisting of only  $k$  non-zero entries in  $x$ .
- We note that  $r^k$  is orthogonal to  $\Phi_{\Lambda^k}$ . This is true due to  $x^k$  being the least squares solution in the update provisional solution step.
- This also ensures that in each iteration a new column from  $\Phi$  indexed by  $\lambda_k$  will be chosen. OMP will never choose the same column again.
- In case  $x$  has a sparsity level less than  $K$  then  $r^k$  will become zero in the middle. At that point we halt. There is no point going forward.
- The algorithm spends most of its time in the update provisional solution step where it needs to compute the least square solution.
- An equivalent formulation of the least squares step is

$$z \leftarrow \underset{v \in \mathbb{R}^k}{\text{minimize}} \| \Phi_{\Lambda^k} v - y \|_2^2$$

followed by

$$x^k(\Lambda^k) \leftarrow z.$$

- Essentially we solve the least squares problem for columns of  $\Phi$  indexed by  $\Lambda^k$  and then assign the  $k$  entries in the resultant  $z$  to the entries in  $x^k$  indexed by  $\Lambda^k$  while keeping other entries as 0.
- Least squares can be accelerated by using  $x^{k-1}$  as the starting estimate of  $x^k$  and carrying out a descent like Richardson's iteration from there.

### 7.3.2. Signal recovery using OMP with random sensing matrices

The objective of this section is to demonstrate that OMP can recover sparse signals from a small set of random linear measurements. In this subsection we discuss the conditions on the random sensing matrix  $\Phi$  under which they are suitable for signal recovery through OMP.

**Definition 7.2** [Admissible sensing matrices for OMP] An **admissible sensing matrix** for  $K$ -sparse signals in  $\mathbb{R}^M$  is an  $M \times N$  random matrix  $\Phi$  with following properties.

**M0:** *Independence:* The columns of  $\Phi$  are stochastically independent.

**M1:** *Normalization:*  $\mathbb{E}(\|\phi_j\|_2^2) = 1$  for  $j = 1, \dots, N$ .

**M2:** *Joint correlation:* Let  $\{u_k\}$  be a sequence of  $K$  vectors whose  $l_2$  norms do not exceed one. Let  $\phi$  be a column of  $\Phi$  that is independent from this sequence. Then

$$\mathbb{P}\left(\max_k |\langle \phi, u_k \rangle| \leq \epsilon\right) \geq 1 - 2K \exp(-c\epsilon^2 M). \quad (7.3.2)$$

**M3:** *Smallest singular value:* Given an  $M \times K$  ( $K < M$ ) submatrix  $Z$  from  $\Phi$ , the  $K$ -th largest singular value  $\sigma_{\min}(Z)$  satisfies

$$\mathbb{P}(\sigma_{\min}(Z) \geq 0.5) \geq 1 - \exp(-cM). \quad (7.3.3)$$

It can be shown Rademacher sensing matrices (section 5.3) and Gaussian sensing matrices (section 5.4) satisfy all the requirements of admissible sensing matrices for sparse recovery using OMP. Some of the proofs are included in the book. You may want to review corresponding sections.

Some remarks are in order to further explain the definition of admissible sensing matrices.

- Usually all the columns of a sensing matrix are drawn from the same distribution. But (M0) doesn't require so. It allows different columns of  $\Phi$  to be drawn from different distributions.
- The joint correlation property (M2) depends on the decay of random variables  $\|\phi_j\|_2$ . i.e. it needs the tails of  $\|\phi_j\|_2$  to be small.
- A bound on the smallest (non-zero) singular value of  $M \times K$ -sub-matrices (M3) controls how much the sensing matrix can shrink  $K$ -sparse vectors.
- I guess that the idea of admissible matrices came as follows. First OMP signal recovery guarantees were developed for Gaussian and Rademacher sensing matrices. Then the proofs were analyzed to identify the minimum requirements they imposed on the structure of random sensing matrices. This was extracted in the form of notion of admissible matrices. Finally the proof was reorganized to work for all random matrices which satisfy the admissibility criteria. It is important to understand this process of abstraction otherwise we just get surprised as to how the ideas like admissible matrices came out of the blue.

### 7.3.3. Signal recovery guarantees with OMP

We now show that OMP can be used to recover the original signal with high probability if the random measurements are taken using an admissible sensing matrix as described in previous section.

Here we consider the case where  $x$  is known to be  $K$ -sparse.

**Theorem 7.5** [OMP with admissible sensing matrices] *Fix some  $\delta \in (0, 1)$ , and choose  $M \geq CK \ln(\frac{N}{\delta})$  where  $C$  is an absolute constant. Suppose that  $x$  is an arbitrary  $K$ -sparse signal in  $\mathbb{R}^N$  and draw a  $M \times N$  admissible sensing matrix  $\Phi$  independent from  $x$ .*

*Given the measurement vector  $y = \Phi x \in \mathbb{R}^M$ , Orthogonal Matching Pursuit can reconstruct the signal with probability exceeding  $1 - \delta$ .*

Some remarks are in order. Specifically we compare OMP here with basis pursuit (BP).

- The theorem provides probabilistic guarantees.
- The theorem actually requires more measurements than the results for BP.
- The biggest advantage is that OMP is a much simpler algorithm than BP and works very fast.
- Results for BP show that a single random sensing matrix can be used for recovering all sparse signals. This theorem says that any sparse signal independent from the sensing matrix can be recovered. Thus this theorem is weaker than the results for BP.
- It can be argued that for practical situations, this limitation doesn't matter much.

**PROOF.** The main challenge here is to handle the issues that arise due to random nature of  $\Phi$ .

We start with setting up some notation for this proof.

We note that the columns that OMP chooses do not depend on the order in which they are stacked in  $\Phi$ . Thus without loss of generality we can assume that the first  $K$  entries of  $x$  are non-zero and rest are zero. If OMP picks up the first  $K$  columns, then OMP has succeeded otherwise it has failed. With this, support of  $x$  given by  $\Lambda_{\text{opt}} = \{1, \dots, K\}$ .

We now partition the sensing matrix  $\Phi$  as

$$\Phi = [\Phi_{\text{opt}} \mid \Psi]$$

where  $\Phi_{\text{opt}}$  consists of first  $K$  columns of  $\Phi$  which correspond to  $\Lambda_{\text{opt}}$ .  $\Psi$  consists of remaining  $(N - K)$  columns of  $\Phi$ .

We recall from the proof of theorem 7.1 that in order for OMP to make absolute progress at step  $k + 1$  we require the *greedy selection ratio*  $\rho(r^k) < 1$  where

$$\rho(r^k) = \frac{\|\Psi^H r^k\|_\infty}{\|\Phi_{\text{opt}}^H r^k\|_\infty} = \frac{\max_{\psi \in \Psi} |\langle \psi, r^k \rangle|}{\|\Phi_{\text{opt}}^H r^k\|_\infty}.$$

The proof is organized as follows:

- We first construct a thought experiment in which  $\Psi$  is not present and OMP is run only with  $y$  and  $\Phi_{\text{opt}}$ .
- We then run OMP with  $\Psi$  present under the condition  $\rho(r^k) < 1$ .
- We show that the sequence of columns chosen and residual obtained in both cases is exactly the same.
- We show that the residuals obtained in the thought experiment are stochastically independent from the columns of  $\Psi$ .
- We then describe the success of OMP as an event in terms of these residuals.
- We compute a lower bound on the probability of the event of OMP success.

For a moment suppose that there was no  $\Psi$  and OMP is run with  $y$  and  $\Phi_{\text{opt}}$  as input for  $K$  iterations. Naturally OMP will choose  $K$  columns in  $\Phi_{\text{opt}}$  one by one.

Let the columns it picks up in each step be indexed by  $\omega_1, \omega_2, \dots, \omega_K$ .

Let the residuals before each step be  $q^0, q^1, q^2, \dots, q^{K-1}$ . Since  $x \in \mathcal{C}(\Phi_{\text{opt}})$ , hence the residual after  $K$  iterations  $q^K = 0$ .

Since OMP is a deterministic algorithm, hence the two sequences are simply functions of  $x$  and  $\Phi_{\text{opt}}$ .

Clearly, we can say that the residual  $q^k$  are stochastically independent of the columns in  $\Psi$  (since columns of  $\Psi$  are independent of the columns of  $\Phi_{\text{opt}}$  ).

We also know that  $q^k \in \mathcal{C}(\Phi_{\text{opt}})$ .

In this thought experiment we made no assumptions about  $\rho(q^k)$  since  $\Psi$  is not present.

We now consider the full matrix  $\Phi$  and execute OMP with  $y$ .

The actual sequence of residuals before each step is  $r^0, r^1, \dots, r^{K-1}$ .

The actual sequence of column indices is  $\lambda_1, \dots, \lambda_K$ .

Clearly OMP succeeds in recovering  $x$  in  $K$  steps if and only if it selects the first  $K$  columns of  $\Phi$  in some order. This can happen if and only if  $\rho(r^k) < 1$  holds.

We are going to show inductively that this can happen if and only if  $\lambda_k = \omega_k$  and  $q^k = r^k$ .

At the beginning of step 1, we have  $r^0 = q^0 = y$ . Now OMP selects one column from  $\Phi_{\text{opt}}$  if and only if  $\rho(r^0) < 1$  which is identical to  $\rho(q^0) < 1$ .

So it remains to show at step 1 that  $\lambda_1 = \omega_1$ .

Because  $\rho(r^0) < 1$ , the algorithm selects the index  $\lambda_1$  of the column from  $\Phi_{\text{opt}}$  whose inner product with  $r^0$  is the largest (in absolute value).

Also since  $\rho(q^0) < 1$  with  $r^0 = q^0$ ,  $\omega_1$  is the index of column in  $\Phi_{\text{opt}}$  whose inner product with  $q^0$  is largest, thus  $\omega_1 = \lambda_1$ .

We now assume that for the first  $k$  iterations, actual OMP chooses the same columns as our imaginary thought experiment. Thus we have

$$\lambda_j = \omega_j \quad \forall 1 \leq j \leq k$$

and

$$r^j = q^j \quad \forall 0 \leq j \leq k.$$

This is valid since the residuals at each step depend solely on the set of columns chosen so far and input  $y$  which are same for both cases.

Clearly OMP choose a column in  $\Phi_{\text{opt}}$  at  $(k+1)$ -th step if and only if  $\rho(r^k) < 1$  which is same as  $\rho(q^k) < 1$ .

Moreover since  $r^k = q^k$  hence the column chosen by maximizing the inner product is same for both situations. Thus

$$\lambda_{k+1} = \omega_{k+1}.$$

Therefore the criteria for success of OMP can be stated as  $\rho(q^k) < 1$  for all  $0 \leq k \leq K - 1$ .

We now recall that  $q^k$  is actually a random variable (depending upon the random vectors which comprise the columns of  $\Phi_{\text{opt}}$ ).

Thus the event on which the algorithm succeeds in sparse recovery of  $x$  from  $y$  is given by

$$E_{\text{succ}} \triangleq \left\{ \max_{0 \leq k < K} \rho(q^k) < 1 \right\}.$$

In a particular instance of OMP execution if the event  $E_{\text{succ}}$  happens, then OMP successfully recovers  $x$  from  $y$ . Otherwise OMP fails. So the probability of success of OMP is same as the probability of event  $E_{\text{succ}}$ . We will be looking for some sort of a lower bound on  $\mathbb{P}(E_{\text{succ}})$ .

We note that we have  $\{q^k\}$  as a sequence of random vectors in the column span of  $\Phi_{\text{opt}}$  and they are stochastically independent from columns of  $\Psi$ .

Its difficult to compute  $\mathbb{P}(E_{\text{succ}})$  directly. We consider another event  $\Gamma = \{\sigma_{\min}(\Phi_{\text{opt}}) \geq 0.5\}$ . Clearly

$$\mathbb{P}(E_{\text{succ}}) \geq \mathbb{P} \left( \max_{0 \leq k < K} \rho(q^k) < 1 \text{ and } \Gamma \right).$$

Using conditional probability we can rewrite

$$\mathbb{P}(E_{\text{succ}}) \geq \mathbb{P} \left( \max_{0 \leq k < K} \rho(q^k) < 1 | \Gamma \right) \mathbb{P}(\Gamma).$$

Since  $\Phi$  is an admissible matrix hence it satisfies (M3) which gives us

$$\mathbb{P}(\Gamma) \geq 1 - \exp(-cM).$$

We just need a lower bound on the conditional probability.

We assume that  $\Gamma$  occurs. For each step index  $k = 0, 1, \dots, K - 1$ , we have

$$\rho(q^k) = \frac{\max_{\psi} |\langle \psi, q^k \rangle|}{\|\Phi_{\text{opt}}^H q^k\|_{\infty}}.$$

Since  $\Phi_{\text{opt}}^H q^k \in \mathbb{R}^K$ , we have

$$\sqrt{K} \|\Phi_{\text{opt}}^H q^k\|_{\infty} \geq \|\Phi_{\text{opt}}^H q^k\|_2.$$

This gives us

$$\rho(q^k) \leq \frac{\sqrt{K} \max_{\psi} |\langle \psi, q^k \rangle|}{\|\Phi_{\text{opt}}^H q^k\|_2}.$$

To simplify this expression, we define a vector

$$u^k \triangleq \frac{0.5 q^k}{\|\Phi_{\text{opt}}^H q^k\|_2}.$$

This lets us write

$$\rho(q^k) \leq 2\sqrt{K} \max_{\psi} |\langle \psi, u^k \rangle|.$$

Thus

$$\mathbb{P}(\rho(q^k) < 1 | \Gamma) \geq \mathbb{P}(2\sqrt{K} \max_{\psi} |\langle \psi, u^k \rangle| < 1 | \Gamma) = \mathbb{P}\left(\max_{\psi} |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}} |\Gamma\right).$$

From the basic properties of singular values we recall that

$$\frac{\|\Phi_{\text{opt}}^H q\|_2}{\|q\|_2} \geq \sigma_{\min}(\Phi_{\text{opt}}) \geq 0.5$$

for all vectors  $q$  in the range of  $\Phi_{\text{opt}}$ .

This gives us

$$\frac{0.5 \|q\|_2}{\|\Phi_{\text{opt}}^H q\|_2} \leq 1.$$

Since  $q^k$  is in the column space of  $\Phi_{\text{opt}}$ , for  $u^k$  defined above we have

$$\|u^k\|_2 \leq 1.$$

From the above we get

$$\mathbb{P} \left( \max_k \rho(q^k) < 1 | \Gamma \right) \geq \mathbb{P} \left( \max_k \max_{\psi} |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}} | \Gamma \right)$$

In the R.H.S. we can exchange the order of two maxima. This gives us

$$\mathbb{P} \left( \max_k \max_{\psi} |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}} | \Gamma \right) = \mathbb{P} \left( \max_{\psi} \max_k |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}} | \Gamma \right).$$

We also note that columns of  $\Psi$  are independent. Thus in above we require that for each column of  $\Psi$   $\max_k |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}}$  should hold independently. Hence we can say

$$\mathbb{P} \left( \max_{\psi} \max_k |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}} | \Gamma \right) = \prod_{\psi} \mathbb{P} \left( \max_k |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}} | \Gamma \right).$$

We recall that event  $\Gamma$  depends only on columns of  $\Phi_{\text{opt}}$ . Hence columns of  $\Psi$  are independent of  $\Gamma$ . Thus

$$\prod_{\psi} \mathbb{P} \left( \max_k |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}} | \Gamma \right) = \prod_{\psi} \mathbb{P} \left( \max_k |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}} \right).$$

Finally since the sequence  $\{u^k\}$  depends only on columns of  $\Phi_{\text{opt}}$ , hence columns of  $\Psi$  are independent of  $\{u^k\}$ , thus we can take help of (M2) to get

$$\prod_{\psi} \mathbb{P} \left( \max_k |\langle \psi, u^k \rangle| < \frac{1}{2\sqrt{K}} \right) \geq \left( 1 - 2K \exp \left( -\frac{cM}{4K} \right) \right)^{N-K}.$$

This gives us the lower bound

$$\mathbb{P} \left( \max_k \rho(q^k) < 1 | \Gamma \right) \geq \left( 1 - 2K \exp \left( -\frac{cM}{4K} \right) \right)^{N-K}.$$

Finally plugging in the lower bound for  $\mathbb{P}(\Gamma)$  we get

$$\mathbb{P}(E_{\text{succ}}) \geq \left( 1 - 2K \exp \left( -\frac{cM}{4K} \right) \right)^{N-K} (1 - \exp(-cM)). \quad (7.3.4)$$

All that is remaining now is to simplify this expression.

We recall that we assumed in the theorem statement

$$\begin{aligned}
 M &\geq CK \ln \left( \frac{N}{\delta} \right) \\
 \implies \frac{M}{CK} &\geq \ln \left( \frac{N}{\delta} \right) \\
 \implies \exp \left( \frac{M}{CK} \right) &\geq \frac{N}{\delta} \\
 \implies \frac{\delta}{N} &\geq \exp \left( -\frac{M}{CK} \right) \\
 \implies \delta &\geq N \exp \left( -\frac{M}{CK} \right).
 \end{aligned} \tag{7.3.5}$$

But we assumed that  $0 < \delta < 1$ , thus

$$N \exp \left( -\frac{M}{CK} \right) < 1.$$

If we choose  $C \geq \frac{4}{c}$  then

$$\begin{aligned}
 -\frac{1}{C} &\geq -\frac{c}{4} \\
 \implies -\frac{M}{CK} &\geq -\frac{cM}{4K} \\
 \implies \exp \left( -\frac{M}{CK} \right) &\geq \exp \left( -\frac{cM}{4K} \right) \\
 \implies N \exp \left( -\frac{M}{CK} \right) &\geq 2K \exp \left( -\frac{cM}{4K} \right) \\
 \implies 1 > \delta &\geq 2K \exp \left( -\frac{cM}{4K} \right)
 \end{aligned} \tag{7.3.6}$$

where we assumed that  $N \gg 2K$ .

We recall that

$$(1 - x)^k \geq 1 - kx \text{ if } k \geq 1 \text{ and } x \leq 1.$$

Applying on (7.3.4) we get

$$\mathbb{P}(E_{\text{succ}}) \geq 1 - 2K(N - K) \exp \left( -\frac{cM}{4K} \right) - \exp(-cM). \tag{7.3.7}$$

We ignored the 4-th term in this expansion.

Now we can safely assume that  $K(N - K) \geq \frac{N^2}{4}$  giving us

$$\mathbb{P}(E_{\text{succ}}) \geq 1 - \frac{N^2}{2} \exp\left(-\frac{cM}{4K}\right) - \exp(-cM).$$

If we choose  $C \geq \frac{8}{c}$  then following (7.3.6) we can get

$$\begin{aligned} N \exp\left(-\frac{M}{CK}\right) &\geq N \exp\left(-\frac{cM}{8K}\right) \\ \implies \delta &\geq N \exp\left(-\frac{cM}{8K}\right) \\ \implies \delta^2 &\geq N^2 \exp\left(-\frac{cM}{4K}\right) \\ \implies 1 - \frac{\delta^2}{2} &\leq 1 - \frac{N^2}{2} \exp\left(-\frac{cM}{4K}\right). \end{aligned} \tag{7.3.8}$$

Thus

$$\mathbb{P}(E_{\text{succ}}) \geq 1 - \frac{\delta^2}{2} - \exp(-cM).$$

Some further simplification can give us

$$\mathbb{P}(E_{\text{succ}}) \geq 1 - \delta.$$

Thus with a suitable choice of the constant  $C$ , a choice of  $M \geq CK \ln\left(\frac{N}{\delta}\right)$  with  $\delta \in (0, 1)$  is sufficient to reduce the failure probability below  $\delta$ .  $\square$

```

Input: Sensing matrix  $\Phi \in \mathbb{R}^{M \times N}$ 
Input: Measurement vector  $y \in \mathbb{R}^M$ 
Input: Sparsity level  $K$ 
Output: A  $K$ -sparse estimate  $\hat{x} \in \Sigma_K \subseteq \mathbb{R}^N$  for the ideal signal  $x$ 
Output: An index set  $\Lambda^K \subset \{1, \dots, N\}$  identifying the support of  $\hat{x}$ 
Output: An approximation  $y^K \in \mathbb{R}^M$  of  $y$ 
Output: A residual  $r^K = y - y^K \in \mathbb{R}^M$ 

// Initialization
 $k \leftarrow 0$ ; // Iteration counter
 $x^0 \leftarrow 0$ ; // Estimate of  $x \in \mathbb{R}^N$ 
 $y^0 \leftarrow \Phi x^0$  // Approximation of  $y$ 
 $r^0 \leftarrow y - y^0$ ; // Residual  $r \in \mathbb{R}^M$ 
 $\Lambda^0 \leftarrow \emptyset$ ; // Solution support  $\Lambda = \text{supp}(\hat{x})$ 

while  $k < K$  do
     $k \leftarrow k + 1$ ; // Increase the iteration count
    // Sweep
     $\lambda_k = \arg \max_{1 \leq j \leq N} |\langle r^{k-1}, \phi_j \rangle|$ ; // maximum inner product
    // Update support
     $\Lambda^k \leftarrow \Lambda^{k-1} \cup \{\lambda_k\}$ ;
    // Update provisional solution
     $x^k \leftarrow \underset{x}{\text{minimize}} \|\Phi x - y\|_2^2$  subject to  $\text{supp}(x) = \Lambda^k$ ;
    // Update residual
     $y^k = \Phi x^k$ ;
     $r^k = y - y^k$ ;
    if  $r^k = 0$  then
        break;
    end
end
 $\hat{x} \leftarrow x^k$ ; // Return estimate

```

FIGURE 7.2. Orthogonal matching pursuit for sparse recovery in CS

#### 7.4. Analysis of OMP using Restricted Isometry Property

In this section we present an alternative analysis of OMP algorithm using the Restricted Isometry Property of the matrix  $\Phi$  [17].

```

 $\hat{\alpha}, r, \hat{\Lambda} = \text{OMP}(\Phi, x);$ 
 $\alpha^0 \leftarrow 0;$ 
 $r^0 \leftarrow x; \quad // r = x - \Phi\alpha^0$ 
 $\Lambda^0 = \emptyset; \quad // \text{Index set of chosen atoms}$ 
 $k \leftarrow 0; \quad // \text{Iteration counter}$ 
repeat
     $h^{k+1} \leftarrow \Phi^H r^k; \quad // \text{Match}$ 
     $\lambda^{k+1} = \arg \max_{j \notin \Lambda^k} |h_j^{k+1}|; \quad // \text{Identify}$ 
     $\Lambda^{k+1} \leftarrow \Lambda^k \cup \{\lambda^{k+1}\}; \quad // \text{Update support}$ 
     $\alpha^{k+1} \leftarrow 0;$ 
     $\alpha_{\Lambda^{k+1}}^{k+1} \leftarrow \Phi_{\Lambda^{k+1}}^\dagger x; \quad // \text{Update representation LS}$ 
     $x^{k+1} = \Phi \alpha^{k+1}; \quad // \text{Update approximation}$ 
     $r^{k+1} \leftarrow x - x^{k+1}; \quad // \text{Update residual}$ 
     $k \leftarrow k + 1; \quad // \text{Update iteration counter}$ 
until halting criteria is satisfied;
 $\hat{\alpha} \leftarrow \alpha^k; \hat{\Lambda} \leftarrow \Lambda^k; r \leftarrow r^k;$ 

```

FIGURE 7.3. Orthogonal matching pursuit [17]

The OMP algorithm is presented again in fig. 7.3. We will follow the notation of section 7.2 with slight changes and additions as explained below.

### 7.4.1. A re-look at the OMP algorithm

Before we get into the RIP based analysis of OMP, it would be useful to get some new insights into the behavior of OMP algorithm. These insights will help us a lot in performing the analysis later.

We will use the symbol  $\Lambda$  to denote the index set of chosen atoms at any stage during the algorithm execution. Earlier we used the symbol  $S$ .

We will assume throughout that whenever  $|\Lambda| \leq K$ , then  $\Phi_\Lambda$  is full rank.

Naturally the pseudo-inverse is given by

$$\Phi_\Lambda^\dagger = (\Phi_\Lambda^H \Phi_\Lambda)^{-1} \Phi_\Lambda^H. \quad (7.4.1)$$

The orthogonal projection operator to the column space for  $\Phi_\Lambda$  is given by

$$P_\Lambda = \Phi_\Lambda \Phi_\Lambda^\dagger. \quad (7.4.2)$$

The orthogonal projection operator onto the orthogonal complement of  $\mathcal{C}(\Phi_\Lambda)$  (column space of  $\Phi_\Lambda$ ) is given by

$$P_\Lambda^\perp = I - P_\Lambda. \quad (7.4.3)$$

Both  $P_\Lambda$  and  $P_\Lambda^\perp$  satisfy the usual properties like  $P = P^H$  and  $P^2 = P$ .

We further define

$$\Psi_\Lambda = P_\Lambda^\perp \Phi. \quad (7.4.4)$$

We are orthogonalizing the columns in  $\Phi$  against  $\mathcal{C}(\Phi_\Lambda)$ , i.e. taking the component of the column which is orthogonal to the column space of  $\Phi_\Lambda$ . Obviously the columns in  $\Psi_\Lambda$  corresponding to the index set  $\Lambda$  would be 0.

We will make some further observations on the behavior of OMP algorithm [17].

Recall that the approximation after the  $k$ -th iteration is given by

$$\alpha_{\Lambda^k}^k = \Phi_{\Lambda^k}^\dagger x \quad \text{and} \quad \alpha_{\Lambda^{k^c}}^k = 0.$$

The residual after  $k$ -th iteration is given by

$$r^k = x - \Phi \alpha^k$$

and by construction  $r^k$  is orthogonal to  $\Phi_{\Lambda^k}$ .

We can actually write

$$\Phi \alpha^k = \Phi_{\Lambda} \alpha_{\Lambda^k}^k + \Phi_{\Lambda^{k^c}} \alpha_{\Lambda^{k^c}}^k = \Phi_{\Lambda^k} \alpha_{\Lambda^k}^k.$$

Thus,

$$\begin{aligned} r^k &= x - \Phi_{\Lambda^k} \alpha_{\Lambda^k}^k \\ &= x - \Phi_{\Lambda^k} \Phi_{\Lambda^k}^\dagger x \\ &= (I - P_{\Lambda^k})x = P_{\Lambda^k}^\perp x. \end{aligned}$$

In summary

$$r^k = P_{\Lambda^k}^\perp x. \quad (7.4.5)$$

This shows that it is not actually necessary to compute  $\alpha^k$  in order to find  $r^k$ . An equivalent way of writing OMP algorithm could be as in fig. 7.4.

```

while halting criteria do
     $h^{k+1} \leftarrow \Phi^H r^k;$  // Match
     $\lambda^{k+1} \leftarrow \arg \max_{j \notin \Lambda^k} |h_j^{k+1}|;$  // Identify
     $\Lambda^{k+1} \leftarrow \Lambda^k \cup \{\lambda^{k+1}\};$  // Update support
     $r^{k+1} \leftarrow P_{\Lambda^{k+1}}^\perp x;$  // Update residual
     $k \leftarrow k + 1;$ 
end
 $\hat{\alpha}_{\Lambda^k} \leftarrow \Phi_{\Lambda^k}^\dagger x;$ 
 $\hat{\alpha}_{\Lambda^{k^c}} \leftarrow 0;$ 

```

FIGURE 7.4. Sketch of OMP without intermediate  $\alpha^k$  computation

In the matching step, we are correlating  $r^k$  with columns of  $\Phi$ . Since  $r^k$  is orthogonal to column space of  $\Phi_{\Lambda^k}$ , hence this correlation is identical to correlating  $r^k$  with  $\Psi_{\Lambda^k}$ .

To see this, observe that

$$r^k = P_{\Lambda^k}^\perp x = P_{\Lambda^k}^\perp P_{\Lambda^k}^\perp x = (P_{\Lambda^k}^\perp)^H P_{\Lambda^k}^\perp x. \quad (7.4.6)$$

Thus,

$$\begin{aligned} h^{k+1} &= \Phi^H r^k = \Phi^H (P_{\Lambda^k}^\perp)^H P_{\Lambda^k}^\perp x \\ &= (P_{\Lambda^k}^\perp \Phi)^H P_{\Lambda^k}^\perp x = (\Psi_{\Lambda^k})^H r^k. \end{aligned} \quad (7.4.7)$$

On similar lines, we can also see that

$$h^{k+1} = \Phi^H r^k = \Phi^H P_{\Lambda^k}^\perp x = \Phi^H (P_{\Lambda^k}^\perp)^H x = (\Psi_{\Lambda^k})^H x.$$

i.e. we have

$$h^{k+1} = (\Psi_{\Lambda^k})^H r^k = (\Psi_{\Lambda^k})^H x. \quad (7.4.8)$$

Thus, we can observe that OMP can be further simplified and we don't even need to compute  $r^k$  in order to compute  $h^{k+1}$ .

There is one catch though. If the halting criterion depends on the need to compute the residual energy, then we certainly need to compute  $r^k$ . If the halting criteria is simply the number of  $K$  iterations, then we don't need to compute  $r^k$ .

The revised OMP algorithm sketch is presented in fig. 7.5.

```
while halting criteria do
     $h^{k+1} \leftarrow (\Psi_{\Lambda^k})^H x;$  // Match
     $\lambda^{k+1} \leftarrow \arg \max_{i \notin \Lambda^k} |h_i^{k+1}|;$  // Identify
     $\Lambda^{k+1} \leftarrow \Lambda^k \cup \{\lambda^{k+1}\};$  // Update support
     $k \leftarrow k + 1;$ 
end
 $\hat{\alpha}_{\Lambda^k} \leftarrow \Phi_{\Lambda^k}^\dagger x;$ 
 $\hat{\alpha}_{\Lambda^k c} \leftarrow 0;$ 
```

FIGURE 7.5. Sketch of OMP without intermediate  $\alpha^k$  computation

With this the OMP algorithm is considerably simplified from the perspective of analyzing its recovery guarantees.

Coming back to  $h^{k+1}$ , note that the columns of  $\Psi_{\Lambda^k}$  indexed by  $\Lambda^k$  are all 0s. Thus

$$h_j^{k+1} = 0 \quad \forall j \in \Lambda^k. \quad (7.4.9)$$

This makes it obvious that  $\lambda^{k+1} \notin \Lambda$  and consequently  $|\Lambda^k| = k$  (inductively).

Lastly for the case of noise free model  $x = \Phi\alpha$ , we may write

$$r^k = P_{\Lambda^k}^\perp x = P_{\Lambda^k}^\perp \Phi\alpha = \Psi_{\Lambda^k}\alpha.$$

Since columns of  $\Psi_{\Lambda^k}$  indexed by  $\Lambda^k$  are 0, hence when  $\text{supp}(\alpha) \subseteq \Lambda^k$ , then  $r^k = 0$ . In this case  $\alpha^k = \alpha$  exactly since it is a least squares estimate over  $\Phi_{\Lambda^k}$ .

For the same reason, if we construct a vector  $\tilde{\alpha}^k$  by zeroing out the entries indexed by  $\Lambda^k$  i.e.

$$\tilde{\alpha}_{\Lambda^k}^k = 0 \quad \text{and} \quad \tilde{\alpha}_{\Lambda^{k^c}} = \alpha_{\Lambda^{k^c}} \quad (7.4.10)$$

then

$$r^k = \Psi_{\Lambda^k}\tilde{\alpha}^k. \quad (7.4.11)$$

If  $\|\alpha\|_0 = K$ , then  $\|\tilde{\alpha}^k\|_0 = K - k$ .

Lastly putting  $r^k$  back in (7.4.8), we obtain

$$h^{k+1} = (\Psi_{\Lambda^k})^H \Psi_{\Lambda^k} \tilde{\alpha}^k. \quad (7.4.12)$$

In this version, we see that  $h^{k+1}$  is computed by applying the matrix  $(\Psi_{\Lambda^k})^H \Psi_{\Lambda^k}$  to the  $(K - k)$  sparse vector  $\tilde{\alpha}^k$ .

We are now ready to carry out RIP based analysis of OMP.

#### 7.4.2. RIP based analysis of OMP

In this section, our analysis will focus on the case for real signals and real dictionaries i.e.  $\Phi \in \mathbb{R}^{N \times D}$  and  $\alpha \in \mathbb{R}^D$ . We will attack the noise free case.

Some results for matrices that satisfy RIP will be useful in the upcoming analysis.

The following **result** applies to approximate preservation of the inner product of sparse signals  $u, v \in \mathbb{R}^D$ .

If  $u, v \in \mathbb{R}^N$  and  $K \geq \max(\|u + v\|_0, \|u - v\|_0)$ . Then

$$|\langle \Phi u, \Phi v \rangle - \langle u, v \rangle| \leq \delta_K \|u\|_2 \|v\|_2. \quad (7.4.13)$$

The next **result** shows that the matrix  $\Psi_\Lambda$  also satisfies a modified version of RIP. Let  $|\Lambda| < K$ . Then

$$\left(1 - \frac{\delta_K}{1 - \delta_K}\right) \|x\|_2^2 \leq \|\Psi_\Lambda x\|_2^2 \leq (1 + \delta_K) \|x\|_2^2 \quad (7.4.14)$$

whenever  $\|x\|_0 \leq K - |\Lambda|$  and  $\text{supp}(x) \cap \Lambda = \emptyset$ .

If  $\Phi$  satisfies RIP of order  $K$ , then  $\Psi_\Lambda$  acts as an approximate isometry on every  $(K - |\Lambda|)$ -sparse vector supported on  $\Lambda^c$ .

From (7.4.11) recall that the residual vector  $r^k$  is formed by applying  $\Psi_{\Lambda^k}$  to  $\tilde{\alpha}^k$  which is a  $K - k$  sparse vector supported on  $\Lambda^{k^c}$ .

Our interest is in combining above two results and get some bound on the inner products  $h_j^{k+1}$ . Exactly what kind of bound? When  $\Lambda^k$  has been identified, our interest is in ensuring that the next index is chosen from the set  $\text{supp}(\alpha) \setminus \Lambda^k$ . A useful way to ensure this would be to verify if the entries in  $h^{k+1}$  are close to  $\tilde{\alpha}^k$ . If they are, then they would be 0 over  $\Lambda^k$ , they would be pretty high over  $\text{supp}(\alpha) \setminus \Lambda^k$  and lastly, very small over  $\text{supp}(\alpha)^c$  which is what we want.

The next result develops these bounds around (7.4.12).

**Lemma 7.6** *Let  $\Lambda \subset \{1, \dots, D\}$  and suppose  $\tilde{\alpha} \in \mathbb{R}^D$  with  $\text{supp}(\tilde{\alpha}) \cap \Lambda = \emptyset$ . Define*

$$h = \Psi_\Lambda^T \Psi_\Lambda \tilde{\alpha}. \quad (7.4.15)$$

*Then if  $\Phi$  satisfies the RIP of order  $K \geq \|\tilde{\alpha}\|_0 + |\Lambda| + 1$  with isometry constant  $\delta_K$ , we have*

$$|h_j - \tilde{\alpha}_j| \leq \frac{\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2 \quad \forall j \notin \Lambda. \quad (7.4.16)$$

Note that  $|\Lambda|$  is the number of entries in the discovered part of the support at any iteration in OMP and  $\|\tilde{\alpha}\|_0$  is the number of entries in not yet discovered part of the support.

**PROOF.** We have  $|\Lambda| < K$  and  $\|\tilde{\alpha}\|_0 < K - |\Lambda|$ . Thus, from (7.4.14), we obtain

$$\left(1 - \frac{\delta_K}{1 - \delta_K}\right) \|\tilde{\alpha}\|_2^2 \leq \|\Psi_\Lambda \tilde{\alpha}\|_2^2 \leq (1 + \delta_K) \|\tilde{\alpha}\|_2^2. \quad (7.4.17)$$

We can make a statement saying  $\Psi_\Lambda$  satisfies a RIP of order  $(\|\tilde{\alpha}\|_0 + |\Lambda| + 1) - |\Lambda| = \|\tilde{\alpha}\|_0 + 1$  with a RIP constant  $\frac{\delta_K}{1 - \delta_K}$ .

By the definition of  $h$ , we have

$$h_j = \langle \Psi_\Lambda \tilde{\alpha}, \Psi_\Lambda e_j \rangle$$

where  $h_j$  is the  $j$ -th entry in  $h$  and  $e_j$  denotes the  $j$ -th vector from the Dirac basis. We already know that  $h_j = 0$  for all  $j \in \Lambda$ .

Consider  $j \notin \Lambda$  and take the two vectors  $\tilde{\alpha}$  and  $e_j$ .

We can easily see that

$$\|\tilde{\alpha} \pm e_j\|_0 \leq \|\tilde{\alpha}\|_0 + 1$$

and

$$\text{supp}(\tilde{\alpha} \pm e_j) \cap \Lambda = \emptyset.$$

Applying (7.4.13) on the two vectors with  $\Psi_\Lambda$  as our RIP matrix, we see that

$$|\langle \Psi_\Lambda \tilde{\alpha}, \Psi_\Lambda e_j \rangle - \langle \tilde{\alpha}, e_j \rangle| \leq \frac{\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2 \|e_j\|_2.$$

But

$$|\langle \Psi_\Lambda \tilde{\alpha}, \Psi_\Lambda e_j \rangle - \langle \tilde{\alpha}, e_j \rangle| = |h_j - \tilde{\alpha}_j|.$$

Noting that  $\|e_j\|_2 = 1$ , we get our desired result.  $\square$

With this bound in place, we can develop a sufficient condition under which the identification step of OMP (which identifies the new index  $\lambda^{k+1}$ ) will succeed.

The following corollary establishes a lower bound on the largest entry in  $\tilde{\alpha}$  which will ensure that OMP indeed chooses the next index  $\lambda^k$  from the support of  $\tilde{\alpha}$ .

**Corollary 7.7.** *Suppose that  $\Lambda$ ,  $\Phi$ ,  $\tilde{\alpha}$  meet the assumptions in lemma 7.6, and let  $h$  be as defined in (7.4.15). If*

$$\|\tilde{\alpha}\|_\infty > \frac{2\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2, \quad (7.4.18)$$

*we are guaranteed that*

$$\arg \max_{j \notin \Lambda} |h_j| \in \text{supp}(\tilde{\alpha}).$$

**PROOF.** If (7.4.16) is satisfied, then for indices  $j \notin \text{supp}(\tilde{\alpha})$ , we will have

$$|h_j| \leq \frac{\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2.$$

We already know that  $h_j = 0$  for all  $j \in \Lambda$ .

If (7.4.18) is satisfied, then there exists  $j \in \text{supp}(\tilde{\alpha})$  with

$$|\tilde{\alpha}_j| > \frac{2\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2.$$

For this particular  $j$ , applying triangular inequality on (7.4.16)

$$\frac{\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2 \geq |h_j - \tilde{\alpha}_j| \geq |\tilde{\alpha}_j| - |h_j|.$$

Thus

$$\begin{aligned} |h_j| &\geq |\tilde{\alpha}_j| - \frac{\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2 \\ &> \frac{2\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2 - \frac{\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2 \\ &= \frac{\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2. \end{aligned}$$

We have established that there exists some  $j \in \text{supp}(\tilde{\alpha})$  for which

$$|h_j| > \frac{\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2$$

and for every  $j \notin \text{supp}(\tilde{\alpha})$

$$|h_j| \leq \frac{\delta_K}{1 - \delta_K} \|\tilde{\alpha}\|_2.$$

Together, they establish that OMP will indeed choose an index from the correct set.  $\square$

All we need to do now is to make sure that (7.4.18) is satisfied by choosing  $\delta_K$  small enough.

**Theorem 7.8** [17] Suppose that  $\Phi$  satisfies the RIP of order  $K+1$  with isometry constant  $\delta < \frac{1}{2\sqrt{K+1}}$ . Then for any  $\alpha \in \mathbb{R}^D$  with  $\|\alpha\|_0 \leq K$ , OMP will recover  $\alpha$  exactly from  $x = \Phi\alpha$  in  $K$  iterations.

The upper bound on  $\delta$  can be simplified as  $\delta < \frac{1}{3\sqrt{K}}$ .

PROOF. The proof works by induction. We show that under the stated conditions,  $\lambda^1 \in \text{supp}(\alpha)$ . Then we show that whenever  $\lambda^k \in \text{supp}(\alpha)$  then  $\lambda^{k+1}$  also  $\in \text{supp}(\alpha)$ .

For the first iteration, we have

$$h^1 = \Phi^T \Phi x.$$

Note that  $\Phi = \Psi_\varnothing$ .

It is given that  $\|\alpha\|_0 \leq K$ . Thus:

$$\|\alpha\|_\infty \geq \frac{\|\alpha\|_2}{\sqrt{K}}.$$

Now  $\delta < \frac{1}{3\sqrt{K}}$  or  $\delta < \frac{1}{2\sqrt{K+1}}$  implies that

$$\frac{2\delta}{1 - \delta} < \frac{1}{\sqrt{K}}. \quad (7.4.19)$$

This can be seen as follows. Assuming  $K \geq 1$ , we have:

$$\begin{aligned} 3\sqrt{K} &\geq 2\sqrt{K} + 1 \\ \implies \frac{1}{3\sqrt{K}} &\leq \frac{1}{2\sqrt{K} + 1} \\ \implies \delta &< \frac{1}{2\sqrt{K} + 1} \\ \implies 2\delta\sqrt{K} + \delta &< 1 \\ \implies 2\delta\sqrt{K} &< 1 - \delta \\ \implies \frac{2\delta}{1 - \delta} &< \frac{1}{\sqrt{K}}. \end{aligned}$$

Therefore

$$\|\alpha\|_\infty > \frac{2\delta}{1 - \delta} \|\alpha\|_2$$

and (7.4.18) is satisfied and  $\lambda^1$  will indeed be chosen from  $\text{supp}(\alpha)$  due to corollary 7.7.

We now assume that OMP has correctly discovered indices up to  $\lambda^1, \dots, \lambda^k$ . i.e.

$$\Lambda^k \subset \text{supp}(\alpha).$$

We have to show that it will also correctly discover  $\lambda^{k+1}$ .

From the definition of  $\tilde{\alpha}$  in (7.4.10), we know that  $\text{supp}(\tilde{\alpha}^k) \cap \Lambda^k = \emptyset$ . Thus

$$\|\tilde{\alpha}^k\|_0 \leq K - k.$$

We also know that  $|\Lambda^k| = k$ . By assumption  $\Phi$  satisfies RIP of order  $K + 1 = (K - k) + k + 1$ . Thus

$$K + 1 \geq \|\tilde{\alpha}^k\|_0 + |\Lambda^k| + 1.$$

**Also:**

$$\|\tilde{\alpha}^k\|_\infty \geq \frac{\|\tilde{\alpha}^k\|_2}{\sqrt{K - k}} \geq \frac{\|\tilde{\alpha}^k\|_2}{\sqrt{K}}.$$

Using (7.4.19), we get

$$\|\tilde{\alpha}^k\|_\infty > \frac{2\delta}{1 - \delta} \|\tilde{\alpha}^k\|_2.$$

This is the sufficient condition for corollary 7.7 in (7.4.18) giving us

$$\lambda^{k+1} = \arg \max_{j \notin \Lambda^k} |h_j^{k+1}| \in \text{supp}(\tilde{\alpha}^k).$$

Hence  $\Lambda^{k+1} \subseteq \text{supp}(\alpha)$ . □

## 7.5. Compressive Sampling Matching Pursuit

We now turn our attention to CoSaMP (Compressive Sampling Matching Pursuit) developed in [29]. This algorithm follows in the tradition of orthogonal matching pursuit while bringing in other ideas from many recent developments in the field leading to an algorithm which is much more robust, fast and provides much stronger guarantees. As noted in section 2.8 compressive sampling is just another name for compressed sensing. We will adapt the discussion on CoSaMP to follow the terminology we have adopted in this book.

This algorithm has many impressive features:

- It can work with a variety of sensing matrices.
- It works with minimal number of measurements (c.f. basis pursuit).
- It is robust against presence of measurement noise (OMP has not such claim).
- It provides optimal error guarantees for every signal (sparse signals, compressible signals, completely arbitrary signals).
- The algorithm is quite efficient in terms of resource (memory, CPU) requirements.

As we move along in this section, we will understand the algorithm and validate all of these claims. Hopefully through this process we will have a very good understanding of how to evaluate the quality of signal recovery algorithm.

```

 $z = \text{CoSaMP}(\Phi, y, K);$ 
Input: Sensing matrix  $\Phi \in \mathbb{C}^{M \times N}$ 
Input: Measurement:  $y \in \mathbb{C}^M$  where  $y = \Phi x + e$ 
Input: Sparsity level:  $K$ 
Output:  $z$ : a  $K$ -sparse approximation of the signal:  $x \in \mathbb{C}^N$ 

// Initialization
 $z^0 = 0;$  // Initial approximation
 $r^0 = y;$  // Residual  $y - \Phi z$ 
 $k = 0;$  // Iteration counter

repeat
     $k \leftarrow k + 1;$ 
     $p \leftarrow \Phi^H r^{k-1};$  // Form signal proxy
     $\Omega = \text{supp}(p|_{2K});$  // Identify  $2K$  large components
     $T \leftarrow \Omega \cup \text{supp}(z^{k-1});$  // Merge supports
     $b_T \leftarrow \Phi_T^\dagger y;$  // Estimation by least-squares
     $b_{T^c} \leftarrow 0;$ 
     $z^k \leftarrow b|_K;$  // Prune to obtain next approximation
     $r^k \leftarrow y - \Phi z^k;$  // Update residual
until halting criteria is true;

```

FIGURE 7.6. CoSaMP for iterative sparse signal recovery

### 7.5.1. Algorithm

The algorithm itself is presented in fig. 7.6. Let us set out the notation before proceeding further. Most of the things are as usual, with few minor updates.

- $x \in \mathbb{C}^N$  represents the signal which is to be estimated through the algorithm.  $x$  is unknown to us within the algorithm.
- As usual  $N$  is the dimension of ambient signal space,  $K \ll N$  is the sparsity level of the approximation of  $x$  that we are estimating and  $M$  is the dimension of measurement space (number of measurements  $K < M \ll N$ ). We note that  $x$  itself may not be  $K$ -sparse. Our algorithm is designed to estimate a  $K$ -sparse approximation of  $x$ .

- $\Phi \in \mathbb{C}^{M \times N}$  represents our sensing matrix. Its known to us.
- $y = \Phi x + e$  represents the measurement vector belonging to  $\mathbb{C}^M$ . This is known to us.
- $e \in \mathbb{C}^M$  represents the measurement noise which is unknown to us within the algorithm.
- $k$  represents the iteration (or step) counter within the algorithm.
- $z \in \mathbb{C}^N$  represents our estimate of  $x$ .  $z$  is updated iteratively.
- $z^k$  represents the estimate of  $x$  at the end of  $k$ -th iteration.
- We start with  $z^k = 0$  and update it in each cycle.
- $p \in \mathbb{C}^N$  represents a proxy for  $x - z^{k-1}$ . We will explain it shortly.
- $T, \Omega$  etc. represent index sets (subsets of  $\{1, 2, \dots, N\}$ ).
- For any  $v \in \mathbb{C}^N$  and any index set  $T \subset \{1, 2, \dots, N\}$ ,  $v_T$  can mean either of the two things:
  - A vector in  $\mathbb{C}^{|T|}$  consisting of only those entries in  $v$  which are indexed by  $T$ .
  - A vector in  $\mathbb{C}^N$  whose entries indexed by  $T$  are same as that of  $v$  while entries indexed by  $\{1, 2, \dots, N\} \setminus T$  are set all to zero.

$$v_T(i) = \begin{cases} v(i) & \text{if } i \in T; \\ 0 & \text{otherwise.} \end{cases} \quad (7.5.1)$$

- For any  $v \in \mathbb{C}^N$  and any scalar  $1 \leq n \leq N$ ,  $v|_n$  means a vector in  $\mathbb{C}^N$  which consists of the  $n$  largest (in magnitude) entries of  $v$  (at the corresponding indices) while rest of the entries in  $v|_n$  are 0.
- With an index set  $T$ ,  $\Phi_T$  means an  $M \times |T|$  matrix consisting of selected columns of  $\Phi$  indexed by  $T$ .
- $r \in \mathbb{C}^M$  represents the difference between the actual measurement vector  $y$  and estimated measurement vector  $\Phi z$ .
- Ideal estimate of  $r$  would be  $e$  itself. But that won't be possible to achieve in general.

- $x|_K$  is the best  $K$ -sparse approximation of  $x$  which can be estimated by our algorithm (definition 2.12).

**Example 7.2: Clarifying the notation in CoSaMP** Let us consider

$$x = (-1, 5, 8, 0, 0, -3, 0, 0, 0, 0)$$

Then

$$x|_2 = (0, 5, 8, 0, 0, 0, 0, 0, 0, 0)$$

Also

$$x|_4 = x$$

since  $x$  happens to be 4-sparse.

$$x_{\{1,2,3,4\}} = (-1, 5, 8, 0, 0, 0, 0, 0, 0, 0)$$

or

$$x_{\{1,2,3,4\}} = (-1, 5, 8, 0)$$

in different contexts. □

**7.5.1.1. The signal proxy.** As we have learnt by now that the most challenging part in a signal recovery algorithm is to identify the support of the  $K$ -sparse approximation. OMP identifies one index in the support at each iteration and hopes that it never makes any mistakes. It ends up taking  $K$  iterations and solving  $K$  least squares problems. If there could be a simple way which could identify the support quickly (even if roughly), that can help in tremendously accelerating the algorithm. The fundamental innovation in CoSaMP is to identify the support through the signal proxy.

If  $x$  is  $K$ -sparse and  $\Phi$  satisfies RIP (see section 3.1) or order  $K$  with the restricted isometry constant  $\delta_K \ll 1$  then it can be argued that  $p = \Phi^H \Phi x$  can serve as a proxy for the signal  $x$ . In particular the largest  $K$  entries of  $p$  point towards the largest  $K$  entries of  $x$ . Although we don't know  $x$  inside the algorithm, yet we have  $y = \Phi x$  (assuming error to be 0), the proxy can be easily estimated by computing  $p = \Phi^H y$ .

This is the key idea in CoSaMP. Rather than identifying just one new index in support of  $x$ , it tries to identify whole of support by picking up the largest  $2K$  entries of  $p$ . It then solves a least squares problem around the columns of  $\Phi$  indexed by this support set. It keeps only the  $K$  largest entries from the least squares solution as an estimate of  $x$ . Of course there is no guarantee that in a single attempt the support of  $x$  will be recovered completely. Hence the residual between actual measurement vector  $y$  and estimated measurement vector  $\Phi z$  is computed and it is used to identify other indices of  $\text{supp}(x)$  iteratively.

**7.5.1.2. Core of algorithm.** There are two things which are estimated in each iteration of algorithm

- $K$ -sparse estimate of  $x$  at  $k$ -th step:  $z^k$ .
- Residual at  $k$ -th step:  $r^k$

We start with a trivial estimate  $z^0 = 0$  and improve it in each iteration.

$r^0$  is nothing but the measurement vector  $y$ .

As explained before  $r^k$  is the difference between actual measurement vector  $y$  and the estimated measurement vector  $\Phi z^k$ . This  $r^k$  is used for computing the signal proxy at  $k$ -th step.

Concretely assuming  $e = 0$

$$p = \Phi^H r^k = \Phi^H(y - \Phi z^k) = \Phi^H \Phi(x - z^k).$$

Thus  $p$  is actually the proxy of the difference between original signal and estimated signal.

During each iteration, the algorithm performs following tasks

**Identification:** The algorithm forms a proxy of the residual  $r^{k-1}$  and locates the  $2K$  largest entries of the proxy.

**Support merger:** The set of newly identified indices is merged with the set of indices that appear in the current approximation  $z^{k-1}$  (i.e.  $\text{supp}(z^{k-1})$ ).

**Estimation:** The algorithm solves a least squares problem to approximate the signal  $x$  on the merged set of indices.

**Pruning:** The algorithm produces a new approximation  $z^k$  by retaining only the largest  $K$  entries in the least squares solution.

**Residual update:** Finally the new residual  $r^k$  between original measurement vector  $y$  and estimated measurement vector  $\Phi z^k$  is computed.

The steps are repeated until the halting criteria is reached. We will discuss the halting criteria in detail later. A possible halting criteria is when the norm of residual  $r^k$  reaches a very small value.

### 7.5.2. CoSaMP analysis sparse case

In this subsection, we will carry out a detailed theoretical analysis of CoSaMP algorithm for sparse signal recovery.

We will make following assumptions in the analysis:

- The sparsity level  $K$  is fixed and known in advance.
- The signal  $x$  is  $K$ -sparse (i.e.  $x \in \Sigma_K \subset \mathbb{C}^N$ ).
- The sensing matrix  $\Phi$  satisfies RIP of order  $4K$  with  $\delta_{4K} \leq 0.1$ .
- Measurement error  $e \in \mathbb{C}^M$  is arbitrary.

For each iteration, we need to define one more quantity: **recovery error**

$$d^k = x - z^k. \quad (7.5.2)$$

$d^k$  captures the difference between actual signal  $x$  and estimated signal  $z^k$  at the end of  $k$ -th iteration. Under ideal recovery,  $d^k$  should become 0 as  $k$  increases. Since we don't know  $x$  within the algorithm, we cannot see  $d^k$  either directly. We do have following equation

$$r^k = y - \Phi z^k = \Phi x + e - \Phi z^k = \Phi(x - z^k) + e = \Phi d^k + e.$$

We will show that in each iteration, CoSaMP reduces the recovery error by a constant factor while adding a small multiple of the measurement

noise  $\|e\|_2$ . As a result, when recovery error is large compared to measurement noise, the algorithm makes substantial progress in each step. The algorithm stops making progress when recovery error is of the order of measurement noise.

We will consider some iteration  $k \geq 1$  and analyze the behavior of each of the steps: identification, support merger, estimation, pruning and residual update one by one. At the beginning of the iteration we have

$$r^{k-1} = \Phi(x - z^{k-1}) + e = \Phi d^{k-1} + e.$$

The quantities of interest are  $z^{k-1}$ ,  $r^{k-1}$  and  $d^{k-1}$  out of which  $d^{k-1}$  is unknowable. In order to simplify the equations below we will write

$$r = r^{k-1}, \quad z = z^{k-1} \quad \text{and} \quad d = d^{k-1}.$$

**7.5.2.1. Identification.** We start our analysis with the identification step in the main loop of CoSaMP (fig. 7.6).

We compute  $p = \Phi^H r$  and consider  $\Omega = \text{supp}(p|_{2K})$  as the index set of largest  $2K$  indices in  $p$ . We will show that most of the energy in  $d$  (the recovery error) is concentrated in the entries indexed by  $\Omega$ . i.e. we will be looking for a bound of the form

$$\|d_{\Omega^c}\|_2 \ll \|d\|_2.$$

Since  $x$  is  $K$ -sparse and  $z$  is  $K$ -sparse hence  $d$  is  $2K$ -sparse. Actually in first iteration where  $z^0 = 0$ , support of  $d^0$  is same as support of  $x$ . In later iterations it may include more indices. Let

$$\Gamma = \text{supp}(d).$$

We remind that  $\Omega$  is known to us while  $\Gamma$  is unknown to us. Ideal case would be when  $\Gamma \subseteq \Omega$ . Then we would have recovered whole of support of  $x$ ; recovering  $x$  would therefore be easier. It so happens, life is not that easy. So let us examine what are the differences between

the two sets. We have  $|\Omega| \leq 2K$  and  $|\Gamma| \leq 2K$ . Moreover,  $\Omega$  indexes  $2K$  largest entries in  $p$ . Thus we have (due to lemma 2.17)

$$\|p_\Gamma\|_2 \leq \|p_\Omega\|_2. \quad (7.5.3)$$

where  $p_\Gamma, p_\Omega \in \mathbb{C}^N$  are obtained from  $p$  as per definition 2.10.

Squaring (7.5.3) on both sides and expanding we get

$$\sum_{\gamma \in \Gamma} |p_\gamma|^2 \leq \sum_{\omega \in \Omega} |p_\omega|^2.$$

Now we do hope that some of the entries are common in both sides. Those entries are indexed by  $\Gamma \cap \Omega$ . The remaining entries on L.H.S. are indexed by  $\Gamma \setminus \Omega$  and on the R.H.S. are indexed by  $\Omega \setminus \Gamma$ . So we have

$$\|p_{\Gamma \setminus \Omega}\|_2^2 \leq \|p_{\Omega \setminus \Gamma}\|_2^2 \implies \|p_{\Gamma \setminus \Omega}\|_2 \leq \|p_{\Omega \setminus \Gamma}\|_2. \quad (7.5.4)$$

For both  $\Gamma \setminus \Omega$  and  $\Omega \setminus \Gamma$  we have

$$|\Gamma \setminus \Omega| \leq 2K \text{ and } |\Omega \setminus \Gamma| \leq 2K.$$

The worst case happens when both sets are totally disjoint.

Let us first examine the R.H.S and find an upper bound for it.

$$\begin{aligned} \|p_{\Omega \setminus \Gamma}\|_2 &= \|\Phi_{\Omega \setminus \Gamma}^H r\|_2 \\ &= \|\Phi_{\Omega \setminus \Gamma}^H (\Phi d + e)\|_2 \\ &\leq \|\Phi_{\Omega \setminus \Gamma}^H \Phi d\|_2 + \|\Phi_{\Omega \setminus \Gamma}^H e\|_2. \end{aligned} \quad (7.5.5)$$

At this juncture, its worthwhile to scan various results in section 3.1 since we are going to need many of them in the following steps.

Let us look at the term  $\|\Phi_{\Omega \setminus \Gamma}^H \Phi d\|_2$ . We have  $|\Omega \setminus \Gamma| \leq 2K$ . Further  $d$  is  $2K$  sparse and sits over  $\Gamma$  which is disjoint with  $\Omega \setminus \Gamma$ . Together  $|\Gamma \cup \Omega| \leq 4K$ . A straightforward application of corollary 3.21 gives us

$$\|\Phi_{\Omega \setminus \Gamma}^H \Phi d\|_2 \leq \delta_{4K} \|d\|_2.$$

Similarly using theorem 3.16 we get

$$\|\Phi_{\Omega \setminus \Gamma}^H e\|_2 \leq \sqrt{1 + \delta_{2K}} \|e\|_2.$$

Combining the two we get

$$\|p_{\Omega \setminus \Gamma}\|_2 \leq \delta_{4K} \|d\|_2 + \sqrt{1 + \delta_{2K}} \|e\|_2. \quad (7.5.6)$$

We now look at L.H.S. and find a lower bound for it.

$$\begin{aligned} \|p_{\Gamma \setminus \Omega}\|_2 &= \|\Phi_{\Gamma \setminus \Omega}^H r\|_2 \\ &= \|\Phi_{\Gamma \setminus \Omega}^H (\Phi d + e)\|_2. \end{aligned} \quad (7.5.7)$$

We will split  $d$  as

$$d = d_{\Gamma \setminus \Omega} + d_\Omega.$$

Further we will use a form of triangular inequality as

$$\|a + b\| \geq \|a\| - \|b\|.$$

We expand L.H.S.

$$\begin{aligned} \|\Phi_{\Gamma \setminus \Omega}^H (\Phi d + e)\|_2 &= \|\Phi_{\Gamma \setminus \Omega}^H (\Phi (d_{\Gamma \setminus \Omega} + d_\Omega) + e)\|_2 \\ &\geq \|\Phi_{\Gamma \setminus \Omega}^H \Phi d_{\Gamma \setminus \Omega}\|_2 - \|\Phi_{\Gamma \setminus \Omega}^H \Phi d_\Omega\|_2 - \|\Phi_{\Gamma \setminus \Omega}^H e\|_2. \end{aligned} \quad (7.5.8)$$

As before

$$\|\Phi_{\Gamma \setminus \Omega}^H \Phi d_{\Gamma \setminus \Omega}\|_2 = \|\Phi_{\Gamma \setminus \Omega}^H \Phi d_{\Gamma \setminus \Omega}\|_2.$$

We can use the lower bound from theorem 3.18 to give us

$$\|\Phi_{\Gamma \setminus \Omega}^H \Phi d_{\Gamma \setminus \Omega}\|_2 \geq (1 - \delta_{2K}) \|d_{\Gamma \setminus \Omega}\|_2$$

since  $|\Gamma \setminus \Omega| \leq 2K$ .

For the other two terms we need to find their upper bounds since they appear in negative sign. Applying corollary 3.21 we get

$$\|\Phi_{\Gamma \setminus \Omega}^H \Phi d_\Omega\|_2 \leq \delta_{2K} \|d\|_2.$$

Again using theorem 3.16 we get

$$\|\Phi_{\Gamma \setminus \Omega}^H e\|_2 \leq \sqrt{1 + \delta_{2K}} \|e\|_2.$$

Putting the three bounds together, we get a lower bound for L.H.S.

$$\|p_{\Gamma \setminus \Omega}\|_2 \geq (1 - \delta_{2K}) \|d_{\Gamma \setminus \Omega}\|_2 - \delta_{2K} \|d\|_2 - \sqrt{1 + \delta_{2K}} \|e\|_2. \quad (7.5.9)$$

Finally plugging the upper bound from (7.5.6) and lower bound from (7.5.9) into (7.5.4) we get

$$\begin{aligned}
 & (1 - \delta_{2K}) \|d_{\Gamma \setminus \Omega}\|_2 - \delta_{2K} \|d\|_2 - \sqrt{1 + \delta_{2K}} \|e\|_2 \leq \delta_{4K} \|d\|_2 + \sqrt{1 + \delta_{2K}} \|e\|_2 \\
 \implies & (1 - \delta_{2K}) \|d_{\Gamma \setminus \Omega}\|_2 \leq (\delta_{2K} + \delta_{4K}) \|d\|_2 + 2\sqrt{1 + \delta_{2K}} \|e\|_2 \\
 \implies & \|d_{\Gamma \setminus \Omega}\|_2 \leq \frac{(\delta_{2K} + \delta_{4K}) \|d\|_2 + 2\sqrt{1 + \delta_{2K}} \|e\|_2}{1 - \delta_{2K}}.
 \end{aligned} \tag{7.5.10}$$

This result is nice but there is a small problem. We would like to get rid of  $\Gamma$  from L.H.S. and get  $d_{\Omega^c}$  instead. Now recall that

$$\Omega^c = (\Gamma \cap \Omega^c) \cup (\Gamma^c \cap \Omega^c)$$

where  $\Gamma \cap \Omega^c$  and  $\Gamma^c \cap \Omega^c$  are disjoint. Thus

$$d_{\Omega^c} = d_{\Gamma \cap \Omega^c} + d_{\Gamma^c \cap \Omega^c}.$$

Since  $d$  is 0 on  $\Gamma^c$  (recall that  $\Gamma = \text{supp}(d)$ ), hence

$$d_{\Gamma^c \cap \Omega^c} = 0.$$

As a result

$$d_{\Omega^c} = d_{\Gamma \cap \Omega^c} = d_{\Gamma \setminus \Omega}.$$

This gives us

$$\|d_{\Omega^c}\|_2 \leq \frac{(\delta_{2K} + \delta_{4K}) \|d\|_2 + 2\sqrt{1 + \delta_{2K}} \|e\|_2}{1 - \delta_{2K}}. \tag{7.5.11}$$

Let us simplify this equation by using  $\delta_{2K} \leq \delta_{4K} \leq 0.1$  assumed at the beginning of the analysis. This gives us

$$1 - \delta_{2K} \geq 0.9, \quad \delta_{2K} + \delta_{4K} \leq 0.2, \quad 2\sqrt{1 + \delta_{2K}} \leq 2\sqrt{1.1} = 2.098.$$

Thus we get

$$\|d_{\Omega^c}\|_2 \leq \frac{0.2\|d\|_2 + 2.098\|e\|_2}{.9}.$$

Simplifying

$$\|d_{\Omega^c}\|_2 \leq 0.223\|d\|_2 + 2.331\|e\|_2. \tag{7.5.12}$$

This result tells us that in the absence of measurement noise, at least 78% of energy in the recovery error is indeed concentrated in the entries

indexed by  $\Omega$ . Moreover, if the measurement noise is small compared to the size of recovery error, then this concentration continues to hold. Thus even if we don't know  $\Gamma$  directly,  $\Omega$  is a close approximation for  $\Gamma$ .

We summarize the analysis for the *identification* step in the following lemma.

**Lemma 7.9** [Identification] *Let  $\Phi$  satisfy RIP of order  $4K$  with  $\delta_{4K} \leq 0.1$ . At every iteration  $k$  in CoSaMP algorithm, let  $d^{k-1} = (x - z^{k-1})$  be the recovery error and  $p = \Phi^H r^{k-1}$  be the signal proxy. The set  $\Omega = \text{supp}(p|_{2K})$  contains at most  $2K$  indices and*

$$\|d_{\Omega^c}^{k-1}\|_2 \leq 0.223\|d^{k-1}\|_2 + 2.331\|e\|_2 \quad (7.5.13)$$

i.e. most of the energy in  $d^{k-1}$  is concentrated in the entries indexed by  $\Omega$  when measurement noise is small.

**7.5.2.2. Support merger.** The next step in a CoSaMP iteration is support merger. Recalling from fig. 7.6, the step involves computing

$$T = \Omega \cup \text{supp}(z)$$

i.e. we merge the support of current signal estimate  $z$  with the newly identified set of indices  $\Omega$ .

In previous lemma we were able to show how well the recovery error is concentrated over the index set  $\Omega$ . But we didn't establish anything concrete about how  $x$  is concentrated. In the support merger step, we will be able to show that  $x$  is highly concentrated over the index set  $T$ . For this we will need to find an upper bound on  $\|x_{T^c}\|_2$  i.e. the energy of entries in  $x$  which are not covered by  $T$ .

We recall that  $|\Omega| \leq 2K$  and since  $z$  is a  $K$ -sparse estimate of  $x$  hence  $|\text{supp}(z)| \leq K$ . Thus

$$|T| \leq 3K.$$

Clearly

$$T^c = \Omega^c \cap \text{supp}(z)^c \subset \Omega^c.$$

Further since  $\text{supp}(z) \subset T$  hence  $z_{T^c} = 0$ . Thus

$$x_{T^c} = (x - z)_{T^c} = d_{T^c}.$$

Since  $T^c \subset \Omega^c$  hence

$$\|d_{T^c}\|_2 \leq \|d_{\Omega^c}\|_2.$$

Combining these facts we can write

$$\|x_{T^c}\|_2 = \|d_{T^c}\|_2 \leq \|d_{\Omega^c}\|_2.$$

We summarize this analysis in following lemma.

**Lemma 7.10** [Support merger] *Let  $\Omega$  be a set of at most  $2K$  entries. The set  $T = \Omega \cup \text{supp}(z^{k-1})$  contains at most  $3K$  entries, and*

$$\|x_{T^c}\|_2 \leq \|d_{\Omega^c}\|_2. \quad (7.5.14)$$

Note that this inequality says nothing about how  $\Omega$  is chosen. But we can use an upper bound on  $\|d_{\Omega^c}\|_2$  from lemma 7.9 to get an upper bound on  $\|x_{T^c}\|_2$ .

**7.5.2.3. Estimation.** Next step is a least square estimate of  $x$  over the columns indexed by  $T$ . Recalling from fig. 7.6 we compute

$$b_T = \Phi_T^\dagger y = \Phi_T^\dagger(\Phi x + e).$$

We also set  $b_{T^c} = 0$ . We have

$$b = b_T + b_{T^c} = b_T.$$

Since  $|T| \leq 3K$ ,  $b$  is a  $3K$ -sparse approximation of  $x$ . What we need here is a bound over the approximation error  $\|x - b\|_2$ . We have already obtained a bound on  $\|x_{T^c}\|_2$ . If an upper bound on  $\|x - b\|_2$  depends on  $\|x_{T^c}\|_2$  and off course measurement noise  $\|e\|_2$  that would be quite reasonable.

We start with splitting  $x$  over  $T$  and  $T^c$ .

$$x = x_T + x_{T^c}.$$

Since  $\text{supp}(b) \subset T$  hence

$$x - b = x_T + x_{T^c} - b_T = (x_T - b_T) + x_{T^c}.$$

This gives us

$$\|x - b\|_2 \leq \|x_T - b_T\|_2 + \|x_{T^c}\|_2.$$

We will now expand the term  $\|x_T - b_T\|_2$ .

$$\|x_T - b_T\|_2 = \|x_T - \Phi_T^\dagger(\Phi x + e)\|_2 = \|x_T - \Phi_T^\dagger(\Phi x_T + \Phi x_{T^c} + e)\|_2$$

Now since  $\Phi_T$  is full column rank (since  $\Phi$  satisfies RIP of order  $3K$ ) hence  $\Phi_T^\dagger \Phi_T = I$ . Also  $\Phi x_T = \Phi_T x_T$  (since column columns indexed by  $T$  are involved). This helps us cancel  $x_T - \Phi_T^\dagger \Phi x_T$ . Thus

$$\|x_T - b_T\|_2 = \|\Phi_T^\dagger(\Phi x_{T^c} + e)\|_2 \leq \|(\Phi_T^H \Phi_T)^{-1} \Phi_T^H \Phi x_{T^c}\|_2 + \|\Phi_T^\dagger e\|_2.$$

Let us look at the terms on R.H.S. one by one. Let  $v = \Phi_T^H \Phi x_{T^c}$ . Then

$$\|(\Phi_T^H \Phi_T)^{-1} \Phi_T^H \Phi x_{T^c}\|_2 = \|(\Phi_T^H \Phi_T)^{-1} v\|_2.$$

This perfectly fits theorem 3.18 with  $|T| \leq 3K$  giving us

$$\|(\Phi_T^H \Phi_T)^{-1} \Phi_T^H \Phi x_{T^c}\|_2 \leq \frac{1}{1 - \delta_{3K}} \|\Phi_T^H \Phi x_{T^c}\|_2.$$

Further, applying corollary 3.21 we get

$$\|\Phi_T^H \Phi x_{T^c}\|_2 \leq \delta_{4K} \|x_{T^c}\|_2$$

since  $|T \cup \text{supp}(x)| \leq 4K$ .

For the measurement noise term, applying theorem 3.17 we get

$$\|\Phi_T^\dagger e\|_2 \leq \frac{1}{\sqrt{1 - \delta_{3K}}} \|e\|_2.$$

Combining the above inequalities we get

$$\|x - b\|_2 \leq \left[ 1 + \frac{\delta_{4K}}{1 - \delta_{3K}} \right] \|x_{T^c}\|_2 + \frac{1}{\sqrt{1 - \delta_{3K}}} \|e\|_2. \quad (7.5.15)$$

Recalling our assumption that  $\delta_{3K} \leq \delta_{4K} \leq 0.1$  we can simplify the constants to get

$$\|x - b\|_2 \leq 1.112 \|x_{T^c}\|_2 + 1.0541 \|e\|_2. \quad (7.5.16)$$

We summarize the analysis in this step in the following lemma.

**Lemma 7.11** [Estimation] Let  $T$  be a set of at most  $3K$  indices, and define the least squares signal estimate  $b$  by the formula

$$b_T = \Phi_T^\dagger y = \Phi_T^\dagger(\Phi x + e) \text{ and } b_{T^c} = 0.$$

If  $\Phi$  satisfies RIP of order  $4K$  with  $\delta_{4K} \leq 0.1$  then the following holds

$$\|x - b\|_2 \leq 1.112\|x_{T^c}\|_2 + 1.0541\|e\|_2. \quad (7.5.17)$$

Note that the lemma doesn't make any assumptions about how  $T$  is arrived at except that  $|T| \leq 3K$ . Also, this analysis assumes that the least squares solution is of infinite precision given by  $\Phi_T^\dagger y$ . The approximation error in an iterative least squares solution needs to be separately analyzed to identify the number of required steps so that the least squares error is negligible.

**7.5.2.4. Pruning.** The last step in the main loop of CoSaMP (fig. 7.6) is pruning. We compute

$$z^k = b|_K$$

as the next estimate of  $x$  by picking the  $K$  largest entries in  $b$ . We note that both  $x$  and  $b|_K$  can be regarded as  $K$  term approximations of  $b$ . As established in lemma 2.17,  $b|_K$  is the best  $K$ -term approximation of  $b$ . Thus

$$\|b - b|_K\|_2 \leq \|b - x\|_2.$$

Now

$$\|x - z^k\|_2 = \|x - b + b - b|_K\|_2 \leq \|x - b\|_2 + \|b - b|_K\|_2 \leq 2\|x - b\|_2.$$

This helps us establish that although the  $3K$ -sparse approximation  $b$  is closer to  $x$  compared to  $b|_K$ , but  $b|_K$  is also not too bad approximation of  $x$  while using only  $K$  entries at most. We summarize it in following lemma.

**Lemma 7.12** *The pruned estimate  $z^k = b|_K$  satisfies*

$$\|x - b|_k\|_2 \leq 2\|x - b\|_2. \quad (7.5.18)$$

**7.5.2.5. The CoSaMP iteration invariant.** Having analyzed each of the steps in the main loop of CoSaMP algorithm, it is time for us to combine the analysis. Concretely, we wish to establish how much progress CoSaMP makes in each iteration. Following theorem provides an upper bound on how the recovery error changes from one iteration to next iteration. This theorem is known as the iteration invariant for the sparse case.

**Theorem 7.13** [CoSaMP iteration invariant for sparse case] *Assume that  $x$  is  $K$ -sparse. Assume that  $\Phi$  satisfies RIP of order  $4K$  with  $\delta_{4K} \leq 0.1$ . For each iteration number  $k \geq 1$ , the signal estimate  $z^k$  is  $K$ -sparse and satisfies*

$$\|x - z^k\|_2 \leq \frac{1}{2}\|x - z^{k-1}\|_2 + 7.5\|e\|_2. \quad (7.5.19)$$

*In particular*

$$\|x - z^k\|_2 \leq 2^{-k}\|x\|_2 + 15\|e\|_2. \quad (7.5.20)$$

This theorem helps us establish that if measurement noise is small, then the algorithm makes substantial progress in each iteration. The proof makes use of the lemmas developed above.

**PROOF.** We run the proof in backtracking mode. We start from  $z^k$  and go back step by step through pruning, estimation, support merger, and identification to connect it with  $z^{k-1}$ .

From lemma 7.12 we have

$$\|x - z^k\|_2 = \|x - b|_K\|_2 \leq 2\|x - b\|_2. \quad (7.5.21)$$

Applying lemma 7.11 for the least squares estimation step gives us

$$2\|x - b\|_2 \leq 2 \cdot (1.112\|x_{T^c}\|_2 + 1.0541\|e\|_2) = 2.224\|x_{T^c}\|_2 + 2.1082\|e\|_2. \quad (7.5.22)$$

Lemma 7.10 (Support merger) tells us that

$$\|x_{T^c}\|_2 \leq \|d_{\Omega^c}^{k-1}\|_2.$$

This gives us

$$\|x - z^k\|_2 \leq 2.224\|d_{\Omega^c}^{k-1}\|_2 + 2.1082\|e\|_2. \quad (7.5.23)$$

From identification step we have lemma 7.9

$$\|d_{\Omega^c}^{k-1}\|_2 \leq 0.223\|d^{k-1}\|_2 + 2.331\|e\|_2.$$

This gives us

$$\begin{aligned} \|x - z^k\|_2 &\leq 2.224(0.223\|d^{k-1}\|_2 + 2.331\|e\|_2) + 2.1082\|e\|_2 \\ &\leq 0.5\|d^{k-1}\|_2 + 7.5\|e\|_2 \\ &= \frac{1}{2}\|x - z^{k-1}\|_2 + 7.5\|e\|_2. \end{aligned} \quad (7.5.24)$$

The constants have been simplified to make them look better.

For the 2nd result in this theorem, we add up the error at each stage as

$$(1 + 2^{-1} + 2^{-2} + \dots + 2^{-(k-1)})7.5\|e\|_2 \leq 2 \cdot 7.5\|e\|_2 = 15\|e\|_2.$$

At  $k = 1$  we have  $z^0 = 0$ . This gives us the result

$$\|x - z^k\|_2 \leq 2^{-k}\|x\|_2 + 15\|e\|_2. \quad (7.5.25)$$

□

### 7.5.3. CoSaMP analysis general case

Having completed the analysis for the sparse case (where the signal  $x$  is  $K$ -sparse) it is time for us to generalize the analysis for the case where  $x$  is assumed to be an arbitrary signal in  $\mathbb{C}^N$ . Although it may look hard at first sight but there is a simple way to transform the problem into the problem of CoSaMP for the sparse case. Essentially we

decompose  $x$  into its  $K$ -sparse approximation and the approximation error. Further, we absorb the approximation error term into the measurement error term. Since sparse case analysis is applicable for an arbitrary measurement error, this approach gives us an upper bound on the performance of CoSaMP over arbitrary signals.

We start with writing

$$x = x - x|_K + x|_K \quad (7.5.26)$$

where  $x|_K$  is the best  $K$ -term approximation of  $x$  (lemma 2.17). Thus we have

$$y = \Phi x + e = \Phi(x - x|_K + x|_K) + e = \Phi x_K + \Phi(x - x_K) + e.$$

We define

$$\hat{e} = \Phi(x - x_K) + e.$$

This lets us write

$$y = \Phi x|_K + \hat{e}. \quad (7.5.27)$$

In this formulation, the problem is equivalent to recovering the  $K$ -sparse signal  $x|_K$  from the measurement vector  $y$ . The results of section 7.5.2 and in particular the iteration invariant theorem 7.13 apply directly. The remaining problem is to estimate the norm of modified error  $\hat{e}$ . We have

$$\|\hat{e}\|_2 = \|\Phi(x - x_K) + e\|_2 \leq \|\Phi(x - x_K)\|_2 + \|e\|_2.$$

Another result for RIP on the energy bound of embedding of arbitrary signals from theorem 3.28 gives us

$$\|\Phi(x - x_K)\|_2 \leq \sqrt{1 + \delta_K} \left[ \|x - x_K\|_2 + \frac{1}{\sqrt{K}} \|x - x_K\|_1 \right]. \quad (7.5.28)$$

Thus we have an upper bound on  $\|\hat{e}\|_2$  given by

$$\|\hat{e}\|_2 \leq \sqrt{1 + \delta_K} \left[ \|x - x_K\|_2 + \frac{1}{\sqrt{K}} \|x - x_K\|_1 \right] + \|e\|_2. \quad (7.5.29)$$

Since we have assumed throughout that  $\delta_{4K} \leq 0.1$ , it gives us

$$\|\hat{e}\|_2 \leq 1.05 \left[ \|x - x_K\|_2 + \frac{1}{\sqrt{K}} \|x - x_K\|_1 \right] + \|e\|_2. \quad (7.5.30)$$

This inequality is able to combine measurement error and approximation error into a single expression. To fix ideas further, we define the notion of unrecoverable energy in CoSaMP.

**Definition 7.3** The **unrecoverable energy** in CoSaMP algorithm is defined as

$$\nu = \left[ \|x - x_K\|_2 + \frac{1}{\sqrt{K}} \|x - x_K\|_1 \right] + \|e\|_2. \quad (7.5.31)$$

This quantity measures the baseline error in the CoSaMP recovery consisting of measurement error and approximation error.

It's obvious that

$$\|\hat{e}\|_2 \leq 1.05\nu.$$

We summarize this analysis in following lemma.

**Lemma 7.14** Let  $x \in \mathbb{C}^N$  be an arbitrary signal. Let  $x|_K$  be its best  $K$ -term approximation. The measurement vector  $y = \Phi x + e$  can be expressed as  $y = \Phi x|_K + \hat{e}$  where

$$\|\hat{e}\|_2 \leq 1.05 \left[ \|x - x_K\|_2 + \frac{1}{\sqrt{K}} \|x - x_K\|_1 \right] + \|e\|_2 \leq 1.05\nu. \quad (7.5.32)$$

Invoking theorem 7.13 for the iteration invariant for recovery of a sparse signal gives us

$$\|x|_K - z^k\|_2 \leq \frac{1}{2} \|x|_K - z^{k-1}\|_2 + 7.5 \|\hat{e}\|_2. \quad (7.5.33)$$

What remains is to generalize this inequality for the arbitrary signal  $x$  itself. We can write

$$\|x|_K - x + x - z^k\|_2 \leq \frac{1}{2} \|x|_K - x + x - z^{k-1}\|_2 + 7.5 \|\hat{e}\|_2.$$

We simplify L.H.S. as

$$\|x|_K - x + x - z^k\|_2 = \|(x - z^k) - (x - x|_K)\|_2 \geq \|x - z^k\|_2 - \|x - x|_K\|_2.$$

On the R.H.S. we expand as

$$\|x|_K - x + x - z^{k-1}\|_2 \leq \|x - z^{k-1}\|_2 + \|x - x|_K\|_2.$$

Combining the two we get

$$\|x - z^k\|_2 \leq 0.5\|x - z^{k-1}\|_2 + 1.5\|x - x|_K\|_2 + 7.5\|\hat{e}\|_2.$$

Putting the estimate of  $\|\hat{e}\|_2$  from lemma 7.14 we get

$$\|x - z^k\|_2 \leq 0.5\|x - z^{k-1}\|_2 + 9.375\|x - x|_K\|_2 + \frac{7.875}{\sqrt{K}}\|x - x_K\|_1 + 7.5\|e\|_2.$$

Now

$$9.375\|x - x|_K\|_2 + \frac{7.875}{\sqrt{K}}\|x - x_K\|_1 + 7.5\|e\|_2 \leq 10 \left( \left[ \|x - x_K\|_2 + \frac{1}{\sqrt{K}}\|x - x_K\|_1 \right] + \|e\|_2 \right).$$

Thus we write a simplified expression

$$\|x - z^k\|_2 \leq 0.5\|x - z^{k-1}\|_2 + 10\nu \quad (7.5.34)$$

where  $\nu$  is the unrecoverable energy (definition 7.3).

We can summarize the analysis for the general case in the following theorem

**Theorem 7.15** Assume that  $\Phi$  satisfies RIP of order  $4K$  with  $\delta_{4K} \leq 0.1$ . For each iteration number  $k \geq 1$ , the signal estimate  $z^k$  is  $K$ -sparse and satisfies

$$\|x - z^k\|_2 \leq \frac{1}{2}\|x - z^{k-1}\|_2 + 10\nu. \quad (7.5.35)$$

In particular

$$\|x - z^k\|_2 \leq 2^{-k}\|x\|_2 + 20\nu. \quad (7.5.36)$$

PROOF. 1st result was developed in this section before the theorem. For the 2nd result in this theorem, we add up the error at each stage as

$$(1 + 2^{-1} + 2^{-2} + \dots + 2^{-(k-1)})10\nu \leq 2 \cdot 10\nu = 20\nu.$$

At  $k = 1$  we have  $z^0 = 0$ . This gives us the result.  $\square$

**7.5.3.1. SNR analysis.** A sensible though unusual definition of *signal-to-noise ratio* (as proposed in [29]) is as follows

$$\text{SNR} = 10 \log \left( \frac{\|x\|_2}{\nu} \right). \quad (7.5.37)$$

Essentially the whole of unrecoverable energy is treated as noise. The signal  $l_2$  norm rather than its square is being treated as the measure of its energy. This is the unusual part. Yet the way  $\nu$  has been developed, this definition is quite sensible.

Further we define the *reconstruction SNR* or *recovery SNR* as the ratio between signal energy and recovery error energy.

$$\text{R-SNR} = 10 \log \left( \frac{\|x\|_2}{\|x - z\|_2} \right). \quad (7.5.38)$$

Both SNR and R-SNR are expressed in dB. Certainly we have

$$\text{R-SNR} \leq \text{SNR}.$$

Let us look closely at the iteration invariant

$$\|x - z^k\|_2 \leq 2^{-k}\|x\|_2 + 20\nu.$$

In the initial iterations,  $2^{-k}\|x\|_2$  term dominates in the R.H.S. Assuming

$$2^{-k}\|x\|_2 \geq 20\nu$$

we can write

$$\|x - z^k\|_2 \leq 2 \cdot 2^{-k}\|x\|_2.$$

This gives us

$$\begin{aligned} \|x - z^k\|_2 &\leq 2^{-k+1}\|x\|_2 \\ \implies \frac{\|x\|_2}{\|x - z\|_2} &\geq 2^{k-1} \\ \implies \text{R-SNR} &\geq 10(k-1)\log 2 \geq 3k-3. \end{aligned}$$

In the later iterations, the  $20\nu$  term dominates in the R.H.S. Assuming

$$2^{-k}\|x\|_2 \leq 20\nu$$

we can write

$$\|x - z^k\|_2 \leq 2 \cdot 20\nu = 40\nu.$$

This gives us

$$\begin{aligned} \|x - z^k\|_2 &\leq 40\nu \\ \implies \frac{\|x\|_2}{\|x - z\|_2} &\geq \frac{1}{40} \frac{\|x\|_2}{\nu} \\ \implies \text{R-SNR} &\geq \text{SNR} - 10 \log 40 \geq \text{SNR} - 16 = \text{SNR} - 13 - 3. \end{aligned}$$

We combine these two results into the following

$$\text{R-SNR} \geq \min\{3k, \text{SNR} - 13\} - 3. \quad (7.5.39)$$

This result tells us that in the initial iterations the reconstruction SNR keeps improving by 3dB per iteration till it hits the noise floor given by  $\text{SNR} - 16$  dB. Thus roughly the number of iterations required for converging to the noise floor is given by

$$k \approx \frac{\text{SNR} - 13}{3}.$$

## 7.6. Digest

## CHAPTER 8

# Shrinkage and Thresholding Algorithms

In this chapter we will review some algorithms based on shrinkage and iterative thresholding techniques which can help us solve the sparse approximation problem and the sparse recovery problem discussed in chapter 2. These algorithms also fall in the general category of greedy algorithms.

The presentation in this chapter is based on a number of sources including [4, 7, 21, 29].

### 8.1. Iterative hard thresholding for signal recovery

In this section we will study an algorithm called Iterative Hard Thresholding (IHT) developed in [7].

#### 8.1.1. Algorithm

The algorithm itself is presented in fig. 8.1.

As usual let us review the notation before proceeding further.

- $x \in \mathbb{C}^N$  represents the signal which is to be estimated through the algorithm.  $x$  is unknown to us within the algorithm.
- As usual  $N$  is the dimension of ambient signal space,  $K \ll N$  is the sparsity level of the approximation of  $x$  that we are estimating and  $M$  is the dimension of measurement space (number of measurements  $K < M \ll N$ ). We note that  $x$  itself may not be  $K$ -sparse. Our algorithm is designed to estimate a  $K$ -sparse approximation of  $x$ .
- $\Phi \in \mathbb{C}^{M \times N}$  represents our sensing matrix. Its known to us.

```

 $z = \text{IHT}(\Phi, y, K);$ 
Input: Sensing matrix  $\Phi$ 
Input: Measurement:  $y$ 
Input: Sparsity level:  $K$ 
Output:  $z$ : A  $K$ -sparse approximation of the target signal:  $x$ 

// Initialization
 $z^0 = 0;$  // Initial approximation
 $r^0 = y;$  // Residual  $y - \Phi z$ 
 $k = 0;$  // Iteration counter

repeat
     $k \leftarrow k + 1;$ 
     $p \leftarrow \Phi^H r^{k-1};$  // Compute residual proxy
     $b \leftarrow z^{k-1} + p;$  // Add residual proxy to previous estimate
     $z^k = b|_K;$  // Prune to obtain next approximation
     $r^k \leftarrow y - \Phi z^k;$  // Update residual
until halting criteria is true;

```

FIGURE 8.1. Iterative hard thresholding for sparse signal recovery

- $y = \Phi x + e$  represents the measurement vector belonging to  $\mathbb{C}^M$ . This is known to us.
- $e \in \mathbb{C}^M$  represents the measurement noise which is unknown to us within the algorithm.
- $k$  represents the iteration (or step) counter within the algorithm.
- $z \in \mathbb{C}^N$  represents our estimate of  $x$ .  $z$  is updated iteratively.  $z^k$  represents the estimate of  $x$  at the end of  $k$ -th iteration.
- $d^k = x - z^k$  denotes the recovery error at the end of  $k$ -th iteration.
- We start with  $z^k = 0$  and update it in each cycle.
- $r \in \mathbb{C}^M$  represents the difference between the actual measurement vector  $y$  and estimated measurement vector  $\Phi z$ .  $r^k$  is updated at the end of each iteration.
- $p \in \mathbb{C}^N$  represents a proxy for the recovery error  $d^{k-1}$ .

- $b$  is the sum of previous estimate  $z^{k-1}$  and proxy for recovery error  $p$ .

The core of this algorithm is so simple that it can be summarized into just one line:

$$z^k = H_K(z^{k-1} + \Phi^H(y - \Phi z^{k-1})) \quad (8.1.1)$$

where  $H_K$  is a hard thresholding operator which keeps the  $K$  largest entries in its input and sets all others to 0. Since we iteratively improve the estimate and apply hard thresholding operator in each iteration, hence the name of the algorithm is iterative hard thresholding.

### 8.1.2. IHTanalysis sparse case

In this subsection, we will carry out a detailed theoretical analysis of IHTalgorithm for sparse signal recovery.

We will make following assumptions in the analysis:

- The sparsity level  $K$  is fixed and known in advance.
- The signal  $x$  is  $K$ -sparse (i.e.  $x \in \Sigma_K \subset \mathbb{C}^N$ ).
- The sensing matrix  $\Phi$  satisfies RIP of order  $3K$  with  $\delta_{3K} \leq \frac{1}{\sqrt{32}}$ .
- Measurement error  $e \in \mathbb{C}^M$  is arbitrary.

As usual we have the equation connecting  $r$  and  $d$

$$r^{k-1} = \Phi d^{k-1} + e. \quad (8.1.2)$$

We will consider some iteration  $k \geq 1$  and analyze the behavior of the core loop of IHT. We will need some additional notation for the analysis:

- We define  $\Lambda$  as support of best  $K$ -term representation of  $x$ .

$$\Lambda = \text{supp}(x|_K) = \text{supp}(x) \quad (8.1.3)$$

since it is assumed that  $x$  is  $K$ -sparse.

- We define  $\Omega$  as the support of  $K$  largest components of  $b$  (in  $k$ -th iteration) i.e.

$$\Omega^k = \text{supp}(b^k|_K). \quad (8.1.4)$$

- Also we define  $S$  and  $T$  index sets as follows

$$S = \Lambda \cup \Omega^{k-1}, \quad T = \Lambda \cup \Omega^k. \quad (8.1.5)$$

- We will also need  $S \setminus T$ . Let us define

$$U = S \setminus T. \quad (8.1.6)$$

A word of caution: in the following analysis when we write  $v_T$  for some vector  $v \in \mathbb{C}^R$  then we will be conveniently switching between the two interpretations where in one we think  $v_T \in \mathbb{C}^{|T|}$  while in other interpretation we say that  $v_T \in \mathbb{C}^R$  with  $v_{T^c}$  set to 0.

Let us examine the support of  $d^k$ :

$$\text{supp}(d^k) = \text{supp}(x - z^k) \subseteq \text{supp}(x) \cup \text{supp}(z^k) = \Lambda \cup \Omega^k = T. \quad (8.1.7)$$

Thus we can safely write

$$d^k = x - z^k = x_T - z_T^k. \quad (8.1.8)$$

Let us now bring  $b$  in picture.

$$\|x_T - z_T^k\|_2 = \|x_T - b_T + b_T - z_T^k\|_2 \leq \|x_T - b_T\|_2 + \|z_T^k - b_T\|_2. \quad (8.1.9)$$

We can consider both  $x_T$  and  $z_T^k$  as  $K$  term approximations of  $b_T$ . But since  $z_T^k$  is the best  $K$ -term approximation of  $b$  hence we have

$$\|z_T^k - b_T\|_2 \leq \|x_T - b_T\|_2. \quad (8.1.10)$$

This gives us

$$\|x - z^k\|_2 \leq 2\|x_T - b_T\|_2. \quad (8.1.11)$$

We now expand  $b$ :

$$b = z^{k-1} + p = z^{k-1} + \Phi^H \Phi d^{k-1} + \Phi^H e.$$

Thus

$$x_T - b_T = x_T - z_T^{k-1} - \Phi_T^H \Phi d^{k-1} - \Phi_T^H e.$$

But

$$x_T - z_T^{k-1} = d_T^{k-1}.$$

Also

$$\text{supp}(d^{k-1}) = \text{supp}(x - z^{k-1}) \subseteq \text{supp}(x) \cup \text{supp}(z^{k-1}) = \Lambda \cup \Omega^{k-1} = S.$$

Thus

$$d^{k-1} = d_S^{k-1}.$$

Hence we can write

$$\Phi d^{k-1} = \Phi_S d_S^{k-1}$$

since only columns of  $\Phi$  indexed by  $S$  will be involved in the product.

Recalling  $U = S \setminus T$ , we can now split

$$d_S^{k-1} = d_T^{k-1} + d_U^{k-1}.$$

This lets us write

$$\Phi d^{k-1} = \Phi_T d_T^{k-1} + \Phi_U d_U^{k-1}.$$

We can combine these observations to write

$$\begin{aligned} x_T - b_T &= d_T^{k-1} - \Phi_T^H \Phi_T d_T^{k-1} - \Phi_T^H \Phi_U d_U^{k-1} - \Phi_T^H e \\ &= (I - \Phi_T^H \Phi_T) d_T^{k-1} - \Phi_T^H \Phi_U d_U^{k-1} - \Phi_T^H e. \end{aligned}$$

Applying triangle equality we get

$$\begin{aligned} \|x - z^k\|_2 &\leq 2\|x_T - b_T\|_2 \\ &\leq 2\|(I - \Phi_T^H \Phi_T) d_T^{k-1}\|_2 + 2\|\Phi_T^H \Phi_U d_U^{k-1}\|_2 + 2\|\Phi_T^H e\|_2. \end{aligned} \tag{8.1.12}$$

Let us look at the terms one by one.  $T = \Lambda \cup \Omega^k$ , hence  $|T| \leq 2K$ .

Applying theorem 3.19 we get

$$\|(I - \Phi_T^H \Phi_T) d_T^{k-1}\|_2 \leq \delta_{2K} \|d_T^{k-1}\|_2.$$

Next since  $T$  and  $U$  are disjoint and  $T \cup U = \Lambda \cup \Omega^k \cup \Omega^{k-1}$  hence application of theorem 3.20 gives us

$$\|\Phi_T^H \Phi_U d_U^{k-1}\|_2 \leq \delta_{3K} \|d_U^{k-1}\|_2.$$

Finally theorem 3.16 gives us

$$\|\Phi_T^H e\|_2 \leq \sqrt{1 + \delta_{2K}} \|e\|_2.$$

Combining these results we get

$$\|x - z^k\|_2 \leq 2\delta_{2K}\|d_T^{k-1}\|_2 + 2\delta_{3K}\|d_U^{k-1}\|_2 + \sqrt{1 + \delta_{2K}}\|e\|_2.$$

Since  $\delta_{2K} \leq \delta_{3K}$  we can merge the first two terms on R.H.S. as

$$\delta_{2K}\|d_T^{k-1}\|_2 + \delta_{3K}\|d_U^{k-1}\|_2 \leq \delta_{3K}(\|d_T^{k-1}\|_2 + \|d_U^{k-1}\|_2).$$

Since  $S$  and  $U$  are disjoint hence  $d_T^{k-1}$  and  $d_U^{k-1}$  are orthogonal. Recalling

$$d^{k-1} = d_S^{k-1} = d_T^{k-1} + d_U^{k-1}.$$

we have

$$\|d^{k-1}\|_2^2 = \|d_T^{k-1}\|_2^2 + \|d_U^{k-1}\|_2^2.$$

Thus using Pythagorean inequality we have

$$\|d_T^{k-1}\|_2 + \|d_U^{k-1}\|_2 \leq \sqrt{2}\|d^{k-1}\|_2.$$

Therefore we get

$$\|x - z^k\|_2 \leq 2\sqrt{2}\delta_{3K}\|d^{k-1}\|_2 + \sqrt{1 + \delta_{2K}}\|e\|_2.$$

Putting  $\delta_{2K} \leq \delta_{3K} \leq \frac{1}{\sqrt{32}}$  we get

$$\|x - z^k\|_2 \leq 0.5\|x - z^{k-1}\|_2 + 2.17\|e\|_2. \quad (8.1.13)$$

Further summing over iterations we get

$$\|x - z^k\|_2 \leq 2^{-k}\|x\|_2 + 4.34\|e\|_2. \quad (8.1.14)$$

### 8.1.3. IHTanalysis general case

## CHAPTER 9

### Union of Orthonormal Bases

In this chapter we consider dictionaries which are made up of multiple orthonormal bases. The discussion is drawn from [24].

As usual we will work with an overcomplete (full rank) dictionary  $\mathcal{D} \in \mathbb{C}^{N \times D}$  with coherence  $\mu = \mu(\mathcal{D})$ . The sparse recovery problems we will attempt to solve are restated below.

Exact sparse problem. Given a signal  $x \in \mathbb{C}^N$  which is known to have a sparse representation in a dictionary  $\mathcal{D}$ , the exact-sparse recovery problem is:

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_0 \text{ subject to } x = \mathcal{D}\alpha. \quad (\text{P}_0)$$

Sparse approximation with sparsity bound. When  $x \in \mathbb{C}^N$  doesn't have a sparse representation in  $\mathcal{D}$ , a  $K$ -sparse approximation of  $x$  in  $\mathcal{D}$  can be obtained by solving the following problem:

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|x - \mathcal{D}\alpha\|_2 \text{ subject to } \|\alpha\|_0 \leq K. \quad (\text{P}_0^K)$$

Here  $x$  is modeled as  $x = \mathcal{D}\alpha + e$  where  $\alpha$  denotes a sparse representation of  $x$  and  $e$  denotes the approximation error.

Sparse approximation with approximation error bound. A different way to formulate the approximation problem is to provide an upper bound to the acceptable approximation error  $\|e\|_2 \leq \epsilon$  and try to find sparsest possible representation within this approximation error bound as

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_0 \text{ subject to } \|x - \mathcal{D}\alpha\|_2 \leq \epsilon. \quad (\text{P}_0^\epsilon)$$

Exact  $l_p$  norm minimization problem. The corresponding problem for  $(P_0)$  with  $l_p$ -pseudo-norm for  $0 \leq p \leq 1$  is

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_p \text{ subject to } x = \mathcal{D}\alpha. \quad (P_p)$$

Basis pursuit. In particular, when  $p = 1$ , we get the basis pursuit problem:

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_1 \text{ subject to } x = \mathcal{D}\alpha. \quad (P_1)$$

### 9.1. Sparse $l_p$ representations

**Theorem 9.1** *Let  $\mathcal{D}$  be a dictionary and  $\Lambda \subseteq \Omega = \{1, \dots, D\}$  a set of indices. For  $0 \leq p \leq 1$  define*

$$P_p(\Lambda, \mathcal{D}) \triangleq \max_{h \in \mathcal{N}(\mathcal{D}), h \neq 0} \frac{\sum_{k \in \Lambda} |h_k|^p}{\sum_k |h_k|^p} \quad (9.1.1)$$

where we use the convention  $0^0 = 0$ .

1. If  $P_p(\Lambda, \mathcal{D}) < \frac{1}{2}$  then for all  $\alpha$  such that  $\text{supp}(\alpha) \subseteq \Lambda$ ,  $\alpha$  is the unique solution to the problem  $(P_p)$ .
2. If  $P_p(\Lambda, \mathcal{D}) = \frac{1}{2}$  then for all  $\alpha$  such that  $\text{supp}(\alpha) \subseteq \Lambda$ ,  $\alpha$  is a solution to the problem  $(P_p)$ .
3. If  $P_p(\Lambda, \mathcal{D}) > \frac{1}{2}$  then there exists  $\alpha$  such that  $\text{supp}(\alpha) \subseteq \Lambda$ , and  $\beta$  (not supported on  $\Lambda$ ) such that  $\|\beta\|_p < \|\alpha\|_p$  and  $\mathcal{D}\alpha = \mathcal{D}\beta$ . Thus  $(P_p)$  will not return a solution supported over  $\Lambda$ .

The quantity  $P_p(\Lambda, \mathcal{D})$  measures the concentration of null space (of  $\mathcal{D}$ ) vectors over the index set  $\Lambda$ .  $P_p(\Lambda, \mathcal{D}) = \frac{1}{2}$  means that there is at least one vector in the null space of  $\mathcal{D}$  whose half of the energy (in terms of  $l_p$  pseudo-norm) is concentrated over the indices in  $\Lambda$ .

PROOF. We start with the case for  $P_p(\Lambda, \mathcal{D}) < \frac{1}{2}$ . We know  $x$  which has been synthesized from a representation  $\alpha$  with  $\text{supp}(\alpha) \subseteq \Lambda$ . We need to show that  $\alpha$  indeed minimizes  $(P_p)$ . Any other feasible vector for  $(P_p)$  is given by  $\alpha + h$  where  $h \in \mathcal{N}(\mathcal{D})$ .

Thus, we need to show that

$$\|\alpha + h\|_p^p > \|\alpha\|_p^p \iff \sum_k |\alpha_k + h_k|^p > \sum_k |\alpha_k|^p$$

holds true for every nonzero  $h \in \mathcal{N}(\mathcal{D})$ . Since  $\alpha$  is supported over  $\Lambda$ , hence this is equivalent to show that

$$\sum_{k \notin \Lambda} |h_k|^p + \sum_{k \in \Lambda} (|\alpha_k + h_k|^p - |\alpha_k|^p) > 0.$$

Although the triangle inequality is not valid for  $0 < p < 1$ , but the quasi-triangle inequality still works which states

$$|a + b|^p \leq |a|^p + |b|^p.$$

With slight manipulation, we can rewrite this as

$$|a + b|^p - |a|^p \geq -|b|^p.$$

Thus,

$$\sum_{k \in \Lambda} (|\alpha_k + h_k|^p - |\alpha_k|^p) \geq - \sum_{k \in \Lambda} |h_k|^p.$$

Thus if

$$\sum_{k \notin \Lambda} |h_k|^p - \sum_{k \in \Lambda} |h_k|^p > 0$$

holds, then it is sufficient condition for

$$\sum_{k \notin \Lambda} |h_k|^p + \sum_{k \in \Lambda} (|\alpha_k + h_k|^p - |\alpha_k|^p) > 0.$$

to hold true also.

This is equivalent to writing

$$\sum_{k \notin \Lambda} |h_k|^p + \sum_{k \in \Lambda} |h_k|^p > 2 \sum_{k \in \Lambda} |h_k|^p$$

or

$$\sum_{k \in \Lambda} |h_k|^p < \frac{1}{2} \sum_k |h_k|^p \iff \frac{\sum_{k \in \Lambda} |h_k|^p}{\frac{1}{2} \sum_k |h_k|^p} < \frac{1}{2}.$$

Since this condition should hold for every nonzero  $h \in \mathcal{N}(\mathcal{D})$ , by maximizing on the L.H.S., the sufficient condition can be written as

$$P_p(\Lambda, \mathcal{D}) < \frac{1}{2}$$

which is exactly the condition assumed. Thus, whenever  $P_p(\Lambda, \mathcal{D}) < \frac{1}{2}$  and  $\alpha$  is supported over  $\Lambda$ , it is the unique solution for  $(\mathbf{P}_p)$ .

Let us revisit the argument with the relaxed requirement that we want  $\alpha$  to be just a solution of  $(\mathbf{P}_p)$  (it need not be unique). Thus, we need to show that

$$\|\alpha + h\|_p^p \geq \|\alpha\|_p^p.$$

Following, the same argument, this results in the sufficient condition:

$$\sum_{k \in \Lambda} |h_k|^p \leq \frac{1}{2} \sum_k |h_k|^p$$

or (considering every nonzero  $h \in \mathcal{N}(\mathcal{D})$ )

$$P_p(\Lambda, \mathcal{D}) \leq \frac{1}{2}.$$

Since, we already know that  $P_p(\Lambda, \mathcal{D}) < \frac{1}{2}$  guarantees uniqueness of  $\alpha$  as the solution, hence  $\alpha$  is a solution of  $(\mathbf{P}_p)$  whenever  $P_p(\Lambda, \mathcal{D}) = \frac{1}{2}$ .

Finally, we need to show that the condition is sharp. Assume that  $P_p(\Lambda, \mathcal{D}) > \frac{1}{2}$ . Thus, there exists some  $h \in \mathcal{N}(\mathcal{D})$  such that

$$\sum_{k \in \Lambda} |h_k|^p > \frac{1}{2} \sum_k |h_k|^p$$

Define  $\alpha$  as  $\alpha_k = -h_k \forall k \in \Lambda$  and  $\alpha_k = 0 \forall k \notin \Lambda$ . Consider  $\beta = \alpha + h$ . This gives us  $\mathcal{D}\beta = \mathcal{D}\alpha + \mathcal{D}h = \mathcal{D}\alpha = x$ . Further,  $\beta_k = 0 \forall k \in \Lambda$  and  $\beta_k = h_k \forall k \notin \Lambda$ . Clearly,

$$\|\alpha\|_p^p = \sum_{k \in \Lambda} |h_k|^p$$

and

$$\|\beta\|_p^p = \sum_{k \notin \Lambda} |h_k|^p = \sum_k |h_k|^p - \sum_{k \in \Lambda} |h_k|^p < \sum_{k \in \Lambda} |h_k|^p = \|\alpha\|_p^p.$$

And  $\beta$  is supported outside  $\Lambda$ . Clearly, the program  $(\mathbf{P}_p)$  will never find  $\alpha$ .  $\square$

So how does theorem 9.1 help us? The theorem presents recovery guarantee for a given index set  $\Lambda$ . If  $K = |\Lambda|$ , then the recovered representation is  $K$ -sparse. But during the signal recovery,  $\Lambda$  is not

known in advance. Hence, we will look for guarantees which work uniformly for all index sets  $\Lambda$  with  $K = |\Lambda|$ .

The guarantees should take the following form. If  $|\Lambda| \leq K_0$  where  $K_0$  is some number depending on the dictionary  $\mathcal{D}$ , then  $P_p(\Lambda, \mathcal{D}) < \frac{1}{2}$ . theorem 9.1 can automatically be invoked along with such a guarantee to ensure that for every  $\alpha$  with  $\|\alpha\|_0 \leq K_0$ , the program  $(P_p)$  will recover it.

**Theorem 9.2** *Let  $\text{spark}_{1/2}(\mathcal{D}) = \left\lceil \frac{\text{spark}(\mathcal{D})}{2} \right\rceil$ . A guarantee of the form*

$$\text{If } |\Lambda| < f(\mathcal{D}), \text{ then } P_0(\Lambda, \mathcal{D}) < \frac{1}{2} \quad (9.1.2)$$

*holds if and only if  $f(\mathcal{D}) \leq \text{spark}_{1/2}(\mathcal{D})$ .*

PROOF. For any nonzero  $h \in \mathcal{N}(\mathcal{D})$  observe that

$$\frac{\sum_{k \in \Lambda} |h_k|^0}{\sum_k |h_k|^0} \leq \frac{|\Lambda|}{\|h\|_0}.$$

We also note that

$$\|h\|_0 \geq \text{spark}(\mathcal{D}).$$

Thus

$$\frac{\sum_{k \in \Lambda} |h_k|^0}{\sum_k |h_k|^0} \leq \frac{|\Lambda|}{\text{spark}(\mathcal{D})}.$$

Finally, taking maximum on both sides

$$P_0(\Lambda, \mathcal{D}) = \max_{h \in \mathcal{N}(\mathcal{D}), h \neq 0} \frac{\sum_{k \in \Lambda} |h_k|^p}{\sum_k |h_k|^p} \leq \frac{|\Lambda|}{\text{spark}(\mathcal{D})}.$$

Note that R.H.S. doesn't depend explicitly on  $h$ . Hence, taking maximum has no effect.

Now, we assume that  $|\Lambda| < f(\mathcal{D}) \leq \text{spark}_{1/2}(\mathcal{D})$ . If  $\text{spark}(\mathcal{D})$  is even, then  $\text{spark}_{1/2}(\mathcal{D}) = \text{spark}(\mathcal{D})/2$ . Thus,

$$\frac{|\Lambda|}{\text{spark}(\mathcal{D})} < \frac{\text{spark}_{1/2}(\mathcal{D})}{\text{spark}(\mathcal{D})} = \frac{1}{2}.$$

If  $\text{spark}(\mathcal{D})$  is odd, then  $\text{spark}_{1/2}(\mathcal{D}) = \frac{\text{spark}(\mathcal{D})+1}{2}$ . Now,

$$|\Lambda| < f \leq \frac{\text{spark}(\mathcal{D}) + 1}{2} \iff |\Lambda| + 1 \leq \frac{\text{spark}(\mathcal{D}) + 1}{2} \iff |\Lambda| \leq \frac{\text{spark}(\mathcal{D}) - 1}{2}.$$

Thus

$$P_0(\Lambda, \mathcal{D}) \leq \frac{|\Lambda|}{\text{spark}(\mathcal{D})} \leq \frac{\frac{\text{spark}(\mathcal{D})-1}{2}}{\text{spark}(\mathcal{D})} \leq \frac{1}{2} - \frac{1}{2\text{spark}(\mathcal{D})} < \frac{1}{2}.$$

Thus, we see that if  $f(\mathcal{D}) \leq \text{spark}_{1/2}(\mathcal{D})$  then a guarantee of the form

$$\text{If } |\Lambda| < f(\mathcal{D}), \text{ then } P_0(\Lambda, \mathcal{D}) < \frac{1}{2} \quad (9.1.3)$$

holds.

We now show the converse. i.e. if  $f(\mathcal{D}) > \text{spark}_{1/2}(\mathcal{D})$  then a guarantee of the form above doesn't hold.

We know that there exists some  $h \in \mathcal{N}(\mathcal{D})$  such that  $\|h\|_0 = \text{spark}(\mathcal{D})$ . Let  $\text{spark}(\mathcal{D})$  be even. Then,  $f(\mathcal{D}) > \text{spark}(\mathcal{D})/2$ . We can pick up some  $\Lambda \subset \text{supp}(h)$  with  $|\Lambda| = \text{spark}(\mathcal{D})/2$ . For this  $|\Lambda| < f(\mathcal{D})$  holds, but

$$\frac{\sum_{k \in \Lambda} |h_k|^0}{\sum_k |h_k|^0} = \frac{\text{spark}(\mathcal{D})/2}{\text{spark}(\mathcal{D})} = \frac{1}{2}$$

leading to

$$P_0(\Lambda, \mathcal{D}) \geq \frac{1}{2}.$$

We can similarly prove this for the case when spark is odd.

□

**Theorem 9.3** *If a guarantee of the form*

$$\text{If } |\Lambda| < f(\mathcal{D}), \text{ then } P_p(\Lambda, \mathcal{D}) < \frac{1}{2} \quad (9.1.4)$$

*holds true for some  $0 \leq p \leq 1$  with some  $f(\mathcal{D})$ , then it also holds true for  $p = 0$ , hence  $f(\mathcal{D}) \leq \text{spark}_{1/2}(\mathcal{D})$ .*

PROOF. Assume that (9.3.1) holds true for some  $0 \leq p \leq 1$ . Let  $\alpha$  be such that  $\|\alpha\|_0 < f(\mathcal{D})$ . Then for all  $\beta \neq \alpha$  such that  $\mathcal{D}\beta = \mathcal{D}\alpha$  we

have that

$$\|\beta\|_p > \|\alpha\|_p$$

i.e.  $\beta$  cannot be the solution of the program  $(P_p)$ . Now for the sake of contradiction assume that there exists some  $\beta$  such that  $\mathcal{D}\beta = \mathcal{D}\alpha$  and  $\|\beta\|_0 \leq \|\alpha\|_0$ . Since  $\|\alpha\|_0 < f(\mathcal{D})$  hence,  $\|\beta\|_0 < f(\mathcal{D})$ .

This means that  $\beta$  is also a unique minimizer of the same program  $(P_p)$ . But, since the minimizer is unique, hence  $\alpha = \beta$ .

Consequently,  $\alpha$  is indeed the unique minimizer of the  $(P_0)$  program. Applying theorem 9.2, we get

$$f(\mathcal{D}) \leq \text{spark}_{1/2}(\mathcal{D}).$$

□

We now develop a bound on sparsity which holds for unique recovery of both  $l_0$  and  $l_1$  minimization problems.

**Theorem 9.4** *For any dictionary  $\mathcal{D}$  with coherence  $\mu$ , if*

$$\|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right) \quad (9.1.5)$$

*then  $\alpha$  is the unique solution to both the  $(P_0)$  and  $(P_1)$  problems.*

PROOF. Let  $\Lambda = \text{supp}(\alpha)$ . We just need to show that

$$\text{If } |\Lambda| < f(\mathcal{D}) = \frac{1}{2} \left( 1 + \frac{1}{\mu} \right), \text{ then } P_1(\Lambda, \mathcal{D}) < \frac{1}{2}$$

holds.

Let

$$\mathcal{D} = [d_1 \ \dots \ d_D].$$

Let  $h \in \mathcal{N}(\mathcal{D})$  and  $h \neq 0$ . Then

$$\begin{aligned} & \sum_{k=1}^D h_k d_k = 0 \\ \iff & h_k d_k = - \sum_{l \neq k} h_l d_l. \end{aligned}$$

Taking inner product on both sides with  $d_k$  (recall that atoms of  $\mathcal{D}$  are unit norm)

$$h_k = - \sum_{l \neq k} h_l d_k^H d_l.$$

Take absolute value on both sides (recall that coherence is highest inner product between atoms)

$$|h_k| = \left| \sum_{l \neq k} h_l d_k^H d_l \right| \leq \sum_{l \neq k} |h_l| |d_k^H d_l| \leq \mu \sum_{l \neq k} |h_l|.$$

Add  $\mu|h_k|$  on both sides.

$$(1 + \mu)|h_k| \leq \mu \sum_{l=1}^D |h_l| = \mu \|h\|_1.$$

Sum the last inequality over  $k \in \Lambda$ . We get

$$\begin{aligned} (1 + \mu) \sum_{k \in \Lambda} |h_k| & \leq \mu |\Lambda| \|h\|_1 = \mu \|\alpha\|_0 \|h\|_1 \\ \iff \frac{\sum_{k \in \Lambda} |h_k|}{\sum_k |h_k|} & \leq \|\alpha\|_0 \frac{\mu}{1 + \mu}. \end{aligned}$$

Maximizing the inequality over all  $h \in \mathcal{N}(\mathcal{D})$  with  $h \neq 0$ , we get

$$P_1(\Lambda, \mathcal{D}) \leq \|\alpha\|_0 \frac{\mu}{(1 + \mu)}.$$

Thus,

$$\text{if } |\Lambda| = \|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right), \text{ then } P_1(\Lambda, \mathcal{D}) < \frac{1}{2}.$$

□

## 9.2. Union of bases

We now consider dictionaries which are a union of  $L$  orthonormal bases. We will denote individual orthonormal bases as  $B_1, B_2, \dots, B_L$ . We will write our dictionary as

$$\mathcal{D} = [B_1 \ B_2 \ \dots \ B_L]. \quad (9.2.1)$$

### 9.2.1. Spark and coherence

We will examine the null space of a union of bases dictionary and develop a bound on the spark in terms of its coherence.

**Theorem 9.5** *Let  $\mathcal{D}$  be a union of  $L$  orthonormal bases. Let  $h = [h^1, \dots, h^L] \in \mathcal{N}(\mathcal{D})$  with  $h^i \in \mathbb{C}^N$  and assume that  $h \neq 0$ . Then*

$$\sum_{l=1}^L \frac{1}{1 + \mu \|h^l\|_0} \leq L - 1. \quad (9.2.2)$$

Consequently,

$$\text{spark}(\mathcal{D}) \geq \left(1 + \frac{1}{L-1}\right) \frac{1}{\mu}. \quad (9.2.3)$$

PROOF. Since  $h \in \mathcal{N}(\mathcal{D})$ , hence

$$\mathcal{D}h = 0 \iff \sum_{l=1}^L B_l h^l = 0.$$

For the  $l$ -th term, we can rewrite this as

$$B_l h^l = - \sum_{k \neq l} B_k h^k.$$

Since  $B_l$  is an ONB, hence  $B_l^H B_l = I$ . Multiplying both sides by  $B_l^H$  we get

$$h^l = - \sum_{k \neq l} B_l^H B_k h^k.$$

Taking absolute values on both sides we get

$$|h^l| = \left| \sum_{k \neq l} B_l^H B_k h^k \right| \preceq \sum_{k \neq l} |B_l^H B_k h^k|.$$

Note that the symbol  $\preceq$  means component wise inequality.

Let us look at the term  $|B_l^H B_k h^k|$  more closely. If we write

$$B_l = \begin{bmatrix} b_{l1} & \dots & b_{lN} \end{bmatrix}$$

where  $b_{li}$  are the column vectors of  $B$ . Then

$$B_l^H B_k = \begin{bmatrix} b_{l1}^H \\ \vdots \\ b_{lN}^H \end{bmatrix} \begin{bmatrix} b_{k1} & \dots & b_{kN} \end{bmatrix} = \left[ b_{li}^H b_{kj} \right]_{1 \leq i, j \leq N}.$$

Thus the column vector

$$|B_l^H B_k h^k| = \left[ \left| \sum_{j=1}^N b_{li}^H b_{kj} h_j^k \right| \right]_{1 \leq i \leq N}.$$

But for each  $1 \leq i \leq N$

$$\left| \sum_{j=1}^N b_{li}^H b_{kj} h_j^k \right| \leq \sum_{j=1}^N |b_{li}^H b_{kj} h_j^k| \leq \mu \sum_{j=1}^N |h_j^k| = \mu \|h^k\|_1.$$

Thus

$$|B_l^H B_k h^k| \preceq \mu \|h^k\|_1 \mathbb{1}_N$$

where  $\mathbb{1}_N$  is a vector of all ones. Putting back we get

$$|h^l| \preceq \sum_{k \neq l} \mu \|h^k\|_1 \mathbb{1}_N.$$

Thus, each of the entries in  $|h^l|$  is upper bounded by  $\mu \sum_{k \neq l} \|h^k\|_1$ .

Now, there are  $\|h^l\|_0$  non-zero entries in  $|h^l|$ . Thus,

$$\|h^l\|_1 \leq \mu \|h^l\|_0 \sum_{k \neq l} \|h^k\|_1.$$

Adding  $\mu \|h^l\|_0 \|h^l\|_1$  on both sides, we get

$$(1 + \mu \|h^l\|_0) \|h^l\|_1 \leq \mu \|h^l\|_0 \sum_k \|h^k\|_1 = \mu \|h^l\|_0 \|h\|_1.$$

We used the fact that  $\|h\|_1 = \sum_k \|h^k\|_1$ .

This gives us

$$\|h^l\|_1 \leq \frac{\mu\|h^l\|_0\|h\|_1}{1 + \mu\|h^l\|_0}.$$

Again summing over  $l$  on both sides, we get

$$\|h\|_1 \leq \sum_{l=1}^L \frac{\mu\|h^l\|_0\|h\|_1}{1 + \mu\|h^l\|_0}.$$

Canceling  $\|h\|_1$  we obtain

$$\sum_{l=1}^L \frac{\mu\|h^l\|_0}{1 + \mu\|h^l\|_0} \geq 1.$$

A small set of steps follow:

$$\begin{aligned} & - \sum_{l=1}^L \frac{\mu\|h^l\|_0}{1 + \mu\|h^l\|_0} \leq -1 \\ \iff & L - \sum_{l=1}^L \frac{\mu\|h^l\|_0}{1 + \mu\|h^l\|_0} \leq L - 1 \\ \iff & \sum_{l=1}^L \left[ 1 - \frac{\mu\|h^l\|_0}{1 + \mu\|h^l\|_0} \right] \leq L - 1 \\ \iff & \sum_{l=1}^L \frac{1}{1 + \mu\|h^l\|_0} \leq L - 1. \end{aligned}$$

This is the desired result in (9.2.2).

To show (9.2.3) we proceed as follows.

We recall that

$$\text{spark}(\mathcal{D}) \geq \|h\|_0 = \sum_{l=1}^N \|h^l\|_0.$$

Consider the function

$$g(y) = \frac{1}{1 + y}.$$

The function is convex over the interval  $(-1, \infty)$ , hence

$$g\left(\frac{x+y}{2}\right) \leq \frac{g(x) + g(y)}{2}.$$

More generally

$$g\left(\frac{\sum_{l=1}^L y_l}{L}\right) \leq \frac{\sum_{l=1}^L g(y_l)}{L}.$$

Choosing  $y_l = \mu \|h^l\|_0$  we get

$$\begin{aligned} g\left(\frac{\sum_{l=1}^L \mu \|h^l\|_0}{L}\right) &= g\left(\frac{\mu \|h\|_0}{L}\right) \leq \frac{\sum_{l=1}^L g(\mu \|h^l\|_0)}{L} \\ &= \frac{1}{L} \sum_{l=1}^L \frac{1}{1 + \mu \|h^l\|_0} \leq \frac{L-1}{L}. \end{aligned}$$

Thus

$$\begin{aligned} \frac{1}{1 + \frac{\mu \|h\|_0}{L}} &\leq \frac{L-1}{L} \\ \iff 1 + \frac{\mu \|h\|_0}{L} &\geq \frac{L}{L-1} \\ \iff \frac{\mu \|h\|_0}{L} &\geq \frac{L}{L-1} - 1 = \frac{1}{L-1} \\ \iff \mu \|h\|_0 &\geq \frac{L}{L-1} = \frac{L-1+1}{L-1} = 1 + \frac{1}{L-1} \\ \iff \|h\|_0 &\geq \left[1 + \frac{1}{L-1}\right] \frac{1}{\mu}. \end{aligned}$$

This gives us the desired result (9.2.3):

$$\text{spark}(\mathcal{D}) \geq \|h\|_0 \geq \left[1 + \frac{1}{L-1}\right] \frac{1}{\mu}.$$

□

**Challenge** Can we obtain tighter bounds using Babel function?

Let us look at the special case for two-ortho bases (with  $L = 2$ ). Our first result is:

$$\sum_{l=1}^2 \frac{1}{1 + \mu \|h^l\|_0} \leq 2 - 1 = 1. \quad (9.2.4)$$

We can rewrite it as

$$\begin{aligned}
 & \frac{1}{1 + \mu\|h^1\|_0} + \frac{1}{1 + \mu\|h^2\|_0} \leq 1 \\
 \iff & 1 + \mu\|h^1\|_0 + 1 + \mu\|h^2\|_0 \leq 1 + \mu\|h^1\|_0 + \mu\|h^2\|_0 + \mu^2\|h^1\|_0\|h^2\|_0 \\
 \iff & 1 \leq \mu^2\|h^1\|_0\|h^2\|_0 \\
 \iff & \|h^1\|_0\|h^2\|_0 \geq \frac{1}{\mu^2} \\
 \iff & \sqrt{\|h^1\|_0\|h^2\|_0} \geq \frac{1}{\mu}.
 \end{aligned}$$

Thus, the null space vectors for two ortho bases satisfy the condition

$$\sqrt{\|h^1\|_0\|h^2\|_0} \geq \frac{1}{\mu}.$$

The condition on spark reduces to

$$\text{spark}(\mathcal{D}) \geq \left(1 + \frac{1}{2-1}\right) \frac{1}{\mu} = \frac{2}{\mu}.$$

There are indeed examples of pairs of bases for which the lower bound of  $\text{spark}(\mathcal{D}) = \frac{2}{\mu}$  is met. Thus, the bound is sharp for  $L = 2$ .

**Challenge** Can we find a set of  $L$  orthonormal bases for which the the lower bound (9.2.3) holds? Starting with  $L = 3$ ?

### 9.2.2. $l_0$ minimization

Let us now extend the argument to identify conditions under which the (P<sub>0</sub>) problem can have a unique solution.

**Theorem 9.6** Let  $\mathcal{D}$  be a union of  $L$  orthonormal bases. If

$$\|\alpha\|_0 < \left(\frac{1}{2} + \frac{1}{2(L-1)}\right) \frac{1}{\mu} \quad (9.2.5)$$

then the unique solution to the (P<sub>0</sub>) is  $\alpha$ .

PROOF. Let

$$f(\mathcal{D}) = \left( \frac{1}{2} + \frac{1}{2(L-1)} \right) \frac{1}{\mu}.$$

Then from theorem 9.5

$$f(\mathcal{D}) \leq \frac{1}{2} \text{spark}(\mathcal{D}) \leq \text{spark}_{1/2}(\mathcal{D})$$

for both even and odd cases. Applying theorem 9.2, we get our result.  $\square$

Looking at the special case  $L = 2$ , the condition reduces to

$$\|\alpha\|_0 < \frac{1}{\mu}$$

then  $\alpha$  is a unique solution to  $(P_0)$  problem.

For  $L = 3$ , we get the upper bound as

$$\|\alpha\|_0 < \frac{3}{4\mu}.$$

So, if coherence  $\mu$  doesn't change, then the level of sparsity for which unique recovery is guaranteed has reduced. We see that the upper bound on sparsity in (9.2.5) gets stricter and stricter as  $L$  (number of orthonormal bases) increases provided the coherence of dictionary remains constant.

Let us compare the bound for general dictionaries in (9.2.7)

$$\|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right)$$

with (9.2.5).

Let us find out the value of  $L$  beyond which the bound in (9.2.5) becomes more restrictive than the general bound in (9.2.7).

$$\begin{aligned}
& \frac{1}{2} \left( 1 + \frac{1}{\mu} \right) \geq \left( \frac{1}{2} + \frac{1}{2(L-1)} \right) \frac{1}{\mu} \\
\iff & 1 + \frac{1}{\mu} \geq \frac{1}{\mu} + \frac{1}{(L-1)\mu} \\
\iff & 1 \geq \frac{1}{(L-1)\mu} \\
\iff & L-1 \geq \frac{1}{\mu} \\
\iff & L \geq 1 + \frac{1}{\mu}.
\end{aligned}$$

Thus for  $L \geq 1 + \frac{1}{\mu}$ , the less restrictive general bound in (9.2.7) should be preferred. At the same time, for smaller values of  $L$ , the specific bound in (9.2.5) should be preferred.

As example, with  $\mu = 0.2$ , for  $L \geq 6$ , the general bound should be used, while for  $1 \leq L \leq 5$  the more specific bound in (9.2.5) should be used.

If the dictionary is even less coherent, with  $\mu = 0.1$ , then for  $L \geq 11$ , the general bound should be used.

### 9.2.3. $l_1$ minimization

We have identified unique recovery conditions for the (P<sub>0</sub>) problem. Let us also identify conditions for the (P<sub>1</sub>) problem with union of bases dictionary. Afterwards we will compare it with the result for general dictionary in theorem 9.4.

Before we prove the main result of this section, a few lemmas are in order which will help establish the main result.

**Lemma 9.7** *For  $p = 1$ ,*

$$P_1(\Lambda, \mathcal{D}) = \max_{h \in \mathcal{N}(\mathcal{D}), \|h\|_1=1} \mathbb{1}_\Lambda^T |h|. \quad (9.2.6)$$

Thus, to show that  $P_1(\Lambda, \mathcal{D}) < \frac{1}{2}$ , it is sufficient to show that

$$\mathbb{1}_\Lambda^T |h| < \frac{1}{2}$$

for all null space vectors with  $\|h\|_1 = 1$ .

PROOF. Let  $h^*$  maximize

$$P_1(\Lambda, \mathcal{D}) = \max_{h \in \mathcal{N}(\mathcal{D}), h \neq 0} \frac{\sum_{k \in \Lambda} |h_k|}{\sum_k |h_k|} = \max_{h \in \mathcal{N}(\mathcal{D}), h \neq 0} \frac{\mathbb{1}_\Lambda^T |h|}{\|h\|_1}.$$

Thus,

$$P_1(\Lambda, \mathcal{D}) = \frac{\mathbb{1}_\Lambda^T |h^*|}{\|h^*\|_1}.$$

Consider  $h' = \frac{h^*}{\|h^*\|_1}$ . Then

$$\|h'\|_1 = \left\| \frac{h^*}{\|h^*\|_1} \right\|_1 = \frac{\|h^*\|_1}{\|h^*\|_1} = 1.$$

Also, note that

$$|h'| = \frac{|h^*|}{\|h^*\|_1}.$$

Thus,

$$\mathbb{1}_\Lambda^T |h'| = \frac{\mathbb{1}_\Lambda^T |h^*|}{\|h^*\|_1} = P_1(\Lambda, \mathcal{D}).$$

Thus, we can write

$$P_1(\Lambda, \mathcal{D}) = \max_{h \in \mathcal{N}(\mathcal{D}), \|h\|_1=1} \mathbb{1}_\Lambda^T |h|.$$

Our result follows. □

We now proceed to the main result of this section.

**Theorem 9.8** *Let  $\mathcal{D}$  be a union of  $L$  orthonormal bases. Denote*

$$\alpha = \begin{bmatrix} \alpha^1 \\ \vdots \\ \alpha^L \end{bmatrix}$$

with  $\alpha^l \in \mathbb{C}^N$ . Without loss of generality, we can assume that the bases  $B_l$  have been arranged so that

$$\|\alpha^1\|_0 \leq \cdots \leq \|\alpha^L\|_0.$$

If

$$\sum_{l \geq 2} \frac{\mu \|\alpha^l\|_0}{1 + \mu \|\alpha^l\|_0} < \frac{1}{2(1 + \mu \|\alpha^1\|_0)} \quad (9.2.7)$$

then  $\alpha$  is the (unique) solution to the (P<sub>1</sub>) problem.

PROOF. Let  $\Lambda = \text{supp}(\alpha)$ . As usual, we will start our work with the analysis of the null space vectors. Let

$$h = \begin{bmatrix} h^1 \\ \vdots \\ h^L \end{bmatrix} \in \mathcal{N}(\mathcal{D})$$

with  $h^l \in \mathbb{C}^N$ . For every  $1 \leq l \leq L$ , we have

$$B_l h^l = - \sum_{k \neq l} B_k h^k.$$

Multiplying with  $B_l^H$  on both sides, we get

$$h^l = - \sum_{k \neq l} B_l^H B_k h^k.$$

Taking absolute values on both sides, we get

$$\begin{aligned} |h^l| &= \left| \sum_{k \neq l} B_l^H B_k h^k \right| \\ &\leq \mu \sum_{k \neq l} \|h^k\|_1 \mathbf{1} \\ &= \mu \sum_{k \neq l} \mathbf{1} \mathbf{1}^T |h^k| \\ &= \mu \sum_{k \neq l} \mathbf{1} |h^k| \end{aligned}$$

where  $\mathbf{1}$  denotes an  $N \times N$  matrix of all ones. The resulting constraint is.

$$|h^l| \leq \mu \sum_{k \neq l} \mathbf{1}_{N \times N} |h^k|. \quad (9.2.8)$$

By definition we have  $|h^l| \succeq 0$ . Since due to lemma 9.7, it is sufficient to focus on unit  $l_1$ -norm vectors in the null space of  $\mathcal{D}$ , let us consider the special case of  $\|h\|_1 = 1$ . We have

$$\|h\|_1 = \sum_{l=1}^L \|h^l\|_1 = 1.$$

We can rewrite this as

$$\sum_{l=1}^L \mathbb{1}_N^T |h^l| = 1. \quad (9.2.9)$$

From (9.2.6) we get

$$P_1(\Lambda, \mathcal{D}) = \max_{h \in \mathcal{N}(\mathcal{D}), \|h\|_1=1} \mathbb{1}_\Lambda^T |h|.$$

To show that  $\alpha$  is the unique solution to (P<sub>1</sub>), we need to show that

$$P_1(\Lambda, \mathcal{D}) < \frac{1}{2}.$$

Let us expand the term  $\mathbb{1}_\Lambda^T |h|$

$$\mathbb{1}_\Lambda^T |h| = \sum_{k \in \Lambda} |h_k| = \sum_{l=1}^L \sum_{k \in \text{supp}(\alpha^l)} |h_k^l| = \sum_{l=1}^L \mathbb{1}_{\text{supp}(\alpha^l)}^T |h^l|.$$

$\mathbb{1}_{\text{supp}(\alpha^l)}$  denotes a vector with ones at the entries indexed by  $\text{supp}(\alpha^l)$  and zeros everywhere else.

Thus, we need to show that under the condition (9.2.7) and the constraints (9.2.8), (9.2.9)

$$\max_{h^1, \dots, h^L} \sum_{l=1}^L \mathbb{1}_{\text{supp}(\alpha^l)}^T |h^l| < \frac{1}{2}$$

holds.

We will restructure the constraints in the form of matrix inequalities and construct a linear program out of it.

Define

$$z \triangleq \begin{bmatrix} |h^1| \\ \vdots \\ |h^L| \end{bmatrix} = |h|..$$

Define

$$v = \begin{bmatrix} -\mathbf{1}_{\text{supp}(\alpha^l)} \\ \vdots \\ -\mathbf{1}_{\text{supp}(\alpha^L)} \end{bmatrix}.$$

Then

$$\sum_{l=1}^L \mathbf{1}_{\text{supp}(\alpha^l)}^T |h^l| = -v^T z.$$

Further,

$$|h| \succeq 0 \iff z \succeq 0.$$

From (9.2.8), we have

$$\begin{aligned} 0 &\preceq -|h^l| + \mu \sum_{k \neq l} \mathbf{1}_{N \times N} |h^k| \\ \iff 0 &\preceq \sum_{k < l} \mu \mathbf{1}_{N \times N} |h^k| - I_{N \times N} |h^l| + \sum_{k > l} \mu \mathbf{1}_{N \times N} |h^k|. \end{aligned}$$

We can put this in matrix form as

$$\begin{bmatrix} -I_{N \times N} & \mu \mathbf{1}_{N \times N} & \dots & \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} \\ \mu \mathbf{1}_{N \times N} & -I_{N \times N} & \dots & \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} & \ddots & -I_{N \times N} & \mu \mathbf{1}_{N \times N} \\ \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} & \ddots & \mu \mathbf{1}_{N \times N} & -I_{N \times N} \end{bmatrix} z \succeq \begin{bmatrix} 0 \cdot \mathbf{1}_N \\ 0 \cdot \mathbf{1}_N \\ \vdots \\ 0 \cdot \mathbf{1}_N \\ 0 \cdot \mathbf{1}_N \end{bmatrix}.$$

$\mathbf{1}_N$  denotes  $N$ -dimensional vector of all ones.

The equality (9.2.9) can be split into two inequalities:

$$\sum_{l=1}^L \mathbf{1}_N^T |h^l| \geq 1$$

and

$$\sum_{l=1}^L \mathbf{1}_N^T |h^l| \leq 1 \iff -\sum_{l=1}^L \mathbf{1}_N^T |h^l| \geq -1.$$

In the matrix form we can write this as

$$\begin{bmatrix} \mathbf{1}_N^T & \dots & \mathbf{1}_N^T \\ -\mathbf{1}_N^T & \dots & -\mathbf{1}_N^T \end{bmatrix} z \succeq \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Define the matrix

$$A \triangleq \begin{bmatrix} -I_{N \times N} & \mu \mathbf{1}_{N \times N} & \dots & \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} \\ \mu \mathbf{1}_{N \times N} & -I_{N \times N} & \dots & \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} & \ddots & -I_{N \times N} & \mu \mathbf{1}_{N \times N} \\ \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} & \ddots & \mu \mathbf{1}_{N \times N} & -I_{N \times N} \\ \mathbf{1}_N^T & \dots & \dots & \dots & \mathbf{1}_N^T \\ -\mathbf{1}_N^T & \dots & \dots & \dots & -\mathbf{1}_N^T \end{bmatrix}$$

and the vector

$$w \triangleq \begin{bmatrix} 0 \cdot \mathbf{1}_N \\ 0 \cdot \mathbf{1}_N \\ \vdots \\ 0 \cdot \mathbf{1}_N \\ 0 \cdot \mathbf{1}_N \\ 1 \\ -1 \end{bmatrix}.$$

We can now combine the constraints (9.2.8) and (9.2.9) into the matrix inequality

$$Az \succeq w.$$

Note that  $z \in \mathbb{R}_+^{NL}$ ,  $A \in \mathbb{R}^{NL+2 \times NL}$  and  $w \in \mathbb{R}^{NL+2}$ .

Let us define the (primal) linear program as

$$\begin{array}{ll} \underset{z}{\text{minimize}} & v^T z \\ \text{subject to} & Az \succeq w, \quad z \succeq 0. \end{array} \quad (\text{primal})$$

We note that optimal value of (primal) program is  $> -\frac{1}{2}$  if and only if  $P_1(\Lambda, \mathcal{D}) < \frac{1}{2}$ .

The corresponding dual linear programming problem is

$$\begin{aligned} & \underset{z}{\text{maximize}} \quad w^T u \\ & \text{subject to} \quad A^T u \preceq v, \quad u \succeq 0. \end{aligned} \tag{dual}$$

Here  $u \in \mathbb{R}^{NL+2}$ . Just for clarify, lets write the problem in expanded form too

$$w^T u = [0 \cdot \mathbf{1}_N^T \ 0 \cdot \mathbf{1}_N^T \ \dots \ 0 \cdot \mathbf{1}_N^T \ 0 \cdot \mathbf{1}_N^T \ 1 \ -1] u$$

$A^T u \preceq v$  is

$$\begin{bmatrix} -I_{N \times N} & \mu \mathbf{1}_{N \times N} & \dots & \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} & \mathbf{1}_N & -\mathbf{1}_N \\ \mu \mathbf{1}_{N \times N} & -I_{N \times N} & \dots & \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} & \mathbf{1}_N & -\mathbf{1}_N \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} & \ddots & -I_{N \times N} & \mu \mathbf{1}_{N \times N} & \mathbf{1}_N & -\mathbf{1}_N \\ \mu \mathbf{1}_{N \times N} & \mu \mathbf{1}_{N \times N} & \ddots & \mu \mathbf{1}_{N \times N} & -I_{N \times N} & \mathbf{1}_N & -\mathbf{1}_N \end{bmatrix} u \preceq \begin{bmatrix} -\mathbf{1}_{\text{supp}(\alpha^1)} \\ -\mathbf{1}_{\text{supp}(\alpha^2)} \\ \vdots \\ -\mathbf{1}_{\text{supp}(\alpha^{L-1})} \\ -\mathbf{1}_{\text{supp}(\alpha^L)} \end{bmatrix} \tag{9.2.10}$$

We know that the optimal value of the two programs are the same. So we need to show that there exists some  $u \succeq 0$  which satisfies  $A^T u \preceq v$  and  $w^T u > -\frac{1}{2}$ .

Due to the structure of matrices  $A$  and vectors  $v, w$ , we will look for a solution in parametric form

$$u \triangleq \begin{bmatrix} a_1 \mathbf{1}_{\text{supp}(\alpha^1)} \\ a_2 \mathbf{1}_{\text{supp}(\alpha^2)} \\ \vdots \\ a_{L-1} \mathbf{1}_{\text{supp}(\alpha^{L-1})} \\ a_L \mathbf{1}_{\text{supp}(\alpha^L)} \\ b \\ c \end{bmatrix}$$

where  $a_l, b, c \geq 0$ . Obviously  $u \succeq 0$  and  $w^T u = b - c$ . Our goal is to choose the  $a_l, b, c$  such that  $b - c > -\frac{1}{2}$  and the constraint  $A^T u \preceq v$  is satisfied.

A straightforward computation over the  $l$ -th (block) row in (9.2.10) gives us (for  $1 \leq l \leq L$ )

$$(b - c)\mathbf{1}_N + \mu \mathbf{1}_{N \times N} \left( \sum_{k \neq l} a_k \mathbf{1}_{\text{supp}(\alpha^k)} \right) - a_l \mathbf{1}_{\text{supp}(\alpha^l)} \preceq -\mathbf{1}_{\text{supp}(\alpha^l)}.$$

Rearranging the terms we get

$$(b - c)\mathbf{1}_N + \mu \mathbf{1}_{N \times N} \left( \sum_{k \neq l} a_k \mathbf{1}_{\text{supp}(\alpha^k)} \right) + (1 - a_l) \mathbf{1}_{\text{supp}(\alpha^l)} \preceq 0.$$

Now

$$\mathbf{1}_{N \times N} \mathbf{1}_{\text{supp}(\alpha^k)} = \mathbf{1}_N \mathbf{1}_N^T \mathbf{1}_{\text{supp}(\alpha^k)} = |\text{supp}(\alpha^k)| \mathbf{1}_N = \|\alpha^k\|_0 \mathbf{1}_N.$$

Thus, we can rewrite as

$$(b - c)\mathbf{1}_N + \mu \left( \sum_{k \neq l} a_k \|\alpha^k\|_0 \mathbf{1}_N \right) + (1 - a_l) \mathbf{1}_{\text{supp}(\alpha^l)} \preceq 0.$$

Or

$$\left( b - c + \mu \sum_{k \neq l} a_k \|\alpha^k\|_0 \right) \mathbf{1}_N + (1 - a_l) \mathbf{1}_{\text{supp}(\alpha^l)} \preceq 0.$$

Or

$$\left( b - c + \mu \sum_{k=1}^L a_k \|\alpha^k\|_0 \right) \mathbf{1}_N - \mu a_l \|\alpha^l\|_0 \mathbf{1}_N + (1 - a_l) \mathbf{1}_{\text{supp}(\alpha^l)} \preceq 0.$$

Or

$$\left( b - c + \mu \sum_{k=1}^L a_k \|\alpha^k\|_0 \right) \mathbf{1}_N \preceq \mu a_l \|\alpha^l\|_0 \mathbf{1}_N + (a_l - 1) \mathbf{1}_{\text{supp}(\alpha^l)} \quad (9.2.11)$$

Consider first the case that  $\|\alpha^l\|_0 \neq 0$ .

Over the indices included in  $\text{supp}(\alpha^l)$ , the inequality is

$$b - c + \mu \sum_{k=1}^L a_k \|\alpha^k\|_0 \leq \mu a_l \|\alpha^l\|_0 + (a_l - 1)$$

Over the indices not included in  $\text{supp}(\alpha^l)$ , the inequality is

$$b - c + \mu \sum_{k=1}^L a_k \|\alpha^k\|_0 \leq \mu a_l \|\alpha^l\|_0 + 0$$

If  $(a_l - 1) < 0$ , then first inequality is tighter. Otherwise, second one is tighter.

Denoting  $x^+ = \max(x, 0)$  and  $x^- = \min(x, 0)$ , we can combine the above two inequalities into

$$b - c + \mu \sum_{k=1}^L a_k \|\alpha^k\|_0 \leq \mu a_l \|\alpha^l\|_0 + (a_l - 1)^-.$$

When  $\|\alpha^l\|_0 = 0$ , then inequality (9.2.11) reduces to

$$b - c + \mu \sum_{k=1}^L a_k \|\alpha^k\|_0 \leq 0$$

for all indices.

□

### 9.3. Digest

#### Problem formulations

Exact  $l_p$  norm minimization problem:

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{C}^D} \|\alpha\|_p \text{ subject to } x = \mathcal{D}\alpha. \quad (\mathbf{P}_p)$$

Concentration coefficient:

$$P_p(\Lambda, \mathcal{D}) \triangleq \max_{h \in \mathcal{N}(\mathcal{D}), h \neq 0} \frac{\sum_{k \in \Lambda} |h_k|^p}{\sum_k |h_k|^p}$$

Recovery conditions based on  $P_p(\Lambda, \mathcal{D})$ :

1. If  $P_p(\Lambda, \mathcal{D}) < \frac{1}{2}$  then for all  $\alpha$  such that  $\text{supp}(\alpha) \subseteq \Lambda$ ,  $\alpha$  is the unique solution to the problem  $(\mathbf{P}_p)$ .
2. If  $P_p(\Lambda, \mathcal{D}) = \frac{1}{2}$  then for all  $\alpha$  such that  $\text{supp}(\alpha) \subseteq \Lambda$ ,  $\alpha$  is a solution to the problem  $(\mathbf{P}_p)$ .
3. If  $P_p(\Lambda, \mathcal{D}) > \frac{1}{2}$  then there exists  $\alpha$  such that  $\text{supp}(\alpha) \subseteq \Lambda$ , and  $\beta$  (not supported on  $\Lambda$ ) such that  $\|\beta\|_p < \|\alpha\|_p$  and  $\mathcal{D}\alpha = \mathcal{D}\beta$ . Thus  $(\mathbf{P}_p)$  will not return a solution supported over  $\Lambda$ .

We are interested in guarantees of the form

$$\text{If } |\Lambda| < f(\mathcal{D}), \text{ then } P_p(\Lambda, \mathcal{D}) < \frac{1}{2}.$$

Half of spark (ceiling to cover cases when spark is odd)

$$\text{spark}_{1/2}(\mathcal{D}) = \left\lceil \frac{\text{spark}(\mathcal{D})}{2} \right\rceil$$

**Spark based guarantee for  $(P_0)$ :** A guarantee of the form

$$\text{If } |\Lambda| < f(\mathcal{D}), \text{ then } P_0(\Lambda, \mathcal{D}) < \frac{1}{2}.$$

holds if and only if

$$f(\mathcal{D}) \leq \text{spark}_{1/2}(\mathcal{D}).$$

**Guarantee for  $l_p$  translating to guarantee for  $l_0$**  Let

$$\text{If } |\Lambda| < f(\mathcal{D}), \text{ then } P_p(\Lambda, \mathcal{D}) < \frac{1}{2} \quad (9.3.1)$$

hold true for some  $0 \leq p \leq 1$  and some  $f(\mathcal{D})$ . Then it also holds for  $p = 0$  and  $f(\mathcal{D}) \leq \text{spark}_{1/2}(\mathcal{D})$ .

**Coherence based sparsity bound for exact recovery of  $l_0$  and  $l_1$  problems**

$$\|\alpha\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right)$$

then  $\alpha$  is the unique solution to both the  $(P_0)$  and  $(P_1)$  problems.

## CHAPTER 10

### Compressed Sensing with Orthogonal Systems

#### 10.1. Digest



## Part 2

# Joint Recovery and Dictionary Learning Problems

## CHAPTER 11

### Joint Sparsity Problems

The objectives of this chapter are

- Identify different types of joint sparsity problems
- Establish a consistent notation to be followed in following chapters
- Discuss a set of mathematical tools which will be useful in following chapters
- Discuss applications of joint sparsity problems

We will start with the simplest joint sparsity problems and make our lives more complicated as we go forward. The notation will also get more sophisticated accordingly.

TABLE 1. Symbols used in this part of the book

Symbol	Purpose
$A$	An arbitrary $m \times n$ complex matrix
$a_{ij}$	The element at $i$ -th row and $j$ -th column of $A$
$a^j$	$j$ -th column column of $A$
$\underline{a}^i$	$i$ -th row vector of $A$
$a_k^j$	$k$ -th entry in the $j$ -th column (vector) of $A$
$\underline{a}_k^i$	$k$ -th entry in the $i$ -th row (vector) of $A$
$a_{(k)}^j$	$k$ -th largest entry (by magnitude) in the $j$ -th column (vector) of $A$
$\underline{a}_{(k)}^i$	$k$ -th largest entry (by magnitude) in the $i$ -th row (vector) of $A$
$A_\Lambda$	A submatrix of $A$ consisting of columns indexed by $\Lambda \subset \{1, \dots, n\}$
$\underline{A}_\Lambda$	A submatrix of $A$ consisting of rows indexed by $\Lambda \subset \{1, \dots, m\}$
$\det(A)$	Determinant of $A$
$ A $	A matrix consisting of absolute values of entries in $A$
$N$	Dimension of ambient space $\mathbb{C}^N$ for signals
$\mathbb{C}^N$	Signal space
$x^s$	A signal belonging to $\mathbb{C}^N$
$D$	Number of dictionary atoms $D \geq N$
$\mathbb{C}^D$	Representation space
$\mathcal{D} \in \mathbb{C}^{N \times D}$	The sparsifying dictionary
$K$	Sparsity level of signals in the dictionary
$S$	Number of signals
$X = \{x^1, \dots, x^S\}$	Set of signals $x^s \in \mathbb{C}^N$
$\mathcal{A} = \{\alpha^1, \dots, \alpha^S\}$	Signal representations in $\mathcal{D}$ , $\alpha^s \in \mathbb{C}^D$
$\Sigma_K$	Set of $K$ -sparse signals over the dictionary $\mathcal{D}$ ; $\alpha^s \in \Sigma_K$

### 11.1. Tools from matrix analysis

In this and following chapters we will be dealing with more complex matrix manipulations. Several tools from matrix algebra and analysis are listed here for reference.

The matrix space is equipped with the usual Hermitian inner product

$$\langle A, B \rangle \triangleq \text{trace}(B^H A). \quad (11.1.1)$$

The Frobenius norm follows from this inner product

$$\|A\|_F^2 \triangleq \langle A, A \rangle. \quad (11.1.2)$$

Note that for  $A \in \mathbb{C}^{m \times n}$

$$\|A\|_F^2 = \sum_{j=1}^n \|a_j\|_2^2. \quad (11.1.3)$$

The dual of a normed linear space  $(\mathbb{C}^M, \|\cdot\|_p)$  is a normed linear space  $(\mathbb{C}^M, \|\cdot\|'_p)$  with the conjugacy relation

$$\frac{1}{p} + \frac{1}{p'} = 1. \quad (11.1.4)$$

The dual space of a normed linear space  $V$  is denoted as  $V^*$ .

Let  $U$  and  $V$  be normed linear spaces of vectors or matrices. Let  $A$  be a matrix representing a linear operator acting on  $U$  producing elements of  $V$ . Then  $A^*$  is a map between the dual spaces  $V^*$  and  $U^*$ . When  $A$  is a matrix then  $A^* = A^H$ .

The operator norm for a linear operator  $A$  mapping from  $U$  to  $V$  is defined as

$$\|A\|_{U \rightarrow V} \triangleq \sup_{x \neq 0} \frac{\|Ax\|_V}{\|x\|_U}. \quad (11.1.5)$$

The operator norm for the adjoint satisfies the identity

$$\|A^*\|_{V^* \rightarrow U^*} = \|A\|_{U \rightarrow V}. \quad (11.1.6)$$

Some specific operator norms of our interest would be norms connecting  $l_p$  spaces from  $(\mathbb{C}^m, \|\cdot\|_p)$  to  $(\mathbb{C}^n, \|\cdot\|_q)$ . They are usually denoted as  $\|A\|_{p \rightarrow q}$ . When  $p = q$ , then we simply denote them as  $\|A\|_p$ .

Of specific interest are norms like

$\|A\|_1$ : the max column sum norm

$\|A\|_2$ : the spectral norm

$\|A\|_\infty$ : the max column sum norm

We develop another set of norms around the row vectors of a matrix.

**Definition 11.1** Let  $A$  be an  $m \times n$  matrix with rows  $\underline{a}^i$  as

$$A = \begin{bmatrix} \underline{a}^1 \\ \vdots \\ \underline{a}^m \end{bmatrix}$$

Then we define

$$\|A\|_{p,\infty} \triangleq \max_{1 \leq i \leq m} \|\underline{a}^i\|_p = \max_{1 \leq i \leq m} \left( \sum_{j=1}^n |\underline{a}_j^i|^p \right)^{\frac{1}{p}} \quad (11.1.7)$$

where  $1 \leq p < \infty$ . i.e. we take  $p$ -norms of all row vectors and then find the maximum.

We define

$$\|A\|_{\infty,\infty} = \max_{i,j} |a_{ij}|. \quad (11.1.8)$$

This is equivalent to taking  $l_\infty$  norm on each row and then taking the maximum of all the norms.

For  $1 \leq p, q < \infty$ , we define the norm

$$\|A\|_{p,q} \triangleq \left[ \sum_{i=1}^m (\|\underline{a}^i\|_p)^q \right]^{\frac{1}{q}}. \quad (11.1.9)$$

i.e., we compute  $p$ -norm of all the row vectors to form another vector and then take  $q$ -norm of that vector.

Note that the norm  $\|A\|_{p,\infty}$  is different from the operator norm  $\|A\|_{p \rightarrow \infty}$ . Similarly  $\|A\|_{p,q}$  is different from  $\|A\|_{p \rightarrow q}$ .

There is a connection between the two norms  $\|A\|_{p,\infty}$  and  $\|A\|_{p \rightarrow \infty}$ . If

$$\frac{1}{p} + \frac{1}{p'} = 1,$$

then

$$\|A\|_{p,\infty} = \|A\|_{p' \rightarrow \infty}. \quad (11.1.10)$$

Some useful properties of operator norms are listed below.

For two matrices  $A, B$  we have

$$\|AB\|_{p \rightarrow q} \leq \|B\|_{p \rightarrow s} \|A\|_{s \rightarrow q}. \quad (11.1.11)$$

For the pseudo-inverse  $A^\dagger$  we have

$$\|A^\dagger\|_2 = \frac{1}{\sigma_{\min}(A)} \quad (11.1.12)$$

where  $\sigma_{\min}(A)$  denotes the smallest non-zero singular value of  $A$ .

Another useful result for two matrices  $A, B$  is

$$\frac{\|AB\|_{p,\infty}}{\|B\|_{p,\infty}} \leq \|A\|_{\infty \rightarrow \infty} = \|A\|_{1,\infty}. \quad (11.1.13)$$

The unit sphere in  $\mathbb{R}^L$  is defined by

$$S^{L-1} \triangleq \{x \in \mathbb{R}^L : \|x\|_2 = 1\}. \quad (11.1.14)$$

The counterpart in complex space is defined by

$$S_C^{L-1} \triangleq \{x \in \mathbb{C}^L : \|x\|_2 = 1\}. \quad (11.1.15)$$

## 11.2. Sparse representation

In this section we generalize the problem of sparse representation of a signal in a redundant dictionary for the joint recovery setting.

In the **single signal setting**, we have a redundant dictionary  $\mathcal{D} \in \mathbb{C}^{N \times D}$  and a signal  $x \in \mathbb{C}^N$  whose representation  $\alpha \in \mathbb{C}^D$  is  $K$ -sparse such that

$$x = \mathcal{D}\alpha + e. \quad (11.2.1)$$

In the **multiple signal setting**, we have  $S$  different signals. The  $S$  signals are put together in a set

$$X = \{x^1, \dots, x^S\}. \quad (11.2.2)$$

We can also form a matrix of size  $N \times S$  by putting together these signals as columns of the matrix. Without any confusion, we will use the symbol  $X$  to represent this matrix as

$$X = \begin{bmatrix} x^1 & \dots & x^S \end{bmatrix} \quad (11.2.3)$$

This matrix is known as the **signal matrix**. Clearly  $X \in \mathbb{C}^{N \times S}$ . Note that we use the superscript  $1 \leq s \leq S$  to refer to  $s$ -th signal in the set or  $s$ -th column in the signal matrix.

Many a times, we need to refer to a particular row inside the signal matrix. We will use the symbol  $\underline{x}^i$  with  $1 \leq i \leq N$  to refer to the  $i$ -th row in  $X$ . This will be a row vector. Thus

$$X = \begin{bmatrix} \underline{x}^1 \\ \vdots \\ \underline{x}^N \end{bmatrix}. \quad (11.2.4)$$

Further note that  $x_n^s$  refers to the  $n$ -th entry in  $x^s$  (column vector). and  $\underline{x}_n^i$  refers to the  $n$ -th entry in  $\underline{x}^i$  (row vector).

$x_{(n)}^s$  refers to the  $n$ -th largest entry (magnitude wise) in  $x^s$ . and  $\underline{x}_{(n)}^i$  refers to the  $n$ -th largest entry (magnitude wise) in  $\underline{x}^i$ .

The columns of  $\mathcal{D}$  (a.k.a. atoms of  $\mathcal{D}$ ) will be denoted as

$$\mathcal{D} = \begin{bmatrix} d^1 & \dots & d^D \end{bmatrix} \quad (11.2.5)$$

The rows of  $\mathcal{D}$  (as a matrix) will be denoted as

$$\mathcal{D} = \begin{bmatrix} \underline{d}^1 \\ \vdots \\ \underline{d}^N \end{bmatrix}. \quad (11.2.6)$$

We denote  $\alpha^s \in \mathbb{C}^D$  as an approximate sparse representation of  $x^s$  in  $\mathcal{D}$  with

$$x^s = \mathcal{D}\alpha^s + e^s \quad \forall 1 \leq s \leq S. \quad (11.2.7)$$

The vector  $e^s \in \mathbb{C}^N$  represents the approximation error for signal  $x^s$ .

We put all  $\alpha^s$  together in a matrix  $\mathcal{A} \in \mathbb{C}^{D \times S}$ . i.e.

$$\mathcal{A} = \begin{bmatrix} \alpha^1 & \dots & \alpha^S \end{bmatrix} \quad (11.2.8)$$

This matrix is known as the **representation matrix**.

Sometimes we will also use the symbol  $\mathcal{A}$  to denote the set of representations

$$\mathcal{A} = \{\alpha^1, \dots, \alpha^S\}. \quad (11.2.9)$$

On the same lines we will use the symbol  $\underline{\alpha}^i$  with  $1 \leq i \leq D$  to refer to the  $i$ -th row in  $\mathcal{A}$ . Thus

$$\mathcal{A} = \begin{bmatrix} \underline{\alpha}^1 \\ \vdots \\ \underline{\alpha}^D \end{bmatrix}. \quad (11.2.10)$$

Similarly, we will refer to individual entries and individual  $n$ -th largest entries in a row or a column.

We put all  $e^s$  together in in a matrix  $E \in \mathbb{C}^{N \times S}$ . i.e.

$$E = \begin{bmatrix} e^1 & \dots & e^S \end{bmatrix} \quad (11.2.11)$$

This matrix is known as the **approximation error matrix**.

The Frobenius norm is a suitable norm for measuring the approximation error for the whole approximation error matrix  $E$ .

Then we have

$$X = \mathcal{D}\mathcal{A} + E. \quad (11.2.12)$$

We will use  $\Omega$  to denote the index set for the atoms in the dictionary  $\mathcal{D}$

$$\Omega = \{1, 2, \dots, D\}. \quad (11.2.13)$$

With this notation, we will say that representation matrices  $\mathcal{A}$  belong to the linear space  $\mathbb{C}^{\Omega \times S}$ .

Suppose that  $\Lambda \subseteq \Omega$ . We will often consider representation matrices in  $\mathbb{C}^{\Lambda \times S}$ . Such matrices can be extended to a matrix in  $\mathbb{C}^{\Omega \times S}$  by introducing zero rows at indices  $\Omega \setminus \Lambda$ . Likewise a representation matrix can be restricted by removing the rows which have only 0 entries.

The support for individual representation  $\alpha^s$  is given by  $\text{supp}(\alpha^s)$ .

**Definition 11.2** The **support** for the representation matrix is defined as

$$\text{supp}(\mathcal{A}) \triangleq \bigcup_{s=1}^S \text{supp}(\alpha^s). \quad (11.2.14)$$

The **joint sparsity** of  $\mathcal{A}$  is defined as

$$|\text{supp}(\mathcal{A})|. \quad (11.2.15)$$

REMARK. In [13] this is referred to as **sparsity rank** of  $\mathcal{A}$ .

REMARK. The support  $\text{supp}(\mathcal{A})$  is same as the number of non-zero rows in  $\mathcal{A}$ .

An alternative definition found in literature [40] is as follows.

**Definition 11.3** We define **row-support** of a representation matrix as the set of indices for its non-zero rows.

$$\text{rowsupp}(\mathcal{A}) \triangleq \{d \in \Omega : \alpha_{ds} \neq 0 \text{ for some } s \in [1, \dots, S]\}. \quad (11.2.16)$$

If we find a representation matrix  $\mathcal{A}$  which has few nonzero rows, then we say that  $\mathcal{A}$  is **row-sparse**.

We can define a row- $l_0$  “norm” of a representation matrix.

**Definition 11.4** The **row- $l_0$  “norm”** of a representation matrix  $\mathcal{A}$  is defined as the number of non-zero rows in  $\mathcal{A}$  non-zero rows.

$$\|\mathcal{A}\|_{\text{row}-0} \triangleq |\text{rowsupp}(\mathcal{A})|. \quad (11.2.17)$$

Clearly

$$\|\mathcal{A}\|_{\text{row}-0} |\bigcup_{s=1}^S \text{supp}(\alpha^s)| = |\text{supp}(\mathcal{A})|. \quad (11.2.18)$$

For a redundant dictionary  $\mathcal{D}$ , the representations  $\mathcal{A}$  are not-unique (even if we have truly sparse signals and  $E = 0$ ). Since we are interested in sparse representations of signals in  $X$ , we need a good measure to penalize representations  $\mathcal{A}$  which are non-sparse. The row- $l_0$  “norm” of  $\mathcal{A}$  is a suitable measure for this purpose. It can be thought of as a **cost of sparse representation** of signals in  $X$ .

We note that an atom ( $d_n$ ) in  $\mathcal{D}$  participates in the representation of one or more signals in  $X$  if and only if the  $d$ -th row in  $\mathcal{A}$  contains at least one non-zero entry.

There are few possibilities at this stage:

- The signals  $x^s$  are truly sparse and there is no approximation error ( $e^s$ ).
- All representations have identical support. i.e.

$$\text{supp}(\alpha^1) = \dots = \text{supp}(\alpha^S).$$

In this case,  $\text{supp}(\mathcal{A})$  is also same as support for individual signals.

- Different representations have different support but  $|\text{supp}(\mathcal{A})| \ll N$ . Thus, overall the joint sparsity level is small.
- Different representations have different support and  $|\text{supp}(\mathcal{A})| \approx N$  or  $|\text{supp}(\mathcal{A})| > N$ . Then the joint sparsity level is too high.

### 11.2.1. Signals with identical support

In this case, we have

$$\text{supp}(\mathcal{A}) = \text{supp}(\alpha^1) = \dots = \text{supp}(\alpha^S). \quad (11.2.19)$$

We will denote this identical support by  $\Lambda$ . i.e.

$$\Lambda = \text{supp}(\mathcal{A}). \quad (11.2.20)$$

### 11.2.2. Generative models for jointly sparse real signals

While arbitrary constructions of  $\mathcal{A}$  are suitable for a deterministic worst case analysis of the recovery algorithm, they usually end up providing quite pessimistic recovery guarantees. Introducing a probabilistic model for the generation of signals in  $X$  can help in developing much more realistic recovery guarantees using average case analysis of the recovery algorithms.

We now present a generative model for the synthesis vectors  $\alpha^s$  with  $1 \leq s \leq S$  adapted from [25]. This particular model is developed for real signals with a real dictionary.

We consider  $\Lambda = \{\lambda_1, \dots, \lambda_K\} \subset \{1, \dots, D\}$  as the index set for entries in  $\alpha^s$  which can be non-zero.

We assume that the entries  $\alpha_{\lambda_k}^s, \lambda_k \in \Lambda$  of the random vector  $\alpha^s$  are independent Gaussian variables of variance  $\sigma_k^2$ . Other entries are 0.

The Gaussian assumption is the most prevalent one used in signal processing. Hence, it can accommodate a wide variety of practical problems. At the same time, incorporating different variances allows us to shape the synthesis coefficients. For example, we could easily impose a statistical decay on the entries  $\alpha_k^s$  using appropriate profile of variances.

Let  $\Sigma$  be a  $K \times K$  diagonal matrix whose diagonal entries are  $\sigma_{\lambda_k}^2$ .

Let  $U$  be a  $K \times S$  random matrix with independent standard Gaussian entries. Then we can write

$$\underline{\mathcal{A}}_\Lambda = \Sigma^{\frac{1}{2}} U \quad (11.2.21)$$

i.e. the rows in  $\mathcal{A}$  corresponding to the index set  $\Lambda$  are given by random matrix  $\Sigma^{\frac{1}{2}}U$  while rest of the rows in  $\mathcal{A}$  are identically zero. We can then write our signals  $X$  as

$$X = \mathcal{D}_\Lambda \Sigma^{\frac{1}{2}} U + E \quad (11.2.22)$$

where  $E$  is a  $N \times S$  matrix collecting innovation (noise) signals on its columns.

### 11.2.3. Sparse recovery problem formulations

Given a signal matrix  $X \in \mathbb{C}^N$  which is known to have a sparse representation in a dictionary  $\mathcal{D}$ , the exact-sparse recovery problem is:

$$\widehat{\mathcal{A}} = \arg \min_{\mathcal{A} \in \mathbb{C}^{D \times S}} \|\mathcal{A}\|_{\text{row-0}} \text{ subject to } X = \mathcal{D}\mathcal{A}. \quad (\text{Joint-P}_0)$$

When  $X \in \mathbb{C}^{N \times S}$  doesn't have a sparse representation in  $\mathcal{D}$ , a  $K$ -sparse approximation of  $X$  in  $\mathcal{D}$  can be obtained by solving the following problem:

$$\widehat{\mathcal{A}} = \arg \min_{\mathcal{A} \in \mathbb{C}^{D \times S}} \|X - \mathcal{D}\mathcal{A}\|_2 \text{ subject to } \|\mathcal{A}\|_{\text{row-0}} \leq K. \quad (\text{Joint-P}_0^K)$$

Here  $X$  is modeled as  $X = \mathcal{D}\mathcal{A} + E$  as discussed above.

A different way to formulate the approximation problem is to provide an upper bound to the acceptable approximation error  $\|E\|_F \leq \epsilon$  and try to find sparsest possible representation within this approximation error bound as

$$\widehat{\mathcal{A}} = \arg \min_{\mathcal{A} \in \mathbb{C}^{D \times S}} \|\mathcal{A}\|_{\text{row-0}} \text{ subject to } \|X - \mathcal{D}\mathcal{A}\|_F \leq \epsilon. \quad (\text{Joint-P}_0^\epsilon)$$

### 11.2.4. Basis pursuit problem formulations

In sparse recovery problem formulation, our goal is to simply minimize the number of non-zero rows. When we change over to basis pursuit, we have some choices. For the row vectors, we can choose some  $l_p$  norm to compute their norms. Then, over the norms from each row, we can compute the  $l_1$  norm.

Thus, the general form of basis pursuit for joint recovery of sparse signal representations becomes:

$$\widehat{\mathcal{A}} = \arg \min_{\mathcal{A} \in \mathbb{C}^{D \times S}} \|\mathcal{A}\|_{p,1} \text{ subject to } X = \mathcal{D}\mathcal{A}. \quad (\text{Joint-P}_1)$$

Different authors choose different values of  $p$ .

The corresponding basis pursuit with inequality constraints problem becomes:

$$\widehat{\mathcal{A}} = \arg \min_{\mathcal{A} \in \mathbb{C}^{D \times S}} \|\mathcal{A}\|_{p,1} \text{ subject to } \|X - \mathcal{D}\mathcal{A}\|_F \leq \epsilon. \quad (\text{Joint-P}_1^\epsilon)$$

### 11.3. Compressed sensing

We switch gears and look at the joint sparsity problems in compressed sensing framework.

A **single measurement vector** problem or in short an SMV problem is posed as

$$y = \Phi x + \eta \quad (11.3.1)$$

where  $\Phi \in \mathbb{C}^{M \times N}$  is the sensing matrix,  $x \in \mathbb{C}^N$  is a sparse signal with  $x \in \Sigma_K$ ,  $y \in \mathbb{C}^M$  is the measurement vector and  $\eta \in \mathbb{C}^M$  is the measurement error. We note that we are assuming that the signal  $x$  is sparse in itself.

We are not considering an orthonormal basis  $\Psi$  or an overcomplete dictionary  $\mathcal{D}$  which could sparsify the signal  $x$ .

In more realistic cases,  $x$  is a compressible signal, while in most general setting  $x$  is an arbitrary signal with a best  $K$ -term approximation given by  $x|_K$ .

The recovery process is defined as

$$\widehat{x} = \Delta(\Phi, y) \quad (11.3.2)$$

where  $\Delta$  is some recovery algorithm which estimates a sparse vector  $\widehat{x}$  hoping that the recovery error  $\|\widehat{x} - x\|_2$  remains small.

We now generalize this situation for the **multiple measurement vector** (a.k.a. MMV) setting. we have  $S$  different signals. As before, the  $S$  signals are put together in a set

$$X = \{x^1, \dots, x^S\}. \quad (11.3.3)$$

The corresponding signal matrix is:

$$X = \begin{bmatrix} x^1 & \dots & x^S \end{bmatrix} \quad (11.3.4)$$

The representation in row vectors of  $X$  is

$$X = \begin{bmatrix} \underline{x}^1 \\ \vdots \\ \underline{x}^N \end{bmatrix}. \quad (11.3.5)$$

We denote  $y^s \in \mathbb{C}^M$  as the measurement vector for  $x^s$  with

$$y^s = \Phi x^s + \eta^s \quad \forall 1 \leq s \leq S. \quad (11.3.6)$$

The vector  $\eta^s \in \mathbb{C}^M$  represents the measurement noise for signal  $x^s$ .

We put all  $y^s$  together in a matrix  $Y \in \mathbb{C}^{M \times S}$ . i.e.

$$Y = \begin{bmatrix} y^1 & \dots & y^S \end{bmatrix} \quad (11.3.7)$$

This matrix is known as the **measurement matrix**.

Sometimes we will also use the symbol  $Y$  to denote the set of measurement vectors

$$Y = \{y^1, \dots, y^S\}. \quad (11.3.8)$$

On the same lines we will use the symbol  $\underline{y}^i$  with  $1 \leq i \leq M$  to refer to the  $i$ -th row in  $Y$ . Thus

$$Y = \begin{bmatrix} \underline{y}^1 \\ \vdots \\ \underline{y}^M \end{bmatrix}. \quad (11.3.9)$$

Similarly, we will refer to individual entries and individual  $n$ -th largest entries in a row or a column.

We put all  $\eta^s$  together in a matrix  $\mathcal{H} \in \mathbb{C}^{M \times S}$ . i.e.

$$\mathcal{H} = \begin{bmatrix} \eta^1 & \dots & \eta^S \end{bmatrix} \quad (11.3.10)$$

This matrix is known as the **measurement error matrix**. The Frobenius norm is a suitable norm for measuring the measurement error for the whole measurement error matrix  $\mathcal{H}$ .

Then we have

$$Y = \Phi X + \mathcal{H}. \quad (11.3.11)$$

The support for individual signals  $x^s$  is given by  $\text{supp}(x^s)$ .

**Definition 11.5** The **support** for the signal matrix is defined as

$$\text{supp}(X) \triangleq \bigcup_{s=1}^S \text{supp}(x^s). \quad (11.3.12)$$

The **joint sparsity** of  $X$  is defined as

$$|\text{supp}(X)|. \quad (11.3.13)$$

**REMARK.** The support  $\text{supp}(X)$  is same as the number of non-zero rows in  $X$ .

There are few possibilities at this stage:

- All signals have identical support. i.e.

$$\text{supp}(x^1) = \dots = \text{supp}(x^S).$$

In this case,  $\text{supp}(X)$  is also same as support for individual signals.

- Different signals have different support but  $|\text{supp}(X)| \ll N$ . Thus, overall the joint sparsity level is small.
- Different signals have different support and  $|\text{supp}(X)| \approx N$ . Then the joint sparsity level is too high.

We will consider the case of compressed sensing with redundant dictionaries later.

### 11.4. Distributed compressed sensing

### 11.5. Miscellaneous results

We collect some miscellaneous results which would be useful in later chapters.

#### 11.5.1. Norm dominance

Let  $\|\cdot\|_\alpha$  and  $\|\cdot\|_\beta$  be two norms defined over  $\mathbb{C}^N$ .

**Definition 11.6** We say that a norm  $\|\cdot\|_\alpha$  is dominated by a norm  $\|\cdot\|_\beta$  in  $\mathbb{C}^N$  if and only if for any  $x, y \in \mathbb{C}^N$

$$\|x\|_\alpha < \|y\|_\alpha \implies \|x\|_\beta < \|y\|_\beta. \quad (11.5.1)$$

**Theorem 11.1** If a norm  $\|\cdot\|_\beta$  dominates norm  $\|\cdot\|_\alpha$ , then there exists a constant  $C > 0$  such that

$$\|x\|_\alpha = C\|x\|_\beta \quad \forall x \in \mathbb{C}^N.$$

**PROOF.** We first show that for all  $x, y \in \mathbb{C}^N$  such that  $\|x\|_\alpha = \|y\|_\alpha$ ,  $\|x\|_\beta = \|y\|_\beta$  also holds. In other words:

$$\|x\|_\alpha = \|y\|_\alpha \implies \|x\|_\beta = \|y\|_\beta. \quad (11.5.2)$$

We start with

$$\|x\|_\alpha = \|y\|_\alpha.$$

For any  $\gamma > 0$ , it is easy to see that

$$\|(1 - \gamma)y\|_\alpha < \|x\|_\alpha < \|(1 + \gamma)y\|_\alpha.$$

Since,  $\|\cdot\|_\beta$  dominates  $\|\cdot\|_\alpha$ , we get

$$\|(1 - \gamma)y\|_\beta < \|x\|_\beta < \|(1 + \gamma)y\|_\beta.$$

Letting  $\gamma \rightarrow 0$ , we obtain the result

$$\|x\|_\beta = \|y\|_\beta.$$

Now we choose some  $x_0 \in \mathbb{C}^N$  such that  $\|x_0\|_\alpha = 1$ . Now for any nonzero  $x \in \mathbb{C}^N$ , we have

$$\frac{\|x\|_\alpha}{\|x\|_\alpha} = \left\| \frac{x}{\|x\|_\alpha} \right\|_\alpha = 1 = \|x_0\|_\alpha.$$

Thus from (11.5.2), we obtain

$$\left\| \frac{x}{\|x\|_\alpha} \right\|_\beta = \|x_0\|_\beta.$$

This gives us

$$\|x\|_\beta = \|x_0\|_\beta \|x\|_\alpha.$$

Since  $x_0$  is fixed, hence  $C = \|x_0\|_\beta$  is the desired constant.  $\square$

## 11.6. Digest

## CHAPTER 12

### Joint Recovery Algorithms

There are several approaches to solving the joint recovery problem. In this chapter, we will study some of the basic algorithms for the same. The algorithms would include greedy algorithms like thresholding, simultaneous orthogonal matching pursuit and convex relaxation methods like basis pursuit.

More advanced algorithms will be the topic of discussion in subsequent chapters.

We recall the basic problem formulations of joint recovery of sparse representations.

The exact-sparse recovery problem is:

$$\widehat{\mathcal{A}} = \arg \min_{\mathcal{A} \in \mathbb{C}^{D \times S}} \|\mathcal{A}\|_{\text{row}-0} \text{ subject to } X = \mathcal{D}\mathcal{A}. \quad (\text{Joint-P}_0)$$

The sparse solution recovery within this approximation error bound is stated as

$$\widehat{\mathcal{A}} = \arg \min_{\mathcal{A} \in \mathbb{C}^{D \times S}} \|\mathcal{A}\|_{\text{row}-0} \text{ subject to } \|X - \mathcal{D}\mathcal{A}\|_F \leq \epsilon. \quad (\text{Joint-P}_0^\epsilon)$$

We **recall** the basis pursuit formulation for joint recovery as:

$$\widehat{\mathcal{A}} = \arg \min_{\mathcal{A} \in \mathbb{C}^{D \times S}} \|\mathcal{A}\|_{p,1} \text{ subject to } X = \mathcal{D}\mathcal{A}. \quad (\text{Joint-P}_1)$$

When  $p = 1$ , then we are essentially minimizing the  $l_\infty$  norm of  $\mathcal{A}$ .

When the value of  $p = 2$  is chosen and it is known as the mixed  $l_{2,1}$ -norm minimization problem.

In the following we first present the thresholding and S-OMP algorithms. We then follow it up with detailed theoretical analyses and experimental results for these algorithms (including basis pursuit).

### 12.1. Thresholding algorithm

At the heart thresholding algorithm is pretty simple. In the single signal setting, thresholding amounts to selecting the atoms from the dictionary which are most correlated with the signal  $x$ .

We compute the vector  $v = \mathcal{D}^H x$  which is a column vector  $\in \mathbb{C}^D$  containing the inner product of  $x$  with each of the atoms in  $\mathcal{D}$ .

Then we identify its  $K$  largest terms in  $v|_K$  (the  $K$  most correlated atoms) and take  $\Lambda' = \text{supp}(v|_K)$  as the index set of the candidate atoms which participate in the construction of  $x = \mathcal{D}\alpha + e$ . Recall that  $\Lambda = \text{supp}(\alpha)$  is the true support of the sparse representation  $\alpha$ .

From here, the recovery of  $\alpha$  takes the simple step of  $\hat{\alpha}_{\Lambda'} = \mathcal{D}_{\Lambda'}^\dagger x$  and setting rest of entries in  $\hat{\alpha}$  as 0.

The choice of  $K$  depends on the sparsity prior assumed for  $\alpha$  i.e.  $\alpha \in \Sigma_K$ . In other words  $K = |\Lambda|$ .

In the multiple signal setting, the main challenge is that we have to combine the correlation of an atom with all the different signals to get a single measure of the correlation of the atom with the signal ensemble. The approach taken in [25] uses an  $l_p$  norm where  $1 \leq p \leq \infty$  for combining the correlations.

Let  $d_j$  be a particular atom in  $\mathcal{D}$  and  $x^1, \dots, x^S$  be different signals in  $X$ . The combined correlation measure is given by

$$\left( \sum_{s=1}^S |\langle x^s, d_j \rangle|^p \right)^{\frac{1}{p}}. \quad (12.1.1)$$

We can see that this is nothing but  $\|d_j^H X\|_p$ .

Thus the correlation vector of correlations of all atoms in  $\mathcal{D}$  with signals in  $X$  is obtained by first computing  $\mathcal{D}^H X$  and then taking  $l_p$  norm for each row.

We call these correlations as  $p$ -correlations as they depend on the choice of  $p$ .

Finally, the  $p$ -thresholding algorithm is simply about selecting a set  $\Lambda'$  of  $K$  atoms whose  $p$ -correlations with  $X$  are among the  $K$  largest. i.e.

$$\|d_k^H X\|_p \geq \|d_l^H X\|_p \quad \forall k \in \Lambda', \forall l \notin \Lambda'. \quad (12.1.2)$$

Once the support has been identified, it is easy to compute the estimated represented matrix  $\widehat{\mathcal{A}}$  as follows.

We compute  $\widehat{\mathcal{A}}_{\Lambda'} = \mathcal{D}_{\Lambda'}^\dagger X$  and set rest of rows in  $\widehat{\mathcal{A}}$  as 0.

Recovering the right support. As we have seen multiple times, the real challenge is to recovery the support of  $\alpha$  correctly. A simple measure of support recovery can be defined as

$$\rho = \frac{|\Lambda \cap \Lambda'|}{|\Lambda|}. \quad (12.1.3)$$

If  $\rho = 1$ , then we have perfect support recovery.

Occasionally, we may also be interested in partial recovery. In these cases we would choose  $K \leq |\Lambda|$ , i.e. we may be okay with choosing lesser number of non-zero entries in  $\Lambda'$ .

## 12.2. Simultaneous orthogonal matching pursuit (S-OMP)

In this section, we develop a greedy pursuit algorithm based on OMP that can be used to solve several different sparse approximation problems.

The S-OMP algorithm is presented in fig. 12.1.

Some remarks are in order to explain the notation:

```

 $\mathcal{A}, R = \text{S-OMP}(\mathcal{D}, X);$ 
Input: An  $N \times S$  signal matrix  $X$ 
Input: A halting criterion
Input:  $\mathcal{D}$ : a signal dictionary of size  $N \times D$ 
Output: A set  $\Lambda^t$  containing  $t$  indices from  $\Omega$  where  $t$  is the number of
iterations completed
Output: a  $D \times S$  approximation matrix  $\mathcal{A}^t$ 
Output: an  $N \times S$  residual matrix  $R^t$ 

// (1) Initialization
 $\mathcal{A}^0 \leftarrow 0;$ 
 $R^0 \leftarrow X ; \quad // R = X - \mathcal{D}\mathcal{A}$ 
 $\Lambda^0 = \emptyset ; \quad // \text{the index set of chosen atoms}$ 
 $t \leftarrow 0 ; \quad // \text{Iteration counter}$ 
repeat
     $t \leftarrow t + 1 ; \quad // \text{Increase iteration counter}$ 
    (2) Find an index  $\lambda^t$  (of the atom most aligned with all residuals) that
    satisfies
        
$$\lambda^t = \arg \max_{\omega \in \Omega} \|R^{t-1}{}^H d_\omega\|_q.$$

    ;
    (3)  $\Lambda^t \leftarrow \Lambda^{t-1} \cup \{\lambda^t\} ; \quad // \text{Update support}$ 
    (4) Calculate the new approximation as follows. First set  $\mathcal{A}^t = 0$ . Then
    compute  $B = \mathcal{D}_{\Lambda^t}^\dagger X$ . Finally assign rows of  $B$  to the rows of  $\mathcal{A}^t$ 
    indexed by  $\Lambda^t$ .
    (5) Calculate the new approximation of  $X$  as  $X^t = \mathcal{D}\mathcal{A}^t$  and the new
    residual as  $R^t = X - X^t ;$ 
until halting criteria is satisfied;

```

FIGURE 12.1. Simultaneous Orthogonal Matching Pursuit

- $X$  is the input signal matrix of dimensions  $N \times S$ . There are  $S$  signals being estimated jointly.
- $\mathcal{D}$  is the sparsifying dictionary in which sparse representations are sought.
- $d_\omega$  denotes the  $\omega$ -th atom in  $\mathcal{D}$ .
- The iteration number within the algorithm is maintained in a variable  $t$ .

- $\mathcal{A}$  denotes the sparse representation matrix.
- $\mathcal{A}^t$  denotes the estimate of  $\mathcal{A}$  at the end of  $t$ -th iteration.
- In every iteration, a new atom is chosen. Thus  $|\Lambda^t| = t$ .
- $R$  represents the measurement residual matrix inside the algorithm. It is computed as  $R = X - \mathcal{D}\mathcal{A}$ .
- At the end of  $t$ -th iteration, we get the  $t$ -th estimate of  $R$  as

$$R^t = X - \mathcal{D}\mathcal{A}^t.$$

- The columns inside  $R$  are referred to as  $r^s$  with  $1 \leq s \leq S$ .
- The value of  $r^s$  at the end of  $t$ -th iteration is denoted as  $r^{s,t}$ .
- Once  $\Lambda^t$  has been identified, we can club those atoms in the matrix  $\Phi_{\Lambda^t}$ .
- $\Phi_{\Lambda^t}$  is  $N \times t$  matrix.  $B = \Phi_{\Lambda^t}^\dagger X$  is  $t \times S$  matrix. The rows of  $B$  correspond to the non-zero indices in the representations in  $\mathcal{A}^t$ . They are assigned to rows in  $\mathcal{A}^t$  indexed by  $\Lambda^t$ .

For the special case of  $S = 1$ , the algorithm reduces to well known OMP.

Step 2 of the algorithm is the **greedy selection** step. We are taking the  $l_q$ -norm of correlations of each atom with all signal residuals and trying to find out the atom with maximum correlation over all signals. The atom so found contributes maximum energy to all signals out of all the remaining atoms (not chosen so far).

Different authors have taken different choices of  $q$  in their construction of simultaneous-OMP algorithm. In [39, 40] a value of  $q = 1$  is taken. Thus, it becomes the absolute sum of correlations of an atom with the residuals at previous iterations.

In [26, 27, 28, 31]  $l_2$  and  $l_\infty$  are proposed for weak matching pursuit for the multiple signal setting.

In [15]  $l_2$  norm is considered.

[13] provides a general analysis applicable for  $q \geq 1$ .

This approach is likely to be most effective when all the input signals are well approximated by the same set of atoms. Otherwise, we may need to consider a different greedy selection criterion.

There are some equivalent ways to represent the greedy selection term

$$\max_{\omega \in \Omega} \|R^{t-1H} d_\omega\|_q = \|R^{t-1H} \mathcal{D}\|_{1 \rightarrow q} = \|\mathcal{D}^H R^{t-1}\|_{q' \rightarrow \infty}.$$

where  $1/q + 1/q' = 1$ .

Essentially each column of  $R^{t-1H} \mathcal{D}$  represents the inner product  $\langle r^{s,t-1}, d_\omega \rangle$  for all signals  $1 \leq s \leq S$ .

For the case of  $q = 1$ , we can expand  $\|R^{t-1H} d_\omega\|_q$  as

$$\|R^{t-1H} d_\omega\|_1 = \sum_{s=1}^S |\langle r^{s,t-1}, d_\omega \rangle|$$

And we can simplify it as

$$\max_{\omega \in \Omega} \|R^{t-1H} d_\omega\|_q = \|R^{t-1H} \mathcal{D}\|_1 = \|\mathcal{D}^H R^{t-1}\|_\infty.$$

Here, the operator-1 norm is nothing but max column sum norm. Similarly operator- $\infty$  norm is the max row sum norm.

Step (4) and (5) are typically implemented using least squares algorithms.

Since the residual  $R^t$  is orthogonal to all atoms chosen in  $\Lambda^t$ , hence no atom in  $\Lambda^t$  can be chosen again in later steps. Thus no atom is chosen twice in this algorithm.

**12.2.0.1. Halting criteria.** There are several obvious possibilities for halting criteria

- Stop the algorithm after a fixed number of iterations (say  $K$ ).
- Wait until the Frobenius norm of the residual  $\|R^t\|_F$  declines to a level  $\epsilon$ .

- Halt the algorithm when the maximum total correlation between an atom and the residual drops below a threshold  $\tau$  i.e.  $\|\mathcal{D}^H R^t\|_{\infty,\infty} \leq \tau$ .

These different halting criteria help solve different flavors of simultaneous sparse approximation problems.

### 12.3. Thresholding recovery guarantees

In this section [25] we will develop some recovery guarantees for the  $p$ -thresholding algorithm. These guarantees will apply for all  $K$ -sparse signal matrices in the presence of noise.

The analysis in this section is a worst case analysis since a) it applies to all signals, b) it provides only sufficient conditions (and not conditions which are both necessary and sufficient) for the recovery.

Recall that our basic signal model is

$$X = \mathcal{D}\mathcal{A} + E$$

where  $\Lambda = \text{supp}(\mathcal{A})$ . We can also write it as

$$X = \mathcal{D}_\Lambda \underline{\mathcal{A}}_\Lambda + E$$

i.e. picking up the columns indexed by  $\Lambda$  in  $\mathcal{D}$  and rows indexed by  $\Lambda$  in  $\mathcal{A}$ .

For economy of expression let us denote  $A = \underline{\mathcal{A}}_\Lambda$ . Thus, we have

$$X = \mathcal{D}_\Lambda A + E.$$

We will also use

$$\Lambda = \{\lambda_1, \dots, \lambda_K\}.$$

Recall from (12.1.2) that an atom  $d_k$  is selected in  $\Lambda'$  if it satisfies

$$\|d_k^H X\|_p \geq \|d_l^H X\|_p \quad \forall k \in \Lambda', \forall l \notin \Lambda'.$$

We choose  $K = |\Lambda|$  such atoms. In case of tie, we can simply choose any of the tied atoms. An easy choice would be to select the atom which comes first in the matrix representation of the dictionary  $\mathcal{D}$ .

A representation  $\mathcal{A}$  from the equation can be recovered correctly (up to the least squares error) if its support has been identified correctly. For this we require that

$$\min_{k \in \Lambda} \|d_k^H X\|_p \geq \max_{l \notin \Lambda} \|d_l^H X\|_p. \quad (12.3.1)$$

We will tighten the recovery condition by taking a lower bound on the L.H.S. and an upper bound on the R.H.S..

On the R.H.S., it is easy to see that

$$\max_{l \notin \Lambda} \|d_l^H X\|_p = \|\mathcal{D}_{\Lambda^c}^H X\|_{p,\infty}. \quad (12.3.2)$$

Recall that by  $(p, \infty)$  norm we mean taking  $l_p$  norm of each row and then finding the maximum.

Applying triangular inequality we get

$$\|\mathcal{D}_{\Lambda^c}^H X\|_{p,\infty} = \|\mathcal{D}_{\Lambda^c}^H (\mathcal{D}_\Lambda A + E)\|_{p,\infty} \leq \|\mathcal{D}_{\Lambda^c}^H \mathcal{D}_\Lambda A\|_{p,\infty} + \|\mathcal{D}_{\Lambda^c}^H E\|_{p,\infty} \quad (12.3.3)$$

We can apply alternative triangular inequality on the L.H.S. as

$$\begin{aligned} \min_{k \in \Lambda} \|d_k^H X\|_p &= \min_{k \in \Lambda} \|d_k^H (\mathcal{D}_\Lambda A + E)\|_p \\ &\geq \min_{k \in \Lambda} \|d_k^H (\mathcal{D}_\Lambda A)\|_p - \max_{k \in \Lambda} \|d_k^H E\|_p \\ &= \min_{k \in \Lambda} \|d_k^H \mathcal{D}_\Lambda A\|_p - \|\mathcal{D}_\Lambda^H E\|_{p,\infty}. \end{aligned} \quad (12.3.4)$$

Thus, the tightened recovery condition can be rewritten as

$$\min_{k \in \Lambda} \|d_k^H \mathcal{D}_\Lambda A\|_p - \|\mathcal{D}_\Lambda^H E\|_{p,\infty} > \|\mathcal{D}_{\Lambda^c}^H \mathcal{D}_\Lambda A\|_{p,\infty} + \|\mathcal{D}_{\Lambda^c}^H E\|_{p,\infty}. \quad (12.3.5)$$

We have replaced the  $\geq$  with  $>$  (further tightening the condition).

Bringing the terms related to the error matrix  $E$  on the L.H.S., we get

$$\|\mathcal{D}_\Lambda^H E\|_{p,\infty} + \|\mathcal{D}_{\Lambda^c}^H E\|_{p,\infty} < \min_{k \in \Lambda} \|d_k^H \mathcal{D}_\Lambda A\|_p - \|\mathcal{D}_{\Lambda^c}^H \mathcal{D}_\Lambda A\|_{p,\infty}. \quad (12.3.6)$$

Let us closely examine the term  $\|d_k^H \mathcal{D}_\Lambda A\|_p$ . Consider the matrix  $\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda A$ . We have

$$\begin{aligned}\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda A &= \begin{bmatrix} d_{\lambda_1} & \dots & d_{\lambda_K} \end{bmatrix}^H \mathcal{D}_\Lambda A \\ &= \begin{bmatrix} d_{\lambda_1}^H \\ \vdots \\ d_{\lambda_K}^H \end{bmatrix}^H \mathcal{D}_\Lambda A \\ &= \begin{bmatrix} d_{\lambda_1}^H \mathcal{D}_\Lambda A \\ \vdots \\ d_{\lambda_K}^H \mathcal{D}_\Lambda A \end{bmatrix}.\end{aligned}$$

Thus, the term  $\|d_{\lambda_k}^H \mathcal{D}_\Lambda A\|_p$  for  $\lambda_k \in \Lambda$  is the  $p$ -norm of  $k$ -th row in  $\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda A$ .

Now we can write

$$\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda A = A + (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I)A.$$

Taking the  $p$ -norm on the  $k$ -th row and applying triangle inequality  $|a+b| \geq |a| - |b|$  on the R.H.S., we obtain

$$\|d_{\lambda_k}^H \mathcal{D}_\Lambda A\|_p \geq \|\underline{a}^k\|_p - \|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I)A\|_{p,\infty}$$

where  $\underline{a}^k$  denotes the  $k$ -th row of  $A$ . Note that rather than taking the  $p$ -norm of  $k$ -th row in  $(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I)A$ , we have taken the maximum  $p$ -norm across all rows in this inequality.

Therefore the inequality (12.3.6) can be satisfied whenever

$$\|\mathcal{D}_\Lambda^H E\|_{p,\infty} + \|\mathcal{D}_{\Lambda^c}^H E\|_{p,\infty} < \min_{k \in \Lambda} \|\underline{a}^k\|_p - \|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I)A\|_{p,\infty} - \|\mathcal{D}_{\Lambda^c}^H \mathcal{D}_\Lambda A\|_{p,\infty}. \quad (12.3.7)$$

Recall that for two arbitrary matrices  $A, B$ , we have the result

$$\frac{\|AB\|_{p,\infty}}{\|B\|_{p,\infty}} \leq \|A\|_{\infty \rightarrow \infty} = \|A\|_{1,\infty}.$$

Thus

$$\|AB\|_{p,\infty} \leq \|A\|_{1,\infty} \|B\|_{p,\infty}.$$

Applying this, we can write

$$\|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I)A\|_{p,\infty} \leq \|\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I\|_{1,\infty} \|A\|_{p,\infty}$$

and

$$\|\mathcal{D}_{\Lambda^c}^H \mathcal{D}_\Lambda A\|_{p,\infty} \leq \|\mathcal{D}_{\Lambda^c}^H \mathcal{D}_\Lambda\|_{1,\infty} \|A\|_{p,\infty}.$$

Any row of  $(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I)$  contains the inner product of an atom  $d_{\lambda_k}$  with all other atoms indexed by  $\Lambda$ . Thus,

$$\|\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I\|_{1,\infty} = \max_{k \in \Lambda} \sum_{j \in \Lambda \setminus \{k\}} |\langle d_k, d_j \rangle| = \mu_1^{\text{in}}(\Lambda).$$

Similarly

$$\|\mathcal{D}_{\Lambda^c}^H \mathcal{D}_\Lambda\|_{1,\infty} = \max_{k \notin \Lambda} \sum_{j \in \Lambda} |\langle d_k, d_j \rangle| = \mu_1(\Lambda).$$

Thus

$$\begin{aligned} \|(\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I)A\|_{p,\infty} + \|\mathcal{D}_{\Lambda^c}^H \mathcal{D}_\Lambda A\|_{p,\infty} &\leq (\|\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda - I\|_{1,\infty} + \|\mathcal{D}_{\Lambda^c}^H \mathcal{D}_\Lambda\|_{1,\infty}) \|A\|_{p,\infty} \\ &= (\mu_1^{\text{in}}(\Lambda) + \mu_1(\Lambda)) \|A\|_{p,\infty}. \end{aligned}$$

Thus, putting back in the inequality (12.3.7), we get a further tightened inequality as

$$\|\mathcal{D}_\Lambda^H E\|_{p,\infty} + \|\mathcal{D}_{\Lambda^c}^H E\|_{p,\infty} < \min_{k \in \Lambda} \|\underline{a}^k\|_p - (\mu_1^{\text{in}}(\Lambda) + \mu_1(\Lambda)) \|A\|_{p,\infty}. \quad (12.3.8)$$

We can capture our analysis so far in the following theorem.

**Theorem 12.1** *If*

$$\|\mathcal{D}_\Lambda^H E\|_{p,\infty} + \|\mathcal{D}_{\Lambda^c}^H E\|_{p,\infty} < \min_{k \in \Lambda} \|\underline{a}^k\|_p - (\mu_1^{\text{in}}(\Lambda) + \mu_1(\Lambda)) \|A\|_{p,\infty} \quad (12.3.9)$$

*then  $p$ -thresholding recovers the support  $\Lambda$  from  $X = \mathcal{D}\mathcal{A} + E$ . Moreover, the reconstructed coefficients satisfy*

$$\|\mathcal{A} - \widehat{\mathcal{A}}\|_{1 \rightarrow 2} \leq (1 + \mu_1^{\text{in}}(\Lambda)) \cdot \|E\|_{1 \rightarrow 2} \quad (12.3.10)$$

*where  $\|\cdot\|_{1 \rightarrow 2}$  is the maximum  $l_2$  norm of any column.*

$\|E\|_{1 \rightarrow 2}$  gives the maximum energy in any of the error vectors.

PROOF. Most of the derivation has been completed in (12.3.8). Once the support has been correctly identified, then from  $X = \mathcal{D}_\Lambda A + E$ , we obtain

$$\widehat{A} = \mathcal{D}_\Lambda^\dagger (\mathcal{D}_\Lambda A + E) = A + \mathcal{D}_\Lambda^\dagger E.$$

Note that

$$\|\mathcal{A} - \widehat{\mathcal{A}}\|_{1 \rightarrow 2} = \|A - \widehat{A}\|_{1 \rightarrow 2}$$

From above, we get

$$\|A - \widehat{A}\|_{1 \rightarrow 2} = \|\mathcal{D}_\Lambda^\dagger E\|_{1 \rightarrow 2}.$$

Now

$$\|\mathcal{D}_\Lambda^\dagger E\|_{1 \rightarrow 2} \leq \|E\|_{1 \rightarrow 2} \|\mathcal{D}_\Lambda^\dagger\|_2$$

using the consistency of operator norms.

We can separately show that (TODO HOW)

$$\|\mathcal{D}_\Lambda^\dagger\|_2 \leq 1 + \mu_1^{\text{in}}(\Lambda).$$

This completes the theorem.  $\square$

#### 12.4. Performance guarantees for S-OMP

In this and following sections we will develop theoretical results for the performance of S-OMP for different types of sparse approximation problems.

The idea of greedy selection ratio (first developed for the analysis of OMP in signal setting [34]) is fundamental to the analysis in many of following results. We develop it here before getting into more specific results.

Let  $X = \mathcal{D}\mathcal{A}^*$  where  $\mathcal{A}^*$  is the optimal solution to the problem (**Joint-P<sub>0</sub>**). Let  $\Lambda = \text{supp}(\mathcal{A}^*)$ .

We introduce the  $N \times K$  matrix  $\mathcal{D}_\Lambda$  which contains the good atoms indexed by  $\Lambda$ .

Let the  $N \times (D - K)$  matrix  $\mathcal{D}_{\Lambda^c}$  contain the remaining atoms not used in the construction of  $X$  from  $\mathcal{A}^*$ .

Recall the definition

$$R^t = X - X^t$$

and consider  $\mathcal{D}_{\Lambda}^H R^t$ . Observe that each row of the  $K \times S$  matrix  $\mathcal{D}_{\Lambda}^H R^t$  lists the inner product between a fixed atom in  $\mathcal{D}_{\Lambda}$  and the  $S$  columns of  $R^t$ .

The rows of the  $(D - K) \times S$  matrix  $\mathcal{D}_{\Lambda^c}^H R^t$  have an analogous interpretation.

The algorithm chooses another optimal atom if and only if the ratio

$$\rho \triangleq \frac{\|\mathcal{D}_{\Lambda^c}^H R^t\|_{q' \rightarrow \infty}}{\|\mathcal{D}_{\Lambda}^H R^t\|_{q' \rightarrow \infty}} \quad (12.4.1)$$

is strictly less than one i.e.  $\rho < 1$ . Here  $q'$  is given by

$$\frac{1}{q} + \frac{1}{q'} = 1.$$

E.g. for  $q = 1$ ,  $q' = \infty$ .

Just like the analysis of single vector setting, we can call this ratio as **greedy selection ratio**.

## 12.5. S-OMP recovery guarantee for Exact sparse problem

We present a result showing that S-OMP for the (Joint-P<sub>0</sub>) problem in terms of ERC.

**Theorem 12.2** [13] *Let  $X = \mathcal{D}\mathcal{A}^*$  where  $\mathcal{A}^*$  is the optimal solution to the problem (Joint-P<sub>0</sub>). Let  $\Lambda = \text{supp}(\mathcal{A}^*)$ .*

*A sufficient condition for S-OMP for  $q \geq 1$  to recover  $\mathcal{A}^*$  is that*

$$\|\mathcal{D}_{\Lambda}^{\dagger} d_j\| < 1 \quad \forall j \notin \Lambda. \quad (12.5.1)$$

Equivalently the sufficient condition in terms of *Exact Recovery Coefficient* is:

$$\text{ERC}(\mathcal{D}, \Lambda) > 0. \quad (12.5.2)$$

**PROOF.** For the exact sparse problem, it is easy to see that residuals in  $R^t$  belong to the column space of  $\mathcal{D}_\Lambda$ . The Projection operator for  $\mathcal{D}_\Lambda$  is

$$\mathcal{D}_\Lambda \mathcal{D}_\Lambda^\dagger = \mathcal{D}_\Lambda (\mathcal{D}_\Lambda^H \mathcal{D}_\Lambda)^{-1} \mathcal{D}_\Lambda^H = \left( \mathcal{D}_\Lambda^\dagger \right)^H \mathcal{D}_\Lambda^H.$$

Clearly

$$R^t = \left( \mathcal{D}_\Lambda^\dagger \right)^H \mathcal{D}_\Lambda^H R^t.$$

Putting it back in (12.4.1), we obtain

$$\rho = \frac{\|\mathcal{D}_{\Lambda^c}^H R^t\|_{q' \rightarrow \infty}}{\|\mathcal{D}_\Lambda^H R^t\|_{q' \rightarrow \infty}} = \frac{\|\mathcal{D}_{\Lambda^c}^H \left( \mathcal{D}_\Lambda^\dagger \right)^H \mathcal{D}_\Lambda^H R^t\|_{q' \rightarrow \infty}}{\|\mathcal{D}_\Lambda^H R^t\|_{q' \rightarrow \infty}}.$$

Using the result

$$\|AB\|_{p \rightarrow \infty} \leq \|A\|_\infty \|B\|_{p \rightarrow \infty}$$

we can write

$$\|\mathcal{D}_{\Lambda^c}^H \left( \mathcal{D}_\Lambda^\dagger \right)^H \mathcal{D}_\Lambda^H R^t\|_{q' \rightarrow \infty} \leq \|\mathcal{D}_{\Lambda^c}^H \left( \mathcal{D}_\Lambda^\dagger \right)^H\|_\infty \|\mathcal{D}_\Lambda^H R^t\|_{q' \rightarrow \infty}$$

Putting it back, we obtain

$$\rho \leq \|\mathcal{D}_{\Lambda^c}^H \left( \mathcal{D}_\Lambda^\dagger \right)^H\|_\infty \leq \|\mathcal{D}_\Lambda^\dagger \mathcal{D}_{\Lambda^c}\|_1$$

It is easy to see that

$$\|\mathcal{D}_\Lambda^\dagger \mathcal{D}_{\Lambda^c}\|_1 = \max_{j \in \Lambda^c} \|\mathcal{D}_\Lambda^\dagger d_j\|.$$

Thus if (12.5.1) holds, then  $\rho < 1$  which is the sufficient condition to ensure that an atom from  $\Lambda$  will be picked up in every iteration of S-OMP.  $\square$

We can now use coherence and Babel function based bounds which ensure that  $\text{ERC}(\mathcal{D}, \Lambda) < 1$ .

**Recall** the relationship between ERC and coherence.

$$\text{ERC}(\Lambda) \geq \frac{1 - (2K - 1)\mu}{1 - (K - 1)\mu}.$$

This gives us the **result** that whenever

$$K < \frac{1}{2} \left( 1 + \frac{1}{\mu} \right)$$

then  $\text{ERC}(\Lambda) > 0$ .

**ERC and Babel function** are related as follows.

$$\text{ERC}(\Lambda) \geq \frac{1 - \mu_1(K - 1) - \mu_1(K)}{1 - \mu_1(K - 1)}.$$

Thus if

$$\mu_1(K - 1) + \mu_1(K) < 1$$

then ERC is positive.

## 12.6. Approximation with a sparsity bound

This flavor of sparse approximation problem can be stated as

$$\min_{\mathcal{A} \in \mathbb{C}^{D,S}} \|X - \mathcal{D}\mathcal{A}\|_F \quad \text{subject to} \quad \|\mathcal{A}\|_{\text{row-0}} \leq K. \quad (\text{SPARSE})$$

$K$  is the upper limit of row- $l_0$  norm allowed on the representation matrix  $\mathcal{A}$  i.e. the maximum number of non-zero rows allowed in  $\mathcal{A}$  is  $K$ .

$\|X - \mathcal{D}\mathcal{A}\|_F$  is the Frobenius norm of approximation error for a specific choice of  $\mathcal{A}$ . Thus we wish to minimize the approximation error subject to maximum number of atoms that can be chosen from the dictionary  $\mathcal{D}$ .

If  $\mathcal{A}_{\text{opt}}$  solves the optimization problem, then the corresponding approximation of the signal matrix  $X$  is given by  $\widehat{X} = \Phi\mathcal{A}_{\text{opt}}$ .

We have the following theoretical guarantee.

**Theorem 12.3** [40] Assume that  $\mu_1(K) < \frac{1}{2}$ . Given an input matrix  $X$ , suppose that  $\mathcal{A}_{opt}$  solves **(SPARSE)** and that  $\widehat{X} = \Phi\mathcal{A}_{opt}$ . After  $K$  iterations, S-OMP will produce an approximation  $X^K$  that satisfies the error bound

$$\|X - X^K\|_F \leq \left[ 1 + SK \frac{1 - \mu_1(K)}{1 - 2\mu_1(K)^2} \right] \|X - \widehat{X}\|_F. \quad (12.6.1)$$

In words, S-OMP is an approximation algorithm for **(SPARSE)**.

We also note that if signals in  $X$  are truly sparse and can be represented exactly by the atoms indexed by  $\Lambda$ , then

$$X = \widehat{X}$$

and S-OMP will indeed recover it under the conditions of theorem 12.3.

**PROOF.** Suppose that some solution of **(SPARSE)** involves  $K$  atoms indexed in  $\Lambda_{opt}$ .

We can rewrite

$$R^t = X - X^t = (X - \widehat{X}) + (\widehat{X} - X^t).$$

We put this in (12.4.1) to obtain

$$\rho = \frac{\|\mathcal{D}_B^H(X - \widehat{X}) + \mathcal{D}_B^H(\widehat{X} - X^t)\|_{q' \rightarrow \infty}}{\|\mathcal{D}_G^H(X - \widehat{X}) + \mathcal{D}_G^H(\widehat{X} - X^t)\|_{q' \rightarrow \infty}} \quad (12.6.2)$$

Since  $\widehat{X}$  is the least squares estimate of  $X$  over the columns in  $\mathcal{D}_G$ , hence

$$\mathcal{D}_G^H(X - \widehat{X}) = 0.$$

This term goes away from the denominator.

Applying the triangular inequality we see that

$$\rho \leq \frac{\|\mathcal{D}_B^H(X - \widehat{X})\|_{\infty \rightarrow \infty}}{\|\mathcal{D}_G^H(\widehat{X} - X^t)\|_{\infty \rightarrow \infty}} + \frac{\|\mathcal{D}_B^H(\widehat{X} - X^t)\|_{\infty \rightarrow \infty}}{\|\mathcal{D}_G^H(\widehat{X} - X^t)\|_{\infty \rightarrow \infty}} \quad (12.6.3)$$

The 2nd fraction in this equation can be bound as follows. Let  $\mathcal{D}_G^\dagger$  be the pseudo-inverse of  $\mathcal{D}_G$ . Then following the steps similar to [34], we can show that

$$\frac{\|\mathcal{D}_B^H(\hat{X} - X^t)\|_{\infty \rightarrow \infty}}{\|\mathcal{D}_G^H(\hat{X} - X^t)\|_{\infty \rightarrow \infty}} \leq \|\mathcal{D}_G^\dagger \mathcal{D}_B\|_{1 \rightarrow 1}. \quad (12.6.4)$$

It can be shown that (HOW? )

$$\frac{\|\mathcal{D}_B^H(X - \hat{X})\|_{\infty \rightarrow \infty}}{\|\mathcal{D}_G^H(\hat{X} - X^t)\|_{\infty \rightarrow \infty}} \leq \frac{\|\mathcal{D}_B^H(X - \hat{X})\|_{\infty \rightarrow \infty}}{\|\mathcal{D}_G^\dagger\|_{2 \rightarrow 1}^{-1} \|(\hat{X} - X^t)\|_F}. \quad (12.6.5)$$

□

## 12.7. Joint $l_1$ minimization recovery guarantee

We now present a recovery guarantee for the equivalence of (**Joint-P<sub>0</sub>**) and (**Joint-P<sub>1</sub>**) problems. This guarantee establishes the conditions under which basis pursuit can be used safely for solving the sparse recovery problem in multiple signal setting.

**Theorem 12.4** [13] *Let  $X = \mathcal{D}\mathcal{A}^*$  where  $\mathcal{A}^*$  is the optimal solution to the problem (**Joint-P<sub>0</sub>**). Let  $\Lambda = \text{supp}(\mathcal{A}^*)$ .*

*A sufficient condition for  $\mathcal{A}^*$  to be the solution of (**Joint-P<sub>1</sub>**) is that*

$$\|\mathcal{D}_\Lambda^\dagger d_j\| < 1 \quad \forall j \notin \Lambda. \quad (12.7.1)$$

*Equivalently the sufficient condition in terms of **Exact Recovery Coefficient** is:*

$$ERC(\mathcal{D}, \Lambda) > 0. \quad (12.7.2)$$

PROOF. Suppose that there exists another feasible representation  $\mathcal{B}$  such that

$$X = \mathcal{D}_\Lambda \underline{\mathcal{A}}_\Lambda^* = \mathcal{D}_{\Lambda'} \underline{\mathcal{B}}_{\Lambda'} \quad (12.7.3)$$

where  $\Lambda \neq \Lambda'$  and  $\Lambda' = \text{supp}(\mathcal{B})$ .

Let us add few more symbols to help out in the proof. We have  $\mathcal{A}, \mathcal{B} \in \mathbb{C}^{D \times K}$ . Let  $K = |\Lambda|$  and  $K' = |\Lambda'|$ . Then  $\mathcal{D}_\Lambda \in \mathbb{C}^{N \times K}$  and  $\mathcal{D}_{\Lambda'} \in \mathbb{C}^{N \times K'}$ . Further,  $\underline{\mathcal{A}}_\Lambda^* \in \mathbb{C}^{K \times S}$  and  $\underline{\mathcal{B}}_{\Lambda'} \in \mathbb{C}^{K' \times S}$ .

Further, let us denote

$$\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_K\}$$

and

$$\Lambda' = \{\omega_1, \omega_2, \dots, \omega_{K'}\}.$$

We note that the set  $\Lambda'$  should contain at least one index different from indices in  $\Lambda$ . Otherwise, it would have been a subset of  $\Lambda$  and that would contradict minimality of  $\Lambda$  as a solution of (Joint-P<sub>0</sub>).

We need to show that under the condition (12.7.1)

$$\|\mathcal{A}^*\|_{p,1} < \|\mathcal{B}\|_{p,1}. \quad (12.7.4)$$

holds. Recall that

$$\|\mathcal{A}^*\|_{p,1} = \sum_{i=1}^D \|\underline{\mathcal{A}}^{*i}\|_p = \sum_{i \in \Lambda} \|\underline{\mathcal{A}}^{*i}\|_p \quad (12.7.5)$$

i.e. it is the sum of  $l_p$  norms of the non-zero rows of  $\mathcal{A}^*$ .

The atoms indexed by  $\Lambda$  are linearly independent, otherwise we could have found a solution with lower row – 0 norm for (Joint-P<sub>0</sub>). Note that we cannot say something like this for  $\Lambda'$ .

Thus, from (12.7.3), we get

$$\underline{\mathcal{A}}_\Lambda^* = (\mathcal{D}_\Lambda^\dagger \mathcal{D}_{\Lambda'}) \underline{\mathcal{B}}_{\Lambda'}. \quad (12.7.6)$$

Here  $(\mathcal{D}_\Lambda^\dagger \mathcal{D}_{\Lambda'})$  is a  $K \times K'$  matrix.

Writing down  $\underline{\mathcal{A}}_\Lambda^*$  as

$$\underline{\mathcal{A}}_\Lambda^* = \begin{bmatrix} \underline{\mathcal{a}}^{\lambda_1} \\ \vdots \\ \underline{\mathcal{a}}^{\lambda_K} \end{bmatrix}$$

and  $\underline{\mathcal{B}}_{\Lambda'}$  as

$$\underline{\mathcal{B}}_{\Lambda'} = \begin{bmatrix} \underline{b}^{\omega_1} \\ \vdots \\ \underline{b}^{\omega_{K'}} \end{bmatrix}$$

we can see that

$$\underline{a}^{\lambda_i} = \sum_{j=1}^{K'} \left( \mathcal{D}_{\Lambda}^{\dagger} \mathcal{D}_{\Lambda'} \right)_{ij} \underline{b}^{\omega_j} \quad (12.7.7)$$

where  $\left( \mathcal{D}_{\Lambda}^{\dagger} \mathcal{D}_{\Lambda'} \right)_{ij}$  is the  $(i, j)$ -th entry of the  $K \times K'$  matrix  $\mathcal{D}_{\Lambda}^{\dagger} \mathcal{D}_{\Lambda'}$ .

In words, a row vector in  $\underline{\mathcal{A}}_{\Lambda}^*$  on the L.H.S. is a linear combination of the row vectors in  $\underline{\mathcal{B}}_{\Lambda'}$  on the R.H.S. where the coefficients of the linear combination come from the corresponding row in the pre-multiplying matrix.

Taking  $l_p$  norm on both sides and applying triangular inequality we get

$$\|\underline{a}^{\lambda_i}\|_p \leq \sum_{j=1}^{K'} \left| \left( \mathcal{D}_{\Lambda}^{\dagger} \mathcal{D}_{\Lambda'} \right)_{ij} \right| \|\underline{b}^{\omega_j}\|_p. \quad (12.7.8)$$

Taking  $\sum_i$  over both sides we get

$$\sum_{i=1}^K \|\underline{a}^{\lambda_i}\|_p \leq \sum_{i=1}^K \sum_{j=1}^{K'} \left| \left( \mathcal{D}_{\Lambda}^{\dagger} \mathcal{D}_{\Lambda'} \right)_{ij} \right| \|\underline{b}^{\omega_j}\|_p. \quad (12.7.9)$$

Recall from (12.7.5) that L.H.S. is nothing but  $\|\mathcal{A}^*\|_{p,1}$ .

Interchanging the order of sums on the R.H.S. we get

$$\|\mathcal{A}^*\|_{p,1} \leq \sum_{j=1}^{K'} \sum_{i=1}^K \left| \left( \mathcal{D}_{\Lambda}^{\dagger} \mathcal{D}_{\Lambda'} \right)_{ij} \right| \|\underline{b}^{\omega_j}\|_p. \quad (12.7.10)$$

Let us look at the term

$$\sum_{i=1}^K \left| \left( \mathcal{D}_{\Lambda}^{\dagger} \mathcal{D}_{\Lambda'} \right)_{ij} \right|.$$

We can write

$$\mathcal{D}_{\Lambda'} = \begin{bmatrix} d_{\omega_1} & \dots & d_{\omega_{K'}} \end{bmatrix}$$

where  $\{\omega_1, \dots, \omega_{K'}\} = \Lambda'$ .

Thus

$$\mathcal{D}_\Lambda^\dagger \mathcal{D}_{\Lambda'} = \begin{bmatrix} \mathcal{D}_\Lambda^\dagger d_{\omega_1} & \dots & \mathcal{D}_\Lambda^\dagger d_{\omega_{K'}} \end{bmatrix}.$$

Consequently

$$\sum_{i=1}^K \left| \left( \mathcal{D}_\Lambda^\dagger \mathcal{D}_{\Lambda'} \right)_{ij} \right| = \|\mathcal{D}_\Lambda^\dagger d_{\omega_j}\|_1.$$

Now, if  $\omega_j \in \Lambda$ , then

$$\|\mathcal{D}_\Lambda^\dagger d_{\omega_j}\|_1 \leq 1$$

and if  $\omega_j \notin \Lambda$ , then due to (12.7.1)

$$\|\mathcal{D}_\Lambda^\dagger d_{\omega_j}\|_1 < 1.$$

We have already established that there is at least one  $\omega_j \in \Lambda'$  such that  $\omega_j \notin \Lambda$ . Thus, putting in (12.7.10), we obtain

$$\|\mathcal{A}^*\|_{p,1} < \sum_{j=1}^{K'} \|\underline{b}^j\|_p = \|\mathcal{B}\|_{p,1}.$$

Thus, if (12.7.1) holds, then  $\mathcal{A}^*$  is necessarily the solution of both (Joint-P<sub>0</sub>) and (Joint-P<sub>1</sub>) problems.  $\square$

We see that the condition (12.7.1) is not different from the condition for the recovery using  $l_1$  minimization for the signal signal setting. We have not been able to show anything so far which can demonstrate that  $l_1$  minimization can take advantage of the presence of multiple signals in the uniqueness guarantees.

From theorem 12.4 it is easy to specialize the results in terms of coherence and Babel function.

## CHAPTER 13

# Rank Aware and MUSIC based Algorithms for Joint Sparse Recovery

This chapter is largely drawn from [13, 19].

### 13.1. $l_0$ norm minimization

We begin our discussion with the problem

$$\widehat{\mathcal{A}} = \arg \min_{\mathcal{A} \in \mathbb{C}^{D \times S}} \|\mathcal{A}\|_{\text{row}-0} \text{ subject to } X = \mathcal{D}\mathcal{A}. \quad (\text{Joint-P}_0)$$

In general, the results will be equally valid for the MMV (multiple measurement vector) problem in compressed sensing setting.

$$\widehat{X} = \arg \min_{X \in \mathbb{C}^{N \times S}} \|X\|_{\text{row}-0} \text{ subject to } Y = \Phi X. \quad (\text{Joint-CS}_0)$$

We will keep switching between the two settings (sparse representation and compressed sensing) during the chapter to have the flavor of both settings. Wherever a result is applicable only for one setting, we will specifically mark them.

#### 13.1.1. Uniqueness of $l_0$ norm minimization

We will assume that the solution to (Joint-P<sub>0</sub>) is unique. This can be observed from the conditions developed for the single signal setting.

**Theorem 13.1** [13] *The matrix  $\widehat{\mathcal{A}}$  is the unique solution to (Joint-P<sub>0</sub>) if  $X = \mathcal{D}\widehat{\mathcal{A}}$  and*

$$\|\widehat{\mathcal{A}}\|_{\text{row}-0} < \frac{1}{2} \text{spark}(\mathcal{D}). \quad (13.1.1)$$

Note: We are not saying that  $\mathcal{A}$  itself is a unique solution. It may happen that  $\mathcal{A}$  is non-sparse or not sufficiently sparse. In general,  $X = \mathcal{D}\mathcal{A}$  has infinite solutions. All, we are saying that a particular solution  $\widehat{\mathcal{A}}$  is unique (in row sparsity sense), if it satisfies (13.1.1). If no solution satisfies (13.1.1), then there is no *unique* (in row sparsity sense) solution to (Joint-P<sub>0</sub>).

PROOF. If (13.1.1) holds, then for every  $\widehat{\alpha}^s$  with  $1 \leq s \leq S$ , we have

$$\|\widehat{\alpha}^s\|_0 < \frac{1}{2} \text{spark}(\mathcal{D}).$$

Since  $x^s = \mathcal{D}\widehat{\alpha}^s$ , the uniqueness-spark condition for single signal setting says that  $\widehat{\alpha}^s$  is unique, establishing the uniqueness of  $\widehat{\mathcal{A}}$ .  $\square$

An alternative proof is presented below.

PROOF. Assume that (13.1.1) holds for two different solutions of (Joint-P<sub>0</sub>) namely  $\mathcal{A}$  and  $\mathcal{B}$ . Then

$$\max(\|\mathcal{A}\|_{\text{row}-0}, \|\mathcal{B}\|_{\text{row}-0}) < \frac{1}{2} \text{spark}(\mathcal{D}).$$

Using triangle inequality for row - 0-“norm”, we have

$$\|\mathcal{A} - \mathcal{B}\|_{\text{row}-0} \leq \|\mathcal{A}\|_{\text{row}-0} + \|\mathcal{B}\|_{\text{row}-0} < \text{spark}(\mathcal{D}). \quad (13.1.2)$$

On the other hand, since  $0 = \mathcal{D}(\mathcal{A} - \mathcal{B})$ , if we consider the first column of  $(\mathcal{A} - \mathcal{B})$  say  $\gamma$ , we have

$$\mathcal{D}\gamma = 0.$$

This means that either  $\gamma = 0$  or

$$\|\gamma\|_0 \geq \text{spark}(\mathcal{D}).$$

Consequently

$$\|\mathcal{A} - \mathcal{B}\|_{\text{row}-0} \geq \|\gamma\|_0 \geq \text{spark}(\mathcal{D})$$

which contradicts (13.1.2). Thus, the only possibility is  $\gamma = 0$  (i.e. the first columns of  $\mathcal{A}$  and  $\mathcal{B}$  are same).

The same logic requires each column of  $(\mathcal{A} - \mathcal{B})$  to be zero leading to  $\mathcal{A} = \mathcal{B}$  and thus providing uniqueness guarantee.  $\square$

So far the uniqueness result is same as the result for single signal setting. What is the advantage that we obtain by having so many signals in the signal matrix  $X$ ?

This is answered in the following uniqueness result which incorporates the rank of  $\mathcal{A}$  as an additional feature. Going forward we will simply characterize the requirements on  $\mathcal{A}$  which ensure its uniqueness.

**Theorem 13.2** [13] *Let  $\text{rank}(X)$  denote the rank of the signal matrix  $X$ . Obviously  $\text{rank}(X) \leq S$ . Matrix  $\mathcal{A}$  will be the unique solution to the problem (Joint-P<sub>0</sub>) if*

$$\|\mathcal{A}\|_{\text{row-0}} < \frac{1}{2} [\text{spark}(\mathcal{D}) - 1 + \text{rank}(X)]. \quad (13.1.3)$$

PROOF. We have  $X \in \mathbb{C}^{N \times S}$ . Suppose, we have  $X = \mathcal{D}\mathcal{A} = \mathcal{D}\mathcal{B}$  where  $\mathcal{A}, \mathcal{B} \in \mathbb{C}^{D \times S}$ , and  $\mathcal{A} \neq \mathcal{B}$ .

Let  $\text{nullity}(A)$  denote the dimension of the null space of a matrix  $A$  and  $\text{rank}(A)$  denote its rank.

We have

$$\text{nullity}(\mathcal{A}) \leq \text{nullity}(X) \quad (13.1.4)$$

and

$$\text{nullity}(\mathcal{B}) \leq \text{nullity}(X). \quad (13.1.5)$$

Similarly, for the ranks, we have the relations

$$\text{rank}(\mathcal{A}) \geq \text{rank}(X) \quad (13.1.6)$$

and

$$\text{rank}(\mathcal{B}) \geq \text{rank}(X). \quad (13.1.7)$$

Let  $\Lambda_a = \text{supp}(\mathcal{A})$  and  $\Lambda_b = \text{supp}(\mathcal{B})$ .  $\Lambda_a$  corresponds to the non-zero rows of  $\mathcal{A}$  and  $\Lambda_b$  corresponds to the non-zero rows of  $\mathcal{B}$ .

Let  $\Lambda_c = \Lambda_a \cap \Lambda_b$ . It corresponds to the rows which are non-zero in both  $\mathcal{A}$  and  $\mathcal{B}$ .

$\Lambda_a \cup \Lambda_b$  identifies all the atoms in  $\mathcal{D}$  which are involved in the constructions  $X = \mathcal{D}\mathcal{A}$  and  $X = \mathcal{D}\mathcal{B}$ .

$\Lambda'_a = \Lambda_a \setminus \Lambda$  identifies atoms which are involved only in the construction  $X = \mathcal{D}\mathcal{A}$ .

$\Lambda'_b = \Lambda_b \setminus \Lambda$  identifies atoms which are involved only in the construction  $X = \mathcal{D}\mathcal{B}$ .

$\Lambda_c$  identifies atoms which are involved in both constructions.

Let us construct a matrix

$$\mathcal{D}' = \begin{bmatrix} \mathcal{D}_{\Lambda'_a} & \mathcal{D}_{\Lambda_c} & \mathcal{D}_{\Lambda'_b} \end{bmatrix} \quad (13.1.8)$$

The submatrix  $\begin{bmatrix} \mathcal{D}_{\Lambda'_a} & \mathcal{D}_{\Lambda_c} \end{bmatrix}$  corresponds to atoms for  $\text{supp}(\mathcal{A})$  and the submatrix  $\begin{bmatrix} \mathcal{D}_{\Lambda_c} & \mathcal{D}_{\Lambda'_b} \end{bmatrix}$  corresponds to atoms for  $\text{supp}(\mathcal{B})$ .

Let  $r_a = \|\mathcal{A}\|_{\text{row}-0}$  and  $r_b = \|\mathcal{B}\|_{\text{row}-0}$ .

Let  $r_c = |\Lambda_c|$  (i.e. the number of atoms common to both supports).

Let  $\mathcal{A}_a$  (resp.  $\mathcal{B}_b$ ) be the submatrix of  $\mathcal{A}$  (resp.  $\mathcal{B}$ ) constructed by picking rows indexed by  $\Lambda'_a$  (resp.  $\Lambda'_b$ ).

Let  $\mathcal{A}_c$  and  $\mathcal{B}_c$  be submatrices of  $\mathcal{A}$  and  $\mathcal{B}$  constructed by picking rows indexed by  $\Lambda_c$ . Then

$$X = \begin{bmatrix} \mathcal{D}_{\Lambda'_a} & \mathcal{D}_{\Lambda_c} \end{bmatrix} \begin{bmatrix} \mathcal{A}_a \\ \mathcal{A}_c \end{bmatrix} = \begin{bmatrix} \mathcal{D}_{\Lambda_c} & \mathcal{D}_{\Lambda'_b} \end{bmatrix} \begin{bmatrix} \mathcal{B}_c \\ \mathcal{B}_b \end{bmatrix} \quad (13.1.9)$$

A simple manipulation gives us

$$0 = \mathcal{D}(\mathcal{A} - \mathcal{B}) = \begin{bmatrix} \mathcal{D}_{\Lambda'_a} & \mathcal{D}_{\Lambda_c} & \mathcal{D}_{\Lambda'_b} \end{bmatrix} \begin{bmatrix} \mathcal{A}_a \\ \mathcal{A}_c - \mathcal{B}_c \\ -\mathcal{B}_b \end{bmatrix}. \quad (13.1.10)$$

From (13.1.10) we have

$$\text{nullity}(\mathcal{D}') \geq \text{rank} \left( \begin{bmatrix} \mathcal{A}_a \\ \mathcal{A}_c - \mathcal{B}_c \\ -\mathcal{B}_b \end{bmatrix} \right). \quad (13.1.11)$$

It is easy to see that

$$\text{rank} \begin{pmatrix} \mathcal{A}_a \\ \mathcal{A}_c - \mathcal{B}_c \\ -\mathcal{B}_b \end{pmatrix} \geq \max(\text{rank}(\mathcal{A}_a), \text{rank}(\mathcal{B}_b)). \quad (13.1.12)$$

Without loss of generality, we consider  $\mathcal{A}_a$  only.

Also, we can see that

$$\text{nullity}(\mathcal{A}_a) \leq \text{nullity}(\mathcal{A}) + r_c. \quad (13.1.13)$$

Consider the linear systems  $\mathcal{A}y = 0$  and  $\mathcal{A}_a y = 0$ . The former has  $r_c$  more constraints. So its solution space (the null space of  $\mathcal{A}$ ) is at most reduced by  $r_c$  dimensions.

This leads to

$$\text{rank}(\mathcal{A}) - r_c \leq \text{rank}(\mathcal{A}_a). \quad (13.1.14)$$

Combining, we have

$$\text{nullity}(\mathcal{D}') \geq \text{rank} \begin{pmatrix} \mathcal{A}_a \\ \mathcal{A}_c - \mathcal{B}_c \\ -\mathcal{B}_b \end{pmatrix} \geq \text{rank}(\mathcal{A}) - r_c \geq \text{rank}(X) - r_c. \quad (13.1.15)$$

Simplifying

$$\text{nullity}(\mathcal{D}') \geq \text{rank}(X) - r_c. \quad (13.1.16)$$

By the definition of spark we have

$$\text{rank}(\begin{bmatrix} \mathcal{D}_{\Lambda'_a} & \mathcal{D}_{\Lambda_c} & \mathcal{D}_{\Lambda'_b} \end{bmatrix}) \geq \text{spark}(\begin{bmatrix} \mathcal{D}_{\Lambda'_a} & \mathcal{D}_{\Lambda_c} & \mathcal{D}_{\Lambda'_b} \end{bmatrix}) - 1 \geq \text{spark}(\mathcal{D}) - 1. \quad (13.1.17)$$

Combining all of the above, we have

$$\begin{aligned} r_a + r_b - r_c &= \#\text{Cols}(\begin{bmatrix} \mathcal{D}_{\Lambda'_a} & \mathcal{D}_{\Lambda_c} & \mathcal{D}_{\Lambda'_b} \end{bmatrix}) \\ &= \text{rank}(\begin{bmatrix} \mathcal{D}_{\Lambda'_a} & \mathcal{D}_{\Lambda_c} & \mathcal{D}_{\Lambda'_b} \end{bmatrix}) + \text{nullity}(\begin{bmatrix} \mathcal{D}_{\Lambda'_a} & \mathcal{D}_{\Lambda_c} & \mathcal{D}_{\Lambda'_b} \end{bmatrix}) \\ &\geq \text{spark}(\mathcal{D}) - 1 + \text{rank}(X) - r_c. \end{aligned}$$

This gives us

$$r_a + r_b \geq \text{spark}(\mathcal{D}) - 1 + \text{rank}(X). \quad (13.1.18)$$

Thus if there are two distinct solutions of (**Joint-P<sub>0</sub>**) problem, then

$$\|\mathcal{A}\|_{\text{row}-0} + \|\mathcal{B}\|_{\text{row}-0} \geq \text{spark}(\mathcal{D}) - 1 + \text{rank}(X).$$

Hence, if

$$\|\mathcal{A}\|_{\text{row}-0} < \frac{1}{2}(\text{spark}(\mathcal{D}) - 1 + \text{rank}(X))$$

then  $\mathcal{A}$  is necessarily the unique solution.  $\square$

**Theorem 13.3** [13] If

$$\|\mathcal{A}\|_{\text{row}-0} < \frac{1}{2}\left(\frac{1}{\mu} + \text{rank}(X)\right) \quad (13.1.19)$$

where  $\mu$  is the coherence of  $\mathcal{D}$ , then  $\mathcal{A}$  is the unique solution to (**Joint-P<sub>0</sub>**).

PROOF. Recall that

$$\text{spark}(\mathcal{D}) \geq 1 + \frac{1}{\mu}.$$

Thus

$$\frac{1}{\mu} \leq \text{spark}(\mathcal{D}) - 1.$$

Thus

$$\|\mathcal{A}\|_{\text{row}-0} < \frac{1}{2}\left(\frac{1}{\mu} + \text{rank}(X)\right) \implies \|\mathcal{A}\|_{\text{row}-0} < \frac{1}{2}(\text{spark}(\mathcal{D}) - 1 + \text{rank}(X)).$$

$\square$

**Theorem 13.4** [13] If

$$\|\mathcal{A}\|_{\text{row}-0} < \mu_{1/2}(G) + \frac{1}{2} \text{rank}(X) \quad (13.1.20)$$

then  $\mathcal{A}$  is the unique solution to (**Joint-P<sub>0</sub>**).

We recall that  $G$  is the Gram-matrix for  $\mathcal{D}$  and  $\mu_{1/2}(G)$  is the smallest number  $m$  such that the sum of magnitudes of a collection of  $m$  off-diagonal entries in a single row or column of the Gram matrix  $G$  is at least  $\frac{1}{2}$ .

PROOF. We recall that

$$\text{spark}(\mathcal{D}) \geq 2\mu_{1/2}(G) + 1.$$

Thus

$$2\mu_{1/2}(G) \leq \text{spark}(\mathcal{D}) - 1.$$

The rest is a simple application of theorem 13.2.  $\square$

An alternative rendition of theorem 13.2 in the context of compressed sensing would be as follows ( $\text{Joint-CS}_0$ )

**Theorem 13.5** [13] *Matrix  $X$  will be the unique solution to the problem ( $\text{Joint-CS}_0$ ) if*

$$\|X\|_{\text{row-0}} < \frac{1}{2} [\text{spark}(\Phi) - 1 + \text{rank}(Y)]. \quad (13.1.21)$$

Let us spend some time understanding this result. In CS setting, we have a flexibility to choose the number of measurements  $M$  in the sensing matrix  $\Phi$ . In general, for well designed sensing matrices, the spark increases as the number of measurements increase. Assuming  $K = \|X\|_{\text{row-0}}$  remains fixed, we see that  $\text{spark}(\Phi)$  can be decreased if  $\text{rank}(Y)$  is high. Thus, with high rank measurement matrices, the number of required measurements can be reduced.

Alternatively, if  $\text{rank}(Y)$  increases and  $\text{spark}(\Phi)$  remains constant (number of measurements doesn't change), then higher levels of sparsity (higher  $K = \|X\|_{\text{row-0}}$ ) can be supported.

In the best case, we would have  $\text{rank}(Y) = K$ . Then

$$K < \frac{\text{spark}(\Phi) - 1 + K}{2}$$

gives us the required condition as  $\text{spark}(\Phi) > K + 1$ . When  $\text{spark}(\Phi)$  also takes up its largest value  $M + 1$  (i.e. the matrix  $\Phi$  is full rank), then we can simplify the condition as  $M \geq K + 1$ . Therefore, in the best case scenario only  $K + 1$  measurements per signal are enough to ensure uniqueness of the sparse signal matrix.

We recall that for the SMV (single measurement vector) setting, the minimum number of required measurements is  $2K$ .

Theorem 13.2 provides a uniqueness condition on  $\|\mathcal{A}\|_{\text{row}-0}$  in terms of spark of the dictionary  $\mathcal{D}$  and rank of the signal matrix  $X$ . It would be interesting to have a sufficient condition in terms of the rank of the representation matrix  $\mathcal{A}$ . Moreover, if we could show that such a condition is also necessary, then we would have established the sharpness of the condition.

The next result shows that we can replace  $\text{rank}(X)$  with  $\text{rank}(\mathcal{A})$  in (13.1.3).

**Theorem 13.6** [19] *The sufficient condition of (13.1.3) in theorem 13.2 is equivalent to*

$$\|\mathcal{A}\|_{\text{row}-0} < \frac{1}{2} [\text{spark}(\mathcal{D}) - 1 + \text{rank}(\mathcal{A})]. \quad (13.1.22)$$

PROOF. Since  $\text{rank}(X) \leq \text{rank}(\mathcal{A})$ , hence if (13.1.3) is true, the (13.1.22) is also true.

We now need to show that (13.1.22) also implies (13.1.3).

We note that  $\text{rank}(\mathcal{A}) \leq \|\mathcal{A}\|_{\text{row}-0}$  i.e. the rank of  $\mathcal{A}$  is not larger than the number of non-zero rows in  $\mathcal{A}$ . Putting this in (13.1.22) we obtain

$$\|\mathcal{A}\|_{\text{row}-0} < \frac{1}{2} [\text{spark}(\mathcal{D}) - 1 + \|\mathcal{A}\|_{\text{row}-0}].$$

Simplifying, we get

$$K = \|\mathcal{A}\|_{\text{row}-0} < \text{spark}(\mathcal{D}) - 1.$$

This means that  $\text{spark}(\mathcal{D}) > K + 1$ . Thus, any set of  $K$  columns is linearly independent. In particular, if  $\Lambda = \text{supp}(\mathcal{A})$ , then the subdictionary  $\mathcal{D}_\Lambda$  with columns indexed by  $\Lambda$  must be full rank. Finally, with  $X = \mathcal{D}_\Lambda \underline{\mathcal{A}}_\Lambda$  gives us  $\text{rank}(X) = \text{rank}(\underline{\mathcal{A}}_\Lambda) = \text{rank}(\mathcal{A})$  since all the rows in  $\mathcal{A}$  not indexed by  $\Lambda$  are zero. Finally  $\text{rank}(X) = \text{rank}(\mathcal{A})$  means that the (13.1.22) implies (13.1.3).  $\square$

The next result shows that the conditions (13.1.3) and (13.1.22) are both necessary and sufficient for uniqueness in (Joint-P<sub>0</sub>) problem.

**Theorem 13.7** [19] *The condition (13.1.22)*

$$\|\mathcal{A}\|_{\text{row-0}} < \frac{1}{2} [\text{spark}(\mathcal{D}) - 1 + \text{rank}(\mathcal{A})]$$

*or equivalently the condition (13.1.3)*

$$\|\mathcal{A}\|_{\text{row-0}} < \frac{1}{2} [\text{spark}(\mathcal{D}) - 1 + \text{rank}(X)]$$

*is both necessary and sufficient condition for  $\mathcal{A}$  to be the unique solution of (Joint-P<sub>0</sub>) problem.*

**PROOF.** We have already proved in theorem 13.2 and theorem 13.6 that the conditions are sufficient.

We now need to show that the conditions are necessary too. For this we need to show that there exists a representation matrix  $\mathcal{A}$  with  $P = \text{rank}(\mathcal{A})$  and  $K = \|\mathcal{A}\|_{\text{row-0}}$  such that if  $2K \geq \text{spark}(\mathcal{D}) - 1 + P$  then  $\mathcal{A}$  cannot be uniquely determined. In other words, for such an  $\mathcal{A}$  there exists another matrix  $\mathcal{B}$  with  $\|\mathcal{B}\|_{\text{row-0}} \leq K$  such that  $X = \mathcal{D}\mathcal{B}$ .

We start with

$$2K \geq \text{spark}(\mathcal{D}) - 1 + P \iff \text{spark}(\mathcal{D}) \leq 2K - P + 1.$$

Define  $T = 2K - P + 1$ . Then  $\text{spark}(\mathcal{D}) \leq T$  means that there exists an index set  $\Gamma$  with  $T = |\Gamma|$  such the columns of  $\mathcal{D}_\Gamma$  are linearly dependent. In other words, there exists a vector  $v \in \mathbb{C}^T$  such that  $\mathcal{D}_\Gamma v = 0$ . Now construct  $V \in \mathbb{C}^{T \times S}$  as

$$V = \begin{bmatrix} v & \dots & v \end{bmatrix}$$

i.e.  $V$  consists of  $S$  repetitions of  $v$ . We construct a representation matrix  $\mathcal{A}$  with  $\text{supp}(\mathcal{A}) \subset \Gamma$  as follows

$$\mathcal{A}_{\Gamma,:} = \begin{bmatrix} V_{1:K-P+1,:} \\ \hline & I_{P-1} & 0 \\ \hline & 0_{K-P+1 \times S} \end{bmatrix}$$

The rows in  $\mathcal{A}$  not indexed by  $\Gamma$  are zero. The rows in  $\mathcal{A}$  indexed by  $\Gamma$  with  $|\Gamma| = T = 2K - P + 1$  are constructed as in the equation above. First  $K - P + 1$  rows are picked directly from  $V$ . The next  $P - 1$  rows consist of an identity matrix of size  $P - 1$  followed by all 0s. The next  $K - P + 1$  rows are all zeros. Thus, total rows are  $K - P + 1 + P - 1 + K - P + 1 = 2K - P + 1 = T$ . Note that since  $K \geq P$ , hence  $K - P + 1 > 0$ .

First  $K - P + 1$  rows of  $\mathcal{A}$  may be non-zero. The next  $P - 1$  rows are definitely non-zero. All other rows are zero. Thus a maximum of  $K$  rows in  $\mathcal{A}$  are non-zero. Hence  $\|\mathcal{A}\|_{\text{row}-0} \leq K$ .

Now construct another matrix  $\mathcal{B}$  as follows. Keep  $\text{supp}(\mathcal{B}) \subset \Gamma$ . Further, construct

$$\mathcal{B}_{\Gamma,:} = X_{\Gamma,:} - V.$$

By construction  $\mathcal{B}$  is also  $K$  row sparse.

Now

$$\mathcal{D}\mathcal{B} = \mathcal{D}_{\Gamma}\underline{\mathcal{B}}_{\Gamma} = \mathcal{D}_{\Gamma}(\underline{\mathcal{A}}_{\Gamma} - V) = \mathcal{D}_{\Gamma}\underline{\mathcal{A}}_{\Gamma} = \mathcal{D}\mathcal{A}.$$

Thus both  $\mathcal{B}$  and  $\mathcal{A}$  are  $K$ -row sparse solutions to the problem  $X = \mathcal{D}\mathcal{A}$ . This means that (13.1.22) is a necessary condition for uniqueness.  $\square$

Let us rewrite the previous two results for the CS setting also.

**Theorem 13.8** [19] *The sufficient condition of (13.1.21) in theorem 13.5 is equivalent to*

$$\|X\|_{row-0} < \frac{1}{2} [\text{spark}(\Phi) - 1 + \text{rank}(X)]. \quad (13.1.23)$$

**Theorem 13.9** [19] *The condition (13.1.23)*

$$\|X\|_{row-0} < \frac{1}{2} [\text{spark}(\Phi) - 1 + \text{rank}(X)]$$

*or equivalently the condition (13.1.21)*

$$\|X\|_{row-0} < \frac{1}{2} [\text{spark}(\Phi) - 1 + \text{rank}(Y)]$$

*is both necessary and sufficient condition for  $\mathcal{A}$  to be the unique solution of (Joint-CS<sub>0</sub>) problem.*

The results above validate the relevance of rank of the representation matrix or the signal matrix or the measurement matrix in the joint sparse recovery problem. Our next goal is to develop practical algorithms which can leverage the rank information and successfully perform joint recovery with lesser number of measurements or higher levels of sparsity.

The first algorithm on this journey will be based on the MUSIC principle.

## 13.2. The MUSIC principle

Before proceeding further, we take a slight detour to an interesting algorithm from signal processing literature.

### 13.2.1. The MUSIC principle

The inverse of the norm of the projection of a vector in the essential range of a Hermitian operator on to its noise subspace is infinite.

MUSIC [32] stands for MUltiple SIgnal Classification.

Let  $A \in \mathbb{C}^{n \times n}$  be a Hermitian matrix with eigen values  $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$  and corresponding (unit norm) eigen vectors  $v_1, v_2, \dots, v_n$ . Let the number of non-zero eigen values be  $m$ . Then the eigen values  $\lambda_{m+1}, \dots, \lambda_n$  are all zero and the eigen vectors  $v_{m+1}, \dots, v_n$  span the null space of  $A$ .

In typical situations, the first  $m$  eigen values would be significant while the rest  $n - m$  would be very small. In this case, we say that the eigen vectors  $v_{m+1}, \dots, v_n$  span the **noise subspace** of  $A$ .

We can now form the projection operator for the noise subspace of  $A$  as

$$P_{\text{noise}} = \sum_{j>m} v_j v_j^H. \quad (13.2.1)$$

We say that the **essential range** of  $A$  is spanned by the vectors  $v_1, \dots, v_m$ . Since  $A$  is Hermitian, hence the noise subspace is orthogonal to the essential range. Therefore, a vector  $x$  is in the essential range if and only if its projection to the noise subspace is zero, i.e. if

$$\|P_{\text{noise}}x\|_2 = 0. \quad (13.2.2)$$

This is equivalent to writing

$$\frac{1}{\|P_{\text{noise}}x\|_2} = \infty. \quad (13.2.3)$$

The equation (13.2.3) can be treated as a characterization of the essential range of  $A$ . This characterization is known as the MUSIC principle.

### 13.2.2. Applications in signal processing

Here is a simple application of the MUSIC principle for harmonic frequency detection.

Consider a signal which is a superposition of two time-harmonic signals of different frequency with noise.

$$x_n = a_1 e^{j\omega_1 n} + a_2 e^{j\omega_2 n} + w_n. \quad (13.2.4)$$

Assume that the random variables  $w_n$  are i.i.d.. The amplitudes  $a_1$  and  $a_2$  are also random variables and independent. The frequencies  $\omega_1$  and  $\omega_2$  are fixed but unknown. Our goal is to find these frequencies.

We can form the correlation matrix:

$$A_{n,m} = \mathbb{E}(x_n \overline{x_m}) \quad (13.2.5)$$

Assume that we have  $N$  samples of  $x_n$ . Since the different terms of  $x_n$  are independent, we can write

$$A = \mathbb{E}(|a_1|^2)s^1 s^{1H} + \mathbb{E}(|a_2|^2)s^2 s^{2H} + \sigma_0^2 I_N \quad (13.2.6)$$

where the  $n$ -th component of the vector  $s^i$  is given by

$$s_n^i = e^{j\omega_i n}$$

and  $\sigma_0^2 = \mathbb{E}(|\omega_n|^2)$  is the noise variance.

The decomposition of  $A$  can be looked in two parts. The first part covers the essential range of  $A$  spanned by the vectors  $s^1$  and  $s^2$ . The second part is the noise subspace with a very small eigen value corresponding to the noise variance.

The correlation matrix  $A$  can be estimated accurately by taking multiple snapshots of  $N$ -sample vectors of  $x_n$ . Once the correlation matrix has been obtained, its eigen value decomposition can be carried out. After that, we can keep the first two eigen vectors (corresponding to the largest eigen values) as part of the essential range and take rest of the eigen vectors as our noise subspace. From them, we can compute the projection operator for the noise subspace  $P_{\text{noise}}$ .

In order to find the actual frequencies  $\omega_1$  and  $\omega_2$ , we plot the following as a function of  $\omega$

$$\frac{1}{\|P_{\text{noise}} s^\omega\|_2} \quad (13.2.7)$$

where  $s^\omega$  is an  $N$ -length vector whose  $n$ -th entry is given by  $e^{j\omega n}$ .

The resulting plot is expected to have two peaks at the frequencies  $\omega_1$  and  $\omega_2$ . This plot is known as the **MUSIC pseudospectrum**.

We note that for this algorithm to work, we need a) large number of samples  $x_n$  and b) large number of frequencies  $\omega$  over which the pseudospectrum is evaluated.

### 13.3. MUSIC based joint recovery

We now explore how we can use the MUSIC principle in joint recovery problem [19].

The MUSIC principle can be directly applied when the rank of  $X$  in  $X = \mathcal{D}\mathcal{A}$  or rank of  $Y$  in  $Y = \Phi X$  equals  $K$  which is the row sparsity level of the representation matrix  $\mathcal{A}$  in the sparse approximation setting or the row sparsity level of the signal matrix  $X$  in CS setting.

In rest of the section, we will work with the CS setting.

Recall that from the relation  $Y = \Phi X$ , we have  $\text{rank}(Y) \leq \text{rank}(X) \leq K$ . The last inequality is due to the fact that  $X$  may have less than  $K$  non-zero rows, and even then we call it  $K$ -row sparse.

Now  $\text{rank}(Y) = K$  implies that

$$\text{rank}(Y) = \text{rank}(X) = K. \quad (13.3.1)$$

With  $\Lambda = \text{supp}(X)$ , we have

$$Y = \Phi_\Lambda \underline{X}_\Lambda \quad (13.3.2)$$

with  $Y \in \mathbb{C}^{M \times S}$ ,  $\Phi_\Lambda \in \mathbb{C}^{M \times K}$  and  $\underline{X}_\Lambda \in \mathbb{C}^{K \times S}$ .

Since  $\text{rank}(\Phi_\Lambda) = K$  too, the column spaces of  $Y$  and  $\Phi_\Lambda$  are same i.e.  $\mathcal{C}(Y) = \mathcal{C}(\Phi_\Lambda)$ . This means that we have complete visibility into the column space of  $\Phi_\Lambda$  through the column space of  $Y$ . In particular, every  $\phi_i$  lies in the column space of  $Y$ .

Assuming that (13.1.23) is satisfied, we know that  $X$  is a unique sparse solution. This means that any  $\phi_i$  with  $i \notin \Lambda$  (i.e. outside  $\Phi_\Lambda$ ) will not

lie entirely in the column space of  $Y$ . If it did, we could easily construct an alternative  $K$ -row sparse solution.

We are now ready to apply the MUSIC principle. Consider the orthogonal complement of the column space of  $Y$ . Clearly, the vectors  $\phi_i$  with  $i \in \Lambda$  fall into its null space and  $\phi_i$  with  $i \notin \Lambda$  have a non-zero projection in this space.

We construct a basis  $U$  for the column space of  $Y$  by orthogonalization  $U = \text{orth}(Y)$ . We then construct the projection operator for the orthogonal complement of the column space of  $Y$  as

$$P = I - UU^H.$$

Based on the discussion above we have the result:

$$\|(I - UU^H)\phi_i\|_2 = 0, \text{ if and only if } i \in \Lambda. \quad (13.3.3)$$

Therefore, if we select  $K$  atoms from  $\Phi$  which minimize  $\|(I - UU^H)\phi_i\|_2$  or alternatively maximize

$$\frac{1}{\|(I - UU^H)\phi_i\|_2}$$

(the MUSIC pseudo-spectrum) we have identified our submatrix  $\Phi_\Lambda$ . After this, the recovery process is a simple least square step given by

$$\underline{X}_\Lambda = \Phi_\Lambda^\dagger Y. \quad (13.3.4)$$

Let us summarize the algorithm

- We construct an orthonormal basis  $U$  for the column space of  $Y$ .
- We construct the projection operator for the orthogonal complete of  $\mathcal{C}(Y)$  as  $I - UU^H$ .
- We select  $K$  atoms from  $\Phi$  whose projection has minimum norm.
- We construct  $\Phi_\Lambda$  and then compute  $X$  using least squares.

We note that this procedure is not iterative at all. In one iteration, we are able to identify the whole of support. Rest is plain old least squares.

With Gaussian sensing matrices having  $\text{spark}(\Phi) = M + 1$  with probability 1, MUSIC algorithm can fully recover  $K$ -sparse signals jointly with as less as  $M = K + 1$  measurements. Moreover this works for any  $X$  as long as  $\text{rank}(X) = K$ . A single SVD of  $Y$  is enough to compute the orthonormal basis  $U$ . Computationally also the algorithm is much simpler than the traditional SMV recovery algorithms.

It is also quite easy to identify, whether we can use MUSIC or not. For this, we simply need to find the rank of  $Y$ . If it equals  $K$ , we can use it. This can be determined as we do the SVD of  $Y$  or we can use any other means for finding the rank.

The only problem is that the MUSIC principle breaks when we have  $\text{rank}(Y) < K$ . We need to develop some other rank aware algorithms for this case. But before we go into that, let us establish that the traditional joint recovery algorithms like S-OMP (OMPMMV), mixed-norm minimization (BP style), thresholding etc. are rank-blind.

### 13.4. Rank blindness in joint recovery algorithms

The typical methods used for joint recovery are S-OMP (OMPMMV),  $p$ -thresholding, and mixed  $l_{p,q}$  norm minimization. They are straightforward generalizations of corresponding SMV algorithms: OMP, thresholding and BP.

In this section, we show that these algorithms are **effectively rank blind**. By being rank blind we mean that

- The algorithms do not allow for perfect recovery in the full rank case.
- The worst case behavior of such algorithms approaches that of the corresponding SMV problem.

### 13.4.1. Greedy methods

Two popular greedy methods are thresholding [25] and SOMP [40].

The thresholding algorithm can be written in following steps. We first compute

$$h = \|\Phi^H Y\|_{\text{row}-q}$$

where  $\text{row} - q$  stands for taking the  $l_q$  norm of each row of the matrix  $\Phi^H Y$ . Thus  $h \in \mathbb{R}^N$ .

The second step is

$$\Lambda = \text{supp}(h|_K)$$

i.e. the indices of  $K$  largest entries in  $h$  are considered as part of the support for  $X$ .

Once the support has been identified, the recovery is done by least squares:

$$\underline{X}_\Lambda = \Phi_\Lambda^\dagger Y. \quad (13.4.1)$$

The S-OMP algorithm has also been previously discussed.

We recall that Chen and Huo [13] showed that  $\text{ERC}(\Phi) < 1$  is a sufficient condition for the success of S-OMP in the worst case. The condition is identical to the OMP recovery guarantee. For the SMV case, Tropp [34] had shown that ERC is also a necessary condition for recovery guarantee. We show next that the necessary condition for recovery using SOMP also approaches ERC condition and this condition is independent of the rank of  $X$ . This implies that the S-OMP algorithm is not able to exploit rank information in the worst case.

**Theorem 13.10** [19] (*S-OMP is not rank aware*) Let  $k$  be such that  $1 \leq k \leq K$ . Let  $\Lambda$  be an index set with  $K = |\Lambda|$  such that

$$\max_{j \notin \Lambda} \|\Phi_\Lambda^\dagger \phi_j\|_2 > 1. \quad (13.4.2)$$

Then, there exists an  $X$  with  $\text{supp}(X) = \Lambda$  and  $\text{rank}(X) = k$  that cannot be recovered by S-OMP from  $Y = \Phi X$ .

(13.4.2) says that ERC condition is not satisfied for the subdicitonary  $\Phi_\Lambda$ . Since,  $k$  varies between 1 to  $K$ , we are saying that SOMP will fail for every rank in the worst case.

**PROOF.** Due to Tropp [34, theorem 3.10], since ERC is not satisfied, there exists a  $K$ -sparse vector  $x$  for the SMV problem for which OMP will fail. Let  $x$  be such a vector with  $\text{supp}(x) = \Gamma$  and  $y = \Phi x$  such that OMP incorrectly selects atom  $j^* \notin \Lambda$  at the first step with

$$|\phi_{j^*}^H y| > \max_{i \in \Lambda} |\phi_i^H y| + \epsilon \quad (13.4.3)$$

for some  $\epsilon > 0$ . In words, the inner product of  $\phi_{j^*}$  with the measurement vector  $y$  is larger than the inner product of  $y$  with all the atoms in  $\Phi_\Lambda$  and the minimum difference is a positive number  $\epsilon$ .

We now construct a rank-1 matrix  $X \in \mathbb{C}^{N \times S}$  as

$$X \triangleq \begin{bmatrix} x & x & \dots & x \end{bmatrix}.$$

It is easy to see that this matrix cannot be recovered by S-OMP from  $Y = \Phi X$  for any choice of  $l_q$  norm is the matching step in S-OMP.

The next step is to show that there exist other matrices with rank between 1 and  $K$  which also cannot be recovered by S-OMP. For this, we will introduce a slight perturbation in  $X$  as follows.

Define

$$\tilde{X} = X + E$$

where  $\text{supp}(E) \subseteq \Lambda$  and  $\max_j \|\phi_j^H \Phi E\|_q \leq S^{\frac{1}{q}} \frac{\epsilon}{2}$  such that  $\tilde{X}$  has rank  $k$  with  $1 \leq k \leq K$ .

Further, define  $\tilde{Y} = \Phi\tilde{X}$ . We now show that S-OMP will pick  $j^*$  as the first atom for this example.

$$\begin{aligned}
\|\phi_{j^*}^H \tilde{Y}\|_q &= \|\phi_{j^*}^H(\Phi X + \Phi E)\|_q \\
&\geq \|\phi_{j^*}^H \Phi X\|_q - \|\phi_{j^*}^H \Phi E\|_q \\
&\geq S^{\frac{1}{q}} |\phi_{j^*}^H \Phi x| - S^{\frac{1}{q}} \frac{\epsilon}{2} \\
&> S^{\frac{1}{q}} \max_{i \in \Lambda} |\phi_i^H \Phi x| + S^{\frac{1}{q}} \frac{\epsilon}{2} \\
&= \max_{i \in \Lambda} \|\phi_i^H \Phi X\|_q + S^{\frac{1}{q}} \frac{\epsilon}{2} \\
&\geq \max_{i \in \Lambda} \{ \|\phi_i^H \Phi X\|_q + \|\phi_i^H \Phi E\|_q \} \\
&\geq \max_{i \in \Lambda} \|\phi_i^H \tilde{Y}\|_q.
\end{aligned} \tag{13.4.4}$$

The steps use the triangle inequality and basic manipulations. Due to this, in the very first iteration itself, S-OMP will choose an incorrect atom. Thus,  $\tilde{X}$  cannot be recovered.  $\square$

### 13.4.2. Mixed $l_q, l_1$ minimization

Another approach based on the generalization of BP is

$$\hat{X} = \arg \min_X \|X\|_{q,1} \quad s.t. \Phi X = Y. \tag{13.4.5}$$

A similar approach can be used to show that this algorithm is also rank-blind.

We recall that for the corresponding SMV program which is basis pursuit or  $l_1$  minimization, the necessary and sufficient condition for the recovery of vectors  $x$  with support  $\Gamma$  is given by the **Null Space Property**

$$\|z_\Gamma\|_1 < \|z_{\Gamma^c}\|_1 \quad \forall z \in \mathcal{N}(\Phi) \tag{13.4.6}$$

**Theorem 13.11** [19] (*The  $l_q, l_1$  minimization is not rank aware.*)  
*Let  $k$  be such that  $1 \leq k \leq K$ . Suppose that there exists  $z \in \mathcal{N}(\Phi)$  such that*

$$\|z_\Lambda\|_1 > \|z_{\Lambda^c}\|_1 \tag{13.4.7}$$

for some support  $\Lambda$ ,  $|\Lambda| = K$ . Then there exists an  $X$  with  $\text{supp}(X) = \Lambda$ ,  $\text{rank}(X) = k$  such that  $l_q, l_1$  minimization program cannot recover  $X$  from  $Y = \Phi X$ .

PROOF. Skipped. □

We have established so far that the traditional joint recovery algorithms are not rank aware. It is now time to explore some rank aware algorithms.

### 13.5. Rank aware algorithms

#### 13.5.1. Rank aware thresholding

In rank blind thresholding [19], we compute

$$h = \|\Phi^H Y\|_{\text{row}-q}$$

The change that we will make here is, we will replace  $Y$  with  $U = \text{orth}(Y)$ . The rest of the steps essentially remain same.

- (1) Compute  $h = \|\Phi^H U\|_{\text{row}-2}$
- (2) Find  $\Lambda = \text{supp}(h|_K)$
- (3) Compute  $\underline{X}_\Lambda = \Phi_\Lambda^\dagger Y$ .
- (4) Fill rest of the rows with zeros.

When  $K$  is unknown, we can choose  $K$  either by applying a fixed cut-off  $\theta$  to the larger entries in  $h$ .

There is one particular difference here. While standard  $q$ -thresholding can work with a variety of choices for  $q$ , only  $l_2$  norm is suitable for rank aware thresholding. This is due to the fact that only  $l_2$  norm is invariant to the (arbitrary) choice of the orthonormal basis,  $U$  for the column space of  $Y$ .

**Theorem 13.12** [19] Let  $Y = \Phi X$  with  $|\text{supp}(X)| = K$ ,  $\text{rank}(X) = K$  and  $K < \text{spark}(\Phi) - 1$ . Then rank aware thresholding is guaranteed to recover  $X$ .

PROOF. Skipped. □

Clearly,  $M = K + 1$  measurements are sufficient to recover  $X$ , as long as  $\Phi$  has full spark.

**Challenge** Is a RIP based analysis of rank aware thresholding possible?

### 13.5.2. Rank aware S-OMP

The S-OMP algorithm can be easily made rank aware by making some changes in the matching and selection of next candidate index.

We recall that in standard S-OMP

$$\lambda^{k+1} = \arg \max_{i \notin \Lambda^k} \|\phi_i^H R^k\|_q$$

and

$$\Lambda^{k+1} = \Lambda^k \cup \{\lambda^{k+1}\}$$

where  $R^k$  is the residual matrix at the end of  $k$ -th iteration.

The essential change we are going to make is how  $\lambda^{k+1}$  is chosen by replacing  $R^k$  with  $U^k = \text{orth}(R^k)$ .

$$\lambda^{k+1} = \arg \max_{i \notin \Lambda^k} \|\phi_i^H U^k\|_2$$

Since  $U$  is an (arbitrary) orthonormal basis, hence the only suitable choice of norm is again  $q = 2$ .

We recall the idea of **greedy selection ratio** which is an essential criterion for deciding whether a pursuit algorithm is proceeding correctly

or not. In our rank aware S-OMP, the ratio takes the form:

$$\rho = \frac{\max_{j \notin \Omega} \|\phi_j^H U\|_2}{\max_{j \in \Omega} \|\phi_j^H U\|_2} < 1. \quad (13.5.1)$$

**Challenge** Can we perform a RIP analysis of rank aware S-OMP?

**Challenge** One of the operations in Joint-CoSaMP is thresholding for choosing the  $2K$  indices in the matching step and  $K$  indices later in the least squares step. The Joint-CoSaMP currently defined uses a rank-blind thresholding scheme. What happens if we replace the rank-blind thresholding scheme with rank-aware one? Would that help in improving the performance of Joint-CoSaMP?

Why this? CoSaMP was developed in the first place since OMP couldn't provide guarantees for noisy recovery. The rank aware thresholding and OMP are also only for noiseless recovery. A rank-aware joint-CoSaMP might provide performance guarantees for noisy recovery also.

### 13.6. $l_1$ norm minimization

**Challenge** Can we do rank aware analysis of BPIC for compressed sensing in joint recovery setting?

**Challenge** Can we do rank aware analysis of **cosamp!** (**cosamp!**) for compressed sensing in joint recovery setting?

### 13.7. Digest

## CHAPTER 14

# Dictionary Learning

### 14.1. Introduction

When designing a dictionary for a particular application, we have several options [30]. On the one hand we can go through the long list of analytically constructed or tunable dictionaries and select one of them as suitable for the application in concern. On the other hand, we can actually take up a number of real example signals from our application and try to construct a dictionary which is optimized for these. **Dictionary learning** (DL) [33, 21] is a process which attempts to solve the problem of constructing a dictionary directly from the set of example signals. The atoms of a learnt dictionary come from the underlying empirical data of training set of example signals for the specific application. While analytically constructed dictionaries are typically meant for only specific applications, the learning method allows one to construct dictionaries for any family of signals which are amenable to the sparse and redundant representations model. This certainly comes at a cost. Learnt dictionaries have to be held completely in memory explicitly as they happen to be usually structure less. Thus they don't provide an efficient implementation of analysis ( $\mathcal{D}^H x$ ) and synthesis ( $\mathcal{D}\alpha$ ) operators. Thus using them in applications leads to more computational costs.

We start by formalizing the notation for DL. We consider a set of  $S$  example signals put together in a signal matrix  $X \in \mathbb{C}^{N \times S}$ . Consider a dictionary  $\mathcal{D} \in \mathbb{C}^{N \times D}$ . Let  $\alpha^i \in \mathbb{C}^D$  be the sparsest possible representation of  $x^i$  in  $\mathcal{D}$  with  $x^i = \mathcal{D}\alpha^i + e^i$ . We put all  $\alpha^i$  together in a matrix  $\mathcal{A} \in \mathbb{C}^{D \times S}$ . Then we have  $X = \mathcal{D}\mathcal{A} + E$  where  $E \in \mathbb{C}^{N \times S}$  represents

approximation error. We are looking for best dictionary from the set of possible dictionaries such that we are able to get sparse representations of  $x^i$  with low approximation error. We can quantify the notion of good approximation by putting an upper bound on the norm of approximation error as  $\|e^i\|_2 \leq \epsilon$ . Combining these ideas, we introduce the notion of a **sparse signal model** denoted as  $\mathcal{M}_{\mathcal{D}, K, \epsilon}$  which consists of a dictionary  $\mathcal{D}$  providing  $K$ -sparse representations for a class of signals (from which the example signals are drawn) with an upper bound on approximation error given by  $\epsilon$ . The DL problem essentially tries to learn best model  $\mathcal{M}$  based on the example signals  $X$ .

We can formulate the DL problem as an optimization problem as follows:

$$\begin{aligned} & \underset{\mathcal{D}, \{\alpha^i\}_{i=1}^S}{\text{minimize}} \quad \sum_{i=1}^S \|\alpha^i\|_0 \\ & \text{subject to} \quad \|x^i - \mathcal{D}\alpha^i\|_2 \leq \epsilon, \quad i = 1, \dots, S. \end{aligned} \tag{14.1.1}$$

In this version, we are trying to minimize total sparsity of  $\alpha^i$  while using the upper bound on approximation error as optimization constraint. We are not enforcing sparsity constraint that  $\|\alpha^i\|_0 \leq K \forall 1 \leq i \leq S$ . Alternatively we can also write:

$$\begin{aligned} & \underset{\mathcal{D}, \{\alpha^i\}_{i=1}^S}{\text{minimize}} \quad \sum_{i=1}^S \|x^i - \mathcal{D}\alpha^i\|_2^2 \\ & \text{subject to} \quad \|\alpha^i\|_0 \leq K, \quad i = 1, \dots, S. \end{aligned} \tag{14.1.2}$$

In this version we are trying to minimize approximation error while keeping  $K$ -sparsity as optimization constraint. We are not enforcing the constraint that  $\|e^i\|_2 \leq \epsilon$ .

Alas, there doesn't exist any computationally tractable algorithm to solve these optimization problems. An alternative is to consider a heuristic iterative approach presented in fig. 14.1.

Some possible halting criteria are:

- Stop after a fixed number of iterations

```

Choose an initial dictionary  $\mathcal{D}^0$  ;
Initialize iteration counter  $k = 1$  ;
repeat
    Obtain representations  $\alpha^i$  for  $x^i$  with the given  $\mathcal{D}^{k-1}$  and sparsity level
     $K$  using some sparse recovery algorithm ;
    Obtain new dictionary  $\mathcal{D}^k$  from the signals  $x^i$  and their representations
     $\alpha^i$  ;
until halting criteria is true;

```

FIGURE 14.1. Dictionary learning: iterative approach

- Stop when  $\|e^i\|_2 \leq \epsilon$  for all examples
- Stop when no further progress in  $\|e^i\|_2$ .

For initialization of the algorithm, we have few options.

- We can start with a randomly generated dictionary.
- We can select  $D$  examples from  $X$  and put them together as our starting.
- Or we can start with some analytical dictionary suitable for the application domain. dictionary

Two popular algorithms implementing this approach are MOD (Method of Optimal Directions) and K-SVD.

In MOD [23], dictionary update is formulated as

$$\mathcal{D}^k = \operatorname{argmin}_{\mathcal{D}} \|X - \mathcal{D}\mathcal{A}^k\|_F^2 \quad (14.1.3)$$

A straight forward least squares solution is obtained as

$$\mathcal{D}^k = X(\mathcal{A}^k)^\dagger. \quad (14.1.4)$$

K-SVD [1] is slightly different. Rather than recomputing whole dictionary at once by solving the LS problem, it updates atoms of  $\mathcal{D}$  one by one. Let  $j$  be the index of atom  $d_j$  being updated. Consider the error

$$E_j = X - \sum_{k \neq j} d_k \beta_k^T$$

$\beta_k$  refers to the  $k$ -th row of  $\mathcal{A}$  i.e. entries for atom  $d_k$  in all examples. Essentially  $E_j$  means the approximation error when the atom  $d_j$  (and corresponding entries in  $\mathcal{A}$ ) has been dropped from consideration. We can see that

$$\|X - \mathcal{D}\mathcal{A}\|_F^2 = \|E_j - d_j\beta_j^T\|_F^2.$$

On the L.H.S. we have total approximation error. On the R.H.S. we have the same expressed in terms of  $E_j$  and  $d_j$  where  $E_j$  doesn't depend on  $d_j$ . Clearly an optimal  $d_j$  is one which can minimize R.H.S. This can easily be obtained by rank-1 approximation of  $E_j$ . The rank-1 approximation gives us both  $d_j$  as well as the entries in  $j$ -th row of  $\mathcal{A}$  given by  $\beta_j$ . There is a small catch though. In general the rank-1 approximation can lead to a dense  $\beta_j$  meaning the atom  $d_j$  gets used in many signal representations. We wish to avoid that. We don't want  $d_j$  to appear in more signal representations. For this, we identify representations  $\alpha^i$  in which atom  $d_j$  appears, and let them be indexed by  $\Gamma$ . We then restrict  $E_j$  to signals indexed by  $\Gamma$  to avoid a dense  $\beta_j$ . Rank-1 approximation can be easily obtained using singular value decomposition. We perform SVD of  $E_{j,\Gamma} = U\Sigma V^H$ . We then pick  $u_1, \sigma_1, v_1$ .  $u_1$  is the new update for  $d_j$ . Further  $\sigma_1 v_1$  is the update for  $\beta_{j,\Gamma}$ . Rest of  $\beta_j$  is left with zero entries. We repeat the process for each of the atoms in  $\mathcal{D}$  to obtain next update of  $\mathcal{D}$ .

## 14.2. Unique dictionary and matrix factorization

Although, most of the methods for dictionary learning are heuristic in nature, there exist some results which provide theoretical guarantees of uniqueness of existence of an overcomplete dictionaries for a given set of training data under certain conditions.

In this section, we study one such result [2]. The dictionary learning process would be modeled as a matrix factorization problem in the following.

### 14.2.1. Assumptions

We will make several assumptions in advance. The first one is:

- We assume that all signals  $x^i$  have a sparse representation in the (unknown) dictionary  $\mathcal{D}$ . Then  $X = \mathcal{D}\mathcal{A}$ . There is no sparse approximation error.

With this assumption, the problem becomes a matrix factorization problem where we see the factorization of  $X$  as

$$X = \mathcal{D}\mathcal{A}$$

In order to simplify our life for the theoretical analysis to follow, we will make many more assumptions

**Support:** Let  $\sigma$  denote the spark of  $\mathcal{D}$  i.e.  $\sigma = \text{spark}(\mathcal{D})$ . We assume that all (unknown) sparse representations satisfy

$$\|\alpha^i\|_0 < \frac{\sigma}{2}.$$

This ensures that the representations  $\alpha^i$  are the unique sparsest representations of  $x^i$  in  $\mathcal{D}$ . Of course both  $\mathcal{D}$  and  $\mathcal{A}$  are unknown.

**Richness:** We know that there are  $\binom{D}{K}$  possible choices of  $K$  atoms out of the (unknown) dictionary  $\mathcal{D}$ . Each of these set of  $K$  elements span a subspace of  $\mathbb{C}^N$ . We will assume that for each such subspace, there are at least  $(K + 1)$  signals in  $X$ . Thus, total number of signals is at least  $(K + 1)\binom{D}{K}$ .

**Non-degeneracy:** Given a group of  $K + 1$  signals built using a particular set of  $K$  atoms, in general their rank (of the matrix composed by putting them together) is expected to be  $K$  or less. We will assume that the rank is exactly  $K$  and is not less. This helps in ensuring that signals from one subspace may not be confused with signals from other subspace. We further assume that if we take  $K + 1$  signals belonging to different (unknown yet) subspaces, then their rank would be

exactly  $K + 1$ . Thus, we are saying that no-degeneracies in the construction of the signals is allowed.

The non-degeneracy assumption helps ensure following. Every representation has exactly  $K$  non-zero entries. Any set of  $K$  representations (of  $K$  signals) belonging to a  $K$ -subspace is necessarily linearly independent. Thus, there are no cases like duplicate signals floating around.

With these assumptions, we will develop a constructive though extremely inefficient procedure for carrying out the matrix factorization and identify our dictionary  $X = \mathcal{D}\mathcal{A}$ .

#### 14.2.2. Equivalent dictionaries

Before moving over to the main result, let us mention certain transformations creating equivalent dictionaries (i.e. the sparsity of representations is not affected).

Suppose we exchange  $p$ -th atom and  $q$ -th atom in  $\mathcal{D}$  to construct a new dictionary  $\mathcal{D}'$ . It is easy to see that if we also exchange  $p$ -th row and  $q$ -th row in  $\mathcal{A}$  to construct a new representation matrix  $\mathcal{A}'$ , then

$$X = \mathcal{D}\mathcal{A} = \mathcal{D}'\mathcal{A}'.$$

Similarly, if we change the sign of  $p$ -th atom in  $\mathcal{D}$  (i.e. replace  $d_p$  with  $-d_p$ ) to construct a new dictionary  $\mathcal{D}'$ , then changing the sign of  $p$ -th row in  $\mathcal{A}$  to construct  $\mathcal{A}'$  provides the new factorization. i.e.:

$$X = \mathcal{D}\mathcal{A} = \mathcal{D}'\mathcal{A}'.$$

Thus, these dictionaries and corresponding sparse representations are equivalent. This transformation can be easily captured using a signed permutation matrix. A signed permutation matrix is a matrix where each row and each column of the matrix has exactly one non-zero entry and that entry has exactly the value of 1 or  $-1$ . Then

$$\mathcal{D}' = \mathcal{D}P.$$

Now, since  $PP^T = I$ , hence

$$\mathcal{A}' = P^T \mathcal{A}.$$

We can now clearly see that

$$X = \mathcal{D}\mathcal{A} = \mathcal{D}PP^T\mathcal{A} = \mathcal{D}'\mathcal{A}'.$$

Note that the desired properties of  $\mathcal{D}$  (atoms being unit norm and spanning the whole of  $\mathbb{C}^N$ , representations being sparse and unique) do not change through these transformations.

So, if a matrix factorization process can find out any  $\mathcal{D}'$  such that  $\mathcal{D}' = \mathcal{D}P$  for some signed permutation matrix  $P$ , then we have achieved our factorization.

### 14.2.3. Uniqueness result

We have the following main result for existence of a matrix factorization of  $X$ .

**Theorem 14.1** *Under the assumptions stated above, the factorization of  $X$  is unique, i.e. the factorization  $X = \mathcal{D}\mathcal{A}$  for which (i)  $\mathcal{D} \in \mathbb{C}^{N \times D}$  with normalized columns; and (ii)  $\mathcal{A} \in \mathbb{C}^{D \times S}$  with  $K$  non-zeros in each column, is unique. This uniqueness is up to a right-multiplication of  $\mathcal{D}$  by a signed permutation matrix, which does not change the desired properties of  $\mathcal{D}$  and  $\mathcal{A}$ .*

**PROOF.** We will present a constructive procedure which constructs a dictionary  $\mathcal{D}'$  and a representation matrix  $\mathcal{A}'$  from the given signal matrix  $X$  subject to  $\mathcal{D}' = \mathcal{D}P$  where  $P$  is a signed permutation matrix.

We note that since all representations are  $K$ -sparse in  $\mathcal{D}$  and  $K < \frac{\sigma}{2}$ , hence once the dictionary  $\mathcal{D}$  (or its permutation  $\mathcal{D}' = \mathcal{D}P$ ) has been found, the corresponding representations  $\mathcal{A}$  (or  $\mathcal{A}'$ ) can be easily found by solving the sparse signal recovery problem.

Our process for recovery of  $\mathcal{D}' = \mathcal{D}P$  from  $X$  will consist of following stages.

**Clustering:** We will divide columns (signals)  $X$  into  $R = \binom{D}{K}$  sets

$$\{G_1, G_2, \dots, G_R\}$$

where each set includes all the signals that share the same support (i.e. they use the same atoms from  $\mathcal{D}$ ).

We will identify the index set of support for  $G_j$  with the symbol  $\Lambda_j \subset \Omega$  where  $\Omega = \{1, 2, \dots, D\}$  is the index set for all atoms.

**Detecting pairs:** We will detect pairs of sets  $G_i$  and  $G_j$  that share exactly one mutual atom. In other words  $\Lambda_i \cap \Lambda_j = \{k\}$  where  $k$  is the index of the only atom shared between the supports of  $G_i$  and  $G_j$ .

**Extracting atoms:** We will extract the mutual atom by analysis of  $G_i$  and  $G_j$ . By analyzing all the pairs  $G_i$  and  $G_j$  we will be able to identify all atoms in  $\mathcal{D}$  subject to a permutation and sign change. This way we will form the complete dictionary  $\mathcal{D}'$ .

**Stage 1: Clustering the signals.** As per our assumptions, any group of  $K$  signals in  $X$  is linearly independent. And any such group spans a  $K$  dimensional subspace of  $\mathbb{C}^N$ .

In this stage our objective is to identify the specific  $R = \binom{D}{K}$   $K$ -dimensional subspaces which can cover the whole set of  $S$  signals. And we will divide the  $S$  signals in  $X$  into groups based on their embedding subspace.

Note that if we randomly select  $K$  signals from  $X$ , then the  $K$ -subspace spanned by them need not be one of the subspaces in  $R$   $K$ -dimensional subspaces covered by atoms in the dictionary.

Consider any set of  $K + 1$  signals from  $X$ . As per our non-degeneracy assumption, if they belong to same subspace, the rank would be  $K$ , otherwise the rank would be  $K + 1$ . There are  $\binom{S}{K+1}$  possible groups of  $K + 1$  signals in  $X$ . We iterate through each such group. If the

rank of the group is  $K$ , then we keep the group, otherwise we discard the group. The moment we find one group with rank  $K$ , we exclude all signals in it from further identification of groups. Thus, after the identification of first group, we are left with  $S - K - 1$  signals from which we now form our  $K + 1$  size groups. Proceeding this way, we would be able to find all the  $\binom{D}{K}$  groups with non-overlapping sets of  $K + 1$  signals. Since it is given that  $S \geq (K+1)\binom{D}{K}$ , at least one group corresponding to each  $K$ -subspace would indeed be found.

If there are more than  $\binom{D}{K}$  groups, then we need to merge groups coming from same subspace. We can do this by iterating through all possible pairs of groups and seeing if their combined rank is  $K$ . If yes, then we merge the groups. Otherwise, we keep them separate.

Finally, we look at all the remaining signals in  $X$ . For each such signal, we see its combined rank with each of the  $R = \binom{D}{K}$  groups identified above. The signal will belong to the subspace spanned by one of the groups. We will merge our signal to that group.

We note that this clustering process is indeed impractical.

**Stage 2: Detecting pairs with mutual atom.** Given the  $R$  sets of signals  $\{G_j\}_{j=1}^R$ , there are  $R(R-1)/2$  pairs of groups. Amongst them we have to identify pairs which share exactly one atom.

Consider any two groups  $G_p$  and  $G_q$ . Each of them have rank  $K$ . The corresponding (unknown) atoms are indexed by index sets  $\Lambda_p$  and  $\Lambda_q$  respectively (both of which are unknown). If, no atoms are common between the supports of the two groups, then  $\Lambda_p \cap \Lambda_q = \emptyset$ . In order to represent, the signals in the merged group, we will need all the  $2K$  atoms. We need to check if these atoms will be linearly independent or not. Since it is given that  $2K < \sigma$  (the spark), hence any set of  $2K$  atoms is linearly independent. Thus, the rank of the merged group  $(G_p, G_q)$  is  $2K$ . It cannot be higher, since the rank of each group is  $K$ . It cannot be lower since the non-degeneracy assumption makes sure

that all the  $2K$  atoms are used in representation of signals and  $2K < \sigma$  ensures that these atoms are independent.

Continuing with the same logic, if  $|\Lambda_p \cap \Lambda_q| = 1$  (i.e. one of the atoms is common between the two groups), then the rank of the merged group  $G_p, G_q$  is  $2K - 1$ . Conversely, if the rank of the merged group is  $2K - 1$ , then it belongs to a  $2K - 1$  dimensional subspace. Thus only  $2K - 1$  atoms in the combined set of  $2K$  atoms are linearly independent. But since  $2K < \sigma$ , hence any set of  $2K$  distinct atoms is linearly independent. Hence, an atom must be common between the pair of groups.

Similarly, we can notice that if rank of the merged group  $G_p, G_q$  is less than  $2K - 1$ , then more than one atoms is common between them.

Thus, we iterate over  $R(R-1)/2$  pairs of atoms and identify those pairs which have exactly one atom in common. The next job is to extract the atom which is common to the support of both groups in such a pair.

**Stage 3: Extracting the common atom.** Let a pair of groups be  $G_p$  and  $G_q$ . All signals in  $G_p$  are from the same  $K$ -dimensional subspace. Due to non-degeneracy assumption, any  $K$  signals from  $G_p$  are linearly independent. Thus, if we pick any  $K$  atoms from  $G$ , they will form a basis. Let us construct one such basis from both  $G_p$  and  $G_q$ . Let those bases be  $B_p$  and  $B_q$  where both  $B_p, B_q \in \mathbb{C}^{N \times K}$ .

Let  $d$  be an atom common to both subspaces. Then, there exists some vector  $v_p$  such that (any vector in a subspace is a linear combination of basis vectors)

$$d = B_p v_p$$

and a vector  $v_q$  such that

$$d = B_q v_q.$$

Thus, we have the relationship

$$B_p v_p = B_q v_q \iff B_p v_p - B_q v_q = 0.$$

Note that at this stage although  $B_p$  and  $B_q$  are known, the vectors  $v_p$  and  $v_q$  are unknown. A slight re-arrangement of this equation gives us

$$\begin{bmatrix} B_p & -B_q \end{bmatrix} \begin{bmatrix} v_p \\ v_q \end{bmatrix} = 0$$

Defining  $v = \begin{bmatrix} v_p \\ v_q \end{bmatrix}$ , and  $B_{p+q} = \begin{bmatrix} B_p & -B_q \end{bmatrix}$  we get

$$B_{p+q}v = 0.$$

Thus, the vector  $v$  belongs to the null space of  $B_{p+q}$ . We further note that the matrix  $B_{p+q}$  has rank  $2L - 1$  since it spans the subspace of the merged group of  $G_p$  and  $G_q$ . Thus, a vector  $v$  from the null space of  $B_{p+q}$  can be obtained by performing SVD of  $B_{p+q}$  and taking the last right singular vector.

Once  $v$  has been obtained, we can break it into  $v_p$  and  $v_q$  easily. Then  $B_p v_p$  gives us a vector parallel to the common atoms (up to a scaling factor and a sign change). We simply normalize this vector and treat this as our common atom we were looking for.

Continuing the same way for the  $R(R - 1)/2$  pairs, we will obtain all our atoms. For each group, there are exactly  $K$  other groups which share an atom with it. Thus, total number of pairs having one atom in common are  $RK/2$ . We will end up extracting  $RK/2$  atoms out of which only  $D$  are unique (subject to a sign change). We scan through the list of atoms. For each atom, we rescan the list and prune those atoms which are parallel to it. Proceeding this way, we will end up with our desired set of  $D$  atoms.

Once the atoms have been obtained, we can put them together into our desired dictionary  $\mathcal{D}'$ . Recall that the obtained dictionary contains the atoms of original dictionary subject to a permutation of order of atoms and changes in the sign of atoms.

After the dictionary has been identified, then finding the representations  $\mathcal{A}$  is a straightforward job of solving the  $l_0$ -“norm” minimization problem.

We note that there are many possibilities in the constructive algorithm which can affect the choice of atoms. For example, in stage 2, the order in which we select pairs of groups can be done in many different ways. Yet, all possible ways will lead to same set of atoms in  $\mathcal{D}$  up to the simple differences of order of atoms and sign.  $\square$

It is possible to ease some of the assumptions made at the beginning of this section. We will not delve into this issue for now.

### 14.3. Digest

## CHAPTER 15

# Distributed Compressed Sensing

### 15.1. Introduction

The material in this chapter is based on [5]. We consider the problem of *joint measurement* for CS in sensor networks that are able to exploit intra signal dependencies as well as inter signal dependencies. We introduce the notion of *joint sparsity* over the *ensemble* of signals obtained from multiple sensors.

#### 15.1.1. Notation

Let  $\Lambda \triangleq \{1, 2, \dots, S\}$  denote the set of indices for the  $S$  signals (from  $S$  sensors) in the ensemble.

We denote each signal in the ensembles as  $x_s$  with  $s \in \Lambda$  and assume that  $x_s \in \mathbb{R}^N$ .

We use  $x_s(n)$  to denote sample  $n$  in signal  $s$ .

Without loss of generality we assume that the signals are sparse in the canonical basis, i.e. the sparsifying dictionary is  $\mathcal{D} = I$ .

We denote by  $\Phi_s$ , the sensing matrix for signal  $s$ ;  $\Phi_s \in \mathbb{R}^{M_s \times N}$  and, in general, the entries of  $\Phi_s$  are different for each signal  $s$ .

We have the measurement vectors as

$$y_s \triangleq \Phi_s x_s \quad (15.1.1)$$

where each  $y_s \in \mathbb{R}^{M_s}$  consists of  $M_s$  random measurements of  $x_s$ . We will be focusing on Gaussian sensing matrices in the sequel. Note that we are not considering measurement noise at the moment.

TABLE 1. Symbols used in this chapter

Symbol	Purpose
$N$	Dimension of ambient space $\mathbb{R}^N$ for signals
$\mathbb{R}^N$	Signal space
$x_s$	A signal belonging to $\mathbb{R}^N$
$K$	Sparsity level when all signals have same sparsity level
$K_c$	Sparsity level of the common component
$K_s$	Sparsity level of the independent component of $s$ -th signal
$S$	Number of sensors
$X$	Combined signal vector $X \in \mathbb{R}^{NS}$
$\Phi_s$	Sensing matrix for $s$ -th signal
$M_s$	Number of measurements for $s$ -th signal
$\bar{M}$	Total number of measurements
$\Phi$	Combined sensing matrix for all signals
$M$	Number of measurements for each signal if same

The total number of measurements are defined as

$$\bar{M} \triangleq \sum_{s=1}^S M_s. \quad (15.1.2)$$

We define  $X \in \mathbb{R}^{SN}$ ,  $Y \in \mathbb{R}^{\bar{M}}$  and  $\Phi \in \mathbb{R}^{\bar{M} \times N}$  as

$$X \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_S \end{bmatrix}, \quad Y \triangleq \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_S \end{bmatrix} \quad \text{and} \quad \Phi \triangleq \begin{bmatrix} \Phi_1 & 0 & \dots & 0 \\ 0 & \Phi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Phi_S \end{bmatrix} \quad (15.1.3)$$

With these definitions we can write

$$Y = \Phi X.$$

We note that  $\Phi$  has a block diagonal structure.

$X$  represents a signal ensemble of the signals  $\{x_1, x_2, \dots, x_S\}$ . Usually we expect all signals in the ensemble to be highly correlated, thus exhibiting high inter-signal dependencies.

## 15.2. Framework for joint sparsity

We now propose a general framework for quantifying the sparsity of a signal ensemble  $X$ .

Consider a signal  $x \in \mathbb{R}^N$  with  $K \ll N$  non-zero entries. There are two components of this representation. The locations of non-zero entries and their values. We can construct a factored representation of  $x$  by writing

$$x = P\theta$$

where  $\theta \in \mathbb{R}^K$  is a vector consisting of non-zero entries of  $x$  and  $P$  is a matrix constructed by choosing the  $K$  columns of an  $N \times N$  identity matrix indexed by the locations of the non zero entries in  $x$ . The matrix  $P$  can thus be considered as an **identity submatrix**.

**Example 15.1: Location-value factored representation of  $x$**  Let  $x = \begin{pmatrix} 0 & -2 & 3 & 0 & 0 & 0 & -1 \end{pmatrix}$ . Clearly

$$\begin{bmatrix} 0 \\ -2 \\ 3 \\ 0 \\ 0 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ 3 \\ -1 \end{bmatrix}$$

.

□

Clearly any  $K$ -sparse signal can be written in this form. We now consider the set of all possible  $K$ -sparse signals with  $1 \leq K \leq N$ . For each value of  $K$  and specific choice of locations in  $x$ , there is a specific identity submatrix.

We now define  $\mathcal{P}$  as the set of all identity submatrices of size  $N \times K$  with  $1 \leq K \leq N$ . This set covers all possible location matrices for all signals with different sparsity levels. It is interesting to note that

this set happens to be a finite set ( $|\mathcal{P}| = 2^N$ ). We refer to  $\mathcal{P}$  as the **sparsity model**.

For a particular signal  $x \in \mathbb{R}^N$ , the set of possible factorizations of  $x$  is given by

$$\{P \in \mathcal{P} | x = P\theta\}.$$

Clearly there are more than one factorizations possible for a signal unless its sparsity level is  $N$  (i.e. completely non-sparse).

**Definition 15.1** Let  $x \in \mathbb{R}^N$  be some signal and  $\mathcal{P}$  be a sparsity model for  $\mathbb{R}^N$ . Let  $\{P \in \mathcal{P} | x = P\theta\}$  be the set of all factorizations of  $x$  in the context of  $\mathcal{P}$ . Among these factorizations, the unique representation with smallest dimensionality of  $\theta$  exists. The **sparsity level** of  $x$  in the context of the sparsity model  $\mathcal{P}$  is defined as the dimensionality of  $\theta$  corresponding to the unique minimal factorization.

For the signal ensemble case with  $X \in \mathbb{R}^{SN}$ , we consider factorizations of the form  $X = P\Theta$  where  $P \in \mathbb{R}^{SN \times \delta}$  is known as the **location matrix** and  $\Theta \in \mathbb{R}^\delta$  is known as the **value vector**.

**Example 15.2:** L et  $x_1 = \begin{pmatrix} 4 & -1 & 0 & 0 \end{pmatrix}$  and  $x_2 = \begin{pmatrix} 4 & 0 & 0 & 5 \end{pmatrix}$ . Then  $X = \begin{pmatrix} 4 & -1 & 0 & 0 & 4 & 0 & 0 & 5 \end{pmatrix}$ . A possible factorization of  $X$  can be

$$\begin{bmatrix} 4 \\ -1 \\ 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 5 \end{bmatrix} = P\Theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \\ 5 \end{bmatrix}$$

Note that in this example  $P$  is not an identity submatrix. We also note that if we take the signals  $x_1$  and  $x_2$  separately then both are 2-sparse

and their overall sparsity is 4. But in the factorization presented above  $\Theta \in \mathbb{R}^3$  thus giving a joint sparsity of 3 to  $X$ .

Also note that  $P$  in this factorization is full rank.  $\square$

**Definition 15.2** A **joint sparsity model** (JSM) is defined in terms of a set  $\mathcal{P}$  of admissible location matrices  $P$  with varying number of columns.

Note that the definition doesn't require  $P$  to be an identity submatrix.

**Definition 15.3** For a given signal ensemble  $X \in \mathbb{R}^{SN}$ , the set of matrices  $P$  belonging to a joint sparsity model  $\mathcal{P}$  for which a factorization  $X = P\Theta$  is possible is known as the set of **feasible location matrices** for  $X$ . The set is denoted by

$$P_F(X) \triangleq \{P \in \mathcal{P} | X = P\Theta\}$$

**Definition 15.4** The **joint sparsity level**  $J$  of the signal ensemble  $X$  is the number of columns of the smallest matrix  $P \in P_F(X)$ .

There are several natural choices for what matrices  $P$  should be considered as members of a joint sparsity model  $\mathcal{P}$ . In this chapter our focus will be on models known as **common / innovation component JSMs**.

In a common / innovation components model, each signal  $x_s$  is generated as a sum of two components:

- A common component  $z_C$  which is present in all signals.
- An innovation component  $z_s$  which is unique to each signal.

$$x_s = z_C + z_s, s \in \Lambda.$$

It is possible that either the common component or the innovation components might be zero in specific situations.

We can now factor the common and innovation components as

$$z_C = P_C \theta_C, \quad z_s = P_s \theta_s, \quad j \in \Lambda$$

where  $\theta_C \in \mathbb{R}^{K_C}$  and each  $\theta_s \in \mathbb{R}^{K_s}$  have non-zero entries. The matrices  $P_C$  and  $P_s$  are identity submatrices.

We can now write

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_S \end{bmatrix} = \begin{bmatrix} P_C & P_1 & 0 & \dots & 0 \\ P_C & 0 & P_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_C & 0 & 0 & \dots & P_S \end{bmatrix} \begin{bmatrix} \theta_C \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_S \end{bmatrix}.$$

**Definition 15.5** The **common / innovations joint sparsity model**  $\mathcal{P}$  for a signal ensemble  $X$  of  $S$  signals belonging to  $\mathbb{R}^N$  is defined as a set of matrices of the form

$$P = \begin{bmatrix} P_C & P_1 & 0 & \dots & 0 \\ P_C & 0 & P_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_C & 0 & 0 & \dots & P_S \end{bmatrix} \quad (15.2.1)$$

where  $P_C \in \mathbb{R}^{N \times K_C}$  and  $P_s \in \mathbb{R}^{N \times K_s}$  are identity sub-matrices of  $N \times N$  identity matrix;  $K_C$  is the sparsity level of the common component  $z_C$  and  $K_s$  are the sparsity levels of innovation components  $z_s$  with  $s \in \Lambda$ .

We note that the factorization of  $X$  is not unique.

**Example 15.3: Common / innovations model** Continuing from previous example, consider

$$z_C = \begin{pmatrix} 4 & 0 & 0 & 0 \end{pmatrix},$$

$$z_1 = \begin{pmatrix} 0 & -1 & 0 & 0 \end{pmatrix}$$

and

$$z_2 = \begin{pmatrix} 0 & 0 & 0 & 5 \end{pmatrix}$$

We see that

$$x_1 = z_C + z_1 \quad x_2 = z_C + z_2.$$

The factorization presented in previous example is a common / innovations factorization. Of course we can choose  $z_C = 0$  leading to a completely different factorization. This is given by

$$\begin{bmatrix} 4 \\ -1 \\ 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 5 \end{bmatrix} = P\Theta = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \\ 4 \\ 5 \end{bmatrix}$$

Again we see that  $P$  is full rank but with  $\Theta \in \mathbb{R}^4$ .

An interesting variation is to consider

$$z_C = \begin{pmatrix} 3 & -1 & 0 & 0 \end{pmatrix},$$

$$z_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$$

and

$$z_2 = \begin{pmatrix} 1 & 1 & 0 & 5 \end{pmatrix}$$

This gives us the factorization

$$\begin{bmatrix} 4 \\ -1 \\ 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 5 \end{bmatrix} = P\Theta = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 5 \end{bmatrix}.$$

This factorization has  $\Theta \in \mathbb{R}^6$ . Interestingly we note that  $P$  is not full rank and its first column can easily be constructed by combining third and fourth columns. This gives us a straight-forward way of reducing the sparsity level of the factorization.  $\square$

If a signal ensemble  $X = P\Theta, \Theta \in \mathbb{R}^\delta$  were to be generated by a selection of  $P_C$  and  $\{P_s\}_{s \in \Lambda}$ , where all the  $S + 1$  identity submatrices share a common column vector, then  $P$  would not be full rank (as seen in previous example). Here we can easily reduce  $P$  by removing one of the columns from any of the identity submatrices under concern.

In other cases, we may observe that  $\Theta$  has some zero-valued entries, i.e. we have may have  $\theta_s(k) = 0$  for some  $s \in \Lambda$  and some  $1 \leq k \leq K_s$ , or  $\theta_C(k) = 0$  for some  $1 \leq k \leq K_C$ . In this case, we can simply remove the corresponding column from  $P$ .

The process of removing columns from  $P$  to get a new matrix  $Q$  such that  $X = Q\Theta'$  where  $\Theta' \in \mathbb{R}^{\delta-1}$  is known as **sparsity reduction**.

**Example 15.4: Sparsity reduction** Continuing with previous example, we first remove third column from  $P$ . This also changes the factorization as

$$\begin{bmatrix} 4 \\ -1 \\ 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 5 \end{bmatrix} = P\Theta = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \\ 0 \\ 1 \\ 5 \end{bmatrix}.$$

We note that a new 0 entry has appeared which can be removed safely giving us

$$\begin{bmatrix} 4 \\ -1 \\ 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 5 \end{bmatrix} = P\Theta = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \\ 1 \\ 5 \end{bmatrix}.$$

□

As we see from this example, *sparsity reduction*, when present, reduces the effective joint sparsity of a signal ensemble.

### 15.2.1. Example joint sparsity models

In this chapter we will consider three specific examples of common / innovation joint sparsity models which arise in real life.

**15.2.1.1. JSM-1: Sparse common component + innovations.** In this model, we suppose that each signal contains a common component  $z_C$  that is *sparse* plus an innovation component  $z_s$  that is also *sparse*.

Thus JSM-1 model  $\mathcal{P}$  consists of all matrices of the form (15.2.1) with  $K_C \ll N$  as well as  $K_s \ll N$  for all  $s \in \Lambda$ .

Assuming that the sparsity reduction is not possible, the joint sparsity is given by

$$J = K_C + \sum_{s \in \Lambda} K_s.$$

**Example 15.5: JSM-1 examples** Consider a group of sensors measuring temperature in an outdoor locality. The temperature readings

$x_s$  have both temporal (intra-signal) as well as spatial (inter-signal) dependencies (correlations).

Global factors such as sun and weather conditions have an effect as  $z_C$  which is common to all sensors as well as structured enough to have a sparse representation in some basis.

Local factors such as shade, water or animals, contribute localized innovations  $z_s$  that are also structured (hence sparse in some basis).  $\square$

**15.2.1.2. JSM-2: Common sparse supports.** In this model, the common component  $z_C$  is zero and each innovation component  $z_s$  is sparse. Further all the innovations share the *same sparse support* but can have entirely different non-zero values.

In this model,  $\mathcal{P}$  is the set of matrices  $P$  given by (15.2.1) where  $P_C = \emptyset$  (i.e.  $K_C = 0$ ) and  $P_s = \bar{P}$  for all  $s \in \Lambda$ .

$\bar{P}$  is an identity submatrix of size  $N \times K$  with  $K \ll N$ . The columns in  $\bar{P}$  correspond to the support common to all signals  $x_s$  in the ensemble  $\{x_s\}_{s \in \Lambda}$ .

We have

$$x_s = \bar{P}\theta_s, \quad \forall s \in \Lambda \text{ with } \theta_s \in \mathbb{R}^K.$$

The matrices  $P$  from JSM-2 are always full rank. Therefore no sparsity reduction is possible. Thus we have the joint sparsity as

$$J = SK.$$

**Example 15.6: JSM-2 examples** In acoustic and RF sensor arrays each sensor acquires a replica of the same Fourier sparse signal but with phase shifts and attenuations caused by signal propagation.

Another example is MIMO communications.  $\square$

This model is also known as *MMV (multiple measurement vectors)* setting (see next chapter) or *simultaneous sparse approximation* problem

**15.2.1.3. JSM-3: Non-sparse common component + sparse innovations.** In this model we assume that each signal consists of an arbitrary common component  $z_C$  which is non-sparse and a sparse innovation component  $z_s$ .

Thus the sparsity level of the common component is  $K_C = N$ . The JSM-3 model  $\mathcal{P}$  is the set of matrices given by (15.2.1) where  $P_C = I$  (the  $N \times N$  identity matrix). Thus we have  $\theta_C \in \mathbb{R}^N$  while  $\theta_s \in \mathbb{R}^{K_s}$ .

Assuming that sparsity reduction is not possible, the joint sparsity is given by

$$J = N + \sum_{s \in \Lambda} K_s.$$

In a special case where the the sparse innovations share a common support, we have

$$P_s = \bar{P}, K_s = K \forall s \in \Lambda.$$

In this case,  $K$  columns out of the first  $N$  columns in  $P$  can be expressed as linear combination of corresponding columns in  $P_s$ . Thus clearly a sparsity reduction is possible, leading to the joint sparsity level given by

$$J = N + (S - 1)K.$$

We note that in this case since each of  $x_s$  are non-sparse (as  $z_c$  is non-sparse) hence separate CS recovery for these signals is not possible if the number of measurements  $M < N$  for any sensor. It turns out that joint CS recovery can indeed take advantage of the common structure and make CS recovery possible with  $M < N$  per sensor.

**Example 15.7: JSM-3 examples** A practical situation is where different sensors are recording several sources with a common background noise. The background noise is not sparse in any basis but individual sources have sparse representations.  $\square$

### 15.3. Theoretical bounds on measurement rates

We recall from (15.1.1) that the number of measurements made by each sensor are different and are given by  $y_s = \Phi_s x_s$  where  $\Phi_s \in \mathbb{R}^{M_s \times N}$ .

We define a tuple of number of measurements from each sensor as

$$\mathcal{M} \triangleq (M_1, M_2, \dots, M_S).$$

We will be looking for conditions on  $\mathcal{M}$  such that perfect recovery of  $X$  is possible given  $Y$  (see (15.1.3)). In this chapter we will be mostly looking at conditions for noiseless recovery.

Recovering  $X$  essentially involves obtaining a factorization of  $X = P\Theta$  from the measurement ensemble  $Y = \Phi X$ . Some observations are in order

- Not every entry in  $\Theta$  affects every measurement in  $Y$ . The common component  $\theta_C$  affects every measurement. But innovation components  $\theta_s$  impact only corresponding measurement vector  $y_s$ .
- Thus if an entry  $\Theta(j)$  (with  $1 \leq j \leq J$ ) doesn't affect any signal coefficient  $x_s(\cdot)$  in sensor  $s$ , then the corresponding measurement vector  $y_s$  provides no information about  $\Theta(j)$ .
- The recovery process must identify a location matrix  $P$  from the set of feasible matrices  $P_F(X)$  for the signal ensemble  $X$  while neither  $X$  nor  $P_F(X)$  are known.

#### 15.3.1. Modeling dependencies using bipartite graphs

We introduce a graphical representation that captures the dependencies between the measurements in  $Y$  and the value vector  $\Theta$  given by

$$Y = \Phi P\Theta.$$

The matrix  $P \in P_F(X)$  defines the sparsities of the common and innovation components  $K_C$  and  $K_s$ ,  $1 \leq s \leq S$ , as well as the joint sparsity  $J = K_C + \sum_{s=1}^S K_s$ .

We define the following vertices for a graphical representation of dependencies.

- The set of **value vertices**  $V_V$  has elements with indices  $j \in \{1, \dots, J\}$  representing the entries of the value vector  $\Theta(j)$ .
- The set of **measurement vertices**  $V_M$  has elements with indices  $(s, m)$  representing the measurements  $y_s(m)$  with  $s \in \Lambda$  and  $m \in \{1, \dots, M_s\}$ .

We have  $|V_V| = J$  and  $|V_M| = \overline{M}$  see (15.1.2).

We now introduce a bipartite graph  $G = (V_V, V_M, E)$ , that represents the relationships between the entries of the value vector and measurements.

The set of edges is defined as follows

- The sensor  $s$  measures every entry in  $\theta_s$  (the innovation component). No other sensor measures it. Hence entries in the measurement vector  $y_s$  must be responsible for the recovery of entries in  $\theta_s$ . Thus the corresponding value vertices are connected to each measurement vertex  $(s, m) \in V_M$  for  $1 \leq m \leq M_s$ . These  $\theta_s(\cdot)$  entries over  $1 \leq s \leq S$  correspond to  $j \in \{K_C + 1, K_C + 2, \dots, J\}$  vertices in the set  $V_V$ .
- For the first  $K_C$  vertices in  $V_V$  which correspond to  $\theta_C$ , we have to be more careful. If for some sensor  $s$ , both  $z_C(n)$  and  $z_s(n)$  are non-zero (i.e. an entry in the innovation component and common component appears at the same index), then that sensor cannot resolve  $z_C(n)$  and  $z_s(n)$  separately. In other words, a sensor  $s$  can contribute to the recovery of a non-zero entry in  $z_C$  at some index  $1 \leq n \leq N$  only if the corresponding entry in the innovation component  $z_j(n) = 0$ . The identity submatrices  $P_C$  and  $P_s$  will have the following difference. The  $n$ -th column from  $I_N$  will be kept in  $P_C$  while it will be dropped in  $P_s$ . Only such dependences are useful in the bipartite graph  $G$ . Formally, for every  $j \in \{1, 2, \dots, K_C\} \subseteq V_V$  and  $s \in \Lambda$  such

that the column  $j$  of  $P_C$  does not also appear as a column of  $P_s$ , we have an edge connecting  $j \in V_V$  to each vertex  $(s, m) \in V_M$  for  $1 \leq m \leq M_s$ .

Essentially,  $y_s(m)$ , the  $m$ -th measurement of sensor  $s$ , measures  $\Theta(j)$  if the vertex  $j \in V_V$  is linked to the vertex  $(s, m) \in V_M$  in the graph  $G$ . An example graph is presented in fig. 15.1. We note that this graph is applicable for a specific factorization of  $X$ . For a different factorization with a different  $P$ , the graph will naturally change accordingly. At the time of signal recovery, the graph is not known to us.

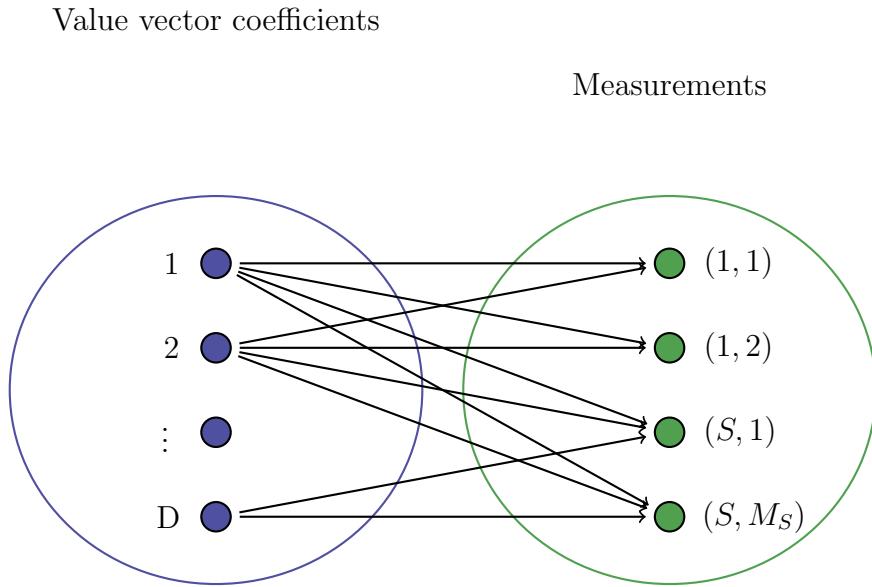


FIGURE 15.1. Bipartite graph for DCS

### 15.3.2. Quantifying redundancies

Our objective is to minimize the number of measurements required to recover the signal perfectly. Thus we would like to exploit all possible redundancies between the locations of the non-zero entries in the common and innovation components.

As discussed before if  $z_C(n) \neq 0$  and  $z_s(n) \neq 0$  for some sensor  $s$  and some  $1 \leq n \leq N$ , then we cannot recover both entries from sensor  $s$ .

Thus we will need to recover  $z_C(n)$  using other sensors which do not feature this overlap.

Now consider a subset of signals  $\Gamma \subset \Lambda$  under a feasible representation given by  $P$  and  $\Theta$ . We quantify the size of this overlap in the following definition. For a particular location matrix  $P$  we also denote the sparsity level of the common component as  $K_C(P)$ .

**Definition 15.6** The **overlap size** for the set of signals  $\Gamma \subset \Lambda$ , denoted  $K_C(\Gamma, P)$ , is the number of indices in which there is overlap between the common and the innovation component supports at all signals  $s \notin \Gamma$ .

$$K_C(\Gamma, P) \triangleq |\{n \in \{1, \dots, N\} : z_C(n) \neq 0 \text{ and } \forall s \notin \Gamma, z_s(n) \neq 0\}|. \quad (15.3.1)$$

We also define  $K_C(\Lambda, P) = K_C(P)$  and  $K_C(\emptyset, P) = 0$ .

Essentially for  $\Gamma \subset \Lambda$ , the overlap size  $K_C(\Gamma, P)$  provides the number of entries in  $\theta_C$  which must be recovered by measurements from sensors in  $\Gamma$  as every sensor  $s$  outside  $\Gamma$  has an overlap with those entries in its innovation component.

For each entry counted in  $K_C(\Gamma, P)$ , some sensor in  $\Gamma$  must take one measurement to account for that entry of the common component.

The definition  $K_C(\Lambda, P) = K_C(P)$  is quite natural since when all sensors are considered in  $\Gamma$  then they together should be able to identify all entries in the common component.

With this, we can now introduce the idea of **conditional sparsity** akin to conditional probability.

**Definition 15.7** The **conditional sparsity** of the set of signals  $\Gamma$  is the number of entries of the vector  $\Theta$  that must be recovered by measurements  $y_s, s \in \Gamma$ :

$$K_{\text{cond}}(\Gamma, P) \triangleq \left( \sum_{s \in \Gamma} K_j(P) \right) + K_C(\Gamma, P). \quad (15.3.2)$$

The joint sparsity gives the number of degrees of freedom for the signals in  $\Lambda$ , while the conditional sparsity gives the number of degrees of freedom for signals in  $\Gamma$  when the signals in  $\Lambda \setminus \Gamma$  are available as side information.

We can also define a joint sparsity for signals in  $\Gamma$  as follows.

**Definition 15.8** The joint sparsity of the set of signals  $\Gamma$  is the number of entries of  $\Theta$  that affect these signals.

$$K_{\text{joint}}(\Gamma, P) \triangleq J - K_{\text{cond}}(\Gamma, P) = \left( \sum_{s \in \Gamma} K_j(P) \right) + K_C(P) - K_C(\Lambda \setminus \Gamma, P). \quad (15.3.3)$$

We note that  $K_{\text{cond}}(\Lambda, P) = K_{\text{joint}}(\Lambda, P) = J$ .

### 15.3.3. Measurement bounds

With the definitions in place, we are now ready to develop some measurement bounds for DCS recovery. The bounds are of two forms. One form states how many measurements  $\mathcal{M}$  are sufficient for guaranteed recovery. The second form states the necessary measurements below which no recovery is possible. The bounds are provided in terms of the subsets  $\Gamma \subseteq \Lambda$ , since the cost of sensing the common components can be amortized across sensors. Essentially it may be possible to reduce the measurement rate at one sensor  $s_1 \in \Gamma$  as long as other sensors in

$\Gamma$  offset the rate reduction. This rate reduction is developed in terms of the notion of conditional sparsity defined above.

**Theorem 15.1** *Achievable known  $P$*  Assume that a signal ensemble  $X$  is obtained from a common / innovation component JSM  $\mathcal{P}$ . Let  $\mathcal{M} = (M_1, \dots, M_S)$  be a measurement tuple, let  $\{\Phi_s\}_{s \in \Lambda}$  be random matrices having  $M_s$  rows of i.i.d. Gaussian entries for each  $s \in \Lambda$ , and write  $Y = \Phi X$ . Suppose there exists a full rank location matrix  $P \in P_F(X)$  such that

$$\sum_{s \in \Gamma} M_s \geq K_{cond}(\Gamma, P) \quad (15.3.4)$$

for all  $\Gamma \subseteq \Lambda$ . Then with probability one over  $\{\Phi_s\}_{s \in \Lambda}$ , there exists a unique solution  $\widehat{\Theta}$  to the system of equations  $Y = \Phi P \widehat{\Theta}$ ; hence, the signal ensemble  $X$  can be uniquely recovered as  $X = P \widehat{\Theta}$ .

**Theorem 15.2** *Assume that a signal ensemble  $X$  and measurement matrices  $\{\Phi_s\}_{s \in \Lambda}$  follow the assumptions of theorem 15.1.* Suppose that there exists a full rank location matrix  $P^* \in \P_F(X)$  such that

$$\sum_{s \in \Gamma} M_s \geq K_{cond}(\Gamma, P^*) + |\Gamma| \quad (15.3.5)$$

for all  $\Gamma \subseteq \Lambda$ . Then  $X$  can be recovered uniquely from  $Y$  with probability one over  $\{\Phi_s\}_{s \in \Lambda}$ .

**Theorem 15.3** *Assume that a signal ensemble  $X$  and measurement matrices  $\{\Phi_s\}_{s \in \Lambda}$  follow the assumptions of theorem 15.1.* Suppose that there exists a full rank location matrix  $P \in P_F(X)$  such that

$$\sum_{s \in \Gamma} M_s < \geq > K_{cond}(\Gamma, P) \quad (15.3.6)$$

for some  $\Gamma \subseteq \Lambda$ . Then there exists a solution  $\widehat{\Theta}$  such that  $Y = \Phi P \widehat{\Theta}$  but  $\widehat{X} \triangleq P \widehat{\Theta} \neq X$ .

**15.4. Practical recover algorithms****15.4.1. Recovery strategies for JSM-1**

For simplicity in this section, we will consider the special case with  $S = 2$ . Thus  $\Lambda = \{1, 2\}$ .

## **Part 3**

### **Inference**

## CHAPTER 16

# Detection with Compressed Measurements

In this chapter we discuss the application of compressed sensing framework for signal detection problems.

We quickly develop the theory of signal detection. We follow this by extending the theory to include compressed measurements of the received signal.

### 16.1. Binary detection theory

The presentation in this section is largely based on [41]. If you are familiar with it, you may skim through and move on to next section.

Let us begin with some examples of signal detection problems in electrical engineering.

- Detection of objects in air (enemy planes, missiles etc.) in a radar problem
- Detection of a string of zeros and ones in a digital communication system
- Detection of pathological tissues in an MRI image.

The transmitted signal undergoes distortion, attenuation and addition of noise as it travels through the channel before reaching the receiver.

The problem at the receiver is to make a sequence of decisions to reconstruct the original signal in the presence of noise, attenuation and distortion of the transmitted signal.

- In the digital communication system, the receiver makes a decision in every bit period whether the bit is 1 or 0.

- In the radar problem, receiver continually makes decisions whether a target is present or not.

### 16.1.1. Binary hypothesis testing

Each of these situations can be modeled in terms of a statistical framework known as binary hypothesis testing problem.

**Definition 16.1** A **binary hypothesis testing problem** for our purposes can be described as follows. The system consists of a source, a medium and a receiver. The source can operate in one of two modes. For each mode, the source generates a different signal. The signal is modified as it travels through the medium and reaches the receiver. The receiver makes observations on the signal. Based on the observations the receiver makes a decision whether the source was operating in one or the other mode.

If we call the modes as mode 0 and mode 1, the receiver has two hypotheses.

- The **null hypothesis** denoted as  $H_0$  stands for the situation where the source is operating in mode 0.
- The **alternate hypothesis** denoted as  $H_1$  stands for the situation where the source is operating in mode 1.

The receiver makes a decision whether null ( $H_0$ ) or the alternate ( $H_1$ ) hypothesis is true.

Since this problem involves only two hypotheses, hence its known as a binary hypothesis testing problem.

### Example 16.1: Binary hypothesis testing problems

- In radar problem, absence of a target is the null hypothesis while the presence of a target is alternate hypothesis.

- In pathological tissue detection problem, the absence of a pathological tissue is the null hypothesis while the presence of such a tissue is the alternate hypothesis.
- In digital communication problem, bit 0 is the null hypothesis while bit 1 is the alternate hypothesis.

□

Usually the most common behavior, or the absence of an anomaly is chosen as the null hypothesis. The opposite hypothesis is chosen as alternate hypothesis.

In digital communication problem, both hypotheses are equivalent, so anyone can be called null. Conventionally bit 0 is called the null hypothesis.

In statistical hypothesis testing, we assume that null hypothesis is true, and estimate the probability of occurrence of the observations accordingly.

If the observations are highly improbable assuming the null hypothesis then we reject the null hypothesis and decide that alternate hypothesis is true.

**Definition 16.2** The **observation space** consists of all possible observations that a receiver can make. We denote the observation space as  $Z$ .

Typically the observation space is  $Z = \mathbb{R}^N$  for some value of  $N$  and each observation is a vector  $y \in Z = \mathbb{R}^N$ .

### Example 16.2: Observation space

- Let bit 0 be coded as 0 volts and bit 1 be coded as 5 volts on an electrical wire. As the signal travels over the wire, at the receiver, the voltage would be some  $x \in \mathbb{R}$  volts (based on attenuation and noise).

- Let a bit be encoded as a square pulse over a bit period of 1 ms. Let 10 samples be made during each bit period. The observation space is  $\mathbb{R}^{10}$  and each observation is a vector  $v \in \mathbb{R}^{10}$  consisting of 10 samples.

□

Since the signal distortion, attenuation and addition of noise are random in nature, it is best to describe the observation vector  $y$  as a random variable  $Y$  which takes values in the observation space  $Z$ .

The r.v.  $Y$  is characterized by two conditional probability density functions based on whether null or alternate hypothesis is true.

- $f_{Y|H_0}(y|H_0)$  denotes the conditional p.d.f. of  $Y$  assuming null hypothesis is true.
- $f_{Y|H_1}(y|H_1)$  denotes the conditional p.d.f. of  $Y$  assuming alternate hypothesis is true.

For every observation vector  $Y$  the receiver decides whether  $H_0$  or  $H_1$  is true.

- $D_0$  denotes the decision that receiver has chosen the null hypothesis  $H_0$  to be true.
- $D_1$  denotes the decision that receiver has chosen the null hypothesis  $H_1$  to be true.

The observation space  $Z$  is partitioned into two regions  $Z_0$  and  $Z_1$  i.e.  $Z = Z_0 \cup Z_1$ .

The receiver operates as follows:

- If  $y \in Z_0$  then decide that  $H_0$  is true (denoted as  $D_0$  decision).
- If  $y \in Z_1$  then decide that  $H_1$  is true (denoted as  $D_1$  decision).

The partitioning is static i.e. it doesn't adaptively change on past observations.

If we look at the whole system, every decision belongs to one of four possible courses of action.

- (1) Receiver decides  $H_0$  is true while at the source  $H_0$  is true ( $D_0|H_0$ ).
- (2) Receiver decides  $H_1$  is true while at the source  $H_0$  is true ( $D_1|H_0$ ).
- (3) Receiver decides  $H_0$  is true while at the source  $H_1$  is true ( $D_0|H_1$ ).
- (4) Receiver decides  $H_1$  is true while at the source  $H_1$  is true ( $D_1|H_1$ ).

1 and 4 are correct decisions while 2 and 3 are incorrect decisions.

The objective of receiver design is to ensure that incorrect decisions are minimized. This essentially translates into a prudent partitioning of observation space  $Z$  into  $Z_0$  and  $Z_1$ .

Some standard terms are used to denote these courses of action in literature.

**Definition 16.3** In radar terminology we either detect a target correctly or miss a target when its present or create a false alarm when the target is not present.

- We say that a **detection** has occurred if receiver decides  $H_1$  when  $H_1$  is true.
- We say that a **miss** has occurred if receiver decides  $H_0$  when  $H_1$  is true.
- We say that a **false alarm** has occurred if receiver decides  $H_1$  when  $H_0$  is true.

We define the a priori probabilities for the null and alternate hypotheses:

- $\mathbb{P}(H_0)$  is the probability of null hypothesis to be true (denoted as  $P_0$  in the sequel).

- $\mathbb{P}(H_1)$  is the probability of alternate hypothesis to be true (denoted as  $P_1$  in the sequel).

Naturally  $P_0 + P_1 = 1$ .

We denote the joint probability of making a decision  $D_i$  when  $H_j$  is true as  $\mathbb{P}(D_i, H_j)$ .

From Bayes' rule we have

$$\mathbb{P}(D_i, H_j) = \mathbb{P}(D_i|H_j)\mathbb{P}(H_j). \quad (16.1.1)$$

The conditional probabilities can be obtained by integrating the conditional p.d.f.s  $f_{Y|H_j}(y|H_j)$  over the partition  $Z_i$  which contributes the decision  $D_i$ .

Thus we have

$$\mathbb{P}(D_i|H_j) = \int_{Z_i} f_{Y|H_j}(y|H_j)dy. \quad (16.1.2)$$

**Definition 16.4** The probabilities  $\mathbb{P}(D_i|H_j)$  have specific names.

- $\mathbb{P}(D_1|H_1)$  is called the **probability of detection** or **detection rate** and is denoted as  $P_D$ .
- $\mathbb{P}(D_1|H_0)$  is called the **probability of false alarm** or **false alarm rate** and is denoted as  $P_F$ .
- $\mathbb{P}(D_0|H_1)$  is called the **probability of miss** or **miss rate** and is denoted as  $P_M$ .

We observe that

$$P_M = 1 - P_D \quad (16.1.3)$$

and

$$\mathbb{P}(D_0|H_0) = 1 - P_F. \quad (16.1.4)$$

which doesn't have a specific name.

**Definition 16.5** We denote the **probability of correct decision** by  $P_c$  and **probability of error** by  $P_e$ .

We have

$$\begin{aligned} P_c &= \mathbb{P}(D_0, H_0) + \mathbb{P}(D_1, H_1) \\ &= \mathbb{P}(D_0|H_0)\mathbb{P}(H_0) + \mathbb{P}(D_1|H_1)\mathbb{P}(H_1) \\ &= (1 - P_F)P_0 + P_D P_1. \end{aligned} \quad (16.1.5)$$

Similarly

$$\begin{aligned} P_e &= \mathbb{P}(D_0, H_1) + \mathbb{P}(D_1, H_0) \\ &= \mathbb{P}(D_0|H_1)\mathbb{P}(H_1) + \mathbb{P}(D_1|H_0)\mathbb{P}(H_0) \\ &= P_M P_1 + P_F P_0. \end{aligned} \quad (16.1.6)$$

In the sequel we develop different approaches which help us come up with an appropriate receiver design which minimizes chances of incorrect decisions.

We note that

$$\int_Z f_{Y|H_j}(y|H_j)dy = 1. \quad (16.1.7)$$

it follows that

$$\int_{Z_0} f_{Y|H_j}(y|H_j)dy + \int_{Z_1} f_{Y|H_j}(y|H_j)dy = 1 \quad (16.1.8)$$

$$\implies \int_{Z_0} f_{Y|H_j}(y|H_j)dy = 1 - \int_{Z_1} f_{Y|H_j}(y|H_j)dy. \quad (16.1.9)$$

### 16.1.2. Bayes' criterion

We start with making some assumptions.

We assume that a priori probabilities  $P_0$  and  $P_1$  are known.

**Example 16.3: A priori probabilities** In a digital communication system typically  $\mathbb{P}(0) = \mathbb{P}(1) = 0.5$ .  $\square$

We assign a cost to each of the four possible courses of action. We say that system incurs a cost of  $C_{ij}$  when receiver decides  $H_i$  is true while at the source  $H_j$  is true.

- $C_{00}$  is the cost for action  $(D_0|H_0)$ .
- $C_{01}$  is the cost for action  $(D_0|H_1)$ .
- $C_{10}$  is the cost for action  $(D_1|H_0)$ .
- $C_{11}$  is the cost for action  $(D_1|H_1)$ .

All costs are 0 or greater. The cost of making an incorrect decision is higher than the cost of making a correct decision. i.e.

$$C_{10} > C_{00} \quad (16.1.10)$$

and

$$C_{01} > C_{11}. \quad (16.1.11)$$

#### Example 16.4: Cost

- In digital communication system, the cost assigned to making a correct decision is 0 while the cost assigned to making a wrong decision is 1. Thus miss and false alarm have same costs.
- In radar problem, the cost of a miss is lower than the cost of false alarm.

□

We develop a notion of risk as the average cost of making the decisions by the receiver.

**Definition 16.6** The **risk** denoted as  $\mathcal{R}$  in a binary hypothesis problem based on Bayes' criterion is the average cost of making a decision defined as

$$\mathcal{R} = \mathbb{E}(\text{Cost}) = \sum_{i=0}^1 \sum_{j=0}^1 C_{ij} \mathbb{P}(D_i, H_j). \quad (16.1.12)$$

In matrix form, let us define the cost matrix as

$$C = \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} \quad (16.1.13)$$

and the joint probability matrix as

$$\mathbb{P}_{DH} = \begin{bmatrix} \mathbb{P}(D_0, H_0) & \mathbb{P}(D_0, H_1) \\ \mathbb{P}(D_1, H_0) & \mathbb{P}(D_1, H_1) \end{bmatrix} = \begin{bmatrix} (1 - P_F)P_0 & (1 - P_D)P_1 \\ P_F P_0 & P_D P_1 \end{bmatrix} \quad (16.1.14)$$

Then the risk  $\mathcal{R}$  is given by

$$\mathcal{R} = \begin{bmatrix} 1 & 1 \end{bmatrix} (C \circ \mathbb{P}_{DH}) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (16.1.15)$$

where  $\circ$  denotes the Hadamard product or element-wise product of two matrices. See ??.

Note that element-wise product is not associative with standard matrix multiplication. Hence we have to keep parentheses in place properly.

**Example 16.5: Risk formula** If we fully expand the risk formula, it looks like

$$\mathcal{R} = (1 - P_F)P_0C_{00} + (1 - P_D)P_1C_{01} + P_F P_0 C_{10} + P_D P_1 C_{11}. \quad (16.1.16)$$

For the digital communication problem cost matrix is given by

$$C = \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (16.1.17)$$

Thus the risk formula reduces to

$$\mathcal{R} = (1 - P_D)P_1 + P_F P_0 = P_M P_1 + P_F P_0. \quad (16.1.18)$$

which is same as the probability of error see [Equation 16.5](#).

A receiver with this kind of cost assignment is known as **minimum probability of error** receiver since minimizing the risk is equivalent to minimizing the probability of error.

Assuming  $P_1 = P_0 = 0.5$ , this further reduces to

$$\mathcal{R} = \frac{1}{2}(P_M + P_F). \quad (16.1.19)$$

□

Let us define the conditional probability matrix as

$$\mathbb{P}_{D|H} = \begin{bmatrix} \mathbb{P}(D_0|H_0) & \mathbb{P}(D_0|H_1) \\ \mathbb{P}(D_1|H_0) & \mathbb{P}(D_1|H_1) \end{bmatrix} = \begin{bmatrix} (1 - P_F) & (1 - P_D) \\ P_F & P_D \end{bmatrix} \quad (16.1.20)$$

The risk expression then becomes

$$\mathcal{R} = \begin{bmatrix} 1 & 1 \end{bmatrix} (C \circ \mathbb{P}_{D|H}) \begin{bmatrix} P_0 \\ P_1 \end{bmatrix}. \quad (16.1.21)$$

After some simplification, we can write the risk expression in terms of integral over  $Z_0$  as

$$\begin{aligned} \mathcal{R} = P_0 C_{10} + P_1 C_{11} + \\ \int_{Z_0} \{ [P_1(C_{01} - C_{11})f_{Y|H_1}(y|H_1)] - [P_0(C_{10} - C_{00})f_{Y|H_0}(y|H_0)] \} dy. \end{aligned} \quad (16.1.22)$$

We note that  $P_0 C_{10} + P_1 C_{11}$  is a constant since all terms are assumed to be known and constant.

The risk depends on how we assign points in  $Z_0$ .

We define

$$L_a(y) = P_1(C_{01} - C_{11})f_{Y|H_1}(y|H_1)$$

and

$$L_b(y) = P_0(C_{10} - C_{00})f_{Y|H_0}(y|H_0).$$

Since  $P_1 \geq 0$ ,  $C_{01} > C_{11}$  and  $f_{Y|H_1}(y|H_1) \forall y \in Z$ , hence  $L_a(y) \geq 0 \forall y \in Z$ .

Similarly  $L_b(y) \geq 0 \forall y \in Z$ .

The risk expression simplifies to

$$\mathcal{R} = P_0 C_{10} + P_1 C_{11} + \int_{Z_0} [L_a(y) - L_b(y)] dy. \quad (16.1.23)$$

Consider a point  $y \in Z$  for which  $L_a(y) > L_b(y)$ .

- If we assign the point  $y$  to  $Z_0$  then the risk increases.
- If we assign the point  $y$  to  $Z_1$  then the risk decreases.

Thus the risk is minimized when we assign only those points  $y \in Z$  to  $Z_0$  for which  $L_a(y) < L_b(y)$ .

Thus our decision rule becomes

$$L_a(y) \stackrel[H_1]{\gtrless}^H L_b(y) \quad (16.1.24)$$

i.e.

- Decide  $H_1$  if  $L_a(y) \geq L_b(y)$ .
- Decide  $H_0$  otherwise.

Rewriting we get

$$P_1(C_{01} - C_{11}) f_{Y|H_1}(y|H_1) \stackrel[H_1]{\gtrless}^H P_0(C_{10} - C_{00}) f_{Y|H_0}(y|H_0) \quad (16.1.25)$$

or

$$\frac{f_{Y|H_1}(y|H_1)}{f_{Y|H_0}(y|H_0)} \stackrel[H_1]{\gtrless}^H \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \quad (16.1.26)$$

**Definition 16.7** The term of the l.h.s. in (16.1.26) is known as **likelihood ratio** which is denoted by

$$\Lambda(y) = \frac{f_{Y|H_1}(y|H_1)}{f_{Y|H_0}(y|H_0)} \quad (16.1.27)$$

where the terms  $f_{Y|H_j}(y|H_j)$  also denote the likelihood of hypothesis  $H_j$  being true given the observation  $y$ .

Thus the likelihood ratio indicates which hypothesis  $H_i$  is more likely to have occurred.

Note that  $\Lambda(y)$  is also a function of  $y$ , thus since  $Y$  is a random variable we can think of  $\Lambda(Y)$  also as a random variable and we can find out its p.d.f. using the standard techniques of finding the p.d.f. of a function of a random variable.

Further we can also find out conditional p.d.f. of  $\Lambda(y)$  given that  $H_0$  or  $H_1$  is true.

Note that while  $Y \in \mathbb{R}^N$  might be a random vector,  $\Lambda(Y)$  is always a scalar random variable. Naturally if  $Y$  is a vector then  $f_{Y|H_j}(y|H_j)$  indicates the joint density of elements of  $Y$  namely  $(Y_1, \dots, Y_N)$ .

We denote the term on the r.h.s. in (16.1.26) as a threshold

$$\eta = \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}. \quad (16.1.28)$$

Note that  $\eta$  is a non-negative quantity.

Thus our test becomes

$$\Lambda(y) \stackrel{H_1}{\underset{H_0}{\gtrless}} \eta. \quad (16.1.29)$$

**Definition 16.8** The equation (16.1.29) is known as **likelihood ratio test** (LRT).

### Example 16.6: Likelihood ratio test

For minimum probability of error receiver

$$\eta = \frac{P_0}{P_1}.$$

For digital communication problem we also have

$$P_1 = P_0 = 0.5$$

Thus  $\eta = 1$ .

The LRT becomes

$$\Lambda(y) \underset{H_0}{\overset{H_1}{\gtrless}} 1.$$

□

Since natural logarithm is a monotonically increasing function hence taking  $\ln$  on both sides, retains the essentials of the rule.

Hence an equivalent decision rule is

$$\ln \Lambda(y) \underset{H_0}{\overset{H_1}{\gtrless}} \ln \eta. \quad (16.1.30)$$

**Definition 16.9** The equation (16.1.30) is known as **log-likelihood ratio test**.

### Example 16.7: Log-likelihood ratio test

For minimum probability of error receiver

$$\ln \eta = \ln(P_0) - \ln(P_1).$$

With

$$P_1 = P_0 = 0.5$$

we have  $\ln \eta = 0$ .

The Log-LRT becomes

$$\ln \Lambda(y) \underset{H_0}{\overset{H_1}{\gtrless}} 0.$$

□

**Example 16.8: Digital communication system with constant voltage output and additive white Gaussian noise** A simple digital communication system is defined as follows.

- Under  $H_0$  source produces a voltage of 0.
- Under  $H_1$  source produces a voltage of  $m$ .

- The channel introduces additive white Gaussian noise denoted by a random variable  $N$  with zero mean and variance  $\sigma^2$ .
- The received signal can be expressed as

$$H_1 : Y = m + N$$

$$H_0 : Y = N$$

where  $Z = \mathbb{R}$ .

Skipping some of the details, the likelihood ratio is given by

$$\Lambda(y) = \exp\left(-\frac{m^2 - 2ym}{2\sigma^2}\right)$$

An equivalent LRT becomes

$$y \stackrel{H_1}{\gtrless} \frac{\sigma^2}{m} \ln \eta + \frac{m}{2} = \gamma$$

□

Essentially we are computing a r.v.  $\Lambda(y)$  from our observation vector  $y$  and comparing it with a threshold  $\eta$  for making our decisions.

Any r.v. which is computed from the observed data is known as a statistic.

**Definition 16.10** A statistic is called a **sufficient statistic** (in our case for the binary hypothesis testing problem) if it provides sufficient information for making the binary decision. i.e. no other statistic from the observations can provide any additional information for making the decision.

Thus  $\Lambda(y)$  is a sufficient statistic for the LRT. When we simplify LRT, we may get simpler sufficient statistics also. In [definition 16.1.2](#) we found that  $y$  itself was a sufficient statistic.

### 16.1.3. Minimax criterion

The Bayes' criterion assigns costs to decisions and assumes that a priori probabilities of null and alternate hypotheses are known in advance. These assumptions may not be valid in many situations. Hence Bayes' criterion won't be applicable.

Let us assume that the costs are known but the a priori probabilities are not known.

Since  $P_0 = 1 - P_1$ , we can write risk as a function of  $P_1$  as follows

$$\begin{aligned} \mathcal{R} = & C_{00}(1 - P_F) + C_{10}P_F + P_1[(C_{11} - C_{00}) \\ & + (C_{01} - C_{11})P_M - (c_{10} - C_{00})P_F]. \end{aligned} \quad (16.1.31)$$

Thus risk is a linear function of  $P_1$  if  $P_M$  and  $P_F$  do not change.

Assuming a fixed value of  $P_1 = P_1^f$  we can design a Bayes' test as developed in the previous section.

This test is given by

$$\Lambda(y) \stackrel{H_1}{\gtrless} \frac{(1 - P_1^f)(C_{10} - C_{00})}{P_1^f(C_{01} - C_{11})}. \quad (16.1.32)$$

This Bayes' test is optimal for  $P_1 = P_1^f$  but it becomes suboptimal for any other value of  $P_1$ .

But this choice of  $P_1 = P_1^f$  freezes the observation partition  $Z_0$  and  $Z_1$  thus the false alarm rate and miss rate also get fixed to  $P_F = P_F^f$  and  $P_M = P_M^f$ .

Thus risk becomes a linear function of  $P_1$ .

In the following, we will denote the optimal risk for a given  $P_1$  as  $\mathcal{R}^o$ .

Let us consider the optimal Bayes' test for extreme values of  $P_1$ .

If we take  $P_1 = 0$ , then the LRT (16.1.29) reduces to

$$\Lambda(y) \stackrel{H_1}{\underset{H_0}{\gtrless}} \infty. \quad (16.1.33)$$

but since  $\Lambda(y)$  is a finite quantity, hence we always decide  $H_0$  to be true.

Thus  $Z = Z_0$ . Hence  $P_M = 1$  and  $P_F = 0$ . i.e. we never make a false alarm and we always miss a target.

Putting these values back in (16.1.31) we get

$$\mathcal{R}^o = C_{00}.$$

If we take  $P_1 = 1$ , then the LRT (16.1.29) reduces to

$$\Lambda(y) \stackrel{H_1}{\underset{H_0}{\gtrless}} 0. \quad (16.1.34)$$

thus we always decide  $H_1$  to be true (since  $\Lambda(y)$  by definition is non-negative).

Thus  $Z = Z_1$ . Hence  $P_M = 0$  and  $P_F = 1$ . i.e. we never miss a target and we always make a false alarm when the target is not present.

Putting these values back in (16.1.31) we get

$$\mathcal{R}^o = C_{11}.$$

Since these Bayes' tests are optimal, hence the optimal risk  $\mathcal{R}^o$  is always lower than the risk decided by the Bayes test for a fixed  $P_1^f$ .

Thus if we consider both optimal and suboptimal risk as functions of  $P_1$ , we have the equation,

$$\mathcal{R}^o(P_1) \leq \mathcal{R}_{P_1^f}(P_1).$$

But  $\mathcal{R}_{P_1^f}$  is a linear function of  $P_1$ , hence the optimal risk  $\mathcal{R}^o$  is a concave function of  $P_1$ .

Thus there exists a priori probability  $P_1 = P_1^*$  at which the optimum risk  $\mathcal{R}^o$  is maximum.

A Bayes test designed with  $P_1 = P_1^*$  thus minimizes the maximum risk.

Since the optimal risk function has a maximum at  $P_1 = P_1^*$  hence, the linear risk function  $\mathcal{R}_{P_1^*}$  has a slope equal to 0.

This gives us the minimax equation:

$$(C_{11} - C_{00}) + (C_{01} - C_{11})P_M - (c_{10} - C_{00})P_F = 0. \quad (16.1.35)$$

If the cost of correct decision is 0 (i.e.  $C_{00} = C_{11} = 0$ ), then the minimax equation for  $P_1 = P_1^*$  reduces to

$$C_{01}P_M = C_{10}P_F. \quad (16.1.36)$$

Furthermore if the cost of wrong decision is 1 (i.e.  $C_{10} = C_{01} = 1$ ), then the probability of false alarm equals the probability of miss i.e.

$$P_F = P_M. \quad (16.1.37)$$

#### 16.1.4. Neyman-Pearson criterion

Finally we come to Neyman-Pearson criterion which makes least of assumptions.

We don't assume any a priori probabilities. We don't assume any cost assignments for the decisions. So neither Bayes' nor minimax criteria are useful.

Since  $P_F$  and  $P_D$  are conditional probabilities, we can still work with them.

In **Neyman-Pearson test** we attempt to maximize  $P_D$  while keeping  $P_F$  bounded by some predefined value  $\alpha$ .

Since  $P_M = 1 - P_D$ , hence maximizing  $P_D$  is equivalent to minimizing  $P_M$ .

Design of the test takes the form of an optimization problem which can be stated as

$$\begin{aligned} & \text{minimize} && P_M \\ & \text{subject to} && P_F \leq \alpha. \end{aligned} \quad (16.1.38)$$

where the objective function is  $P_M$  and the (only) constraint is  $P_F \leq \alpha$ .

We construct the Lagrangian for this optimization problem as

$$J = P_M + \lambda(P_F - \alpha) \quad (16.1.39)$$

where  $\lambda \geq 0$  is the Lagrange multiplier.

Since  $P_F - \alpha \leq 0$  and  $\lambda \geq 0$ , hence for any feasible solution to this optimization problem we have

$$J \leq P_M.$$

After some algebraic manipulations, we can rewrite it as

$$\begin{aligned} J &= \int_{Z_0} f_{Y|H_1}(y|H_1) dy + \lambda \left[ 1 - \int_{Z_0} f_{Y|H_0}(y|H_0) dy - \alpha \right] \\ &= \lambda(1 - \alpha) + \int_{Z_0} [f_{Y|H_1}(y|H_1) - \lambda f_{Y|H_0}(y|H_0)] dy \end{aligned} \quad (16.1.40)$$

Clearly  $J$  is minimized when observations for which

$$f_{Y|H_1}(y|H_1) > \lambda f_{Y|H_0}(y|H_0)$$

are assigned to  $Z_1$ .

Thus the decision rule becomes

$$\Lambda(y) = \frac{f_{Y|H_1}(y|H_1)}{f_{Y|H_0}(y|H_0)} \stackrel{H_1}{\gtrless} \lambda. \quad (16.1.41)$$

This looks pretty familiar and is similar to the LRT derived for Bayes' criterion.

Only thing remaining is to find the value of Lagrangian multiplier  $\lambda$ .

We note that fixing a value of  $\lambda$  fixes the LRT and hence fixes the value of  $P_F$  also.

We recall that

$$\begin{aligned} P_F &= \mathbb{P}(D_1|H_0) = \mathbb{P}(\Lambda(y) \geq \lambda|H_0) \\ &= \int_{\lambda}^{\infty} f_{\Lambda(Y)|H_0}(\Lambda(y)|H_0)d\lambda \end{aligned} \quad (16.1.42)$$

Thus we choose  $\lambda$  by solving the following optimization problem:

$$\begin{aligned} &\text{maximize} && \lambda \\ &\text{subject to} && \lambda \geq 0 \\ & && P_F = \int_{\lambda}^{\infty} f_{\Lambda(Y)|H_0}(\Lambda(y)|H_0)d\lambda \leq \alpha. \end{aligned} \quad (16.1.43)$$

## 16.2. Detection with compressed measurements

We now turn our attention to the problem of signal detection with compressed measurements. The presentation in this section is largely based on [18].

We pick up one of the simplest detection problems.

We wish to detect the presence of a signal  $s \in \mathbb{R}^N$  in the presence of AWG noise.

The null and alternate hypotheses are described below

$$\begin{aligned} H_0 : Y &= \Phi G \\ H_1 : Y &= \Phi(s + G) \end{aligned} \quad (16.2.1)$$

where  $G \sim \mathcal{N}(0, \sigma^2 I_N)$  is i.i.d. Gaussian noise vector and  $\Phi \in \mathbb{R}^{M \times N}$  is a known measurement matrix.

$Y \in \mathbb{R}^M$  is the measurement random vector.

One particular realization of the random vectors will be given by

$$\begin{aligned} H_0 : y &= \Phi g \\ H_1 : y &= \Phi(s + g) \end{aligned} \quad (16.2.2)$$

with  $g \in \mathbb{R}^N$  and  $y \in \mathbb{R}^M$ .

If  $s$  is known in advance, then the optimal choice for  $\Phi = s^T \in \mathbb{R}^{1 \times N}$ . This is known as the **matched filter**.

This design has limitations in some situations.

- The design is matched to a specific signal  $s$ . If the signal changes, the detector design becomes useless. Essentially the detector assumes a lot of a priori information about the signal vector  $s$ .
- The matched filter design doesn't work if we wish to design a detector which could work for a large class of signals.
- The design cannot handle distortions in the signal well.
- The design is focused on solving only one problem i.e. the detection of presence or absence of signal. The design cannot be used for any other inferencing applications.

An alternative approach is to use a sensing matrix  $\Phi \in \mathbb{R}^{M \times N}$  which computes random projections of the received signal  $s + N$ .

This design provides a number of additional features.

- Same sensing matrix can be used for a large class of signals  $s$ . Thus  $\Phi$  is universal.
- $\Phi$  is agnostic to the choice of  $s$ , as it makes very weak assumptions about  $s$ .
- Thus the detection hardware becomes highly flexible and reusable for a number of situations. It allows us to evolve  $s$  over time.
- The design becomes much more robust w.r.t. distortions in the signal.
- The observation vector  $Y \in \mathbb{R}^M$  can be used for other inferencing tasks apart from the detection problem also.

**Implementation note:**

- In the matched filter design, we can assume that a continuous signal  $s(t) + G(t)$  has been uniformly sampled and then its dot product with the stored discretized version of  $s^T$  is computed. Alternatively the matched filter can be implemented in analog domain directly. Naturally analog implementation is extremely tied to the shape of  $s(t)$ .
- In the CS case, the design would involve some kind of random demodulator, which will compute  $\Phi(s + G)$  directly during sampling. Thus we don't go through the two steps of first sampling and then computing the matrix vector product. Note that the random demodulator design will depend on  $\Phi$  and will not depend on specific signal  $s(t)$ .

### 16.2.1. Theory

We note that  $Y$  is a Gaussian random vector under both hypotheses. In particular

- $Y \sim \mathcal{N}(0, \sigma^2 \Phi \Phi^T)$  under  $H_0$ .
- $Y \sim \mathcal{N}(\Phi s, \sigma^2 \Phi \Phi^T)$  under  $H_1$ .

Let us define

$$K = \sigma^2 \Phi \Phi^T. \quad (16.2.3)$$

Thus the conditional density functions are given by

$$f_{Y|H_0}(y|H_0) = \frac{1}{(2\pi)^{M/2} \sqrt{|K|}} \exp\left(-\frac{1}{2} y^T K^{-1} y\right). \quad (16.2.4)$$

and

$$f_{Y|H_1}(y|H_1) = \frac{1}{(2\pi)^{M/2} \sqrt{|K|}} \exp\left(-\frac{1}{2} (y - \Phi s)^T K^{-1} (y - \Phi s)\right). \quad (16.2.5)$$

### 16.2.2. Neyman-Pearson criterion

In this section, we will develop NP criterion for solving the detection problem in a CS setting. Basic theory for NP criterion was developed in subsection 16.1.4.

We assume that costs of making detection decisions cannot be assigned and a priori probabilities  $P_1$  and  $P_0$  are not known.

We will maximize  $P_D$  subject to the constraint  $P_F \leq \alpha$ .

The likelihood ratio test (LRT) for NP criterion is

$$\Lambda(y) = \frac{f_{Y|H_1}(y|H_1)}{f_{Y|H_0}(y|H_0)} \stackrel{H_1}{\gtrless} \lambda \quad (16.2.6)$$

where  $\lambda$  is obtained by solving the inequality

$$P_F = \int_{\Lambda(y) > \lambda} f_{Y|H_0}(y|H_0) dy \leq \alpha. \quad (16.2.7)$$

By taking the logarithm and simplifying an equivalent test is given by

$$y^T(\Phi\Phi^T)^{-1}\Phi s \stackrel{H_1}{\gtrless} \sigma^2 \log(\lambda) + \frac{1}{2}s^T(\Phi\Phi^T)^{-1}\Phi s \equiv \gamma. \quad (16.2.8)$$

where we have defined  $\gamma$  as a simplified threshold parameter.

We define our compressed detector as

$$t \triangleq y^T(\Phi\Phi^T)^{-1}\Phi s. \quad (16.2.9)$$

Thus the LRT test simplifies to

$$t \stackrel{H_1}{\gtrless} \gamma. \quad (16.2.10)$$

It can be shown that  $t$  is a *sufficient statistic* for this binary hypothesis problem.

We note that both  $t$  and  $\gamma$  depend on the signal vector  $s$ .

Assuming that  $\Phi$  is a full rank matrix (thus  $\Phi\Phi^T$  is invertible), we now define

$$P_{\Phi^T} = \Phi^T(\Phi\Phi^T)^{-1}\Phi. \quad (16.2.11)$$

Clearly

$$P_{\Phi^T}^2 = P_{\Phi^T}$$

and

$$P_{\Phi^T}^T = P_{\Phi^T}$$

hence  $P_{\Phi^T}$  is an orthogonal projection operator (see ??) on the row space of  $\Phi$ .

The row space of  $\Phi$  is given by

$$\mathcal{R}(\Phi) = \{\Phi^T x \mid x \in \mathbb{R}^M\}. \quad (16.2.12)$$

In particular

$$\Phi P_{\Phi^T} = \Phi\Phi^T(\Phi\Phi^T)^{-1}\Phi = \Phi$$

Alternatively

$$P_{\Phi^T}\Phi^T = \Phi^T.$$

Thus if we rewrite  $\Phi$  as

$$\begin{bmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_M^T \end{bmatrix} \quad (16.2.13)$$

where  $\phi_i \in \mathbb{R}^N$  are the  $M$  row vectors of  $\Phi$ , then we have

$$\begin{bmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_M^T \end{bmatrix} P_{\Phi^T} = \begin{bmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_M^T \end{bmatrix}. \quad (16.2.14)$$

Thus

$$\phi_i^T P_{\Phi^T} = \phi_i^T.$$

Taking transpose on both sides we get

$$P_{\Phi^T} \phi_i = \phi_i.$$

Thus  $P_{\Phi^T}$  preserves the row space of  $\Phi$ .

With this notation in place we have

$$\begin{aligned} s^T \Phi^T (\Phi \Phi^T)^{-1} \Phi s &= s^T P_{\Phi^T} s = s^T P_{\Phi^T}^2 s = s^T P_{\Phi^T} P_{\Phi^T} s \\ &= s^T P_{\Phi^T}^T P_{\Phi^T} s = (P_{\Phi^T} s)^T (P_{\Phi^T} s) = \|P_{\Phi^T} s\|_2^2. \end{aligned} \quad (16.2.15)$$

We now look back at our sufficient statistic  $t$  as defined in (16.2.9).

Under  $H_0$  we have

$$t = (\Phi g)^T (\Phi \Phi^T)^{-1} \Phi s = g^T \Phi^T (\Phi \Phi^T)^{-1} \Phi s. = \langle g, P_{\Phi^T} s \rangle \quad (16.2.16)$$

Thus  $t$  is an instance of a Gaussian r.v.  $T$  with

$$\mathbb{E}(T) = 0.$$

and

$$\text{Var}(T) = \sigma^2 \|P_{\Phi^T} s\|_2^2.$$

Under  $H_1$  we have

$$\begin{aligned} t &= (\Phi(s + g))^T (\Phi \Phi^T)^{-1} \Phi s \\ &= g^T \Phi^T (\Phi \Phi^T)^{-1} \Phi s + s^T \Phi^T (\Phi \Phi^T)^{-1} \Phi s \\ &= g^T \Phi^T (\Phi \Phi^T)^{-1} \Phi s + \|P_{\Phi^T} s\|_2^2 \\ &= \langle g, P_{\Phi^T} s \rangle + \|P_{\Phi^T} s\|_2^2. \end{aligned} \quad (16.2.17)$$

Thus  $t$  is an instance of a Gaussian r.v.  $T$  with

$$\mathbb{E}(T) = \|P_{\Phi^T} s\|_2^2.$$

and

$$\text{Var}(T) = \sigma^2 \|P_{\Phi^T} s\|_2^2.$$

In summary

$$T \sim \begin{cases} \mathcal{N}(0, \sigma^2 \|P_{\Phi^T} s\|_2^2) & \text{under } H_0 \\ \mathcal{N}(\|P_{\Phi^T} s\|_2^2, \sigma^2 \|P_{\Phi^T} s\|_2^2) & \text{under } H_1 \end{cases} \quad (16.2.18)$$

Thus we have the false alarm rate given by

$$P_F = \mathbb{P}_{T|H_0}(t > \gamma | H_0) = Q\left(\frac{\gamma}{\sigma \|P_{\Phi^T} s\|_2}\right) \quad (16.2.19)$$

and the detection rate given by

$$P_D = \mathbb{P}_{T|H_1}(t > \gamma | H_1) = Q\left(\frac{\gamma - \|P_{\Phi^T} s\|_2^2}{\sigma \|P_{\Phi^T} s\|_2}\right) \quad (16.2.20)$$

The threshold  $\gamma$  is given by

$$\gamma = \sigma \|P_{\Phi^T} s\|_2 Q^{-1}(\alpha) \quad (16.2.21)$$

Putting it back we get

$$P_D = Q\left(Q^{-1}(\alpha) - \frac{\|P_{\Phi^T} s\|_2}{\sigma}\right). \quad (16.2.22)$$

For the special case when  $\Phi = s^T$ , we have

$$P_{\Phi^T} = P_s = s(s^T s)^{-1} s^T. \quad (16.2.23)$$

Thus

$$P_s s = s(s^T s)^{-1} s^T s = s. \quad (16.2.24)$$

Thus for the matched filter:

$$P_F = Q\left(\frac{\gamma}{\sigma \|s\|_2}\right) \quad (16.2.25)$$

$$P_D = Q\left(Q^{-1}(\alpha) - \frac{\|s\|_2}{\sigma}\right). \quad (16.2.26)$$

Thus we see that  $P_D$  for the general  $\Phi$  varies w.r.t. the matched filter case based on the difference between  $\|s\|_2$  and  $\|P_{\Phi^T} s\|_2$ .

Since  $P_{\Phi^T} s$  is nothing but the orthogonal projection of  $s$  on to the row space of  $\Phi$ , hence

$$\|P_{\Phi^T} s\|_2 \leq \|s\|_2 \quad (16.2.27)$$

Thus if  $\|P_{\Phi^T} s\|_2$  is close to  $\|s\|_2$  then, the performance would be quite good but if  $\|P_{\Phi^T} s\|_2 \ll \|s\|_2$ , then the performance would be poor.

Thus in general performance can be quite good or poor depending on  $\Phi$ .

However if  $\Phi$  is a random matrix, then  $\|P_{\Phi^T} s\|_2$  strongly concentrates around  $\sqrt{\frac{M}{N}} \|s\|_2$ .

Before we quickly take a detour into the notion of stable embeddings in section 3.3.

Let us define

$$\text{SNR} \triangleq \frac{\|s\|_2^2}{\sigma^2}. \quad (16.2.28)$$

We can bound the performance of compressed detector as follows.

**Theorem 16.1** Suppose that  $\sqrt{\frac{N}{M}} P_{\Phi^T}$  provides a  $\delta$ -stable embedding of  $(S, \{0\})$ . Then for any  $s \in S$ , we can detect  $s$  with error rate

$$P_D(\alpha) \leq Q \left( Q^{-1}(\alpha) - \sqrt{1+\delta} \sqrt{\frac{M}{N}} \sqrt{\text{SNR}} \right) \quad (16.2.29)$$

and

$$P_D(\alpha) \geq Q \left( Q^{-1}(\alpha) - \sqrt{1-\delta} \sqrt{\frac{M}{N}} \sqrt{\text{SNR}} \right). \quad (16.2.30)$$

PROOF. By assumption  $\sqrt{\frac{N}{M}} P_{\Phi^T}$  provides a  $\delta$ -stable embedding of  $(S, \{0\})$ . Thus as per [definition 3.4](#) we have:

$$\sqrt{1-\delta} \|s\|_2 \leq \sqrt{\frac{N}{M}} \|P_{\Phi^T} s\|_2 \leq \sqrt{1+\delta} \|s\|_2 \quad \forall s \in S. \quad (16.2.31)$$

This implies

$$\sqrt{1-\delta} \sqrt{\frac{M}{N}} \sqrt{\text{SNR}} \leq \frac{\|P_{\Phi^T} s\|_2}{\sigma} \leq \sqrt{1+\delta} \sqrt{\frac{M}{N}} \sqrt{\text{SNR}} \quad \forall s \in S.$$

since  $Q$ -function is a decreasing function, substituting these bounds in (16.2.22), we get the result.

□

The natural question at this moment is how do we find matrices for which  $\sqrt{\frac{N}{M}} P_{\Phi^T}$  provides a  $\delta$ -stable embedding?

Consider a random  $M$  dimensional subspace of  $\mathbb{R}^N$  and consider  $\Phi$  having orthonormal rows spanning this subspace i.e.  $\Phi$  represents a **random orthogonal projection**.

Then  $\Phi\Phi^T = I$ .

Thus

$$P_{\Phi^T} = \Phi^T \Phi.$$

Hence

$$\|P_{\Phi^T} s\|_2^2 = \|\Phi^T \Phi s\|_2^2 = (\Phi^T \Phi s)^T \Phi^T \Phi s = \|\Phi s\|_2^2.$$

Thus

$$\|P_{\Phi^T} s\|_2 = \|\Phi s\|_2.$$

For random orthogonal projections, it is known that

$$(1 - \delta) \frac{M}{N} \|s\|_2^2 \leq \|P_{\Phi^T} s\|_2^2 \leq (1 + \delta) \frac{M}{N} \|s\|_2^2 \quad (16.2.32)$$

with probability exceeding  $1 - 2 \exp(-cM\delta^2)$ .

# APPENDIX A

## Useful MATLAB Functions

### A.1. General purpose utilities

#### normalizeColumns

```
function [ A ] = normalizeColumns( A )
2 %NORMALIZECOLUMNS Normalizes all columns of A
    columnNorms = columnWiseNorm(A);
4 numColumns = size(A, 2);
    for i=1:numColumns
6        columnNorm = columnNorms(i);
        if 0 == columnNorm
8            continue
        end
10       A(:, i) = A(:, i) / columnNorm;
    end
```

LISTING A.1. normalizeColumns.m

### A.2. Functions for generating signal patterns

#### simpleSparseVector

```
function [ x ] = simpleSparseVector( N, K )
2 %SIMPLESPARSEVECTOR Constructs a simple sparse vector
    % N : number of elements in the vector
4 % K : number of non-zero elements
    % x : resultant vector
6
    % Let us construct a zero vector
8 x = zeros(N,1);
    % let us generate a random permutation of numbers from 1 to n
10 q = randperm(N);
```

```
12 % let us put a value in first k positions as identified
12 % in the random permutation
13 x(q(1:K)) = sign(randn(K,1));
14 end
```

LISTING A.2. simpleSparseVector.m

**Example 1.1: Using simpleSparseVector**

```
>> x = simpleSparseVector(10, 5);
>> x'
-1     0     0     0    -1     1    -1     0     1     0
```



## Bibliography

- [1] Michal Aharon, Michael Elad, and Alfred M Bruckstein. K-svd and its non-negative variant for dictionary design. In *Optics & Photonics 2005*, pages 591411–591411. International Society for Optics and Photonics, 2005.
- [2] Michal Aharon, Michael Elad, and Alfred M Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear algebra and its applications*, 416(1):48–67, 2006.
- [3] A Bandeira, Edgar Dobriban, D Mixon, and W Sawin. Certifying the restricted isometry property is hard. 2012.
- [4] Richard Baraniuk, M Davenport, M Duarte, and Chinmay Hegde. An introduction to compressive sensing. *Connexions e-textbook*, 2011.
- [5] Dror Baron, Marco F Duarte, Michael B Wakin, Shriram Sarvotham, and Richard G Baraniuk. Distributed compressive sensing. *arXiv preprint arXiv:0901.3403*, 2009.
- [6] E. van den Berg, M. P. Friedlander, G. Hennenfent, F. Herrmann, R. Saab, and Ö. Yılmaz. Sparco: A testing framework for sparse reconstruction. Technical Report TR-2007-20, Dept. Computer Science, University of British Columbia, Vancouver, October 2007.
- [7] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.
- [8] Emmanuel J Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

- [9] Emmanuel J Candes and Justin Romberg. Practical signal recovery from random projections. *Wavelet Applications in Signal and Image Processing XI Proc. SPIE Conf.* 5914., 2004.
- [10] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.
- [11] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.
- [12] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.
- [13] Jie Chen and Xiaoming Huo. Theoretical results on sparse representations of multiple-measurement vectors. *Signal Processing, IEEE Transactions on*, 54(12):4634–4643, 2006.
- [14] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.
- [15] Shane F Cotter, Bhaskar D Rao, Kjersti Engan, and Kenneth Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *Signal Processing, IEEE Transactions on*, 53(7):2477–2488, 2005.
- [16] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of the johnson-lindenstrauss lemma. *International Computer Science Institute, Technical Report*, pages 99–006, 1999.
- [17] Mark A Davenport and Michael B Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *Information Theory, IEEE Transactions on*, 56(9):4395–4401, 2010.
- [18] Mark A Davenport, Petros T Boufounos, Michael B Wakin, and Richard G Baraniuk. Signal processing with compressive measurements. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):445–460, 2010.

- [19] Mike E Davies and Yonina C Eldar. Rank awareness in joint sparse recovery. *Information Theory, IEEE Transactions on*, 58(2):1135–1146, 2012.
- [20] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via  $l_1$  minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.
- [21] Michael Elad. *Sparse and redundant representations*. Springer, 2010.
- [22] Michael Elad and Alfred M Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *Information Theory, IEEE Transactions on*, 48(9):2558–2567, 2002.
- [23] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE, 1999.
- [24] Rémi Gribonval and Morten Nielsen. Sparse representations in unions of bases. *Information Theory, IEEE Transactions on*, 49(12):3320–3325, 2003.
- [25] Rémi Gribonval, Holger Rauhut, Karin Schnass, and Pierre Vandergheynst. Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms. *Journal of Fourier analysis and Applications*, 14(5-6):655–687, 2008.
- [26] Dany Leviatan and Vladimir N Temlyakov. Simultaneous greedy approximation in banach spaces. *Journal of Complexity*, 21(3):275–293, 2005.
- [27] Dany Leviatan and Vladimir N Temlyakov. Simultaneous approximation by greedy algorithms. *Advances in Computational Mathematics*, 25(1-3):73–90, 2006.
- [28] Adam Lutoborski and Vladimir N Temlyakov. Vector greedy algorithms. *Journal of Complexity*, 19(4):458–473, 2003.

- [29] Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.
- [30] Ron Rubinstein, Alfred M Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.
- [31] Vladimir N Temlyakov. A remark on simultaneous greedy approximation. *East journal on approximations*, 10(1):17–25, 2004.
- [32] Charles W Therrien. *Discrete random signals and statistical signal processing*. Prentice Hall PTR, 1992.
- [33] Ivana Tasic and Pascal Frossard. Dictionary learning. *Signal Processing Magazine, IEEE*, 28(2):27–38, 2011.
- [34] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, 2004.
- [35] JOEL A TROPP. Just relax: convex programming methods for subset selection and sparse approximation. 2004.
- [36] JOEL A Tropp. *Topics in sparse approximation*. PhD thesis, The University of Texas at Austin, August 2004.
- [37] Joel A Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *Information Theory, IEEE Transactions on*, 52(3):1030–1051, 2006.
- [38] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.
- [39] Joel A Tropp, Anna C Gilbert, and Martin J Strauss. Simultaneous sparse approximation via greedy pursuit. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP'05). IEEE International Conference on*, volume 5, pages v–721. IEEE, 2005.
- [40] Joel A Tropp, Anna C Gilbert, and Martin J Strauss. Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. *Signal Processing*, 86(3):572–588, 2006.

- [41] Harry L Van Trees. *Detection, estimation, and modulation theory*. Wiley-Interscience, 2004.

## Index

- ( $\mathcal{D}, K$ )-EXACT-SPARSE, 51, 53
- ( $\mathcal{D}, K$ )-SPARSE approximation, 51, 53
- ( $\mathcal{D}, K$ )-sparse, 49
- $K$ -term approximation, 59
- $\mu_{1/2}(G)$ , 162
- $p$ -compressible signal, 64
- Admissible sensing matrix, 309
- Alternate hypothesis, 473
- Analysis matrix, 53
- Analysis operator, 26
- Antipodal convex hull, 172
- Atom, 48
- Babel function, 81
- Babel function upper bound, 83
- Bayes risk, 479
- Bi-Lipschitz mapping, 156
- Binary hypothesis testing, 473
- Coherence of a matrix, 74
- Coherence of dictionary, 72
- Common / innovations joint sparsity model, 458
- complementary p-Babel function, 166
- Complete dictionary, 49
- Conditional sparsity, 468
- Desirability, 13
- Detection rate, 477
- Dictionary, 48
- Dirac basis, 27
- Dirac Fourier basis, 42
- Dirac-DCT dictionary, 209
- Exact recovery coefficient, 168
- Exact recovery condition for OMP, 295
- Explanation signal, 181
- False alarm rate, 477
- Feasible location matrices, 457
- Fourier basis, 28
- Gaussian sensing matrix, 215
- Gram matrix, 72
- Grassmannian frame, 74
- Greedy selection ratio, 295
- Incoherent dictionary, 73
- Johnson-Lindenstrauss theorem, 147
- Joint correlation, 309
- Joint sparsity, 391, 397
- Joint sparsity level, 457
- Joint sparsity model, 457
- Joint sparsity of subset of signals, 468
- Largest entries approximation, 60

- Likelihood ratio, 483
- Likelihood ratio test, 483
- Log likelihood ratio test, 484
- Low distortion embedding, 147
  - Maximum correlation, 259
  - Measurement space, 91
  - Measurement vector, 91
  - Minimum probability of error receiver, 480
  - Miss rate, 477
  - Mutual coherence, 72
  - Mutual coherence of two orthonormal bases, 31
  - Norm dominance, 398
  - normalizeColumns, 499
  - Null hypothesis, 473
  - Null space property, 184
  - Observation space, 474
  - OMP, 288
  - Optimal RIP constant, 145
  - Orthogonal Matching Pursuit, 288
  - Orthonormal analysis equation, 26
  - Orthonormal synthesis equation, 26
  - Over-complete dictionary, 49
  - Overlap size, 467
  - p-Babel function, 165, 166
  - Probability of correct decision, 478
  - Probability of detection, 477
  - Probability of error, 478
  - Probability of false alarm, 477
  - Probability of miss, 477
  - Proximity of two orthonormal bases, 31
  - Quasi-incoherent dictionary, 90
  - Rademacher sensing matrix, 210
  - Redundant dictionary, 49
  - Regularization, 13
  - Representation matrix support, 391
  - Restricted almost orthonormal system, 111
  - Restricted Isometry Constant, 111
  - Restricted Isometry Property, 111
  - Restriction of a matrix on an index set, 62
  - Restriction of a signal on an index set, 58
  - Row column norms, 387
  - Row support, 391
  - Row- $l_0$ -“norm”, 391
  - Sensing matrix, 91, 92
  - Sensing vector, 92
  - Sign vector, 54
  - Signal matrix support, 397
  - Signal space, 91
  - simpleSparseVector, 500
  - Smallest singular value, 309
  - Spark, 68, 238
  - Spark lower bound, 74
  - Sparse approximation, 50
  - Sparse signal, 9
  - Sparsity level, 456
  - Sparsity rank, 391
  - Stability of recovery algorithm, 189
  - Stable embedding, 156
  - Sub-dictionary, 67
  - Sufficient statistic, 485
  - Synthesis matrix, 52
  - Synthesis operator, 26
  - Target detection, 476
  - Target false alarm, 476
  - Target miss, 476
  - Two-ortho basis, 44

Uncertainty principle for two ortho  
bases, **30**

Uniqueness of a sparse  
representation, **47**

Uniqueness Spark, **70**

Uniqueness-Coherence, **76**

Unrecoverable energy in CoSaMP,  
**347**