An Introduction to Compressed Sensing

Mathukumalli Vidyasagar

March 8, 2016

Contents

Chapter 1

Compressed Sensing: Problem Formulation

Compressed sensing refers to the problem of recovering a relatively sparse entity from a limited number of measurements. Examples of "sparse entities" include high-dimensional vectors with very few nonzero elements, high-dimensional but low rank matrices, images with very few "frequency" components, and others. The term "compressed sensing" (or its alternate "compressive sensing") is of relatively recent origin, as are the various problem formulations studied in this book. However, as is often the case in mathematical research, compressed sensing theory as presented here has antecedents that go back a couple of decades if not earlier. In the initial phases, the research literature in compressed sensing drew upon relatively advanced mathematics. However, in recent years, the proofs of the basic results have been simplified to such an extent that practically all of compressed sensing theory can be presented with just undergraduate linear algebra as a prerequisite. This is the approach adopted in this book.

In this chapter, we give a precise formulation of two basic problems in compressed sensing, namely robust signal recovery and the matrix completion problem. The contents of the book go beyond these two basic problem formulations; yet these two are the core problems, and the rest can be thought of as variations of the basic problems. In order to state these problems precisely, it is necessary first to introduce appropriate notation, as well as various concepts from linear algebra. Some of these concepts are very basic and can be found in almost any book on the topic, while others are of recent origin and require formal statements and proofs. Thus the reader is urged to read this chapter carefully, and ensure that the foundations are in place before moving on to the remaining chapters. In particular, if the reader is not familiar with some basic material from linear algebra, then those topics should be mastered before proceeding. After the background in linear algebra, we state two specific problems that form the subject of study for the bulk of the book. After this, some illustrative applications of compressed sensing are presented, to demonstrate the utility of the approach. The theoretical justification for these applications (and others) are given in subsequent chapters.

1.1 Mathematical Preliminaries

In this section we will begin by mentioning various elementary concepts from linear algebra that are used in the book. That part of the section is *not intended* as a detailed discussion. Rather, it is a catalog of topics with which the reader is expected to be familiar. Those who are encountering these topics for the first time should consult some standard references in linear algebra, for example [24]. Once the standard material ends, we will also introduce some fairly recent material on matrix norms etc. that would be used in subsequent chapters.

Throughout we consider the linear vector space $\mathbb{R}^n$. Much of the discussion below carries over with very little modification for the linear vector space $\mathbb{C}^n$. However, we do not use this generality because, for the most part, we will be working in $\mathbb{R}^n$. There are some situations, for example when a signal is sampled in the frequency domain instead of the time domain, when it may become necessary to deal with complex matrices. These situations are highlighted as and when they arise.

Definition 1.1. A function $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}_+$ is said to be a **norm** if it satisfies the following conditions:

N1. $\|x\| \geq 0 \; \forall x \in \mathbb{R}^n$, and $\|x\| = 0 \iff x = 0$.

N2. $\|\alpha x\| = |\alpha| \cdot \|x\|, \; \forall \alpha \in \mathbb{R}, \; \forall x \in \mathbb{R}^n$.

N3. $\|x + y\| \leq \|x\| + \|y\|, \; \forall x, y \in \mathbb{R}^n$.

Property 3 above is usually referred to as the "triangle inequality."

Among the most commonly used norms are the so-called ℓ_p-norms, defined by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \tag{1.1}$$

where $p \in [1, \infty)$. An important special case corresponds to $p = 2$, namely the norm

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}, \tag{1.2}$$

which is commonly known as the **Euclidean norm**. Note that it is not possible directly to substitute the value $p = \infty$ into the formula (1.1). However, it is not difficult to show that

$$\lim_{p \to \infty} \|x\|_p = \max_i |x_i|.$$

Accordingly we define

$$\|x\|_\infty = \max_i |x_i|, \tag{1.3}$$

so that the ℓ_p-norms are well-defined for all $p \in [1, \infty]$.

Figure 1.1 shows the "unit sphere" in $\mathbb{R}^2$ for $p = 1, 2, \infty$.

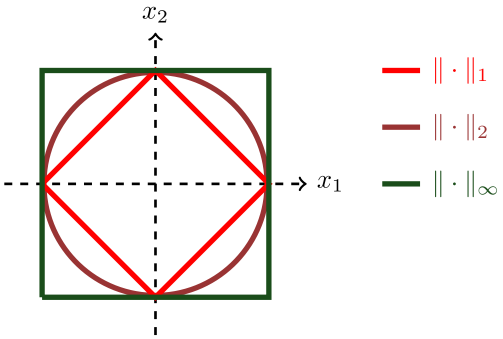

Figure 1.1: Unit Spheres in the ℓ_p-Norm for Various p.

Another useful notation is the symbol $\|x\|_0$ which is defined next. Given a vector $x \in \mathbb{R}^n$, define its **support** as the set of nonzero components. Thus

$$\text{supp}(x) := \{i : x_i \neq 0\}.$$

With this convention, we define

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

In words, $\|x\|_0$ equals the cardinality of the support of x. This usage is justified on the basis that

$$\lim_{p \to 0^+} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} = |\text{supp}(x)|.$$

Some authors refer to $\|x\|_0$ as the "ℓ_0-norm," though the quantity $\|x\|_0$ is not at all a norm. It does satisfy two out the three requirements of a norm. First,

$$\|x\|_0 = 0 \iff x = 0.$$

Second, $\| \cdot \|_0$ satisfies a triangle inequality, in that

$$\|x + z\|_0 \leq \|x\|_0 + \|z\|_0.$$

However, it is easy to see that for every $\alpha \neq 0$, we have

$$\|\alpha x\|_0 = \|x\|_0.$$

Therefore $\| \cdot \|_0$ is not a norm. Nevertheless, the symbol $\|x\|_0$ is a useful symbol to describe how sparse the vector x is.

Definition 1.2. A map $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is said to be an **inner product** if it satisfies the following conditions:

IP1. $\langle y, x \rangle = \langle x, y \rangle$.[1]

IP2. For all $\alpha, \beta \in \mathbb{R}$, $x, y, z \in \mathbb{R}^n$, we have that $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$.

IP3. $\langle x, x \rangle \geq 0 \; \forall x \in \mathbb{R}^n$.

IP4. $\langle x, x \rangle = 0 \iff x = 0$.

If $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ satisfies all of the above conditions, then the quantity defined by

$$\|x\| = \sqrt{\langle x, x \rangle}$$

can be shown to be a norm. This norm is said to be "induced" by the inner product. Specifically, if we define

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i,$$

then it can be verified that conditions (IP1) through (IP4) are satisfied, so that it is a valid inner product. Moreover, the norm induced by this inner product is the ℓ_2-norm.

Definition 1.3. Suppose $\| \cdot \|$ is a norm on $\mathbb{R}^n$ and $\langle \cdot, \cdot \rangle$ is an inner product. Then the **dual norm** of $\| \cdot \|$ with respect to the inner product $\langle \cdot, \cdot \rangle$ is denoted by $\| \cdot \|_d$, and is defined by

$$\|y\|_d := \max_{\|x\| \leq 1} |\langle y, x \rangle| = \max_{\|x\| \leq 1} \langle y, x \rangle. \tag{1.4}$$

Strictly speaking we should also include the inner product in the notation, because the dual norm of $\| \cdot \|$ under a different inner product could be different. However this is not done to keep the notation simple.

[1] In case the underlying linear vector space is $\mathbb{C}^n$ and not $\mathbb{R}^n$, this condition is replaced by $\langle y, x \rangle = \overline{\langle x, y \rangle}$.

Fact 1.1. *Under the standard inner product defined by $\langle x, y\rangle = y^t x$, the dual norm of $\| \cdot \|_p$ is the norm $\| \cdot \|_q$ where*

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{1.5}$$

If p, q satisfy (1.5), *then they are said to be* ***conjugate indices*** *of each other.*

It is obvious that (1.5) is symmetric in p and q. If $p \in (1, \infty)$, then (1.5) can be rewritten as $q = p/(p-1)$. In particular, if $p = 2$, then $q = 2$. Therefore the Euclidean norm is its own dual. If $p = \infty$, then $q = 1$ and vice versa.

A ready consequence of the duality of the norms $\| \cdot \|_p$ and $\| \cdot \|_q$ is given next.

Fact 1.2. *If $x, y \in \mathbb{R}^n$, then*

$$|\langle x, y\rangle| \leq \|x\|_p \cdot \|y\|_q, \ \forall x, y \in \mathbb{R}^n, \tag{1.6}$$

where p and q are conjugate indices. In particular,

$$|\langle x, y\rangle| \leq \|x\|_2 \cdot \|y\|_2, \ \forall x, y \in \mathbb{R}^n. \tag{1.7}$$

Equation (1.6) *is known as* ***Hölder's inequality****, whereas* (1.7) *is known as* ***Schwarz' inequality****.*

Note that some authors refer to (1.7) as the **Cauchy-Schwarz** inequality on the grounds that the great mathematician Augustin-Louis Cauchy had already discovered this inequality before Schwarz. However, in these notes we will stick with the phraseology "Schwarz' inequality."

Fact 1.3. *Suppose $x \in \mathbb{R}^n$. Then*

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2 \leq n\|x\|_\infty. \tag{1.8}$$

Proof: Suppose $\|x\|_\infty = |x_i|$. In other words, suppose that the i-th component of x has the largest magnitude; note that the index i need not be unique. Then

$$\|x\|_\infty = |x_i| = \sqrt{|x_i|^2} \leq \left(|x_i|^2 + \sum_{j \neq i} |x_j|^2 \right)^{1/2} = \|x\|_2.$$

This proves the left-most inequality. Next, let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ denote the canonical basis for $\mathbb{R}^n$. Thus $\mathbf{e}_i$ has 1 as its i-th component and 0 as the other components. Then we can reason as follows: Note that

$$x = \sum_{i=1}^n x_i \mathbf{e}_i.$$

Therefore, from the triangle inequality, it follows that

$$\|x\|_2 \leq \sum_{i=1}^n |x_i| \cdot \|\mathbf{e}_i\|_2 = \sum_{i=1}^n |x_i| = \|x\|_1.$$

Next, define $z \in \{-1, 1\}^n$ by $z_i = \text{sign}(x_i)$, and if $x_i = 0$, define z_i arbitrarily as either $+1$ or -1. Note that

$$\|x\|_1 = \sum_{i=1} x_i z_i.$$

Then it follows by Schwarz' inequality that

$$\|x\|_1 \leq \|x\|_2 \cdot \|z\|_2 = \sqrt{n}\|x\|_2.$$

Finally, observe that $|x_i| \leq \|x\|_\infty$ for all indices i. Therefore

$$\|x\|_2 = \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2} \leq \left(\sum_{i=1}^n \|x\|_\infty^2 \right)^{1/2} = \sqrt{n}\|x\|_\infty.$$

This is the right-most inequality. □

The next fact and subsequent material deals with a very useful concept known as the singular value decomposition.

Fact 1.4. *Suppose $A \in \mathbb{R}^{m \times n}$, and has rank r. Then there exist unit ℓ_2-norm vectors $u_1, \ldots, u_r \in \mathbb{R}^m$, $v_1, \ldots, v_r \in \mathbb{R}^n$, and numbers $\sigma_1 \geq \ldots \geq \sigma_r > 0$, such that (i) the vectors $u_1, \ldots, u_r$ are pairwise orthogonal, (ii) the vectors $v_1, \ldots, v_r$ are pairwise orthogonal, and (iii)*

$$A = \sum_{i=1}^r \sigma_i u_i v_i^t. \tag{1.9}$$

Remarks:

1. The numbers $\sigma_1, \ldots, \sigma_r$ are called the **singular values** of the matrix A, the vectors $u_1, \ldots, u_r$ the **row singular vectors**, and the vectors $v_1, \ldots, v_r$ the **column singular vectors**.

2. The relationship (1.9) can be expressed in a couple of different ways. First, we can define

   $$U = [\ u_1 \ | \ \ldots \ | \ u_r \ ] \in \mathbb{R}^{m \times r}, V = [\ v_1 \ | \ \ldots \ | \ v_r \ ] \in \mathbb{R}^{n \times r}, \Sigma = \text{Diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r},$$

   in which case (1.9) can be expressed as

   $$A = U \Sigma V^t. \tag{1.10}$$

   Alternatively, one can "pad" U and V into $m \times m$ and $n \times n$ orthogonal matrices, define $k = \min\{m, n\}$, extend the singular values by defining $\sigma_{r+1} = \ldots = \sigma_k = 0$ if $r < k$, and define $D = \text{Diag}(\sigma_1, \ldots, \sigma_k)$. Then one can write $A = U \Sigma V^t$, where

   $$\Sigma = D \text{ if } m = n, \tag{1.11}$$

   $$\Sigma = [\ D \quad \mathbf{0}_{m \times (n-m)} \ ] \text{ if } m < n, \tag{1.12}$$

   $$\Sigma = \begin{bmatrix} D \\ \mathbf{0}_{(m-n) \times n} \end{bmatrix} \text{ if } m > n. \tag{1.13}$$

   Here and elsewhere a bold-faced symbol $\mathbf{0}$ denotes a vector or matrix of zeros, while a non-bold symbol 0 denotes the scalar.

3. The above decomposition is known as a **singular value decomposition** of the matrix A. Some authors refer to (1.11) through (1.13) as the SVD and to (1.10) as the "reduced" SVD. Both representations are useful. Therefore this is the nomenclature adopted in this book.

Proof: The proof is by induction on r, the rank of A. If $r = 0$, then A is the zero matrix; but then the sum in (1.9) is empty and thus also equals zero. Suppose $r \geq 1$, and define $A_1 = A$. Choose a vector $v_1 \in \mathbb{R}^n$ such that $\|v_1\|_2 = 1$ and $\|A_1 v_1\|_2$ is maximized. Call the maximum value σ_1, and define

$$u_1 = \frac{1}{\sigma_1} A_1 v_1, A_2 = A_1 - \sigma_1 u_1 v_1^t. \tag{1.14}$$

Then $\|u_1\|_2 = 1$ by construction. If $A_2 = 0$ then we are done, so suppose A_2 is not the zero matrix. Choose $v_2 \in \mathbb{R}^n$ such that $\|v_2\|_2 = 1$ and $\|A_2 v_2\|_2$ is as large possible. Call the maximum σ_2, and define $u_2 = (1/\sigma_2) A_2 v_2$. Then we make the following claims:

1. $\text{rank}(A_2) = \text{rank}(A_1) - 1$.
2. $\sigma_2 \leq \sigma_1$.
3. $v_1^t v_2 = 0$.
4. $u_1^t u_2 = 0$.

Each of these claims is proved in turn.

Once v_1, v_1 are defined, choose any appropriate matrices $U_2 \in \mathbb{R}^{(m-1)\times m}, V_2 \in \mathbb{R}^{(n-1)\times n}$ such that $U = [u_1|U_2]$ and $V = [v_1|V_2]$ are orthogonal. Then

$$\begin{aligned} U^t A_1 V &= \begin{bmatrix} u_1^t \\ U_2^t \end{bmatrix} \begin{bmatrix} A_1 v_1 & A_1 V_2 \end{bmatrix} = \begin{bmatrix} u_1^t \\ U_2^t \end{bmatrix} \begin{bmatrix} \sigma_1 u_1 & A_1 V_2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 & b_1^t \\ \mathbf{0} & C_2 \end{bmatrix} =: M_1, \end{aligned} \tag{1.15}$$

because $U_2^t u_1 = \mathbf{0}$. Now it is shown that $b_1 = \mathbf{0}$. Suppose to the contrary that $b_1 \neq \mathbf{0}$, and define $v = [\sigma_1 \ \ b_1^t]^t$. Then

$$M_1 v = M_1 \begin{bmatrix} \sigma_1 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 + b_1^t b_1 \\ C_2 b_1 \end{bmatrix}.$$

Therefore

$$\|M_1 v\|_2 \geq \sigma_1^2 + b_1^t b_1,$$

while

$$\|v\|_2 = \sqrt{\sigma_1^2 + b_1^t b_1}.$$

Therefore

$$\frac{\|M_1 v\|_2}{\|v\|_2} \geq \sqrt{\sigma_1^2 + b_1^t b_1} > \sigma_1.$$

On the other hand, M_1 is obtained by pre- and post-multiplying A_1 by orthogonal matrices, and multiplying a vector by an orthogonal matrix does not change its Euclidean norm. Therefore

$$\|M_1 v\|_2 = \|U^t A_1 V v\|_2 = \|A_1 V v\|_2,$$

and as a result $\|M_1 v\|_2 \leq \sigma_1 \|Vv\|_2 \leq \sigma_1 \|v\|_2$ for all v, which contradicts the previous inequality. Therefore $b_1 = \mathbf{0}$, and it follows that

$$U^t A_1 V = \begin{bmatrix} \sigma_1 & \mathbf{0} \\ \mathbf{0} & C_2 \end{bmatrix} =: M_1. \tag{1.16}$$

Claim 1: Multiplying a matrix by an orthogonal matrix does not change its rank. Therefore the matrix M_1 has rank r, from which it follows that C_2 has rank $r - 1$.

Claim 2: Note that

$$U^t A_2 V = U^t (A_1 - \sigma_1 u_1 v_1^t) V = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & C_2 \end{bmatrix}.$$

Suppose $\sigma_2 > \sigma_1$. Then there exists a vector $z \in \mathbb{R}^{n-1}$ such that $\|z\|_2 = 1$ and $\|C_2 z\|_2 = \sigma_2 > \sigma_1$. Now define $\bar{z} = [0 \ \ z^t]^t$, and $v = V\bar{z}$. Observing that multiplication by an orthogonal matrix does not change the Euclidean norm, we reason as follows:

$$\|A_1 v\|_2 = \|U A_1 v\|_2 = \|U A_1 V^t \bar{z}\|_2 = \|C_2 z\|_2 = \sigma_2 > \sigma_1.$$

But this contradicts the assumption that σ_1 is the maximum value. Therefore $\sigma_2 \leq \sigma_1$.

Claim 3: By the manner in which v_2 is chosen, it follows that

$$\|A_2 v_2\|_2 = \left\| U \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & C_2 \end{bmatrix} V^t v_2 \right\|_2 = \left\| \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & C_2 \end{bmatrix} V^t v_2 \right\|_2 = \sigma_2.$$

This means that the first component of $V^t v_2$ must be zero, because a nonzero component increases the norm of $V^t v_2$ without increasing the norm of Av_2. Therefore

$$v_2 = V \begin{bmatrix} 0 \\ z_2 \end{bmatrix}$$

for some $z_2 \in \mathbb{R}^{n-1}$. Thus

$$v_1^t v_2 = v_1^t V \begin{bmatrix} 0 \\ z_2 \end{bmatrix} = [\ 1 \quad \mathbf{0}\ ] \begin{bmatrix} 0 \\ z_2 \end{bmatrix} = 0.$$

Claim 4: From the definition of u_2, it follows that

$$\begin{aligned} u_2 &= \frac{1}{\sigma_2} A_2 v_2 = \frac{1}{\sigma_2} U \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & C_2 \end{bmatrix} V^t v_2 \\ &= \frac{1}{\sigma_2} U \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & C_2 \end{bmatrix} \begin{bmatrix} 0 \\ z_2 \end{bmatrix} = \frac{1}{\sigma_2} [u_1 | U_2] \begin{bmatrix} 0 \\ C_2 z_2 \end{bmatrix} \\ &= \frac{1}{\sigma_2} U_2 C_2 z_2. \end{aligned}$$

Therefore

$$u_1^t u_2 = \frac{1}{\sigma_2} u_1^t U_2 C_2 z_2 = 0$$

because $u_1^t U_2 = 0$.

Now that all claims have been proved, the theorem follows by induction. □

Thus far we have discussed norms of vectors. Now we discuss various norms of matrices, and their inter-relationships.

Definition 1.4. Suppose $A, B \in \mathbb{R}^{m \times n}$, where m need not equal n. The **Frobenius inner product** on $\mathbb{R}^{m \times n}$ is denoted by $\langle \cdot, \cdot \rangle_F$ and is defined as follows:

$$\langle A, B \rangle_F = \mathrm{tr}(A^t B) = \mathrm{tr}(AB^t),$$

where $\mathrm{tr}(\cdot)$ denotes the trace, or sum of diagonal elements, of a square matrix. An equivalent definition is

$$\langle A, B \rangle_F = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} b_{ij}.$$

Therefore the Frobenius inner product is just the normal inner product of the two nm-dimensional vectors consisting of the elements of A and B written out as vectors. The associated matrix norm

$$\|A\|_F = \mathrm{tr}(A^t A) = \mathrm{tr}(AA^t) = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} |a_{ij}|^2 \right)^{1/2}$$

is called the **Frobenius norm** of a matrix.

Fact 1.5. *Suppose $A \in \mathbb{R}^{m \times n}$, and let $k = \min\{m, n\}$. Suppose $A = U \Sigma V^t$ is a singular value decomposition of A. Then*

$$\|A\|_F = \|\boldsymbol{\sigma}\|_2, \tag{1.17}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$.[2]

[2] Note that if A has rank $r < k$, then the last $k - r$ singular values would be zero.

Proof: It is easy to verify that, if U, V are *any* orthogonal matrices of compatible dimensions, then

$$\langle A, B \rangle_F = \langle U^t AV, U^t BV \rangle_F$$

This is because

$$\langle U^t AV, U^t BV \rangle_F = \mathrm{tr}(U^t AVV^t B^t U) = \mathrm{tr}(U^t AB^t U) = \mathrm{tr}(AB^t),$$

where in the first step we use the fact that $VV^t = I$, and in the second step we use the fact that an orthogonal transformation does not change the eigenvalues of a matrix and therefore does not change its trace. Now apply the above observation with U, V satisfying $U^t AV = \Sigma$. Then it readily follows that

$$\|A\|_F = \sqrt{\mathrm{tr}(AA^t)} = \sqrt{\mathrm{tr}(\Sigma\Sigma^t)} = \left(\sum_{i=1}^{k} \sigma_i^2 \right)^{1/2} = \|\boldsymbol{\sigma}\|_2.$$

This is the desired conclusion. □

Next we introduce a family of matrix norms known as "induced" norms.

Definition 1.5. Suppose $p, q \in [1, \infty]$, and let $A \in \mathbb{R}^{m \times n}$. Then the **induced matrix norm** $A_{p \to q}$ is defined by

$$\|A\|_{p \to q} := \max_{\|x\|_p \leq 1} \|Ax\|_q. \tag{1.18}$$

Some alternate but equivalent definitions of the induced matrix norm are given in Problem 1.3.

The exact computation of the induced matrix norm is not always easy, and depends on the values of p and q. The table below shows the explicit formulas, where available, for the common situations where p and q are 1, 2, or ∞. Note that throughout the book, whenever n is an integer, we use the notation $[n]$ to denote the set $\{1, \ldots, n\}$.

$p \setminus q$	1	2	∞
1	$\max_{j \in [n]} \sum_{i \in [m]} \lvert a_{ij} \rvert$	$\max_{j \in [n]} \|\mathbf{a}_j\|_2$	$\max_{i \in [m], j \in [n]} \lvert a_{ij} \rvert$
2	??	$\sqrt{\lambda_{\max}(A^t A)}$	$\max_{j \in [n]} \|\mathbf{a}_j\|_2$
∞	NP-hard	??	$\max_{i \in [m]} \sum_{j \in [n]} \lvert a_{ij} \rvert$

In the above table, $\mathbf{a}_j$ denotes the j-th column of the matrix A, and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a positive semidefinite matrix. The computation of the induced norm $\|A\|_{\infty \to 1}$ is NP-hard.

One of the main advantages of induced norms over other types of matrix norms is that they are *submultiplicative.* In other words, for matrices A, B of compatible dimensions, and any indices $p, q, r \in [1, \infty]$, we have that

$$\|AB\|_{p \to r} \leq \|A\|_{p \to q} \cdot \|B\|_{q \to r}. \tag{1.19}$$

See Problem 1.4.

Fact 1.6. *Suppose $A \in \mathbb{R}^{m \times n}$, and let $k = \min\{m, n\}$. Let $A = U\Sigma V^t$ denote the singular value decomposition of A. Then*

$$\|A\|_{2 \to 2} = \sigma_1 = \|\boldsymbol{\sigma}\|_\infty, \tag{1.20}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$.

Proof: It is first shown that the norm $\|A\|_{2 \to 2}$ remains invariant if A is replaced by a product UAV^t where U and V are orthogonal matrices of compatible dimensions. Observe that multiplication by an orthogonal matrix preserves the Euclidean norm, that is, if $\|x\|_2 = 1$, then $\|Vx\|_2 = 1$ whenever V is orthogonal. Similarly, $\|Ax\|_2 = \|U^t Ax\|_2$ whenever U is orthogonal. Therefore

$$\max_{\|x\|_2 \leq 1} \|Ax\|_2 = \max_{\|x\|_2 \leq 1} \|U^t AVx\|_2 = \max_{\|x\|_2 \leq 1} \|\Sigma x\|_2.$$

Now (1.20) follows readily. □

The SVD allows us to define yet another norm on matrices.

Definition 1.6. Suppose $A \in \mathbb{R}^{m\times n}$, and let $k = \min\{m, n\}$. Let $\sigma_1, \ldots, \sigma_k$ denote its singular values. Then the **nuclear norm** of A is denoted by $\|A\|_N$ and is defined by

$$\|A\|_N = \sum_{i=1}^{k} \sigma_i = \|\boldsymbol{\sigma}\|_1, \tag{1.21}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$.

It is easy to see that $\|\cdot\|_N$ satisfies two of the three conditions to be a norm. That is, $\|A\|_N \geq 0$, and $\|A\|_N = 0$ if and only if A is the zero matrix; also, $\|\alpha A\|_N = |\alpha| \|A\|_N$ for all constants α. The proof that $\|\cdot\|_N$ satisfies the triangle inequality is based on the following fact, which is also required later in this section.

Fact 1.7. *Suppose $A, B \in \mathbb{R}^{m\times n}$, and let $k = \min\{m, n\}$. Let $\sigma_i(A), \sigma_i(B), \sigma_i(A - B), i = 1, \ldots, k$ denote the singular values of $A, B, A - B$ respectively. Then, for every $l \in [k]$, we have that*

$$\sum_{i=1}^{l} |\sigma_i(A) - \sigma_i(B)| \leq \sum_{i=1}^{l} \sigma_i(A - B). \tag{1.22}$$

The proof of Fact 1.7 requires several additional concepts that are not used again in the book. For this reason, the proof of the triangle inequality for the nuclear norm is omitted, and the reader is directed to [24, 22] for the proof.

Thus far we have introduced three different kinds of norms on matrices, namely: ℓ_2-induced matrix norm, Frobenius norm, and nuclear norm, which are equal to $\|\boldsymbol{\sigma}\|_\infty$, $\|\boldsymbol{\sigma}\|_2$, and $\|\boldsymbol{\sigma}\|_1$ respectively, where $\boldsymbol{\sigma}$ is the vector of singular values of the matrix A. Given that, under the standard inner product, the ℓ_1-norm and the ℓ_∞-norm are duals of each other, it is perhaps not surprising that, under the Frobenius inner product, the ℓ_2-induced norm and nuclear norm are duals of each other.

Theorem 1.1. *Suppose $A \in \mathbb{R}^{m\times n}$. Then*

$$\|A\|_N := \max_{\|B\|_{2\to 2}\leq 1} |\langle A, B\rangle|. \tag{1.23}$$

Proof: Let $A = U\Sigma V^t$ be a singular value decomposition of A, and define $B \in \mathbb{R}^{m\times n}$ as follows: If $m = n$, then $B = UV^t$. If $m < n$, then $B = U[\ I_m \quad \mathbf{0}_{m\times(n-m)}\ ]V^t$. Finally, if $m > n$, then

$$B = U \begin{bmatrix} \Sigma \\ \mathbf{0}_{(m-n)\times n} \end{bmatrix} V^t.$$

We will do the proof in the first case in detail and leave the other two cases to the reader, as they are entirely similar. Suppose A is square, and $B = UV^t$; then

$$\mathrm{tr}(AB^t) = \mathrm{tr}(AVU^t) = \mathrm{tr}(U^tAV) = \mathrm{tr}(\Sigma) = \|A\|_N.$$

It is clear that $\|B\|_{2\to 2} = 1$ because both U and V^t are orthogonal. Therefore it follows that

$$\max_{\|B\|_{2\to 2}\leq 1} |\langle A, B\rangle| \geq \|A\|_N.$$

The proof in the other direction, namely that

$$\max_{\|B\|_{2\to 2}\leq 1} |\langle A, B\rangle| \leq \|A\|_N,$$

requires the concepts of duality in semidefinite programming and the absence of a duality gap, which are beyond the scope of the book. Therefore the reader is referred to [31, Proposition 2.1] for this part of the proof. □

Theorem 1.2. *Suppose $A \in \mathbb{R}^{m \times n}$ has rank r. Then*

$$\|A\|_{2\to 2} \leq \|A\|_F \leq \|A\|_N \leq \sqrt{r}\|A\|_F \leq r\|A\|_{2\to 2}. \tag{1.24}$$

Proof: If the matrix A has rank $r < k = \min\{m,n\}$, then the vector $\boldsymbol{\sigma}$ has r nonzero entries. In addition, it also has $k-r$ zero entries, which obviously do not affect any of the above norms of this vector. Now it is already known from Fact 1.3, specifically (1.8) that

$$\|\boldsymbol{\sigma}\|_\infty \leq \|\boldsymbol{\sigma}\|_2 \leq \|\boldsymbol{\sigma}\|_1 \leq \sqrt{r}\|\boldsymbol{\sigma}\|_2 \leq r\|\boldsymbol{\sigma}\|_\infty.$$

In view of identities already established, this set of inequalities is the same as 1.24. □

A ready consequence of (1.23) is that

$$|\langle A, B\rangle_F| \leq \|B\|_{2\to 2} \cdot \|A\|_N. \tag{1.25}$$

This is a matrix analog of Hölder's inequality.

The next result shows the importance of the singular value decomposition in constructing low-rank approximations of a given matrix.

Fact 1.8. *Suppose $A \in \mathbb{R}^{m\times n}$, and let $k = \min\{m,n\}$. Let $A = U\Sigma V^t$ be a singular value decomposition of A, and let $\sigma_1, \ldots, \sigma_k$ denote the singular values of A. Let $\mathcal{M}(i)$ denote the set of all $m \times n$ matrices of rank i or less. Then for each index i between 0 and $k-1$, we have that*

$$\min_{B\in\mathcal{M}(i)} \|A-B\|_{2\to 2} = \sigma_{i+1}, \tag{1.26}$$

$$\min_{B\in\mathcal{M}(i)} \|A-B\|_F = \left(\sum_{j=i+1}^{k} \sigma_j \right)^{1/2}, \tag{1.27}$$

$$\min_{B\in\mathcal{M}(i)} \|A-B\|_N = \sum_{j=i+1}^{k} \sigma_j. \tag{1.28}$$

In each case, the minimum is achieved by the choice $B = UD_iV^t$, where

$$D_i = \begin{bmatrix} \mathrm{Diag}(\sigma_1, \ldots, \sigma_i) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{1.29}$$

Proof: Note that all three of these norms are invariant if we make the substitutions $A \leftarrow U^tAV$ and $B \leftarrow UBV^t$. Similarly, the set $\mathcal{M}(i)$ is also invariant under a transformation $B \leftarrow UBV^t$. Therefore we can replace A by Σ in all three inequalities.

We will prove the inequalities in the opposite order. The proofs of (1.27) and (1.28) are quite short, while that of (1.26) is relatively longer.

To prove (1.28), we appeal to Fact 1.7, and in particular (1.21). Suppose B has rank i or less. Now (1.22) implies that

$$\|A-B\|_N = \sum_{j=1}^{k} \sigma_j(A-B) \geq \sum_{j=1}^{k} |\sigma_j(A) - \sigma_j(B)| \geq \sum_{j=i+1}^{k} \sigma_j(A),$$

because $\sigma_j(B) = 0$ for $j \geq i+1$. Therefore

$$\min_{B\in\mathcal{M}(i)} \|A-B\|_N \geq \sum_{j=i+1}^{k} \sigma_j,$$

that is, the right side of (1.28) is a lower bound. It is now easy to verify that the matrix $B = UD_iV^t$ achieves this bound.

To prove (1.27), we have already shown that

$$\min_{B \in \mathcal{M}(i)} \|A - B\|_F = \min_{B \in \mathcal{M}(i)} \|\Sigma - B\|_F,$$

where $A = U\Sigma V^t$ is the singular value decomposition of A. Now we recall that $\| \cdot \|_F$ is just the standard Euclidean norm of the nm-dimensional vector $\Sigma - B$ written out row-wise or column-wise. This shows that whenever a component of Σ is zero, the corresponding component of B must also be zero, in order to minimize $\|\Sigma - B\|_F$. Thus B must also be a diagonal matrix. Moreover, if more than i entries of this diagonal matrix are positive, then the rank of B would exceed i. Therefore B can have no more than i nonzero entries. Further, these must match the i largest entries of Σ in order to achieve the minimum.

Now it remains only to establish (1.26). It is easy to verify that

$$\|\Sigma - D_i\|_{2\to 2} = \sigma_{i+1},$$

where the matrix D_i is defined in (1.29). Therefore

$$\|A - UD_iV^t\|_{2\to 2} = \|U(\Sigma - D_i)V^t\|_{2\to 2} = \|\Sigma - D_i\|_{2\to 2} = \sigma_{i+1},$$

and as a result

$$\min_{B \in \mathcal{M}(i)} \|A - B\|_{2\to 2} \leq \sigma_{i+1}.$$

To show that the above minimum is in fact equal to σ_{i+1}, we assume that the minimum is less than σ_{i+1} and show that this leads to a contradiction. Let B be a matrix that achieves the minimum in (1.26). We use the following fact from linear algebra: If S, T are subspaces of $\mathbb{R}^n$, and $\dim(S) + \dim(T) > n$, then $S \cap T$ contains at least one nonzero vector. Define two subspaces S, T in $\mathbb{R}^n$ as follows:

$$S := \{w \in \mathbb{R}^n : \|\Sigma w\|_2 < \sigma_{i+1}\|w\|_2\} \cup \{\mathbf{0}\}, T := \{w \in \mathbb{R}^n : \|\Sigma w\|_2 \geq \sigma_{i+1}\|w\|_2\}.$$

Because the two conditions contradict each other, it is clear that $S \cap T = \{\mathbf{0}\}$. Let $\mathcal{N}(B) \subseteq \mathbb{R}^n$ denote the **null space** of B, that is, the set of vectors $w \in \mathbb{R}^n$ such that $Bw = \mathbf{0}$. Then $\mathcal{N}(B)$ has dimension no smaller than $n - i$, because B has n columns and rank no larger than i. To show why this is a contradiction, we show first that $\mathcal{N}(B) \subseteq S$. Suppose $Bw = 0$. Then

$$\|\Sigma w\|_2 = \|(\Sigma - B)w\|_2 \leq \|\Sigma - B\|_{2\to 2}\|w\|_2 < \sigma_{i+1}\|w\|_2.$$

This shows that $\mathcal{N}(B) \subseteq S$, and in turn that $\dim(S) \geq n - i$. Next, let W denote the $(i + 1)$-dimensional subspace in $\mathbb{R}^n$ spanned by the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_{i+1}$, where $\mathbf{e}_j$ is the j-canonical basis vector. In other words, W consists of n-vectors where the first $i + 1$ components could be nonzero but the rest are zero. Then for all $w \in W$, we have that

$$\|\Sigma w\|_2^2 = \sum_{j=1}^{i+1} \sigma_j^2 |w_j|^2 \geq \sigma_{i+1}^2 \|w\|_2^2,$$

which shows that $w \in T$. Finally, observe that

$$\dim(S) + \dim(T) \geq n - i + i + 1 = n + 1,$$

which means that $S \cap T$ contains a nonzero vector, which is a contradiction. Therefore (1.26) is true. It has already been shown that the choice $B = UD_iV^t$ achieves this minimum. □

We conclude this section by showing that the ℓ_1-norm and the nuclear norm are "decomposable." Though these results are easy to state and relatively easy to prove, they have very profound consequences in compressed sensing theory.

Fact 1.9. *The ℓ_1-norm is **decomposable** in the following sense: If $x, z \in \mathbb{R}^n$, and $supp(x) \cap supp(z) = \emptyset$, then*

$$\|x + z\|_1 = \|x\|_1 + \|z\|_1. \tag{1.30}$$

Proof: Because x and z have disjoint support sets, the index set $[n]$ can be partitioned into two sets S and T such that $S \cup T = [n]$, $\mathrm{supp}(x) \subseteq S$, and $\mathrm{supp}(z) \subseteq T$. Therefore

$$\|x + z\|_1 = \sum_{i=1}^{n} |x_i + z_i| = \sum_{i \in S} |x_i + z_i| + \sum_{i \in T} |x_i + z_i| = \sum_{i \in S} |x_i| + \sum_{i \in T} |z_i| = \|x\|_1 + \|z\|_1,$$

which is the desired result. □

The matrix analog of Fact 1.9 is given next.

Fact 1.10. *Suppose $A, B \in \mathbb{R}^{m \times n}$ and that $AB^t = \mathbf{0}_{m \times m}$, and $A^t B = \mathbf{0}_{n \times n}$. Then*

$$\|A + B\|_N = \|A\|_N + \|B\|_N. \tag{1.31}$$

Proof: Put A, B in *reduced* singular value form, so that

$$A = U_A \Sigma_A V_A^t, B = U_B \Sigma_B V_B^t.$$

Note that U_A has full column rank, V_A has full row equal to $\mathrm{rank}(A)$, and Σ_A is nonsingular. Similarly U_B has full column rank, V_B has full row rank equal to $\mathrm{rank}(B)$, and Σ_B is nonsingular. Now

$$AB^t = \mathbf{0} \iff U_A \Sigma_A V_A^t V_B \Sigma_B U_B^t = \mathbf{0} \iff V_A^t V_B = \mathbf{0},$$

because $U_A \Sigma_A$ has full column rank and $\Sigma_B U_B^t$ has full row rank. Similarly

$$A^t B = \mathbf{0} \iff U_A^t U_B = \mathbf{0}.$$

Note that we can write

$$A + B = [U_A \;\; U_B] \begin{bmatrix} \Sigma_A & \mathbf{0} \\ \mathbf{0} & \Sigma_B \end{bmatrix} \begin{bmatrix} V_A^t \\ V_B^t \end{bmatrix}.$$

Because the two extreme matrices meet the criteria of having pairwise orthogonal and normalized columns and rows, the above is a reduced singular value decomposition for $A + B$, except for the possibility that the diagonal elements might not be in nonascending order. This is easily taken care of by symmetric row and column permutations. Therefore the desired conclusion follows. □

Problem 1.1. Show that if $\| \cdot \|_a$ and $\| \cdot \|_b$ are norms, and $\alpha, \beta > 0$ are positive constants, then $x \mapsto \alpha\|x\|_a + \beta\|x\|_b$ is also a norm. State and prove a result for any finite number of norms.

Problem 1.2. Prove the **parallelogram identity**: If $\langle \cdot, , \rangle \cdot$ is an inner product, and $\| \cdot \|$ is the associated norm, then

$$\langle u, v \rangle = \frac{\|u + v\|^2 - \|u - v\|^2}{4}.$$

Problem 1.3. Show that the definition (1.18) is equivalent to the following definition:

$$\|A\|_{p \to q} = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_q}{\|x\|_p}.$$

Problem 1.4. Show that, for any indices $p, q, r \in [1, \infty]$ and any matrices A, B of compatible dimensions, it is the case that

$$\|AB\|_{p \to r} \leq \|A\|_{p \to q} \cdot \|B\|_{q \to r}.$$

1.2 Problem Formulations

The basic objective of compressed sensing is to reconstruct a "large but sparse" vector or matrix from a limited number of nonadaptive linear measurements. Of course, what constitutes "sparsity" varies from one application domain to another. We begin with the problem of vector recovery and introduce appropriate terminology and problem formulation(s). With slight changes, the same will apply to the problem of matrix recovery as well.

Suppose that $x \in \mathbb{R}^n$ is an unknown vector, but it is known beforehand that the number of nonzero components of x is no larger than k, which is an integer specified beforehand. The question is: Is it possible to reconstruct x *exactly* by taking $m \ll n$ measurements of x? The question can be made more precise. For notational convenience, given an integer n, let the symbol $[n]$ denote the index set $\{1, \ldots, n\}$. For a vector $x \in \mathbb{R}^m$, define

$$\text{supp}(x) := \{i \in [n] : x_i \neq 0\}$$

to be the **support** of the vector x. Further, given integers n and $k \ll n$, define

$$\Sigma_k := \{x \in \mathbb{R}^n : |\text{supp}(x)| \leq k\}$$

to be the set of **k-sparse** vectors in $\mathbb{R}^n$. Now suppose that x is k-sparse, but is otherwise unknown.

Definition 1.7. Given integers n and $k < n$, a **measurement matrix** $A \in \mathbb{R}^{m \times n}$, and a **demodulation map** $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ are said to achieve **exact recovery of k-sparse vectors** if

$$\Delta(Ax) = x, \ \forall x \in \Sigma_k.$$

The above equation states that, whenever x is k-sparse but otherwise arbitrary, one can recover x exactly by taking m linear measurements of x (namely Ax), and then applying the demodulation map Δ that converts the m-dimensional measurement vector $y = Ax$ back into an n-dimensional vector. Of course, as of now there is no guarantee that such a construction is possible. However, this formulation is useful as a precise statement of what one wishes to achieve.

In the above problem formulation, the measurement map A is assumed to be linear, but the demodulation map Δ need not be. To see why this is so, suppose that we use both a linear measurement map (that is, a matrix) $A \in \mathbb{R}^{m \times n}$, as well as a linear demodulation map $B \in \mathbb{R}^{n \times m}$. Then the process of first measuring an unknown vector x and then applying demodulation results in the recovered vector BAx. Now let us ask: Is it possible that $BAx = x$ for all $x \in \Sigma_k$? If $m < n$, then the matrix BA has rank no larger than m and is thus rank-deficient. Therefore it is not possible for BAx to equal x for *all* k-sparse vectors x. This shows that, in order to achieve exact recovery of all k-sparse vectors, either the measurement map, or the demodulation map, or both, need to be nonlinear. For historical reasons, compressed sensing theory has focused on the situation where the measurement map is linear and the demodulation map is not. The other two possibilities, namely a nonlinear measurement map together with a demodulation map that can be either linear or nonlinear, has not been the subject of much study. Note too that the above measurement scheme is *nonadaptive.* In other words, *the same* measurement map is applied to *every* k-sparse vector.

The above problem formulation, while precise, is also somewhat unrealistic. In the "real world," no vector is *actually* sparse; rather, it is reasonable to assume that the unknown vector x is "nearly sparse." It is also reasonable to assume that some measurement errors might creep in, so that the measured vector y equals $Ax + \eta$ where A is a known measurement matrix, and η is an unknown noise signal. In such a case, one can no longer aspire to recover the unknown vector x exactly; rather, one can at best aspire to recover a good approximation to x. To make these ideas precise, we introduce a few definitions.

Definition 1.8. Suppose integers n and $k < n$, and a norm $\|\cdot\|$ on $\mathbb{R}^n$ are specified. Let $x \in \mathbb{R}^k$ be arbitrary. Then the quantity

$$\sigma_k(x, \|\cdot\|) := \min_{z \in \Sigma_k} \|x - z\|$$

is called the **k-sparsity index** of the vector x with respect to the norm $\|\cdot\|$.

Thus the k-sparsity index of a vector x is the closest distance in the sense of the norm $\| \cdot \|$ to the set of k-sparse vectors. Note that the k-sparsity index of a vector x depends not only on the integer k, but also on the underlying norm. It is easy to see that

$$x \in \Sigma_k \iff \sigma_k(x, \| \cdot \|) = 0$$

for any norm $\| \cdot \|$ on $\mathbb{R}^n$.

If $\| \cdot \|$ is one of the ℓ_p-norms, then computing the k-sparsity index is straight-forward. Suppose $S \subseteq [n]$ and let $x \in \mathbb{R}^n$. Then the vector $x_S \in \mathbb{R}^n$ is defined by

$$(x_S)_i = \begin{cases} x_i, & \text{if } i \in S, \\ 0, & \text{if } i \notin S. \end{cases}$$

One can think of x_S as the projection of x onto the subspace of $\mathbb{R}^n$ spanned by vectors whose support is contained in the set S. Now suppose $x \in \mathbb{R}^n$ and that it is desired to determine $\sigma_k(x, \| \cdot \|_p)$ where $p \in [1, \infty]$. Let $\Lambda_0 \subseteq [n]$ denote the index set corresponding to the k largest components by magnitude of x. In case there are ties, choose the components of Λ_0 in some arbitrary fashion. For instance, if $n = 7, k = 2$ and

$$x = [\ 0 \quad 2 \quad -4 \quad -1 \quad 3 \quad 1 \quad 2\ ],$$

then

$$\Lambda_0 = \{3, 5\}, \Lambda_o^c = \{1, 2, 4, 6, 7\}.$$

On the other hand, if

$$x = [\ 0 \quad 2 \quad -4 \quad -3 \quad 3 \quad 1 \quad 2\ ],$$

then $\Lambda_0 = \{3, 4\}$ and $\Lambda_0 = \{3, 5\}$ would both be acceptable choices.

The definition of the index set Λ_0 ensures that

$$\min_{i \in \Lambda_0} |x_i| \geq \max_{j \in \Lambda_0^c} |x_j|.$$

As mentioned above, the set Λ_0 might not be unique, but *any* choice that makes the above statement true would be acceptable. With these definitions, it is easy to see that

$$\sigma_k(x, \| \cdot \|_p) = \|x_{\Lambda_0^c}\|_p, \ \forall p \in [1, \infty].$$

In other words, the vector $z = x_{\Lambda_0}$ is the (or a) best approximation of x in Σ_k, for *every* value of p. However, the k-sparsity index $\sigma_k(x, \| \cdot \|_p)$, which equals the ℓ_p-norm of the residual vector $x - z = x_{\Lambda_0^c}$, could be different for different values of p.

It is easy to see that

$$\|x\|_p \leq \|x\|_1, \ \forall p \geq 1.$$

This is because

$$\|x\|_p = \left\| \sum_{i=1}^n x_i \mathbf{e}_i \right\| \leq \sum_{i=1}^n |x_i| \cdot \|\mathbf{e}_i\|_p = \|x\|_1.$$

Consequently, it follows that

$$\sigma_k(x, \| \cdot \|_p) = \|x_{\Lambda_0^c}\|_p \leq \|x_{\Lambda_0^c}\|_1 = \sigma_k(x, \| \cdot \|_1), \ \forall k \in [n], \ \forall x \in \mathbb{R}^n. \tag{1.32}$$

An alternative upper bound is given next.

Lemma 1.1. *Suppose $x \in \mathbb{R}^n$, and that $1 \leq p \leq q$. Define $r = q/p \geq 1$. Then for each $k \in [n]$, we have that*

$$\sigma_k(x, \| \cdot \|_q) \leq \left[\frac{1}{r} \cdot \left(1 - \frac{1}{r} \right)^{r-1} \cdot \frac{1}{k^{r-1}} \right]^{1/q} \|x\|_p \tag{1.33}$$

$$\leq \frac{\|x\|_p}{k^{1/p-1/q}}. \tag{1.34}$$

In particular,

$$\sigma_k(x, \| \cdot \|_2) \leq \frac{\|x\|_1}{2\sqrt{k}}, \ \forall k \in [n]. \tag{1.35}$$

Proof: Given $x \in \mathbb{R}^n$, without loss of generality assume that the elements of x are sorted in nonascending order by magnitude. Define $z \in \mathbb{R}^n_+$ by

$$z_i = \frac{|x_i|^p}{\|x\|_p^p}, \tag{1.36}$$

and note that $\|z\|_1 = 1$. Next, define $f : \mathbb{R}^n_+ \to \mathbb{R}_+$ by

$$f(z) = \sum_{i=k+1}^{n} z_i^r.$$

If $z \in \mathbb{R}^n_+$ is defined as in (1.36), then

$$f(z) = \frac{1}{\|x\|_p^q} \sum_{i=k+1}^{n} |x_i|^q = \left[\frac{\sigma_k(x, \| \cdot \|_q)}{\|x\|_p} \right]^q.$$

Now we seek to maximize $f(z)$ as z varies over the set S defined by

$$S = \left\{ z \in \mathbb{R}^n_+ : z_1 \geq z_2 \geq \ldots \geq z_n \geq 0, \sum_{i=1}^{n} z_i = 1 \right\}.$$

The set S is polyhedral, and is generated by the $n+1$ vectors $v_0 = \mathbf{0}$, v_1 through v_n defined by

$$(v_l)_i = \begin{cases} 1/l, & 1 \leq i \leq l, \\ 0 & i > l. \end{cases}$$

The objective function $f(\cdot)$ is convex because $r \geq 1$. Therefore the maximum of $f(\cdot)$ over the polyhedral set S occurs at one of the $l+1$ generators. (The maximum might also occur at some other point, but we are not bothered about that; we are interested only in the maximum value, not where it occurs.) Therefore we evaluate $f(v_l)$ for $l = 0, \ldots, 1$. Now $f(v_0) = 0$, so this need not be considered further. For $l \geq 1$, we have

$$f(v_l) = \sum_{i=k+1}^{n} [(v_l)_i]^r = \begin{cases} 0, & \text{if } l \leq k, \\ g(l), & \text{if } l \geq k+1, \end{cases}$$

where

$$g(l) = \sum_{i=k+1}^{l} (1/l)^r = \frac{l-k}{l^r}.$$

To generate an upper bound on $f(v_l) = g(l)$, we ignore the fact that l is an integer, and treat it as a real variable. The resulting maximum may not be achieved by an integer l, but would provide an upper bound, which is the object of our interest. Thus

$$g'(l) = \frac{l^r - rl^{r-1}(l-k)}{l^{2r}}.$$

So $g(l)$ attains its maximum at

$$l^* = \frac{kr}{r-1}.$$

Note that

$$l^* - k = k \left(\frac{r}{r-1} - 1 \right) = \frac{k}{r-1}.$$

Therefore

$$\begin{aligned} g(l^*) &= \frac{k}{r-1} \cdot \frac{(r-1)^r}{(rk)^r} = \frac{1}{r} \cdot \left(1 - \frac{1}{r} \right)^{r-1} \cdot \frac{1}{k^{r-1}} \\ &= C(r) \cdot \frac{1}{k^{r-1}}, \end{aligned} \tag{1.37}$$

where

$$C(r) := \frac{1}{r} \cdot \left(1 - \frac{1}{r} \right)^{r-1} \leq 1, \tag{1.38}$$

because $r \geq 1$.

Because (1.37) provides an upper bound for $f(z)$, we have shown that

$$\left[\frac{\sigma_k(x, \| \cdot \|_q)}{\|x\|_p} \right]^q \leq C(q/p) \cdot \frac{1}{k^{q/p-1}}.$$

Raising both sides to the power $1/q$ leads to

$$\sigma_k(x, \| \cdot \|_q) \leq [C(q/p)]^{1/q} \cdot \frac{1}{k^{1/p-1/q}} \cdot \|x\|_p,$$

which is the desired conclusion.

To prove (1.35), simply substitute $p = 1, q = 2, r = 2$ into (1.33). □

With these preliminaries, we are now ready to give a precise definition of compressed sensing in vector recovery.

Definition 1.9. Suppose we are given integers n and $k < n$, norms $\| \cdot \|_p$ and $\| \cdot \|_q$ on $\mathbb{R}^n$, a **measurement matrix** $A \in \mathbb{R}^{m \times n}$, and a **demodulation map** $\Delta : \mathbb{R}^m \to \mathbb{R}^n$. Then various properties of the pair (A, Δ) are defined as follows:

- The pair (A, Δ) is said to achieve **exact k-sparse recovery** if

$$\Delta(Ax) = x, \ \forall x \in \Sigma_k. \tag{1.39}$$

- The pair (A, Δ) is said to achieve **stable k-sparse recovery** if there exists a constant C_1 such that

$$\|\Delta(Ax) - x\|_q \leq C_1 \sigma_k(x, \| \cdot \|_p). \tag{1.40}$$

- The pair (A, Δ) is said to achieve **robust k-sparse recovery** if there exist constants C_1 and C_2 such that, for all $\eta \in \mathbb{R}^m$ with $\|\eta\|_q \leq \epsilon$, it is the case that

$$\|\Delta(Ax + \eta) - x\|_q \leq C_1 \sigma_k(x, \| \cdot \|_p) + C_2 \epsilon. \tag{1.41}$$

In the above definition, strictly speaking we should say "stable (or robust) k-sparse recovery *with respect to the norms* $\| \cdot \|_p$ *and* $\| \cdot \|_q$." We do not do this in the interests of brevity. As we shall see, $p = 1$ and $q = 1$ or $q = 2$ are the most-studied cases. Where necessary, the values of p and q are displayed explicitly.

Also, the constants C_1 and C_2 could depend on the measurement matrix A and the demodulation map Δ, but they do not depend on the unknown vector x or the unknown measurement noise η.

The reader is cautioned that the above terminology is not at all standard. In [22], the authors use the adjectives "stable" and "robust," but in connection with something called the "null space property," which is discussed later in this book. However, when restricted to that context, the above terminology is consistent with that in [22]. Another phrase that is sometimes used in compressed sensing is "near-ideal behavior," see for example [9].

It is obvious that robust k-sparse recovery implies stable k-sparse recovery, which in turn implies exact k-sparse recovery. To see this, suppose $x \in \Sigma_k$ and that $\eta = 0$. Then

$$\sigma_k(x, \| \cdot \|_p) = 0, \|\eta\|_q = 0.$$

Substituting this into the above shows that

$$\|\Delta(Ax) - x\|_q = 0,$$

which in turn implies that $\Delta(Ax) = x$. Because this holds for *every* k-sparse vector x, the pair (A, Δ) achieves exact k-sparse recovery.

Now suppose that $x \in \mathbb{R}^n$ is k-sparse, and it is desired to recover x from a noisy measurement $y = Ax + \eta$. Let J denote the support set of x. By assumption, $|J| \leq k$. Suppose that an "oracle" not only knows that x is k-sparse, but also knows the support set J. In this case the measured vector y can be expressed as

$$y = Ax + \eta = A_J x_J + \eta,$$

where $A_J \in \mathbb{R}^{m \times |J|}$ denotes the submatrix of A consisting of the columns in the index set J, and, in a deviation from the usual notation, $x_J \in \mathbb{R}^{|J|}$ denotes the components of x in the set J. With this convention, if the matrix A_J has full column rank, then the best estimate $\hat{x}_J$ of x_J in terms of minimizing the error $\|\hat{x}_J - x_J\|_2$ is given by the familiar least squares procedure, namely

$$\hat{x}_J = (A_J^t A_J)^{-1} A_J^t y,$$

while the error itself is given by

$$\hat{x}_J - x_J = (A_J^t A_J)^{-1} A_J^t \eta.$$

Therefore the least-squares error achieved by an oracle equals

$$\begin{aligned} \|\hat{x}_J - x_J\|_2 &= \|(A_J^t A_J)^{-1} A_J^t \eta\|_2 \\ &\leq \|(A_J^t A_J)^{-1} A_J^t\|_{2 \to 2} \|\eta\|_2 \\ &= c\|\eta\|_2 \end{aligned}$$

for some constant c that depends on A but not on η. On the other hand, if the pair (A, Δ) achieves robust k-sparse recovery with $q = 2$, then whenever x is sparse (so that $\sigma_k(x, \| \cdot \|_p) = 0$, it follows from (1.41) that

$$\|\Delta(Ax + \eta) - x\|_2 \leq C_2 \|\eta\|_2.$$

In other words, the Euclidean norm of the residual error achieved by the pair (A, Δ) is within a constant of the residual error achieved by an oracle that knows the support set of x and uses that information in least-squares error minimization. Finally, if the vector x is not exactly k-sparse, and one attempts to recover x via a noisy measurement $y = Ax + \eta$, then (1.41) states that the norm of the residual error $\Delta(y) - x$ is bounded by a sum of two terms: One that involves the sparsity index of x and another that involves the norm of the measurement error η.

In the definition of robust k-sparse recovery, the error model is that the measurement error η satisfies $\|\eta\|_q \leq \epsilon$. However, there is another error model that is sometimes used, which we will call a "Dantzig error model." Again this terminology is not standard. The phrase "Dantzig selector" is introduced in [12] for the noise model introduced here, but the noise model itself has no name. Therefore the present usage is consistent with that in [12]. To motivate it, note that one can write

$$Ax = \sum_{j \in [n]} \mathbf{a}_j x_j .$$

In other words, Ax is a linear combination of the columns of A. Now suppose that the measurement error η has the property that $A^t\eta$ has a very small norm, perhaps even zero. In such a case, the error η is nearly (or perhaps actually) orthogonal to all columns of A, in which case an additive noise of η should not interfere too much with the recovery process. With this in mind, we define the measurement error η to satisfy a **Dantzig error model** if there exists a constant α such that $\|A^t\eta\|_\infty \leq \alpha$. In subsequent chapters we will see that algorithms that achieve robust k-sparse recovery continue to do so when the noise model $\|\eta\|_q \leq \epsilon$ is replaced by $\|A^t\eta\|_\infty \leq \alpha$. In other words, with the Dantzig error model, (1.41) is replaced by the following requirement: There exist constants C_1 and C_2 such that, for all $\eta \in \mathbb{R}^m$ with $\|A^t\eta\|_\infty \leq \alpha$,

$$\|\Delta(Ax + \eta) - x\|_q \leq C_1 \sigma_k(x, \| \cdot \|_p) + C_2 \alpha. \tag{1.42}$$

Up to now we have discussed the problem of recovering an unknown vector. Now we discuss the problem of recovering an unknown matrix, which contains as a special case a problem known as "matrix completion." Suppose $X \in \mathbb{R}^{k \times l}$ is an unknown matrix, and that it is desired to recover X from measurements of the form $\mathcal{A}(X)$ where $\mathcal{A} : \mathbb{R}^{k \times l} \to \mathbb{R}^m$ is a linear map. In the case where we deal with a vector x, the assumption is that x is sparse. However, when we deal with a matrix, sparsity (in the sense that most components of X are zero) is not always the appropriate assumption. Instead, the assumption is that X has low rank. Suppose $X \in \mathbb{R}^{k \times l}$, and that $r \ll \min\{k, l\}$. If X has exactly rank r or less, then the singular values s_i would be zero for $i \geq r + 1$. This corresponds to the notion of a k-sparse vector. The notion of a "nearly" k-sparse vector is replaced by defining the quantity

$$\sigma_r(X, \| \cdot \|_N) := \sum_{i=r+1}^{\min\{k,l\}} \sigma_i(X), \tag{1.43}$$

This is the analog of the k-sparsity index for a vector. In particular, as shown in Fact 1.8 and specifically (1.28), we know that

$$\min_{B \in \mathcal{M}(r)} \|X - B\|_N = \sum_{i=r+1}^{\min\{k,l\}} \sigma_i(X),$$

where as before $\mathcal{M}(r)$ denotes the set of matrices of rank r or less. Thus the right side of (1.43) is the closest distance, in the sense of the nuclear norm, between the given matrix X and the set $\mathcal{M}(r)$ of matrices of rank r or less. Therefore we are justified in using the symbol σ_r. We can think of a matrix $X \in \mathbb{R}^{k \times l}$ as having "nearly rank r" if $\sigma_r(X, \| \cdot \|_N)$ is small. We aspire to recover a good approximation to X from possibly noise-corrupted linear measurements of the matrix X.

The most common situation is that $\mathcal{A}(X)$ consists of the values X_{ij} for all index pairs $(i, j) \in S$, where $S \subseteq [k] \times [l]$, and $|S| \ll kl$. In other words, one measures a small fraction of the components of the matrix X, and from these measurements, wishes to recover X. This is the motivation behind the nomenclature "matrix completion" – knowing only a few components of an unknown matrix X, but knowing that it has low rank, we wish to complete the matrix by finding the remaining components. However, the theory is not any more difficult if one were to use more general linear measurement maps $\mathcal{A}$. In other words, matrix recovery using arbitrary linear maps is not much more difficult than matrix completion.

Definition 1.10. Suppose we are given integers k, l, and $r \ll \min\{k, l\}$. Suppose we are given a linear **measurement map** $\mathcal{A} \in \mathbb{R}^{k \times l} \to \mathbb{R}^m$, and a **demodulation map** $\Delta : \mathbb{R}^m \to \mathbb{R}^{k \times l}$. Then various properties of the pair (A, Δ) are defined as follows:

- The pair (A, Δ) is said to achieve **exact r-rank recovery** if

$$\Delta(\mathcal{A}(X)) = X, \text{ if } \mathrm{rank}(X) \leq r. \tag{1.44}$$

- The pair (A, Δ) is said to achieve **stable r-rank recovery** if there exists a constant C_1 such that

$$\|\Delta(\mathcal{A}(X)) - X\|_F \leq C_1 \sigma_r(X, \|\cdot\|_N). \tag{1.45}$$

- The pair (A, Δ) is said to achieve **robust r-rank recovery** if there exist constants C_1 and C_2 such that, for all $\eta \in \mathbb{R}^m$ with $\|\eta\|_2 \leq \epsilon$, it is the case that

$$\|\Delta(Ax + \eta) - x\|_F \leq C_1 \sigma_r(X, \|\cdot\|_N) + C_2 \epsilon. \tag{1.46}$$

There is one more issue that needs to be mentioned. The ℓ_2-norm of a vector is invariant under an orthogonal change of basis. Other norms are not invariant, but there are upper and lower bounds on how much a change of basis will change the norm. In contrast, "sparsity" is highly dependent on the underlying basis. A sparse vector can become completely full, or vice versa, under a change of basis. In Definition 1.9, the presumption is that the vector under study is sparse or nearly sparse in the canonical basis. This is why the set Σ_k is defined in terms of the cardinality of the support set. Clearly, the set Σ_k, viewed as a subset of $\mathbb{R}^n$, will change if the underlying basis is changed. Perhaps the most common situation is when we change from the time domain to the frequency domain, or vice versa. A signal can be very sparse in the frequency domain and its time domain representation may be full; such an example is given in the next section. In such a case, it is necessary to carry out the change of basis first, before applying a compressed sensing algorithm. The question therefore arises as to *which* change of basis is to be performed. In general, there is no obvious answer, and the user simply has to make a guess. For instance, given a time-domain signal, it would be reasonable to guess that it is sparse in the frequency domain, and proceed accordingly. But given an image, it is not clear-cut in which transform domain the signal is sparse, or nearly so. There is some research on "adaptive" choice of basis, whereby the user guesses the basis in which the unknown vector has a sparse representation, and then *updates the basis*. The phrase "basis" connotes that the underlying representation vectors are linearly independent. If this requirement is removed, then the representation vectors are referred to as a "dictionary." Decide how much of the k-SVD algorithm you wish to cover.

Problem 1.5. Suppose $p, q \in [1, \infty]$, and that $\alpha, \beta > 0$. Let $\|\cdot\|$ denote $\alpha\|\cdot\|_p + \beta\|\cdot\|_q$. Show that

$$\sigma_k(x, \|\cdot\|) = \alpha\sigma_k(x, \|\cdot\|_p) + \beta\sigma_k(x, \|\cdot\|_q).$$

1.3 Some Historical Background

The problem of determining an unknown vector from possibly noise-corrupted linear measurements is very old, dating back at least to Legendre and Gauss, with contributions also by Laplace. To state these classical results using modern terminology, suppose $x \in \mathbb{R}^n$ is an unknown vector, $A \in \mathbb{R}^{m \times n}$ is a measurement matrix, and $\eta \in \mathbb{R}^m$ is a measurement noise. The measured vector y equals $Ax + \eta$, but there is no assumption here (as in the compressed sensing problem) that $m < n$. The objective is to formulate an optimization problem, the solution of which would lead to an estimate $\hat{x}$ for the unknown vector x.

This is a good place to highlight the manner in which persons working in this area use the phrases "approach" (or "formulation") versus "algorithm." In the prevailing parlance, an "approach" is a problem formulation, the solution of which would lead to an estimate for the unknown vector. In contrast, an "algorithm" is a systematic procedure, often iterative, that would eventually lead to the solution to the formulated problem.

In the traditional least squares setting, it is the case that $m > n$, so that the set of equations is over-determined. Assuming that the matrix A has full column rank n, if the estimate $\hat{x}$ is chosen such that

$$\hat{x} = \operatorname*{argmin}_z \|y - Az\|_2^2, \tag{1.47}$$

then the solution for $\hat{x}$ can be written down in closed form as

$$\hat{x}_{\rm LS} := (A^t A)^{-1} A^t y.$$

However, in the case where $m < n$, the set of equations is under-determined. Consequently (if A has full row rank), there are infinitely many choices of $\hat{x}$ such that $A\hat{x}$ *exactly equals* the measured vector y. Thus, unless some additional "structure" is imposed on the problem, there is not a unique solution in general.

Among the first methods to cope with under-determined problems is **regularization**, usually credited to the Russian mathematician Tikhonov [34]. In this approach, one chooses a "regularizer" or "regularizing function" $\mathcal{R} : \mathbb{R}^n \rightarrow \mathbb{R}_+$, which is a continuous function with the property that

$$\mathcal{R}(z) \rightarrow \infty \text{ as } \|z\| \rightarrow \infty.$$

One consequence of this property is that, for every number $c \in R$, the so-called "level set"

$$L_c(\mathcal{R}) := \{z \in \mathbb{R}^n : \mathcal{R}(z) \leq c\}$$

is closed and compact. In other words, the regularizer penalizes very large norms of the vector x. By choosing $\hat{x}$ by solving the optimization problem

$$\hat{x}_{\rm Reg} := \operatorname*{argmin}_z \|y - Az\|_2^2 + \alpha \mathcal{R}(z), \tag{1.48}$$

it is often possible to obtain a unique solution. Here the parameter α is a "tuning" parameter that provides a trade-off between the importance of minimizing the fitting error and that of penalizing a large norm of z. The original regularizer suggested by Tikhonov is

$$\mathcal{R}(z) = \|z\|_2^2,$$

though in principle any convex function with compact level sets could be used. To be completed.

1.4 Some Illustrative Applications

Three applications: One each in signal reconstruction from the discrete cosine transform, image reconstruction from wavelets or some other orthonormal transformation, and matrix completion.

Chapter 2

Sufficient Conditions for Recovery

In this chapter we will present several general conditions for compressed sensing. While some of these are both necessary and sufficient, the majority are sufficient conditions. The first set of conditions, presented in Section 2.1, belong to a class known as "null space-based conditions." As the name would suggest, these conditions are stated in terms of the null space of the measurement matrix A. These conditions are not always easy to verify directly. Nevertheless, it is worthwhile to have general necessary and sufficient conditions, even if they are not readily verifiable. In this way, the structure of the solution becomes very clear. In Section 2.1, we introduce a property known as the "restricted isometry property" (RIP), which in turn implies an appropriate set of null space-based conditions. In some sense, this merely transfers the problem, from one of constructing a measurement matrix that satisfies null space-based conditions, to one of constructing a measurement matrix that satisfies the RIP. Chapter 3 provides explicit procedures for actually constructing the measurement matrix.

In the first two sections of the chapter, the focus is on vector recovery. In Section 2.3, it is shown that, within broad limits, any sufficient condition for *vector* recovery can be extended to a sufficient condition for *matrix* recovery.

2.1 Null Space-Based Conditions

As stated in Section 1.2, the problem at hand is to reconstruct a sparse (or nearly sparse) vector x from possibly a noise-corrupted measurement of the form $y = Ax + \eta$, where A is the measurement matrix and η is the measurement noise. It is assumed that a prior upper bound ϵ is available on $\|\eta\|_2$.

As the title of the chapter suggests, throughout this chapter a central role is played by the null space of the matrix A. This is defined next.

Definition 2.1. Suppose $A \in \mathbb{R}^{m \times n}$ where $m < n$. Then the **null space** of the matrix A, denoted by $\mathcal{N}(A)$, is defined by

$$\mathcal{N}(A) := \{z \in \mathbb{R}^n : Az = 0\}.$$

Note that some authors refer to $\mathcal{N}(A)$ as the "kernel" of A. The phrase "kernel-based methods" has quite a distinct interpretation in machine learning, wherein a support vector machine is replaced by a more general "kernel-based" classifier. For this reason, the phrase "null space" is preferred here.

We begin with a study of the problem in the case where $\eta = 0$, that is, the case of noise-free measurements. Naturally, signal recovery in this case is easier than in the case of noisy measurements. As we shall see, while it is possible to state and prove necessary and sufficient conditions for noise-free signal recovery, these conditions are rather "fragile" in that they do not readily extend to the case of noisy measurements. Nevertheless, they are useful as an indication of a bare minimum set of conditions that the measurement matrix A needs to satisfy. Then we move on to the case of noisy measurements.

2.1.1 Exact Sparse Recovery

In this section, we will study the problem of recovering sparse, or nearly sparse, vectors from *noise-free* measurements. Specifically, suppose $x \in \mathbb{R}^n$ is unknown, and that we can measure $y = Ax$ where we are free to choose the measurement matrix $A \in \mathbb{R}^{m \times n}$. From y, we attempt to recover x. Because $m < n$, this set of equations is under-determined, whence it is not possible to determine x uniquely from y, unless we impose some conditions on x. However, simple linear algebra tells us that the set of *all* solutions to $Az = y = Ax$ is given by[1]

$$A^{-1}(y) = A^{-1}(Ax) = x + \mathcal{N}(A) = \{x + z : z \in \mathcal{N}(A)\}.$$

One possibility is to choose the *most sparse* preimage of y under A, by solving the optimization problem

$$\min \|z\|_0 \text{ s.t. } Az = y.$$

For an arbitrary matrix A and arbitrary vector y, this problem is NP-hard [28]. However, as we shall see below, under certain conditions the problem becomes tractable.

Let us now turn the problem around and ask: What conditions must the matrix A satisfy, in order that *every* k-sparse vector x can be exactly recovered from $y = Ax$? This is possible provided that, amongst the infinitely many vectors in the preimage set $A^{-1}(Ax) = x + \mathcal{N}(A)$, the only k-sparse vector is x itself. Thus, in order to be able to recover every k-sparse vector x exactly from $y = Ax$, a necessary and sufficient condition is that

$$A^{-1}(Ax) \cap \Sigma_k = \{x\}, \ \forall x \in \Sigma_k. \tag{2.1}$$

With this observation, we can now prove the followig:

Theorem 2.1. *The following statements are equivalent:*

1. *The matrix A satisfies* (2.1).
2. *$\mathcal{N}(A)$ does not contain any nontrivial $2k$-sparse vectors; that is*
$$\mathcal{N}(A) \cap \Sigma_{2k} = \{0\}. \tag{2.2}$$
3. *Every set of $2k$ or fewer columns of A is linearly independent.*

Proof: The equivalence of Items (2) and (3) is obvious; so it remains only to establish that Items (1) and (2) are equivalent. This too turns out to be straight-forward.

To show that Item (1) implies Item (2), we shall actually prove the contrapositive, i.e., that if Item (2) is false, then Item (1) is also false. Accordingly, suppose $z \in \Sigma_{2k} \setminus \{0\}$ and that $Az = 0$. Because $z \in \Sigma_{2k}$, it can be expressed as $x_1 - x_2$ where $x_1, x_2 \in \Sigma_k$. Further, because $z \neq 0$, it follows that $x_1 \neq x_2$. Next, because $z \in \mathcal{N}(A)$, it follows that $Ax_1 = Ax_2$. Therefore

$$x_2 \in A^{-1}(Ax_1) \cap \Sigma_k$$

and of course $x_2 \neq x_1$. Therefore Item (1) is false.

The proof in the other direction, namely that Item (2) implies Item (1), is almost the same. Suppose (2.2) holds, and let $x_1 \in \Sigma_k$ be arbitrary. Suppose $x_2 \in A^{-1}(Ax_1) \cap \Sigma_k$. Then $Ax_2 = Ax_1$ implies that $z = x_1 - x_2 \in \mathcal{N}(A)$, while $x_1, x_2 \in \Sigma_k$ implies that $z = x_1 - x_2 \in \Sigma_{2k}$. Therefore it follows from (2.2) that $z = 0$, i.e., that $x_2 = x_1$. This is the same as (2.1). □

One consequence of Theorem 2.1 is that, in order for Item (3) of the theorem to hold, it is necessary that $m \geq 2k$. Therefore, in order to achieve exact k-sparse recovery with noise-free measurements, which is in some sense the easiest problem, the minimum number of measurements is $2k$.

Another consequence of Theorem 2.1 is that, in a compressed sensing problem of k-sparse vector recovery, the measurement matrix A *cannot be too sparse*! If the matrix A has too many zero entries, then Item (3)

[1] Strictly speaking we should write $A^{-1}(\{y\})$ and not $A^{-1}(y)$. The latter symbol is used in the interests of simplicity.

would not hold. As an extreme illustration of this, let us understand why the "vector completion" problem does not make any sense, while the "matrix completion" problem makes sense (though this hasn't been established as yet). Recall that in the matrix completion problem, one attempts to reconstruct a low-rank matrix $X \in \mathbb{R}^{k \times l}$ from measurements of a small number of individual components of X. An analogous "vector completion" problem would be this: Given a k-sparse vector $x \in \mathbb{R}^n$, measure $m \ll n$ components of x, and from these, reconstruct x. In this formulation, the measurement matrix A would be a submatrix of the identity matrix. Specifically, if one were to measure components $x_{i_1}, \ldots, x_{i_m}$, then the matrix A would consist of rows i_1 through i_m of the $n \times n$ identity matrix. Consequently $n - m$ columns of A would be identically zero. This implies that Item (3) is satisfied only if $n - m < 2k$, or $m > n - 2k$, which is clearly not a useful result. Thus, in the recovery of sparse vectors, though the number of measurements is small, each measurement must take into account multiple components of x. We shall see this later in Chapter 3.

Now the reader might ask: If it is not possible to achieve *vector* completion unless one measures practically every component, how is *matrix* completion possible? The reason is that, in the matrix completion problem, the requirement that the unknown matrix X has low rank imposes algebraic constraints on the components of the matrix. There are no such restrictions in the vector case.

Theorem 2.1 suggests a simple recipe for constructing a measurement matrix A that permits exact recovery of k-sparse vectors. All one has to do is to ensure that every set of $2k$ columns of A is linearly independent. One way is to choose A to be a **Vandermonde matrix**, as follows: Let $\lambda_1, \ldots, \lambda_n$ be pairwise distinct real numbers, and define $a_{ij} = \lambda_j^{i-1}$. Thus A has the following form:

$$A = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \lambda_1 & \lambda_2 & \ldots & \lambda_n \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{m-1} & \lambda_2^{m-1} & \ldots & \lambda_n^{m-1} \end{bmatrix}.$$

Then, whenever $m \geq 2k$, Item (3) of Theorem 2.1 is satisfied, and as a result, one can always exactly recover every k-sparse vector x from the m-dimensional measurement Ax. However, this result is only of theoretical interest. Vandermonde matrices are notoriously ill-conditioned.

In any case, Theorem 2.1 addresses the most elementary form of compressed sensing, namely the exact recovery of sparse vectors. Now we study the recovery of vectors that may not be exactly sparse. In such a case, one can at best aspire to the recovery also being nearly exact.

Definition 2.2. Suppose $1 \leq p \leq q \leq \infty$. We say that a pair (A, Δ) where $A \in \mathbb{R}^{m \times n}$ and $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ is **(ℓ_q, ℓ_p)-instance optimal** of order k with constant C if

$$\|x - \Delta(Ax)\|_q \leq C\sigma_k(x, \| \cdot \|_p), \ \forall x \in \mathbb{R}^n. \tag{2.3}$$

The above definition of instance-optimality is an adaptation of that introduced in [13]. If a pair (A, Δ) satisfies (2.3), then for every vector $x \in \mathbb{R}^n$, by applying the demodulation map Δ to the measurement Ax, one "recovers" a vector $\hat{x} := \Delta(Ax)$ whose ℓ_q-distance from the unknown vector x is bounded by some universal constant times the k-sparsity index of x with respect to the ℓ_p-norm. Note that there is no measurement noise in this model. Thus the demodulation map Δ is applied to a *noise-free* measurement $y = Ax$.

Note that in (2.3), we explicitly permit the indices p and q to be different. Indeed, we will see later that it is essential to do so. For instance, if we let $p = q = 2$, then *any* pair (A, Δ) satisfying (2.3) must satisfy $m \geq cn$, for some constant n. In other words, as the dimension of the unknown vector x grows, the minimum number of measurements required also grows in a *linear* fashion with respect to n. In general, this is not considered a satisfactory version of "compressed" sensing; rather, one would prefer that the number of measurements m grows sublinearly with respect to n. We will return to this topic later.

The following theorem gives an abstract necessary and sufficient condition for (2.3) to hold, in terms of the so-called null space property, which is now defined.

Definition 2.3. Suppose p, q satisfying $1 \leq p \leq q \leq \infty$ and $A \in \mathbb{R}^{m \times n}$ are specified. Then A is said to satisfy the **null space property** with respect to the norms $\| \cdot \|_q$ and $\| \cdot \|_p$ of order k with constant C if

$$\|v\|_q \leq C\sigma_k(v, \| \cdot \|_p), \ \forall v \in \mathcal{N}(A). \tag{2.4}$$

Theorem 2.2. *Suppose p, q satisfying $1 \leq p \leq q \leq \infty$ and $A \in \mathbb{R}^{m \times n}$ are specified.*

1. *Suppose there exists a demodulation map $\Delta : \mathbb{R}^m \to \mathbb{R}^m$ such that the pair (A, Δ) is (ℓ_q, ℓ_p)-instance optimal of order k with constant C. Then A satisfies the null space property of order $2k$ with constant C with respect to $\| \cdot \|_q$ and $\| \cdot \|_p$; that is*

$$\|v\|_q \leq C\sigma_{2k}(v, \| \cdot \|_p), \ \forall v \in \mathcal{N}(A). \tag{2.5}$$

2. *Conversely, suppose A satisfies the null space property of order $2k$ with constant C with respect to $\| \cdot \|_q$ and $\| \cdot \|_p$; that is,* (2.5) *holds. Then there exists a demodulation map $\Delta : \mathbb{R}^m \to \mathbb{R}^m$ such that the pair (A, Δ) is (ℓ_q, ℓ_p)-instance optimal of order k with constant $2C$; that is,* (2.3) *holds with C replaced by $2C$.*

Proof: We will first prove Item 1. Accordingly, suppose that there exists a demodulation map $\Delta : \mathbb{R}^m \to \mathbb{R}^m$ such that the pair (A, Δ) is (ℓ_q, ℓ_p)-instance optimal of order k with constant C. For an arbitrary $v \in \mathcal{N}(A)$, define Λ_0 to be the index set corresponding to the k-largest entries by magnitude of v. Now instance optimality implies the exact recovery of k-sparse signals. Therefore $-v_{\Lambda_0} = \Delta(-Av_{\Lambda_0})$. Next, $v \in \mathcal{N}(A)$ implies that

$$0 = Av = Av_{\Lambda_0} + Av_{\Lambda_0^c}, \text{ or } -Av_{\Lambda_0} = Av_{\Lambda_0^c}.$$

Combining with the previous equality shows that

$$-v_{\Lambda_0} = \Delta(-Av_{\Lambda_0}) = \Delta(Av_{\Lambda_0^c}).$$

Therefore

$$\|v\|_q = \|v_{\Lambda_0} + v_{\Lambda_0^c}\|_q = \|v_{\Lambda_0^c} - \Delta(Av_{\Lambda_0^c})\|_q. \tag{2.6}$$

Now instance optimality implies that

$$\|v_{\Lambda_0^c} - \Delta(Av_{\Lambda_0^c})\|_q \leq C\sigma_k(v_{\Lambda_0^c}, \| \cdot \|_p). \tag{2.7}$$

Note that $v_{\Lambda_0^c}$ already has zeros in the set Λ_0. Let Λ_1 denote the index set of the k largest elements by magnitude of $v_{\Lambda_0^c}$. Then clearly $\Lambda_0 \cup \Lambda_1$ corresponds to the index set of the $2k$ largest elements by magnitude of v. In other words,

$$\sigma_k(v_{\Lambda_0^c}, \| \cdot \|_p) = \sigma_{2k}(v, \| \cdot \|_p).$$

Substituting this into (2.7) leads to

$$\|v_{\Lambda_0^c} - \Delta(Av_{\Lambda_0^c})\|_q \leq C\sigma_{2k}(v, \| \cdot \|_p).$$

Substituting this into (2.6) leads to

$$\|v\|_q \leq C\sigma_{2k}(v, \| \cdot \|_p),$$

which is the same as (2.5).

Now we establish Item 2. Accordingly, suppose (2.5) holds. Define the map $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ by

$$\Delta(y) := \operatorname*{argmin}_z \sigma_k(z, \| \cdot \|_p) \text{ s.t. } Az = y. \tag{2.8}$$

It is shown now that the pair (A, Δ) is (ℓ_q, ℓ_p)-instance optimal. Note that $\Delta(y) \in A^{-1}(y)$, so that $A\Delta(y) = y$ for all $y \in \mathbb{R}^m$. So let $x \in \mathbb{R}^n$ be arbitrary, and let $v = x - \Delta(Ax)$. Then

$$Av = Ax - A\Delta(Ax) = Ax - Ax = 0.$$

Therefore $v \in \mathcal{N}(A)$. As a consequence of the null space property in (2.5)

$$\|v\|_q \leq C\sigma_{2k}(v, \|\cdot\|_p),$$

$$\|x - \Delta(Ax)\|_q \leq C\sigma_{2k}(x - \Delta(Ax), \|\cdot\|_p). \tag{2.9}$$

The next step is to show that, if $u_1, u_2 \in \mathbb{R}^n$, then

$$\sigma_{2k}(u_1 - u_2, \|\cdot\|_p) \leq \sigma_k(u_1, \|\cdot\|_p) + \sigma_k(u_2, \|\cdot\|_p).$$

To see this, let $\bar{u}_1, \bar{u}_2 \in \Sigma_k$ be optimal k-sparse approximations of u_1, u_2 respectively. Then $\bar{u}_1 - \bar{u}_2 \in \Sigma_{2k}$. Therefore

$$\begin{aligned} \sigma_{2k}(u_1 - u_2, \|\cdot\|_p) &\leq \|u_1 - u_2 - (\bar{u}_1 - \bar{u}_2)\|_p \\ &\leq \|u_1 - \bar{u}_1\|_p + \|u_2 - \bar{u}_2\|_p \\ &= \sigma_k(u_1, \|\cdot\|_p) + \sigma_k(u_2, \|\cdot\|_p \end{aligned}$$

Substituting this into (2.9) leads to

$$\|x - \Delta(Ax)\|_q \leq C[\sigma_k(x, \|\cdot\|_p) + \sigma_k(\Delta(Ax), \|\cdot\|_p)]. \tag{2.10}$$

The last step is to observe that

$$\sigma_k(\Delta(Ax), \|\cdot\|_p) \leq \sigma_k(x, \|\cdot\|_p). \tag{2.11}$$

To see this, recall the definition of the map Δ, which leads to

$$\Delta(Ax) = \operatorname*{argmin}_z \sigma_k(z, \|\cdot\|_p) \text{ s.t. } Az = Ax.$$

However, since x itself is feasible for this optimization problem, (2.11) follows. Substituting from (2.11) into (2.10) leads to

$$\|x - \Delta(Ax)\|_q \leq 2C\sigma_k(x, \|\cdot\|_p),$$

which is the desired conclusion. □

The proof of Item 2 in Theorem 2.2 suggests that, if the matrix A satisfies the null space condition (2.4), then stable k-sparse recovery can be achieved by defining the demodulation map Δ as in (2.8). While this is true in principle, it is now shown that the function $x \mapsto \sigma_k(x, \|\cdot\|)$ is in general *nonconvex*. Therefore in general it is very difficult to compute the demodulation map Δ defined in (1.22), because that involves solving a nonconvex optimization problem.

Lemma 2.1. *Suppose $n \geq k+1$, and that $p \in [1, \infty]$. Then the map $x \mapsto \sigma_k(x, \|\cdot\|)$ is not convex.*

Proof: Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **convex** if

$$f(\lambda x + (1-\lambda)z) \leq \lambda f(x) + (1-\lambda)f(z), \ \forall \lambda \in [0,1], \ \forall x, z \in \mathbb{R}^n.$$

It is easy to see that

$$\sigma_k(\alpha x, \|\cdot\|_p) = \alpha \sigma_k(x, \|\cdot\|_p), \ \forall \alpha \geq 0, \ \forall x \in \mathbb{R}^n.$$

Therefore the function $x \mapsto \sigma_k(x, \|\cdot\|)$ is convex if and only if

$$\sigma_k(x + z, \|\cdot\|_p) \leq \sigma_k(x, \|\cdot\|_p) + \sigma_k(z, \|\cdot\|_p), \ \forall x, z \in \mathbb{R}^n. \tag{2.12}$$

It is now shown that (2.12) is false if $n \geq k+1$. Choose

$$x = [\ \bar{x} \quad \mathbf{0}_{n-k}\ ], z = [\ \mathbf{0}_{n-k} \quad \bar{z}\ ],$$

where every component of $\bar{x}, \bar{z} \in \mathbb{R}^k$ is nonzero. Then both x and z are k-sparse, whence

$$\sigma_k(x, \| \cdot \|_p) = \sigma_k(z, \| \cdot \|_p) = 0,$$

and as a result

$$\sigma_k(x, \| \cdot \|_p) + \sigma_k(z, \| \cdot \|_p) = 0.$$

However, $x + z$ has at least $k + 1$ nonzero components because $n \geq k + 1$, whence $\sigma_k(x + z, \| \cdot \|_p) > 0$. Therefore (2.12) is false. □

In the definition of instance optimality, we have deliberately permitted two different norms: One to measure the sparsity index, and another to measure the residual error. We will see in subsequent chapters that the most common situations are $p = 1$ (to measure the sparsity index), and $q = 1$ or $q = 2$ to measure the residual error. The next theorem shows that if we set $p = q = 2$, that is, if the ℓ_2-norm is used to measure both the residual error and the sparsity index, then compressed sensing is not possible, in that the number of measurements m must grow at least linearly with respect to the dimension n.

Theorem 2.3. *If a pair (A, Δ) is (ℓ_2, ℓ_2)-instance optimal of order $k \geq 1$ with constant C, then*

$$m \geq \gamma n \text{ where } \gamma = 1 - \sqrt{\frac{C^2 - 1}{C^2}}. \tag{2.13}$$

Proof: By Theorem 2.2, if the pair (A, Δ) is (ℓ_q, ℓ_p)-instance optimal, then the matrix A satisfies

$$\|v\|_q \leq C\sigma_{2k}(v, \| \cdot \|_p), \ \forall v \in \mathcal{N}(A).$$

Now suppose that the pair (A, Δ) is (ℓ_2, ℓ_2)-instance optimal. Apply the above condition with $p = q = 2$, and with $k = 1$. This implies that

$$\|v\|_2^2 \leq C^2(\|v\|_2^2 - |v_j|^2), \ \forall j \in [n], \ \forall v \in \mathcal{N}(A),$$

or equivalently

$$|v_j|^2 \leq \frac{C^2 - 1}{C^2} \|v\|_2^2, \ \forall j \in [n], \ \forall v \in \mathcal{N}(A).$$

Now let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ denote the canonical orthonormal basis for $\mathbb{R}^n$. Then the previous inequality can be written as

$$|\langle v, \mathbf{e}_j \rangle| \leq C' \|v\|_2, \ \forall j \in [n], \ \forall v \in \mathcal{N}(A),$$

where

$$C' = \sqrt{\frac{C^2 - 1}{C^2}}.$$

Now let P denote the orthogonal projection matrix of $\mathbb{R}^n$ onto $\mathcal{N}(A)$. Then P has l eigenvalues of one and $n - l$ eigenvalues of zero, where l is the dimension of $\mathcal{N}(A)$. Therefore $\text{tr}(P) = l$. On the other hand, because A has dimensions $m \times n$, it follows that $l \geq n - m$. Therefore

$$n - m \leq \text{tr}(P) = \sum_{j=1}^{n} \langle P\mathbf{e}_j, \mathbf{e}_j \rangle \leq \sum_{j=1}^{n} C' \|P\mathbf{e}_j\|_2 \leq nC',$$

because $\|P\mathbf{e}_j\|_2 \leq \|\mathbf{e}_j\|_2 = 1$ for all j. This can be rearranged as

$$m \geq (1 - C')n,$$

which is the same as (2.13). □

2.1.2 Robust Sparse Recovery

In the previous section the focus is on the *exact* recovery problem. Though Theorem 2.2 gives conditions that are almost necessary and sufficient (with a gap of a factor of two between the two conditions), it turns out that this approach does not work too well when there are noisy measurements. Also, the demodulation map Δ suggested in the proof of Theorem 2.2, namely to minimize $\sigma_k(z, \| \cdot \|_p)$ subject to the condition that $Az = y = Ay$, is not always easy to implement, because it involves solving a nonconvex optimization problem. Therefore in the present section we explore a different direction.

Even before "compressed sensing" became a separate discipline, the use of ℓ_1-norm minimization to recover sparse signals was popular under the name of "basis pursuit." Subsequently, several useful results were derived by adapting basis pursuit to the problem of vector recovery. Accordingly, the remainder of this chapter is devoted to a study of ℓ_1-norm minimization and its properties. In other words, throughout the remainder of the chapter, the demodulation map $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ is given by one of two forms. In case $y = Ax$ so that there is no measurement error, we define

$$\Delta(y) = \hat{x} := \operatorname*{argmin}_z \|z\|_1 \text{ s.t. } Az = y. \tag{2.14}$$

In case $y = Ax + \eta$ where $\|\eta\| \leq \epsilon$, where $\| \cdot \|$ is a specified norm and ϵ is a specified *a priori* upper bound on the magnitude of the measurement noise, we define

$$\Delta(y) = \hat{x} := \operatorname*{argmin}_z \|z\|_1 \text{ s.t. } \|y - Az\| \leq \epsilon. \tag{2.15}$$

Note that the optimization problem in (2.14) is a linear programming problem. This is because it is possible reformulate the problem equivalently as

$$\min_{w \in \mathbb{R}^n} \sum_{i=1}^n w_i \text{ s.t. } w_i \geq z_i, w_i \geq -z_i \ \forall i, Az = y.$$

The constraints on the vector w ensure that $w_i \geq |z_i|$ for all i. Moreover, the objective function ensures that w_i precisely equals $|z_i|$ for all i, because if $w_i > |z_i|$ for some i, then the objective function would increase. Therefore the ℓ_1-norm minimization problem can be solved effectively for very large values of n. The optimization problem in (2.15) is no longer a linear program, but it is still a convex programming problem. Note that, if x, A, y are all complex-valued, then we would not be able to reformulate the optimization problem in (2.14) as a linear program. But it would still be a convex programming problem.

Observe that if $\epsilon = 0$ then the map in (2.15) reduces to that in (2.14). Therefore in principle we could straightaway analyze the map in (2.15). However, we lead up to that in stages by first analyzing the map in (2.14). In either case, the question under study is: What conditions must the measurement matrix A satisfy, in order that the pair (A, Δ) achieves exact k-sparse, stable k-sparse, or robust k-sparse recovery? As will be seen below, the decomposability of $\| \cdot \|_1$, as brought out in Fact 1.9, plays an important role in all the proofs.

Definition 2.4. A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy the **exact null space (ENS) property** of order k if

$$\|v\|_1 < 2\sigma_k(v, \| \cdot \|_1), \ \forall v \in \mathcal{N}(A). \tag{2.16}$$

Comparing Definition 2.4 with Definition 2.3, we can see that the exact null space property is a special case of the null space property with $p = q = 1$, $C = 2$, and the inequality replaced by a strict inequality. There are several equivalent formulations of the exact null space property, as brought out next.

Lemma 2.2. *The following are equivalent:*

1. *$A \in \mathbb{R}^{m \times n}$ satisfies the exact null space property.*

2. *Let $v \in \mathcal{N}(A) \setminus \{\mathbf{0}\}$ be arbitrary, and let $\Lambda_0 \subseteq [n]$ denote the index set consisting of the k largest components by magnitude of v. Then*
$$\|v_{\Lambda_0}\|_1 < \|v_{\Lambda_0^c}\|_1. \tag{2.17}$$

3. *Let $v \in \mathcal{N}(A) \setminus \{\mathbf{0}\}$ be arbitrary, and let $S \subseteq [n]$ satisfy $|S| \leq k$. Then*
$$\|v_S\|_1 < \|v_{S^c}\|_1. \tag{2.18}$$

Proof: (1) $\Longrightarrow$ (2): Suppose $v \in \mathcal{N}(A) \setminus \{\mathbf{0}\}$, and that (2.16) holds. Observe that
$$\|v\|_1 = \|v_{\Lambda_0}\|_1 + \|v_{\Lambda_0^c}\|_1,$$
and that
$$\|v_{\Lambda_0^c}\|_1 = \sigma_k(v, \|\cdot\|_1).$$
Therefore
$$\|v_{\Lambda_0}\|_1 = \|v\|_1 - \|v_{\Lambda_0^c}\|_1 < 2\sigma_k(v, \|\cdot\|_1) - \sigma_k(v, \|\cdot\|_1) = \sigma_k(v, \|\cdot\|_1) = \|v_{\Lambda_0^c}\|_1,$$
which is Item (2).

(2) $\Longrightarrow$ (3): Suppose $v \in \mathcal{N}(A) \setminus \{\mathbf{0}\}$, and that $|S| \leq k$. By the definition of the set Λ_0, it follows that $\|v_S\|_1 \leq \|v_{\Lambda_0}\|_1$, whence
$$\|v_{S^c}\|_1 = \|v\|_1 - \|v_S\|_1 \geq \|v\|_1 - \|v_{\Lambda_0}\|_1 = \|v_{\Lambda_0^c}\|_1.$$
Substituting this into (2.17) leads to
$$\|v_S\|_1 \leq \|v_{\Lambda_0}\|_1 < \|v_{\Lambda_0^c}\|_1 \leq \|v_{S^c}\|_1.$$
The two extreme terms in the inequality are Item (3).

(3) $\Longrightarrow$ (1). Suppose $v \in \mathcal{N}(A) \setminus \{\mathbf{0}\}$. Apply (2.18) with $S = \Lambda_0$. Then
$$\|v\|_1 = \|v_{\Lambda_0}\|_1 + \|v_{\Lambda_0^c}\|_1 < 2\|v_{\Lambda_0^c}\|_1 = 2\sigma_k(v, \|\cdot\|_1),$$
which is Item (1). □

The significance of the exact null space property is brought out in the next theorem.

Theorem 2.4. *Suppose $A \in \mathbb{R}^{m \times n}$ and define the map Δ as in* (2.14)*. Then the pair (A, Δ) achieves exact k-sparse recovery if and only if A satisfies the exact null space condition.*

Remark: It is important to understand what this theorem does *not* say. The theorem says that *the particular* demodulation map defined by (2.14) works if and only if the matrix A satisfies the exact null space property. As shown in Theorem 2.2, the demodulation map Δ defined by (2.8) achieves exact k-sparse recovery whenever the matrix A satisfies the null space property defined in Definition 2.3, which is weaker than the exact null space property defined in Definition 2.4. Therefore Theorem 2.4 is more about the efficacy of ℓ_1-norm minimization as a recovery procedure, and less about when recovery is possible.

Proof: "If": Suppose A satisfies the exact null space property in the form (2.18), and that $x \in \Sigma_k$. It is to be shown that $\Delta(Ax) = x$. Let $S = \text{supp}(x)$, and note that $|S| \leq k$. Now let $z \in \mathbb{R}^n$ satisfy $Az = Ax$, and suppose $z \neq x$. It is shown that $\|z\|_1 > \|x\|_1$, establishing that $\hat{x} = x$ is the unique minimizer of the optimization problem in (2.14).

Define $v = z - x$, and observe that $v \in \mathcal{N}(A) \setminus \{\mathbf{0}\}$. Therefore $\|v_S\|_1 < \|v_{S^c}\|_1$ by (2.18). Now
$$\begin{aligned} \|x\|_1 &\leq \|x - z_S\|_1 + \|z_S\|_1 \\ &= \|-v_S\|_1 + \|z_S\|_1 \\ &< \|v_{S^c}\|_1 + \|z_S\|_1 \\ &= \|z_{S^c}\|_1 + \|z_S\|_1 = \|z\|_1, \end{aligned}$$

where we use the fact that $v_{S^c} = z_{S^c}$ because $x_{S^c} = \mathbf{0}$. Therefore $\hat{x} = x$ is the unique minimizer of the optimization problem in (2.14), and $\Delta(Ax) = x \; \forall x \in \Sigma_k$.

"Only If": Suppose the pair (A, Δ) achieves exact k-sparse recovery, and let $v \in \mathcal{N}(A), S \subseteq [n]$ with $|S| \leq k$ be arbitrary. Then $v_S \in \Sigma_k$, so that, by assumption $\Delta(Av_S) = v_S$. In other words, for any $z \in \mathbb{R}^n$ that satisfies $Az = Av_S$, it must be that $\|z\|_1 > \|v_S\|_1$. In particular, let $z = -v_{S^c}$. Then $v \in \mathcal{N}(A)$ implies that

$$Av = 0 \iff Av_S + Av_{S^c} = 0 \iff A(-v_{S^c}) = Av_S.$$

Therefore $z = -v_{S^c}$ is feasible for the optimization problem. The optimality of v_S thus implies that $\|v_S\|_1 < \|v_{S^c}\|_1$. It has thus been established that (2.18) holds, which by Lemma 2.2 is equivalent to the null space property. □

To extend Theorem 2.4 to the case where the vector to be recovered is not necessarily sparse, we replace the exact null space property by a stronger requirement.

Definition 2.5. A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy the **stable null space (SNS) property** of order k with constant $\rho \in (0, 1)$ if

$$\|v\|_1 \leq (1 + \rho)\sigma_k(v, \| \cdot \|), \; \forall v \in \mathcal{N}(A). \tag{2.19}$$

Because $\rho < 1$, it is obvious that the stable null space property implies the exact null space property. As with the exact null space property, it is possible to give several equivalent characterizations of this property.

Lemma 2.3. *The following statements are equivalent.*

1. *The matrix A satisfies* (2.19)*.*
2. *Let $v \in \mathcal{N}(A)$ be arbitrary, and let $\Lambda_0 \subseteq [n]$ denote the index set consisting of the k largest components by magnitude of v. Then*
$$\|v_{\Lambda_0}\|_1 \leq \rho \|v_{\Lambda_0^c}\|_1. \tag{2.20}$$
3. *Let $v \in \mathcal{N}(A)$ be arbitrary, and let $S \subseteq [n]$ satisfy $|S| \leq k$. Then*
$$\|v_S\|_1 \leq \rho \|v_{S^c}\|_1. \tag{2.21}$$
4. *Let $v \in \mathcal{N}(A)$ be arbitrary, and let $S \subseteq [n]$ satisfy $|S| \leq k$. Then*
$$\|v\|_1 \leq \frac{1 + \rho}{1 - \rho}(\|v_{S^c}\|_1 - \|v_S\|_1). \tag{2.22}$$

Proof: The equivalence of Items (1), (2) and (3) is established just as in the case of Lemma 2.2. The required modifications are minor and thus left to the reader. Thus the only new item is No. (4). To prove that Items (3) and (4) are equivalent, recall that $\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1$, from Fact 1.9. Consequently

$$\{\|v_S\|_1 \leq \rho \|v_{S^c}\|_1\} \iff \{\|v\|_1 \leq (1 + \rho)\|v_{S^c}\|_1\}. \tag{2.23}$$

On the other hand,

$$\begin{aligned} \{\|v_S\|_1 \leq \rho \|v_{S^c}\|_1\} \quad &\iff \quad \{\|v_{S^c}\|_1 - \|v_S\|_1 \geq (1 - \rho)\|v_{S^c}\|_1\} \\ &\iff \quad \left\{ \|v_{S^c}\|_1 \leq \frac{1}{1 - \rho}(\|v_{S^c}\|_1 - \|v_S\|_1) \right\}. \end{aligned} \tag{2.24}$$

Combining the two statements shows that

$$\{\|v_S\|_1 \leq \rho \|v_{S^c}\|_1\} \iff \left\{ \|v\|_1 \leq \frac{1 + \rho}{1 - \rho}(\|v_{S^c}\|_1 - \|v_S\|_1) \right\}. \tag{2.25}$$

Hence Items (3) and (4) are equivalent. □

Now we state the main result concering stable k-sparse recovery.

Theorem 2.5. *Suppose $A \in \mathbb{R}^{m \times n}$ and define the map Δ as in* (2.14). *Suppose A satisfies the stable null space property of order k with constant $\rho \in (0, 1)$. Then the pair (A, Δ) achieves stable k-sparse recovery as defined in Definition 1.9 with the constant C_1 defined in* (1.40) *given by*

$$C_1 = 2\frac{1+\rho}{1-\rho}. \tag{2.26}$$

Proof: Suppose that $x \in \mathbb{R}^n$ is the vector that we are trying to recover, and that $z \in \mathbb{R}^n$ is arbitrary. Let $v = z - x$. It is shown now that

$$\|v_{S^c}\|_1 - \|v_S\|_1 \leq \|z\|_1 - \|x\|_1 + 2\|x_{S^c}\|_1. \tag{2.27}$$

To establish (2.27), we proceed as follows:

$$\|v_{S^c}\|_1 = \|z_{S^c} - x_{S^c}\|_1 \leq \|z_{S^c}\|_1 + \|x_{S^c}\|_1,$$

while

$$\|v_S\|_1 = \|z_S - x_S\|_1 \geq \|x_S\|_1 - \|z_S\|_1 = \|x\|_1 - \|x_{S^c}\|_1 - \|z_S\|_1.$$

Subtracting the second inequality from the first gives

$$\begin{aligned} \|v_{S^c}\|_1 - \|v_S\|_1 &\leq \|z_{S^c}\|_1 + \|x_{S^c}\|_1 - \|x\|_1 + \|x_{S^c}\|_1 + \|z_S\|_1 \\ &= \|z\|_1 - \|x\|_1 + 2\|x_{S^c}\|_1. \end{aligned} \tag{2.28}$$

This holds for *any* $x, z \in \mathbb{R}^n$. Now substitute $z = \hat{x}$, the solution to the optimization problem in (2.14). Then $\|\hat{x}\|_1 \leq \|x\|_1$ by definition. Also, $\|x_{S^c}\|_1 = \sigma_k(x, \|\cdot\|_1)$. Substituting these into (2.28) leads to

$$\|v_{S^c}\|_1 - \|v_S\|_1 \leq 2\sigma_k(x, \|\cdot\|_1).$$

Finally, substituting this into (2.22) gives

$$\|\hat{x} - x\|_1 = \|v\|_1 \leq 2\frac{1+\rho}{1-\rho}\sigma_k(x, \|\cdot\|_1).$$

Hence (1.40) is satisfied with the constant C_1 given by (2.26). □

Finally, to handle the case of noisy measurements, we replace the stable null space property with another, still stronger, requirement.

Definition 2.6. A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy the **robust null space (RNS) property** of order k and norm $\|\cdot\|$, with constants $\rho \in (0, 1)$ and $\tau \in \mathbb{R}_+$, if for all $h \in \mathbb{R}^n$ and all $S \subseteq [n]$ with $|S| \leq k$, it is true that

$$\|h_S\|_1 \leq \rho\|h_{S^c}\|_1 + \tau\|Ah\|. \tag{2.29}$$

Remarks:

1. If indeed $h \in \mathcal{N}(A)$, then $Ah = \mathbf{0}$, and (2.29) reduces to (2.19), with v replaced by h. Therefore the robust null space property implies the stable null space property.

2. Note that, unlike the definitions of the exact null space property and the stable null space property which involve only vectors in the null space of the matrix A, the robust null space property involves an inequality that needs to be satisfied by *all* vectors in $\mathbb{R}^n$. Therefore, it is not a "null space property" at all! Nevertheless, we will use the phrase "robust *null space* property" to highlight the point that it is a logical extension of the earlier two properties, which are indeed "null space properties." To highlight the fact that the robust null space property has to hold for all vectors in $\mathbb{R}^n$, the symbol v is replaced by h.

As with the exact and stable null space properties, it is possible to give an alternate characterization of the robust null space property.

Lemma 2.4. *The following two statements are equivalent:*

1. *The matrix A satisfies the robust null space property of order k and norm $\| \cdot \|$ with constants $\rho \in (0,1)$ and $\tau \in \mathbb{R}_+$.*
2. *Let $h \in \mathbb{R}^n$ be arbitrary, and let S_0 denote the index set consisting of the k largest components of h by magnitude. Then*
$$\|h_{S_0}\|_1 \leq \rho \|h_{S_0^c}\|_1 + \tau \|Ah\|. \tag{2.30}$$

The proof is very similar to that of Lemma 2.3 and is thus omitted.

Theorem 2.6. *Suppose $A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n$, and $y = Ax + \eta$ where $\|\eta\| \leq \epsilon$. Define $\hat{x} = \Delta(y)$ as in (2.15). Suppose that A satisfies the robust null space property of order k and norm $\| \cdot \|$, with constants $\rho \in (0,1)$ and $\tau \in \mathbb{R}_+$. Then*
$$\|\hat{x} - x\|_1 \leq 2 \frac{1+\rho}{1-\rho} \sigma_k(x, \| \cdot \|_1) + \frac{4\tau}{1-\rho} \epsilon. \tag{2.31}$$

Proof: Let Λ_0 denote the index set of the k largest components by magnitude of x. Let $h = \hat{x} - x$. Then the optimality of $\hat{x}$ implies that
$$\|\hat{x}\|_1 = \|x + h\|_1 \leq \|x\|_1.$$
Using the decomposability of $\| \cdot \|_1$, we get
$$\|x_{\Lambda_0} + h_{\Lambda_0}\|_1 + \|x_{\Lambda_0^c} + h_{\Lambda_0^c}\|_1 \leq \|x_{\Lambda_0}\|_1 + \|x_{\Lambda_0^c}\|_1.$$
Applying the triangle inequality twice to the left side gives
$$\|x_{\Lambda_0}\|_1 - \|h_{\Lambda_0}\|_1 - \|x_{\Lambda_0^c}\|_1 + \|h_{\Lambda_0^c}\|_1 \leq \|x_{\Lambda_0}\|_1 + \|x_{\Lambda_0^c}\|_1.$$
Cancelling the common term $\|x_{\Lambda_0}\|_1$, and noting that $\|x_{\Lambda_0^c}\|_1 = \sigma_k(x, \| \cdot \|_1)$ or σ_k for short, we can rewrite the above inequality as
$$\|h_{\Lambda_0^c}\|_1 - \|h_{\Lambda_0}\|_1 \leq 2\sigma_k. \tag{2.32}$$
Now let S_0 denote the index set of the k largest components of h by magnitude, and observe that
$$\|h_{S_0^c}\|_1 \leq \|h_{\Lambda_0^c}\|_1, \|h_{\Lambda_0}\|_1 \leq \|h_{S_0}\|_1.$$
Therefore (2.32) implies that
$$\|h_{S_0^c}\|_1 - \|h_{S_0}\|_1 \leq 2\sigma_k. \tag{2.33}$$
This is the first of two inequalities that we will require.

To derive the second inequality, note that
$$\|Ah\| = \|(A\hat{x} - y) - (Ax - y)\| \leq \|A\hat{x} - y\| + \|Ax - y\| \leq 2\epsilon,$$
because both $\hat{x}$ and x are feasible for the optimization problem in (2.15). Now apply the robust null space property with $S = S_0$. This implies that
$$\|h_{S_0}\|_1 \leq \rho \|h_{S_0^c}\|_1 + 2\tau\epsilon, \tag{2.34}$$
after replacing $\|Ah\|$ by its upper bound 2ϵ.

The two inequalities (2.33) and (2.34) can be expressed in matrix-vector form as
$$\begin{bmatrix} 1 & -1 \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} \|h_{S_0^c}\|_1 \\ \|h_{S_0}\|_1 \end{bmatrix} \leq \begin{bmatrix} 2\sigma_k \\ 2\tau\epsilon \end{bmatrix}. \tag{2.35}$$

Let M denote the coefficient matrix on the left side, and note that $\det(M) = 1 - \rho > 0$. Therefore all elements of M^{-1} are positive, and it is possible to multiply both sides of (2.35) by M^{-1}. This gives

$$\begin{aligned} \begin{bmatrix} \|h_{S_0^c}\|_1 \\ \|h_{S_0}\|_1 \end{bmatrix} &\leq \begin{bmatrix} 1 & -1 \\ -\rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2\sigma_k \\ 2\tau\epsilon \end{bmatrix} \\ &= \frac{1}{1-\rho} \begin{bmatrix} 1 & 1 \\ \rho & 1 \end{bmatrix} \begin{bmatrix} 2\sigma_k \\ 2\tau\epsilon \end{bmatrix} \\ &= \frac{1}{1-\rho} \begin{bmatrix} 2(\sigma_k + \tau\epsilon) \\ 2(\rho\sigma_k + \tau\epsilon) \end{bmatrix}. \end{aligned} \tag{2.36}$$

Finally, using the triangle inequality gives

$$\begin{aligned} \|h\|_1 &\leq \|h_{S_0^c}\|_1 + \|h_{S_0}\| \\ &= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \|h_{S_0^c}\|_1 \\ \|h_{S_0}\|_1 \end{bmatrix} \\ &\leq \frac{1}{1-\rho}[2(1+\rho)\sigma_k + 4\tau\epsilon], \end{aligned}$$

which is the same as (2.31). □

Theorem 2.6 provides an upper bound for only the ℓ_1-norm of the residual error $h = \hat{x} - x$. To obtain upper bounds for $\|h\|_p$ for values of $p > 1$, we replace the robust null space property by a stronger assumption.

Definition 2.7. A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy the ℓ_q **robust null space (ℓ_q-RNS) property** of order k and norm $\| \cdot \|$, with constants $\rho \in (0,1)$ and $\tau \in \mathbb{R}_+$, if for all $h \in \mathbb{R}^n$ and all $S \subseteq [n]$ with $|S| \leq k$, it is true that

$$\|h_S\|_q \leq \frac{\rho}{k^{1-1/q}} \|h_{S^c}\|_1 + \frac{\tau}{k^{1-1/q}} \|Ah\|, \tag{2.37}$$

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

Remarks:

1. It is a ready consequence of Hölder's inequality that

$$\|h_S\|_p \leq k^{1/p - 1/q} \|h_S\|_q, \ \forall p \in [1, q]. \tag{2.38}$$

In particular,

$$\|h_S\|_1 \leq k^{1-1/q} \|h_S\|_q. \tag{2.39}$$

Therefore the ℓ_q-RNS property implies the RNS property of Definition 2.6, with the same values for ρ and τ. Consequently Theorem 2.6 continues to hold if the RNS assumption is replaced by the ℓ_q-RNS assumption.

2. As in Lemma 2.4, (2.37) is equivalent to the following statement: Let $h \in \mathbb{R}^n$ be arbitrary, and let S_0 denote the index set consisting of the k largest components of h by magnitude. Then

$$\|h_{S_0}\|_q \leq \frac{\rho}{k^{1-1/q}} \|h_{S_0^c}\|_1 + \frac{\tau}{k^{1-1/q}} \|Ah\|. \tag{2.40}$$

3. Note that some authors omit the "scaling factor" $1/(k^{1-1/q})$ in front of the constant τ.

With this definition, we can extend Theorem 2.6 as follows:

Theorem 2.7. *Suppose $A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n$, and $y = Ax + \eta$ where $\|\eta\| \leq \epsilon$. Define $\hat{x} = \Delta(y)$ as in (2.15). Suppose that A satisfies the ℓ_q-robust null space property of order k and norm $\| \cdot \|$, with constants $\rho \in (0,1)$ and $\tau \in \mathbb{R}_+$. Then, for all $p \in [1, q]$, we have that*

$$\|\hat{x} - x\|_p \leq \frac{1}{k^{1-1/p}} \cdot \frac{2}{1-\rho}[(1+2\rho)\sigma_k(x, \| \cdot \|_1) + 3\tau\epsilon]. \tag{2.41}$$

Proof: As before, let $h := \hat{x} - x$, and let S_0 denote the index set of the k largest components by magnitude of h. Then

$$\|h\|_p \leq \|h_{S_0}\|_p + \|h_{S_0^c}\|_p. \tag{2.42}$$

We will find upper bounds for each term on the right side. First, note that $\|h_{S_0^c}\|_p = \sigma_k(h, \|\cdot\|_p)$. Therefore, by Lemma 1.1,

$$\|h_{S_0^c}\|_p \leq \frac{1}{k^{1-1/p}} \|h\|_1.$$

However, because the ℓ_q-robust null space property implies the robust null space property, the bounds derived in Theorem 2.6 continue to apply. Consequently

$$\|h\|_1 \leq \frac{2}{1-\rho}[(1+\rho)\sigma_k + 2\tau\epsilon],$$

where σ_k is a shorthand for $\sigma_k(x, \|\cdot\|_1)$. Therefore

$$\|h_{S_0^c}\|_p \leq \frac{1}{k^{1-1/p}} \cdot \frac{2}{1-\rho}[(1+\rho)\sigma_k + 2\tau\epsilon]. \tag{2.43}$$

Next, by Hölder's inequality,

$$\|h_{S_0}\|_p \leq k^{1/p-1/q} \|h_{S_0}\|_q, \ \forall p \in [1, q].$$

Now apply the ℓ_q-robust null space property, and observe that $\|Ah\| \leq 2\epsilon$ as before. This gives

$$\|h_{S_0}\|_p \leq \frac{k^{1/p-1/q}}{k^{1-1/q}} (\rho \|h_{S_0^c}\|_1 + 2\tau\epsilon). \tag{2.44}$$

However, (2.36) provides an upper bound for $\|h_{S_0^c}\|_1$, namely

$$\|h_{S_0^c}\|_1 \leq \frac{2}{1-\rho}(\sigma_k + \tau\epsilon).$$

Substituting this into (2.44), and clearing fractions, gives

$$\begin{aligned}
\|h_{S_0}\|_p &\leq \frac{1}{k^{1-1/p}} \left[\frac{2\rho}{1-\rho}(\sigma_k + \tau\epsilon) + 2\tau\epsilon \right] \\
&= \frac{1}{k^{1-1/p}} \cdot \frac{2}{1-\rho}[\rho(\sigma_k + \tau\epsilon) + (1-\rho)\tau\epsilon] \\
&= \frac{1}{k^{1-1/p}} \cdot \frac{2}{1-\rho}(\rho\sigma_k + \tau\epsilon). \qquad (2.45)
\end{aligned}$$

Combining (2.43) and (2.45) gives

$$\|h\|_p \leq \frac{1}{k^{1-1/p}} \cdot \frac{2}{1-\rho}[(1+2\rho)\sigma_k + 3\tau\epsilon],$$

which is the desired bound. □

2.2 Restricted Isometry Property

In the previous chapter we saw that if the demodulation map is defined via ℓ_1-norm minimization, then robust k-sparse recovery results whenever the measurement matrix A satisfies the robust null space property as in Definition 2.6 or Definition 2.7. However, the definition of the robust null space property is very abstract, and it is not *a priori* clear how one may go about choosing a matrix A to have this property. In the present chapter, a property known as the restricted isometry property (RIP) is introduced, and it is shown that RIP implies RNSP. The question then becomes: How can one choose a matrix A so that the RIP is satisfied? There are broadly two classes of approaches, namely: deterministic and probabilistic. Both approaches are described in this chapter.

Definition 2.8. A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy the **restricted isometry property (RIP)** of order k with constant $\delta_k \in (0,1)$ if

$$(1-\delta_k)\|u\|_2^2 \leq \|Au\|_2^2 \leq (1+\delta_k)\|u\|_2^2, \ \forall u \in \Sigma_k. \tag{2.46}$$

The RIP condition can be interpreted in a few different ways. First, suppose that $u \in \Sigma_k$, and let $J = \{j_1, \ldots, j_k\}$ denote the support set of u. Suppose that the corresponding k columns of A, namely $a_{j_1}, \ldots, a_{j_k} \in \mathbb{R}^n$, form an orthonormal set. Then it is easy to see that $\|Au\|_2^2 = \|u\|_2^2$. However, because A has only m rows, it is not possible for more than m columns of A to be *exactly* orthonormal. But it is possible for more than m columns of A to be *nearly* orthonormal. The RIP property means that every set of k or fewer columns of the matrix A are nearly orthonormal, in the sense of (2.46). Another, equivalent, way of expressing (2.46) is the following: For every subset $J \subseteq [n]$ with $|J| \leq k$, the spectrum of the matrix $A_J^t A_J$ lies in the interval $[1-\delta_k, 1+\delta_k]$.

The utility of the RIP is brought out in the next several theorems, which are the central results of this chapter. These theorems are found in [7], and improve upon earlier results in [5, 4, 3]. Stated briefly, taken together these theorems imply the following: First, if a matrix A satisfies the RIP of order tk with constant $\delta_{tk} < \sqrt{(t-1)/t}$, then A also satisfies the ℓ_2-robust null space property, as defined in Definition 2.7. Therefore, as a consequence of Theorem 2.7, it follows that demodulation via ℓ_1-norm minimization as in (2.15) leads to robust k-sparse signal recovery. Moreover, explicit upper bounds can be obtained for the ℓ_p-norm of the residual error for all $p \in [1,2]$. Second, if $t \geq 4/3$, then there exist a matrix A and an integer k such that ℓ_1-norm minimization *fails* to achieve robust (or even exact) k-sparse recovery. Therefore the bound $\delta_{tk} < \sqrt{(t-1)/t}$ is sharp whenever $t \geq 4/3$.

Theorem 2.8. *Suppose that, for some number $t > 1$, the matrix A satisfies the RIP of order tk with constant $\delta_{tk} =: \delta < \sqrt{(t-1)/t}$. Define*

$$\mu := \sqrt{t(t-1)} - (t-1) \in (0, 0.5), \tag{2.47}$$

$$a := [\mu(1-\mu) - \delta(0.5 - \mu + \mu^2)]^{1/2}, b := \mu(1-\mu)\sqrt{1+\delta}, c := \left[\frac{\delta\mu^2}{2(t-1)}\right]^{1/2}. \tag{2.48}$$

Then A satisfies the ℓ_2-robust null space property. Specifically, A satisfies (2.37) *with $q = 2$, and*

$$\rho = \frac{c}{a} < 1, \tau = \frac{b\sqrt{k}}{a^2}. \tag{2.49}$$

Remark: Note that an alternate and equivalent expression for a is

$$a = \frac{[(1-\delta) - (1-2\mu)^2(1+\delta)]^{1/2}}{2}. \tag{2.50}$$

Theorem 2.9. *Suppose that, for some number $t > 1$, the matrix A satisfies the RIP of order tk with constant $\delta_{tk} =: \delta < \sqrt{(t-1)/t}$. Define constants a, b, c, ρ, τ as in* (2.48) *and* (2.49)*. Suppose $x \in \mathbb{R}^n$ and that $y = Ax + \eta$ where $\|\eta\|_2 \leq \epsilon$. Define*

$$\hat{x} = \operatorname*{argmin}_z \|z\|_1 \ \ s.t. \ \|y - Az\|_2 \leq \epsilon. \tag{2.51}$$

Then

$$\|\hat{x} - x\|_1 \leq \frac{2(1+\rho)}{1-\rho}\sigma_k(x, \|\cdot\|_1) + \frac{4\tau}{1-\rho}\epsilon. \tag{2.52}$$

For all $p \in (1,2]$ we have

$$\|\hat{x} - x\|_p \leq \frac{1}{k^{1-1/p}} \cdot \frac{2}{1-\rho}[(1+2\rho)\sigma_k(x, \|\cdot\|_1) + 3\tau\epsilon]. \tag{2.53}$$

In particular,

$$\|\hat{x} - x\|_2 \leq \frac{2}{\sqrt{k}(1-\rho)}[(1+2\rho)\sigma_k(x, \|\cdot\|_1) + 3\tau\epsilon]. \tag{2.54}$$

The next theorem is applicable to the case where the measurement noise η satisfies the so-called "Dantzig Selector" bound $\|A^t\eta\|_\infty \leq \zeta$.

Theorem 2.10. *Suppose that, for some number $t > 1$, the matrix A satisfies the RIP of order tk with constant $\delta_{tk} =: \delta < \sqrt{(t-1)/t}$. Define constants a, b, c, ρ, τ as in* (2.48) *and* (2.49)*. Suppose $x \in \mathbb{R}^n$ and that $y = Ax + \eta$ where $\|A^t\eta\|_\infty \leq \zeta$. Define*

$$\hat{x}_{\text{DS}} = \underset{z}{\operatorname{argmin}} \|z\|_1 \text{ s.t. } \|A^t(y - Az)\|_2 \leq \zeta. \tag{2.55}$$

Define constants a, c, ρ, τ as in (2.48) *and* (2.49)*, and further define the constant*

$$d := \frac{\mu(1-\mu)}{\sqrt{t}}. \tag{2.56}$$

Then

$$\|\hat{x}_{\text{DS}} - x\|_1 \leq \frac{2(1+\rho)}{1-\rho}\sigma_k(x, \|\cdot\|_1) + \frac{4d}{(1-\rho)a^2}\zeta. \tag{2.57}$$

For all $p \in (1, 2]$ we have

$$\|\hat{x}_{\text{DS}} - x\|_p \leq \frac{1}{k^{1-1/p}} \cdot \frac{2}{1-\rho}\left[(1+2\rho)\sigma_k(x, \|\cdot\|_1) + \frac{3d}{a^2}\zeta\right]. \tag{2.58}$$

In particular,

$$\|\hat{x}_{\text{DS}} - x\|_2 \leq \frac{2}{\sqrt{k}(1-\rho)}\left[(1+2\rho)\sigma_k(x, \|\cdot\|_1) + \frac{3d}{a^2}\zeta\right]. \tag{2.59}$$

Theorem 2.11. *Suppose $t \geq 4/3$. Then for all $\xi > 0$ and all $k \geq 5/\xi$, there exists a matrix A that satisfies the RIP of order tk with constant $\delta_{tk} < \sqrt{(t-1)/t} + \xi$, and a vector $x \in \Sigma_k$ such that*

1. *With the noise-free measurement $y = Ax$, the demodulation map Δ defined in* (2.14) *fails to recover x.*
2. *With a noisy measurement $y = Ax + \eta$ where $\|\eta\|_2 \leq \epsilon$, the demodulation map Δ defined in* (2.15) *fails to recover x.*

Now we present the proofs of these theorems. Before proving Theorem 2.8, we establish two claims that are implicit in the theorem statement, namely that $\mu < 0.5$ and that $c < a$.

Lemma 2.5. *The number μ defined in* (2.47) *satisfies $\mu \in (0, 0.5)$.*

Proof: It is obvious that $\mu > 0$ because $t > 1$. To show that $\mu < 0.5$, we use the obvious inequality that $\sqrt{1+x} < 1 + x/2$ for all $x > 0$, and reason as follows:

$$\begin{aligned} \mu &= \sqrt{t(t-1)} - (t-1) \\ &= \sqrt{(t-1+1)(t-1)} - (t-1) \\ &= (t-1)\left[\sqrt{1 + \frac{1}{t-1}} - 1\right] \\ &< (t-1)\frac{1}{2(t-1)} = 0.5. \end{aligned}$$

This is the desired result. □

Lemma 2.6. *With a, c defined as in* (2.48)*, we have that $c < a$.*

Proof: It is enough to show that $a^2 < c^2$, that is

$$\frac{\delta\mu^2}{2(t-1)} < \mu(1-\mu) - \delta(0.5 - \mu + \mu^2).$$

This can be equivalently rewritten as

$$\delta\left[\frac{\mu^2}{2(t-1)} + 0.5 - \mu + \mu^2\right] < \mu(1-\mu),$$

and again as

$$\delta < \mu(1-\mu)\left[\frac{\mu^2}{2(t-1)} + 0.5 - \mu + \mu^2\right]^{-1}. \tag{2.60}$$

Now tedious but routine computations show that

$$\mu(1-\mu) = (2t-1)\sqrt{t(t-1)} - 2t(t-1),$$

and

$$\frac{\mu^2}{2(t-1)} + 0.5 - \mu + \mu^2 = (2t-1)t - 2t\sqrt{t(t-1)} = \mu(1-\mu)\sqrt{\frac{t}{t-1}}.$$

Therefore

$$\mu(1-\mu)\left[\frac{\mu^2}{2(t-1)} + 0.5 - \mu + \mu^2\right]^{-1} = \sqrt{\frac{t-1}{t}}.$$

Therefore (2.60) is equivalent to $\delta < \sqrt{(t-1)/t}$. Consequently $c < a$ whenever $\delta < \sqrt{(t-1)/t}$. □

The proof of Theorem 2.8 makes use of the following decomposition lemma, which is also proved in [7].

Lemma 2.7. *([7, Lemma 1.1]) Given a positive number α and a positive integer k, define*

$$T(\alpha,k) := \{v \in \mathbb{R}^n : \|v\|_\infty \leq \alpha, \|v\|_1 \leq k\alpha\}.$$

There there exist an integer N and vectors $u_1, \ldots, u_N \in \Sigma_k$ such that

1. *supp$(u_i) \subseteq$ supp(v) for all $i \in [N]$,*
2. *$\|u_i\|_\infty \leq \alpha$ for all $i \in [N]$,*
3. *v is a convex combination of $u_i, i \in [N]$.*

Proof: **"Only if":** Suppose that v is a convex combination of vectors $u_1, \ldots, u_N$ that are k-sparse, such that $\|u_i\|_\infty \leq \alpha$. Now, the fact that each u_i is k-sparse implies that

$$\|u_i\|_1 \leq k\|u_i\|_\infty \leq k\alpha, \ \forall i \in [N].$$

Next, the fact that v is a convex combination of the u_i implies, via the triangle inequality, that

$$\|v\|_\infty \leq \|u\|_\infty \leq \alpha, \|v\|_1 \leq \|u\|_1 \leq k\alpha.$$

Therefore v belongs to the set $T(\alpha,k)$.

"If": This part of the proof is more involved, and is by induction on $|\text{supp}(v)|$. Suppose that the theorem is true whenever $v \in T(\alpha,k)$ with $|v| \leq l-1$, and suppose now that $|\text{supp}(v)| = l$. If $l \leq k$ then we can take $N = 1$ and $u_1 = v$, so suppose $l > k$. Through the argument below, the vector v is decomposed into a convex combination of vectors $v_j, \ldots, v_l \in T(\alpha,k)$ where $j \geq 1$ and $|\text{supp}(v_i)| = l-1$. By the inductive hypothesis, each of these vectors can be expressed as a convex combination of appropriately chosen k-sparse vectors, which in turn implies that the original vector v is also such a convex combination, and completes the proof by induction.

Accordingly, suppose $v \in T(\alpha, k)$ with $|\text{supp}(v)| = l$. By permuting the indices as required, it can be assumed that

$$|v_1| \geq \ldots \geq |v_l| > 0, v_i = 0 \;\forall i \geq l+1.$$

Let e_i denote a variant of the i-elementary basis vector, which has $\text{sign}(v_i)$ in the i-th component and zeros elsewhere. Then it is clear that

$$v = \sum_{i=1}^{l} |v_i| e_i.$$

For clarity of notation let a_i denote $|v_i|$ for all i, so that $v = \sum_{i=1}^{l} a_i e_i$.

The fact that $v \in T(\alpha, k)$ implies that

$$\|v\|_\infty = \max_{1 \leq i \leq l} a_i \leq \alpha, \|v\|_1 = \sum_{i=1}^{l} a_i \leq k\alpha.$$

Now define the index set

$$D = \{j \in [l-1] : \sum_{i=j}^{l} a_i \leq (l-j)\alpha\}.$$

The set D is not empty because $1 \in D$. This is because, with $j = 1$, we have that

$$\sum_{i=1}^{l} a_i = \|v\|_1 \leq k\alpha \leq (l-1)\alpha,$$

where the last step follows from the fact that $k < l$. Because the set D is nonempty, it has a largest element; call it j. Then

$$\sum_{i=j}^{l} a_i \leq (l-j)\alpha, \tag{2.61}$$

$$\sum_{i=j+1}^{l} a_i > (l-j-1)\alpha. \tag{2.62}$$

With the symbol j thus defined, let

$$c = \frac{1}{l-j} \sum_{i=j}^{l} a_i, b_i = c - a_i, j \leq i \leq l.$$

Note that, as a consequence of (2.72), we have that $c \leq \alpha$. Also, each b_i is positive, as shown next. Because the a_i are in nonascending order, it is easily seen that the b_i are in nondescending order; therefore it is enough to show that $b_j > 0$. Now (2.73) implies that

$$\begin{aligned} (l-j)b_j &= (l-j)c - (l-j)a_j = \sum_{i=j}^{l} a_i - (l-j)a_j \\ &= \sum_{i=j+1}^{l} a_i - (l-j-1)a_j \geq \sum_{i=j+1}^{l} a_i - (l-j-1)\alpha > 0, \end{aligned} \tag{2.63}$$

where the last step follows from (2.73). Next, define the vector $u = \sum_{i=j}^{l} e_i$, and note that the vector u has a component of $\text{sign}(v_i)$ in position i for $j \leq i \leq l$, and zeros elsewhere. Consequently the vector $u - e_i$ has a component of $\text{sign}(v_s)$ in position s for $j \leq s \leq l$, except that it also has a component of zero in position i. Therefore $\|u\|_1 = l - j + 1$ and $\|u - e_i\|_1 = l - j$ for all i.

With this preparation, define the vectors

$$v_i = \sum_{i=1}^{j-1} a_i e_i + c(u - e_i), j \leq i \leq l.$$

Then each of these vectors has $l - 1$ nonzero components, because $u - e_i$ has a zero in position i. Moreover, from the definition of the constant c, it follows that

$$\|v_i\|_1 = \sum_{i=1}^{j-1} a_i + c(l - j) = \sum_{i=1}^{l} a_i = \|v\|_1 \leq k\alpha.$$

Next

$$\|v_i\|_\infty = \max\{\max_{1 \leq i \leq j-1} a_i, c\} \leq \alpha.$$

Therefore each vector v_i belongs to the set $T(\alpha, k)$. To show that v is a convex combination of the v_i, define the weights

$$\lambda_i = \frac{b_i}{\sum_{s=j}^{l} b_s}, j \leq i \leq l.$$

Observe that

$$\sum_{i=j}^{l} b_i = (l - j + 1)c - \sum_{i=j}^{l} a_i = (l - j + 1)c - (l - j)c = c.$$

Therefore we can also define $\lambda_i = b_i/c$. To conclude the proof, we observe that

$$\sum_{i=j}^{l} \lambda_i v_i = \sum_{s=1}^{j-1} a_s e_s + \sum_{i=j}^{l} \lambda_i c(u - e_i),$$

because the first summation does not depend on i and $\sum_{i=j}^{l} \lambda_i = 1$. So it is enough to show that

$$\sum_{i=j}^{l} \lambda_i c(u - e_i) = v - \sum_{s=1}^{j-1} a_s e_s = \sum_{i=j}^{l} a_i e_i.$$

This last step follows because

$$\begin{aligned}
\sum_{i=j}^{l} \lambda_i c(u - e_i) &= \sum_{i=j}^{l} b_i(u - e_i) = \left(\sum_{i=j}^{l} b_i \right) u - \sum_{i=j}^{l} b_i e_i \\
&= cu - \sum_{i=j}^{l} (c - a_i) e_i \\
&= cu - c \sum_{i=j}^{l} e_i + \sum_{i=j}^{l} a_i e_i = \sum_{i=j}^{l} a_i e_i,
\end{aligned}$$

because $u = \sum_{i=j}^{l} e_i$. □

Proof of Theorem 2.8: Suppose $A \in \mathbb{R}^{m \times n}$ satisfies the RIP of order tk with constant $\delta_{tk} =: \delta$. This implies that tk is an integer. Suppose further that $\delta < \sqrt{(t-1)/t}$. The objective is to establish the inequality (2.40) with the various quantities defined in (2.48). Suppose $h \in \mathbb{R}^n$ is arbitrary, and let S_0 denote the index set of the k largest components of h. Define

$$\gamma := \frac{\|h_{S_0^c}\|_1}{k}. \tag{2.64}$$

Now partition S_0^c as a disjoint union $S_1 \cup S_2$, where

$$S_1 := \{i \in S_0^c : |h_i| > \gamma/(t-1)\},$$

$$S_2 := \{i \in S_0^c : |h_i| \leq \gamma/(t-1)\}.$$

For brevity, denote $h_{S_0}, h_{S_1}, h_{S_2}$ as h_0, h_1, h_2 respectively. Let $r = |S_1|$. Then $\|h_1\|_1 \geq r\gamma/(t-1)$. But since $\|h_1\|_1 \leq \|h_{S_0^c}\|_1 = k\gamma$, it follows that

$$\frac{r\gamma}{t-1} \leq k\gamma, \text{ or } r \leq k(t-1).$$

Next, by the decomposability of the ℓ_1-norm, we have that

$$\|h_2\|_1 = \|h_{S_0^c}\|_1 - \|h_1\|_1 \leq k\gamma - \frac{r\gamma}{t-1} = [k(t-1)-r]\frac{\gamma}{t-1}.$$

Moreover, by the definition of the set S_2, it follows that

$$\|h_2\|_\infty \leq \frac{\gamma}{t-1}.$$

Now apply the decomposition lemma (Lemma 2.7) with k replaced by $k(t-1)-r$, and $\alpha = \gamma/(t-1)$. Then it follows that there exist an integer N, and vectors $u_1, \ldots, u_N \in \Sigma_{k(t-1)-r}$ whose support lies in S_2 such that

$$\|u_i\|_\infty \leq \|h_2\|_\infty \leq \frac{\gamma}{t-1},$$

and in addition, h_2 is a convex combination of $u_i, i \in [N]$. Suppose to be specific that $h = \sum_{i=1}^N \lambda_i u_i$.

The next step is to choose vectors $x_i, z_i \in \Sigma_{tk}$ such that

$$\sum_{i=1}^N \lambda_i(x_i + z_i) = (1-\mu)(h_0+h_1), \sum_{i=1}^N \lambda_i(x_i - z_i) = \mu h = \mu(h_0+h_1) + \mu \sum_{i=1}^N \lambda_i u_i.$$

One possible choice is

$$x_i + z_i = (1-\mu)(h_0+h_1), x_i - z_i = \mu(h_0+h_1+u_i), \; \forall i \in [N].$$

The solution to these equations is

$$z_i = \frac{1-2\mu}{2}(h_0+h_1) - \frac{\mu}{2}u_i, x_i = \frac{1}{2}(h_0+h_1) + \frac{\mu}{2}u_i.$$

With this choice, define vectors

$$\alpha = x_i + z_i = (1-\mu)(h_0+h_1)$$

which does not depend on i, and

$$\beta_i = x_i - z_i = \mu(h_0+h_1+u_i), \; \forall i \in [N].$$

Then

$$\sum_{i=1}^N \lambda_i \langle A\alpha, A\beta_i \rangle = \left\langle A\alpha, A\sum_{i=1}^N \lambda_i \beta_i \right\rangle = \mu(1-\mu)\langle A(h_0+h_1), Ah \rangle.$$

However, for each index i, we have that

$$\langle A\alpha, A\beta_i \rangle = \langle Ax_i + Az_i, Ax_i - Az_i \rangle = \|Ax_i\|_2^2 - \|Az_i\|_2^2.$$

Therefore it follows that

$$\sum_{i=1}^{N} \lambda_i \left(\|Ax_i\|_2^2 - \|Az_i\|_2^2 \right) = \mu(1-\mu)\langle A(h_0 + h_1), Ah \rangle,$$

or equivalently

$$\sum_{i=1}^{N} \lambda_i \|Ax_i\|_2^2 = \sum_{i=1}^{N} \lambda_i \|Az_i\|_2^2 + \mu(1-\mu)\langle A(h_0 + h_1), Ah \rangle. \tag{2.65}$$

Now we make use of the RIP of the matrix A, and observe that x_i, z_i are tk-sparse for all i. Substituting this observation into (2.65), and using δ as a shorthand for δ_{tk}, gives

$$(1-\delta) \sum_{i=1}^{N} \lambda_i \|x_i\|_2^2 \leq (1+\delta) \sum_{i=1}^{N} \lambda_i \|z_i\|_2^2 + \mu(1-\mu)\langle A(h_0 + h_1), Ah \rangle. \tag{2.66}$$

Since h_0, h_1, u_i all have disjoint support sets, it follows that

$$\|x_i\|_2^2 = 0.25(\|h_0 + h_1\|_2^2 + \mu^2 \|u_i\|_2^2),$$
$$\|z_i\|_2^2 = 0.25[(1-2\mu)^2 \|h_0 + h_1\|_2^2 + \mu^2 \|u_i\|_2^2],$$

for all $i \in [N]$. Substituting these relationships into (2.66), multiplying both sides by 4, and noting that $\sum_{i=1}^{N} \lambda_i = 1$ leads to

$$\begin{aligned} (1-\delta)\left[\|h_0 + h_1\|_2^2 + \mu^2 \sum_{i=1}^{N} \|u_i\|_2^2 \right] &\leq (1+\delta)\left[(1-2\mu)^2 \|h_0 + h_1\|_2^2 + \mu^2 \sum_{i=1}^{N} \|u_i\|_2^2 \right] \\ &+ 4\mu(1-\mu)\langle A(h_0 + h_1), Ah \rangle, \end{aligned}$$

or upon rearranging terms

$$\begin{aligned} \|h_0 + h_1\|_2^2 [(1-\delta) - (1-2\mu)^2(1+\delta)] &\leq 2\delta\mu^2 \sum_{i=1}^{N} \|u_i\|_2^2 \\ &+ 4\mu(1-\mu)\langle A(h_0 + h_1), Ah \rangle, \end{aligned} \tag{2.67}$$

Now we make two observations. First

$$\begin{aligned} \|u_i\|_2 &\leq \sqrt{k(t-1) - r} \cdot \|u_i\|_\infty \leq \sqrt{k(t-1)} \cdot \|u_i\|_\infty \\ &\leq \sqrt{k(t-1)} \frac{\gamma}{t-1} = \sqrt{k(t-1)} \frac{\|h_{S_0^c}\|_1}{k(t-1)} = \frac{\|h_{S_0^c}\|_1}{\sqrt{k(t-1)}}. \end{aligned} \tag{2.68}$$

Second, by Schwarz' inequality

$$\begin{aligned} \langle A(h_0 + h_1), Ah \rangle &\leq \|A(h_0 + h_1)\|_2 \cdot \|Ah\|_2 \\ &\leq \sqrt{1+\delta} \|h_0 + h_1\|_2 \cdot \|Ah\|_2. \end{aligned} \tag{2.69}$$

These bounds can be substituted into (2.67). Note that, because $\sum_{i=1}^{N} \lambda_i = 1$, this term is transparent. Therefore (2.67) implies that

$$\begin{aligned} \|h_0 + h_1\|_2^2 [(1-\delta) - (1-2\mu)^2(1+\delta)] &\leq 2\delta\mu^2 \frac{\|h_{S_0^c}\|_1^2}{k(t-1)} \\ &+ 4\mu(1-\mu)\sqrt{1+\delta} \|h_0 + h_1\|_2 \cdot \|Ah\|_2. \end{aligned} \tag{2.70}$$

Now we invoke the definitions of the constants a, b, c from (2.48), and denote $\|h_0 + h_1\|_2$ by g for brevity. Then the above inequality can be written as

$$4a^2 g^2 \leq 4c^2 \frac{\|h_{S_0^c}\|_1^2}{k} + 4bg\|Ah\|_2,$$

or after dividing both sides by 4 and rearranging,

$$a^2 g^2 - bg\|Ah\|_2 \leq c^2 \frac{\|h_{S_0^c}\|_1^2}{k}.$$

The next step is to "complete the square" on the left side. The above inequality implies that

$$a^2 g^2 - bg\|Ah\|_2 + \frac{b^2}{4a^2}\|Ah\|_2^2 \leq \frac{b^2}{4a^2}\|Ah\|_2^2 + c^2 \frac{\|h_{S_0^c}\|_1^2}{k},$$

or equivalently

$$[ag - (b/2a)\|Ah\|_2]^2 \leq \frac{b^2}{4a^2}\|Ah\|_2^2 + c^2 \frac{\|h_{S_0^c}\|_1^2}{k}.$$

Taking the square root of both sides, and using the obvious inequality $\sqrt{x^2+y^2} \leq x + y$ whenever $x, y \geq 0$, leads to

$$ag - (b/2a)\|Ah\|_2 \leq (b/2a)\|Ah\|_2 + c\frac{\|h_{S_0^c}\|_1}{\sqrt{k}},$$

or upon rearranging and replacing g by $\|h_0 + h_1\|_2$,

$$a\|h_0 + h_1\|_2 \leq (b/a)\|Ah\|_2 + c\frac{\|h_{S_0^c}\|_1}{\sqrt{k}}.$$

Dividing both sides by a and observing that $\|h_0\|_2 \leq \|h_0 + h_1\|_2$ leads to

$$\|h_0\|_2 \leq \|h_0 + h_1\|_2 \leq \frac{b}{a^2}\|Ah\|_2 + \frac{c}{a}\frac{\|h_{S_0^c}\|_1}{\sqrt{k}}. \tag{2.71}$$

By Lemma 2.6, we have that $c < a$ whenever $\delta = \delta_{tk} < \sqrt{(t-1)/t}$. Therefore (2.71) shows that A satisfies the robust null space property with $\rho = c/a$ and $\tau = (b\sqrt{k})/a^2$. □

Proof of Theorem 2.9: The various estimates follow from Theorem 2.8 and Theorem 2.7.

Proof of Theorem 2.10: With $\hat{x}_{\rm DS}$ defined as in (2.55), define $h := \hat{x}_{\rm DS} - x$. Then the computations in the proof of Theorem 2.8 continue to apply until (2.67). Moreover, the bound for $\|u_i\|_2$ in (2.68) continues to apply. However, the bound in (2.69) is replaced by another, as follows. Note that $h_0 + h_1 \in \Sigma_{tk}$. Also, both $\hat{x}_{\rm DS}$ and x are feasible for the optimization problem in (2.55). Therefore it follows that

$$\|A^t A h\|_\infty = \|A^t(y - Ax_{\rm DS}) - A^t(y - Ax)\|_\infty \leq 2\zeta.$$

Now we can write

$$\begin{aligned} \langle A(h_0 + h_1), Ah\rangle &= \langle h_0 + h_1, \rangle \\ &\leq \|h_0 + h_1\|_1 \cdot \|A^t A h\|_\infty \\ &\leq \frac{\|h_0 + h_1\|_2}{\sqrt{tk}} \cdot 2\zeta \end{aligned} \tag{2.72}$$

Substituting from (2.68) and (2.72) into (2.67) gives

$$\begin{aligned} \|h_0 + h_1\|_2^2[(1-\delta) - (1-2\mu)^2(1+\delta)] &\leq 2\delta\mu^2 \frac{\|h_{S_0^c}\|_1^2}{k(t-1)} \\ &+ \frac{8\mu(1-\mu)\zeta}{\sqrt{tk}}\|h_0 + h_1\|_2. \end{aligned}$$

Recall now the definitions of the constants a, c from (2.48), ρ from (2.49), and d from (2.56). Then, after dividing both sides by 4 and denoting $\|h_0 + h_1\|_2$ by g as before, the above inequality becomes

$$a^2 g^2 - \frac{2d\zeta}{\sqrt{k}} g \leq c^2 \frac{\|h_{S_0^c}\|_1^2}{k}.$$

Completing the square and taking the square root of both sides leads to

$$\left(ag - \frac{d\zeta}{a\sqrt{k}} \right)^2 \leq \frac{d^2\zeta^2}{ak} + c^2 \frac{\|h_{S_0^c}\|_1^2}{k},$$

$$\begin{aligned} ag - \frac{d\zeta}{a\sqrt{k}} &\leq \left[\frac{d^2\zeta^2}{ak} + c^2 \frac{\|h_{S_0^c}\|_1^2}{k} \right]^{1/2} \\ &\leq \frac{d\zeta}{a\sqrt{k}} + c \frac{\|h_{S_0^c}\|_1}{\sqrt{k}}, \end{aligned}$$

and finally

$$\|h_0 + h_1\|_2 = g \leq \frac{2d}{a^2\sqrt{k}}\zeta + \frac{c}{a}\frac{\|h_{S_0^c}\|_1}{\sqrt{k}} = \frac{2d}{a^2\sqrt{k}}\zeta + \rho \frac{\|h_{S_0^c}\|_1}{\sqrt{k}}.$$

Next, by Schwarz' inequality, we get

$$\|h_0\|_1 \leq \sqrt{k}\|h_0\|_2 \leq \sqrt{k}\|h_0 + h_1\|_2 \leq \frac{2d}{a^2}\zeta + \rho \|h_{S_0^c}\|_1. \tag{2.73}$$

As in the proof of Theorem 2.7, specifically (2.33), the fact that $\|x_{\mathrm{S}}\|_1 \leq \|x\|_1$ leads to

$$\|h_{S_0^c}\|_1 - \|h_0\|_1 \leq 2\sigma_k, \tag{2.74}$$

where it is to be noted that h_0 is the same as h_{S_0}, and σ_k is a shorthand for $\sigma_k(x, \|\cdot\|_1)$. Therefore we can repeat the reasoning in the proof of Theorem 2.6. Equations (2.73) and (2.74) can be written as

$$\begin{bmatrix} 1 & -1 \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} \|h_{S_0^c}\|_1 \\ \|h_{S_0}\| \end{bmatrix} \leq \begin{bmatrix} 2\sigma_k \\ 2d\zeta/a^2 \end{bmatrix}.$$

Because $\rho < 1$, this leads to

$$\begin{bmatrix} \|h_{S_0^c}\|_1 \\ \|h_{S_0}\| \end{bmatrix} \leq \frac{1}{1-\rho} \begin{bmatrix} 1 & 1 \\ \rho & 1 \end{bmatrix} \begin{bmatrix} 2\sigma_k \\ 2d\zeta/a^2 \end{bmatrix}. \tag{2.75}$$

Combining these shows that

$$\|h\|_1 \leq \|h_{S_0^c}\|_1 + \|h_{S_0}\| \leq \frac{2(1+\rho)}{1-\rho}\sigma_k + \frac{4d}{(1-\rho)a^2}\zeta,$$

which is the same as (2.57).

To prove (2.58), we mimic the proof of Theorem 2.7. First, by Lemma 1.1 and in particular (1.35), we have that, for each $p \in (1, 2]$,

$$\begin{aligned} \|h_{S_0^c}\|_p &\leq \frac{1}{k^{1-1/p}}\|h\|_1 \\ &\leq \frac{1}{k^{1-1/p}} \left[\frac{2(1+\rho)}{1-\rho}\sigma_k + \frac{4d}{(1-\rho)a^2}\zeta \right]. \end{aligned} \tag{2.76}$$

Next, by Hölder's inequality, (2.73) and (2.75), we get

$$\begin{aligned}\|h_{S_0}\|_p &\leq k^{1/p-1/2}\|h_{S_0}\|_2 \leq k^{1/p-1/2}\|h_0+h_1\|_2 \\ &\leq \frac{1}{k^{1-1/p}}\left[\frac{2d}{a^2\sqrt{k}}\zeta + \rho\frac{\|h_{S_0^c}\|_1}{\sqrt{k}}\right] \\ &\leq \frac{1}{k^{1-1/p}}\left[\frac{2d}{a^2\sqrt{k}}\zeta + \rho\left(\frac{2(1+\rho)}{1-\rho}\sigma_k + \frac{4d}{(1-\rho)a^2}\zeta\right)\right]. \end{aligned} \tag{2.77}$$

Adding (2.76) and (2.77) gives (2.58), and substituting $p = 2$ gives (2.59). □

Proof of Theorem 2.11: To be filled in later.

2.3 Extension to the Matrix Recovery Problem

The results from [30]

Notes and References

The connection between the null space of the measurement matrix and signal recovery has a fairly long history. Early versions of the null space condition can be found in [19, 20, 21, 18]. The contents of Section 2.1.1 mostly follow [13], which also introduces the phrase "null space condition." The phrases "stable null space condition" and "robust null space condition" are introduced in [22]. Theorem 2.6 is stated in this form in [22, Theorem 4.19]; however, the proof based on matrix inequalities is introduced here. This approach leads to the present Theorem 2.7 giving slightly sharper bounds compared to [22, Theorem 4.25].

The restricted isometry property (RIP) is in some sense the culmination of a variety of conditions that were tried out in the literature. Precursors of this property include "uniform uncertainty principle" [11] which also introduces the phrases "restricted isometry" and "restricted orthogonality." The latter property is not used here, as it has been subsumed by subsequent developments in the literature. The RIP is also used in [10].

There are several papers that show that robust k-sparse recovery is possible under an appropriate RIP condition. Previously known sufficient conditions on the RIP constant include $\delta_{2k} < \sqrt{2}-1$ [8] (see also [14]), $\delta_{2k} < 0.472$ [4], $\delta_k < 0.307$ [3], $\delta_k < 1/3$ together with $\delta_{2k} < 0.5$ [6], and some others. The proof of the bound $\delta_{tk} < \sqrt{(t-1)/t}$ for all $t > 1$ is given in [7]. However, the proof that the above condition on δ_{tk} leads to an ℓ_2-robust null space property is original. In the other direction, the proof that, if $t \geq 4/3$, then the bound on δ_{tk} is tight is also found in [7]. A precursor to this result is found in [15], where it is shown that the bound $\delta_{2k} < 1/\sqrt{2}$ is tight, which is a special case of the result in [7].

Chapter 3

Construction of Measurement Matrices

The results presented in Chapter 2 show the central role played by measurement matrices satisfying the restricted isometry property (RIP). In the present chapter, several methods are presented for constructing a matrix A such that A satisfies the RIP of order k with constant δ_k, where both k and δ_k are specified by the user. Note that the integer n, corresponding to the dimension of the unknown vector, is fixed. The challenge therefore is to determine an integer m and a matrix $A \in \mathbb{R}^{m \times n}$ such that

$$(1 - \delta_k)\|u\|_2^2 \leq \|Au\|_2^2 \leq (1 + \delta_k)\|u\|_2^2, \ \forall u \in \Sigma_k.$$

Broadly speaking, there are two approaches to constructing A, namely deterministic and probabilistic. The deterministic approach consists of constructing matrices with low "coherence" (defined in Section 3.1). It turns out that $m \times n$ matrices with minimum coherence can be constructed only if $n \approx m^2/2$ or less. Some authors refer to this as the "quadratic bottleneck." In this approach, the number of measurements scales as the maximum of k^2/δ_k^2 and $n^{1/2}$. Therefore, when $k \leq n^{1/4}$, the number of measurements m scales as the square root of the dimension n. In contrast, in the probabilistic approach, the measurement matrix A consists of independent samples of a common underlying random variable, which has to be "sub-Gaussian." This property is defined in Section 3.2. In this approach, the resulting matrix A satisfies the RIP with high probability that can be made as close to 1 as the user wishes, but cannot be made exactly equal to 1. The main advantage of using probabilistic methods is that the number of measurements m scales as $O(k \log n)$. Thus, *in principle* the probabilistic approach results in smaller values of m. It turns out that the O symbol hides a *huge* constant, making this approach less attractive compared to the deterministic approach unless n is of the order of a billion or so. However, if a *normal random variable* is used (instead of a generic sub-Gaussian random variable), then there are special properties of Gaussian matrices that make this approach attractive.

One of the attractions of using the deterministic approach is that the measurement matrix A can be made both sparse as well as "multiplication-free." In other words, most entries of A can be chosen to be zero, and all nonzero entries can be chosen to be $\pm c$ for some fixed constant c. In this manner, the actual computational burden of solving various optimization problems in order to recover the vector x from measurements of the form Ax or $Ax + \eta$ can be greatly reduced. One of the major drawbacks of using actual normal random variable to construct A is that, with probability 1, *every single entry* of the matrix A is nonzero. Moreover, the substantial reduction in m with normal random variables is achieved due to subtle cancellations in "truly" Gaussian matrices. Therefore these cancellations would be destroyed if the Gaussian samples were to be implemented with finite precision. In this respect, deterministic approaches are more robust against implementation issues, especially when A is multiplication-free in the sense described above. Finally, it is also possible to construct multiplication-free matrices using a probabilistic approach; this topic is discussed in the sequel.

3.1 Matrices with Low Coherence

The deterministic approach to construcing matrices that have the RIP is based on the Gerschgorin circle theorem [23], [35, Theorem 1.11]. Suppose $A \in \mathbb{C}^{n \times n}$. Define the n circles

$$\mathcal{C}_i = \{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \}, i \in [n].$$

Then every eigenvalue of A lies in one of the circles $\mathcal{C}_i$. To put it in another way, the spectrum of A is contained in the union $\cup_i \mathcal{C}_i$.

The next concept is used to construct matrices with the RIP. Throughout this chapter, we say that a matrix $A \in \mathbb{C}^{m \times n}$ is **column-normalized** if $\|a_j\|_2 = 1$ for all $j \in [n]$, or in words, every column of A has unit ℓ_2-norm.

Definition 3.1. Suppose $A \in \mathbb{C}^{n \times n}$ is column-normalized. The **one-column coherence** $\mu_1(A)$ is defined as

$$\mu_1(A) := \max_{i \in [n]} \max_{j \in [n] \setminus \{i\}} |\langle a_i, a_j \rangle|. \tag{3.1}$$

The **k-column coherence** $\mu_k(A)$ is defined as

$$\mu_k(A) := \max_{i \in [n]} \max_{S \subseteq [n] \setminus \{i\}, |S| \leq k} \sum_{j \in S} |\langle a_i, a_j \rangle|. \tag{3.2}$$

It is easy to see that (3.1) is a special case of (3.2) when $k = 1$. Note that the above terminology is not entirely standard. In [22], the quantity $\mu_s(A)$ is referred to as the "ℓ_1-coherence function," presumably because the right side of (3.2) is the ℓ_1-norm of the k-dimensional vector $[\langle a_i, a_j \rangle, j \in S]$. In principle one could define the coherence as *any* ℓ_p-norm of this vector, and any such definition would reduce to (3.1) when $k = 1$. However, in practice the ℓ_1-norm of the vector is the only used to define coherence (at least thus far); so we opt to simplify terminology by referring to the right side of (3.2) simply as the "coherence" without any reference to the ℓ_1-norm.

It is easy to see that

$$\mu_{k+1}(A) \geq \mu_k(A),$$

and also that

$$\mu_k(A) \leq k \mu_1(A), \tag{3.3}$$

because for each $i \in [n]$ and each $j \in [n] \setminus \{i\}$, we have that $|\langle a_i, a_j \rangle| \leq \mu_1(A)$. Therefore, whenever $S \subseteq [n] \setminus \{i\}$, we have that

$$\sum_{j \in S} |\langle a_i, a_j \rangle| \leq |S| \cdot \mu_1(A).$$

Note that a column-normalized matrix A has small coherence if all its columns are nearly orthogonal to each other. Therefore it is not surprising that there is a relationship between coherence and the RIP.

Theorem 3.1. *Suppose $A \in \mathbb{C}^{m \times n}$ is column-normalized. Let $k \leq m$ be a fixed integer. Suppose $S \subseteq [n]$ and that $|S| \leq k$. Then*

$$[1 - \mu_{k-1}(A)] \|u\|_2^2 \leq \|Au\|_2^2 \leq [1 + \mu_{k-1}(A)] \|u\|_2^2, \ \forall u \in \Sigma_k. \tag{3.4}$$

Proof: We will actually prove the following result, which will in turn imply (3.4). Let $S \subseteq [n]$ with $|S| = k$. Let $A_S \in \mathbb{C}^{m \times k}$ denote the submatrix of A corresponding to the columns in S. Then

$$1 - \mu_{k-1}(A) \leq \lambda_{\min}(A_S^t A_S) \leq \lambda_{\max}(A_S^t A_S) \leq 1 + \mu_{k-1}(A). \tag{3.5}$$

To prove (3.5), observe that

$$(A_S^t A_S)_{ji} = |\langle a_j, a_i \rangle|.$$

Because all columns of A are ℓ_2-normalized, it follows that the diagonal elements of $A_S^t A_S$ are all equal to one. Now, for each index $i \in [n]$, the sum of the moduli of the off-diagonal elements of the i-th row of $A_S^t A_S$ is

$$\sum_{j \in S \setminus \{i\}} |\langle a_j, a_i \rangle| \leq \mu_{k-1}(A).$$

Therefore each of the k Gerschgorin circles $\mathcal{C}_i$ corresponding to the matrix $A_S^t A_S$ is centered at one and has radius less than or equal to $\mu_{k-1}(A)$. Thus the union of all Gerschborin circles is also contained in the circle of radius $\mu_{k-1}(A)$ centered at one. But since $A_S^t A_S$ is symmetric, all of its eigenvalues are real, which means that all of its eigenvalues lie in the interval $[1 - \mu_{k-1}(A), 1 + \mu_{k-1}(A)]$. This is the same as (3.5). □

Lemma **??** shows the importance of constructing matrices with low coherence. This raises the question of how low the coherence of a matrix can be. Now we present two very important *lower bounds* on the coherence of a matrix. The bounds are stated for complex matrices for greater generality; therefore the bounds apply also to real matrices.

Lemma 3.1. *Suppose $A \in \mathbb{C}^{m \times n}$ has ℓ_2-normalized columns. Then*

$$\mu_1(A) \geq \sqrt{\frac{n-m}{m(n-1)}} =: c(m,n). \tag{3.6}$$

Moreover, equality holds in (3.6) if and only if (i) $|\langle a_j, a_k \rangle| = c(m,n)$ for all $j, k \in [n]$ with $j \neq k$, and (ii) $AA^* = \lambda I_n$ for some constant $\lambda > 0$.

The above bound for the coherence of a matrix is called the “Welch bound.” A matrix satisfying condition (i) of the theorem is called “equiangular,” while a matrix satisfying condition (ii) is called a “tight frame,” according to [22].

Proof: Define the Gram matrix $G \in \mathbb{C}^{n \times n}$ by $G = A^* A$. Then clearly $g_{jk} = \langle a_j, a_k \rangle$ for all $j, k \in [n]$. Moreover, sice A is column-normalized, it follows that $g_{jj} = 1$ for all $j \in [n]$. Now define $H = AA^* \in \mathbb{C}^{m \times m}$, and note that H is the Gram matrix of the *rows* of A. Now the identity $\text{tr}(AB^*) = \text{tr}(A^* B)$ implies that $\text{tr}(H) = \text{tr}(G) = n$. Next, we apply Schwarz's inequality to the Frobenius inner product $\langle \cdot, \cdot \rangle_F$, and reason as follows:

$$\begin{aligned} n^2 &= [\text{tr}(H)]^2 = [\text{tr}(HI_m)]^2 \\ &= |\langle H, I_m \rangle|^2 \leq \|H\|_F^2 \|I_m\|_F^2 = m\|H\|_F^2 \end{aligned} \tag{3.7}$$

Next, observe that

$$\begin{aligned} \|H\|_F^2 &= \text{tr}(HH^*) = \text{tr}(AA^*AA^*) \\ &= \text{tr}(A^*AA^*A) = \text{tr}(GG^*). \end{aligned}$$

Moreover, for each $j \in [n]$, we have that

$$\begin{aligned} (GG^*)_{jj} &= \sum_{k=1}^{n} |g_{jk}|^2 = |g_{jj}|^2 + \sum_{k \neq j} |g_{jk}|^2 \\ &= 1 + \sum_{k \neq j} |\langle a_j, a_k \rangle|^2 \leq 1 + (n-1)\mu_1^2(A), \end{aligned} \tag{3.8}$$

because $|\langle a_j, a_k \rangle| \leq \mu_1(A)$ whenever $j \neq k$. Therefore

$$\begin{aligned} \|H\|_F^2 &= \text{tr}(GG^*) = \sum_{j=1}^{n} (GG^*)_{jj} \\ &\leq n[1 + (n-1)\mu_1^2(A)]. \end{aligned}$$

Substituting this bound for $\|H\|_F$ into (3.7) gives

$$n^2 \leq mn[1 + (n-1)\mu_1^2(A)].$$

After dividing both sides by n and rearranging, we get

$$\mu_1^2(A) \geq \frac{n-m}{m(n-1)},$$

which is the same as (3.6).

In order for equality to hold in (3.6), both the inequalities above must be equalities. In particular, equality must hold in Schwarz's inequality in (3.7), which is possible if and only if H is "aligned" with the identity matrix, that is, $H = \lambda I_m$ for some constant λ. Since H is Hermitian, λ must be real; since H is positive semidefinite, $\lambda \geq 0$, and since $H \neq 0$, $\lambda > 0$. Similarly, equality holds in (3.8) only if $|\langle a_j, a_k \rangle| = c(m,n)$ for *every* $j, k \neq j$. □

Lemma 3.2. *Suppose* $A \in \mathbb{C}^{m \times n}$ *has* ℓ_2*-normalized columns Then for all* $k \leq \sqrt{n-1}$*, we have*

$$\mu_k(A) \geq kc(m,n) = k\sqrt{\frac{n-m}{m(n-1)}}. \tag{3.9}$$

Equality holds only if conditions (i) and (ii) of Lemma 3.1 hold.

Proof: To be filled in later.

Lemma 3.2 shows that, if we try to construct matrices having low coherence and use the Gerschgorin circle theorem to estimate the RIP constant, then the best estimate we can get from (3.4) is

$$\delta_k \geq (k-1)\sqrt{\frac{n-m}{m(n-1)}} \approx \frac{k-1}{\sqrt{m}} \text{ if } n \gg m.$$

Since we wish to have $n \gg m$ in order to have "compressed" sensing, the above bound is asymptotically tight. Turning this around, we see that, in order to achieve an RIP constant of δ_k with order k, we need to take

$$m \geq \left(\frac{k-1}{\delta_k}\right)^2$$

measurements, if $n \gg m$. Potentially there is room for improvement, because the Gerschgorin circle theorem provides only a sufficient condition, even though the bound for $\mu_k(A)$ is tight in view of Lemma 3.2. Be that as it may, in the constructions below that make use of matrices with low coherence, we treat $1/\sqrt{m}$ as the (asymptotic) lower bound for $\mu_1(A)$, and consider a construction procedure to be "optimal" if it achieves this lower bound.

Recall from Theorem 2.8 that, in order for the RIP to lead to robust k-sparse vector recovery, it is sufficient for the matrix A to satisfy the inequality

$$\delta_{tk} < \sqrt{\frac{t-1}{t}}$$

for some $t > 1$. Now suppose we use the approach of constructing matrices with low coherence in order to satisfy the above sufficient condition. In this approach, the integer k is fixed, as that is the sparsity count of the unknown vector x or at least, the integer with respect to which we wish to compute the k-sparsity index $\sigma_k(x, \|\cdot\|_1)$. Now the bound for δ_{tk} obtained by applying the Gerschgorin circle theorem is given in (3.4). Combining (3.4) with the above requirement on δ_{tk} leads to

$$(tk-1)\mu_1(A) < \sqrt{\frac{t-1}{t}}.$$

Note that $tk - 1 < tk$. Therefore, if

$$tk\mu_1(A) < \sqrt{\frac{t-1}{t}}, \text{ or } k\mu_1(A) < \sqrt{\frac{t-1}{t^3}},$$

then automatically it follows that

$$\delta_{tk} \leq (tk-1)\mu_1(A) < \sqrt{\frac{t-1}{t}}.$$

Therefore this raises the question as to what an "optimal" choice of t might be. Note that t need not be an integer, so long as tk is an integer. Elementary calculus shows that $\sqrt{(t-1)/t^3}$ is maximized when $t = 1.5$, and equals $1/\sqrt{3} \approx 0.577$. Of course, it is possible to maximize the ratio $\sqrt{(t-1)/(t(tk-1)^2)}$ with respect to t for a given value of k, but the optimal value of t would not be much different from $1/\sqrt{3}$. Therefore we could potentially use the bound

$$\delta_{1.5k} < \sqrt{1/3} \approx 0.577$$

as the requirement for the RIP. However, the closer $\delta_{1.5k}$ is to this number, the larger will be the bounds on the residual error. So it is advisable to choose a bound that is sufficiently far from this limit, say 0.5. Therefore, in all the examples in this chapter based on constructing matrices with low coherence, we will strive to ensure that $\delta_{1.5k} < 0.5$.

3.2 Deterministic Approaches

In this section, we present several methods for constructing matrices with known (low) coherence. These methods are all deterministic, in contrast to the probabilistic methods to be discussed in the next section.

First we present a method of constructing a matrix of dimensions $p^2 \times p^{r+1}$, where p is a prime number and $r \geq 1$ is an integer, such that its one-colmn coherence is $1/p$. Since the matrix has p^2 rows, this is the asymptotically optimal lower bound. In addition, the matrix is extremely sparse and multiplication-free. This construction is due to [17].

The construction is as follows: Let p be any prime number. Then the quotient $\mathbb{Z}_p := \{0, 1, \ldots, p-1\}$, with addition and multiplication defined modulo p, is a field. More generally, if p is a *power of a prime number*, then it is possible to define a field with p elements, which is again denoted by $\mathbb{Z}_p$. The definition of arithmetic operations in $\mathbb{Z}_p$ when p is a power of a prime is more subtle than in the case where p is a prime. The reader is referred to standard texts for these details; see [26, 25, 27] as examples of such texts, though many others could be cited. Suppose a is a polynomial of degree r or less with coefficients in $\mathbb{Z}_p$, and define its "graph" as the set of all pairs $(x, a(x))$ as x varies over $\mathbb{Z}_p$. To illustrate, suppose $p = 3$, $r = 2$, and that $a(x) = 1 + 2x + x^2$. Then $a(0) = 0$, $a(1) = 1$, and $a(2) = 0$. Now define a $p \times p$ matrix M, where $M_{ij} = 1$ if $a(i) = j$, and $M_{ij} = 0$ otherwise. Thus for $p = 3$ and $a(x) = 1 + 2x + x^2$, we get

$$M = \begin{array}{c c} & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \end{array} & \left[\begin{array}{ccc} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{array} \right] \end{array}$$

Now write out this $p \times p$ matrix as a $p^2 \times 1$ column vector, proceeding row by row and then taking the transpose. In this example, the corresponding 9×1 column vector (written as a row vector for convenience) is

$$u_a = [\ 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0\ ]^t.$$

Observe now that each block of p elements contains *exactly* one 1 with the remaining $p - 1$ elements being equal to zero. Therefore $\|u_a\|_2 = p$. Let $\Pi_r(\mathbb{Z}_p)$ denote the set of all polynomials of degree r or less with coefficients in $\mathbb{Z}_p$. In other words,

$$\Pi_r(\mathbb{Z}_p) := \left\{ a(x) = \sum_{i=0}^{r} a_i x^i, a_i \in \mathbb{Z}_p \right\}.$$

Note that $\Pi_r(\mathbb{Z}_p)$ contains precisely p^{r+1} polynomials, because each of the $r+1$ coefficients can assume p different values.[1] Now define

$$A' := [u_a, a \in \Pi_r(\mathbb{Z}_p)], A = \frac{1}{\sqrt{p}} A'. \tag{3.10}$$

Note that $A \in \{0,1\}^{p^2 \times p^{r+1}}$ and that A is column-normalized.

Theorem 3.2. *The matrix A defined in* (3.10) *satisfies*

$$\mu_1(A) \leq \frac{r}{p}. \tag{3.11}$$

Therefore A satisfies the RIP of order k with constant

$$\delta_k = \frac{(k-1)r}{p}, \ \forall k \leq 1 + \frac{p}{r}. \tag{3.12}$$

Proof: Suppose $a, b \in \Pi_r(\mathbb{Z}_p)$ are distinct polynomials, and consider the inner product

$$\langle u_a, u_b \rangle = \sum_{l=1}^{p^2} (u_a)_l \cdot (u_b)_l.$$

The product $(u_a)_l \cdot (u_b)_l = 0$ unless $(u_a)_l = (u_b)_l = 1$. Let us restore the double subscript notation and write $l = (i,j)$. Then $(u_a)_l = (u_b)_l = 1$ for $l = (i,j)$if and only if $a(i) = b(i) = j$, or equivalently, $a(i) - b(i) = 0$. The polynomial $a - b$ has degree no larger than r. Therefore it can have no more than r zeros. Therefore the relationship $(u_a)_l = (u_b)_l = 1$ can hold for no more than r distinct values of l. In other words,

$$0 \leq \langle u_a, u_b \rangle \leq r, \ \forall a, b \in \Pi_r(\mathbb{Z}_p), a \neq b.$$

Now (3.11) follows from the fact that $A = A'/\sqrt{p}$. Now (3.12) is a consequence of Lemma **??**. □

Note that the matrix A has dimensions $p^2 \times p^{r+1}$. Therefore the bound in (3.11) becomes

$$\mu_1(A) = \frac{r}{\sqrt{m}}. \tag{3.13}$$

In other words, the bound for $\mu_1(A)$ is worse by a factor of r compared to $1/\sqrt{m}$, which is the (approximately) optimal lower bound due to Welch. Specifically, if we choose $r = 3$ so that $n = p^4 = m^2$, then the bound for $\mu_1(A)$ is worse by a factor of 3, whereas if we choose $r = 2$ so that $n = p^3$, then the bound is worse by a factor of 2. On the other hand, this matrix two clear advantages. First, it is *extremely sparse*, with only a fraction $1/p$ of the elements being nonzero. It is also "multiplication-free," in the sense that every nonzero element has the same value, and equals $1/\sqrt{p}$.

Example 3.1. To illustrate the construction of measurement matrices using Theorem 3.2 and (3.4), suppose that $n = 10,000$ and $k = 6$. To satisfy the RIP, we would like to ensure that

$$\delta_{tk} < \sqrt{\frac{t-1}{t}}$$

[1] If the leading coefficient of a polynomial is zero, then the degree would be less than r.

for some $t > 1$. Moreover, as discussed previously, an “optimal” choice for t is 1.5, and it is reasonable to ensure that

$$\delta_{1.5k} \leq 0.5,$$

or equivalently that

$$\mu_1(A) \leq \frac{0.5}{1.5k}.$$

Given that r/p is an upper bound for $\mu_1(A)$, and that we must have $n \leq p^r$ (because the matrix A has dimensions $p^2 \times p^r$), the prime or prime power p must satisfy

$$\frac{r}{p} \leq \frac{0.5}{1.5k} \text{ and } n \leq p^r,$$

or equivalently

$$p \geq \max\{3kr, n^{1/r}\}.$$

We can experiment with different values for the integer r, starting with $r = 3$ (because $r = 2$ would lead to $n = m$, or a square matrix A). Choosing $r = 3$ leads to $p \geq 54$. Hence we select $p = 59$, which is the smallest prime power greater than or equal to 54, which in turn leads to $m = 59^2 = 3,481$. Now $p^3 = 205,379 > n$. Therefore the full construction using Theorem 3.2 would lead to a matrix with 3,481 rows and 205,379 columns, whereas we require only 10,000 columns. It should be clear that if A is a matrix that satisfies the RIP of order k with constant δ_k, then any subset of columns of A also satisfies the RIP with the same k and δ_k (and possibly a larger δ_k). So we can stop the construction procedure once we read 10,000 columns.

Now we present another construction due to [36], that takes a partial Fourier matrix and establishes bounds on its coherence. Recall that, if n is any integer, then the corresponding Fourier matrix $\mathcal{F}_n \in \mathbb{C}^{n\times n}$ is a matrix with entries defined as follows:

$$(\mathcal{F}_n)_{jk} = (1/\sqrt{n}) \exp 2\pi \mathbf{i} jk/n, j, k \in [n], \tag{3.14}$$

where $\mathbf{i} = \sqrt{-1}$. There is another way to represent the matrix $\mathcal{F}_n$. Define $\omega_n := \exp(2\pi\mathbf{i}/n)$ to be the primitive n-th root of unity on the unit circle in the complex plane. Let $\lambda_i = \omega_n^i, i \in [n]$. Then $\sqrt{n}\mathcal{F}_n$ is just the Vandermonde matrix corresponding to the constants $\lambda_i, i \in [n]$, with a cyclic row and column permutation. Further, it can be shown that $\mathcal{F}_n^* \mathcal{F}_n = I_n$, so that the matrix $\mathcal{F}_n$ is unitary. If $x \in \mathbb{C}^n$ is any n-dimensional vector, then its **discrete Fourier transform** is the vector $\tilde{x} := \mathcal{F}_n x$. Because $\mathcal{F}_n$ is unitary, the inverse discrete Fourier transform of $\tilde{x}$ is given simply by $\mathcal{F}_n^* \tilde{x}$. For further details, the reader can consult any standard text, such as [33, 2] for example.

Note that the elements of $\mathcal{F}_n$ are *complex* numbers. If one wishes to avoid complex numbers and stay with the realm of real numbers, then the discrete *Fourier* transform can be replaced by the discrete *cosine* transform, which is discussed in [2]. There are several variants of the discrete cosine transform. The one used here is the matrix $\mathcal{C}_n \in \mathbb{R}^{n\times n}$ defined by

$$(\mathcal{C}_n)_{jk} = w_j f_{jk}, \tag{3.15}$$

where

$$w_j = \left\{ \begin{array}{ll} 1/\sqrt{n}, & k = 1, \\ \sqrt{2/n}, & k = 2, \ldots, n. \end{array} \right. , f_{jk} = \cos\left(\frac{\pi}{2n}(2n-1)(k-1)\right), j, k \in [n]. \tag{3.16}$$

The matrix $\mathcal{C}_n$ belongs to $\mathbb{R}^{n\times n}$ and is orthogonal. Therefore, given a real vector $x \in \mathbb{R}^n$, its discrete cosine transform $\tilde{x}_c \in \mathbb{R}^n$ is given by $\mathcal{C}_n x$.

Until now we have been concerned with the case where the unknown vector x is k-sparse, or nearly so. To put it another way, the unknown vector x has a sparse representation with respect to the standard orthonormal basis $\{\mathbf{e}_i, i \in [n]\}$, where the i-th component of $\mathbf{e}_i$ equals one and the rest equal zero. However, it is possible that the vector x is not sparse with respect to the standard basis, but has a sparse representation with respect to another basis. A common situation occurs when the discrete Fourier transform or discrete cosine transform of x is sparse; to put it another way, the "time domain" representation of x is not sparse, whereas the "frequency domain" representation of x is sparse. Suppose to be specific that $\tilde{x}$ is the discrete Fourier transform of x, and suppose further that $\tilde{x}$ is k-sparse or nearly so. In this case the approach is to select m rows of the matrices $\mathcal{F}_n$ or $\mathcal{C}_n$, that is, to measure just m out of the n components of the discrete Fourier or discrete Cosine transform of x. Then one attempts to recover x from these m measurements. Suppose to be specific that $S \subseteq [n]$ with $|S| = m$ is the index set of the frequency components selected, and let F^S denote the $m \times n$ submatrix of $\mathcal{F}_n$; define C^S analogously. The objective is to choose the integer m and the index set S in such a way that the matrix F^S satisfies the RIP.

Now we change notation slightly. The problem under study is to choose $m < n$ rows of $\mathcal{F}_n$ such that the resulting matrix $\Phi \in \mathbb{C}^{m \times n}$ satisfies the RIP of order k and constant δ, both of which are specified. The construction is quite elaborate. In fact two separate constructions are given in [36]. In the first, a prime number p is chosen, along with an integer a. Letting $q = p^a$, the procedure results in a matrix $\Phi \in \mathbb{C}^{q \times (q^s - 1)}$, which satisfies

$$\mu_1(\Phi) \leq \frac{s-1}{\sqrt{q}}. \tag{3.17}$$

Hence the one-colulmn coherence is suboptimal by a factor of $s - 1$ compared to the Welch bound. In particular, if we choose $q = 2$ so that the matrix Φ has dimensions $q \times q^2 - 1$, then the coherence is optimal. In the second construction, the number of rows m again equals $q = p^a$, a power of a prime, while the number of columns n equals

$$n = \frac{q^s - 1}{p^b - 1},$$

where b is another integer that divides a. Clearly, if we take $b = a$, then we get back the previous construction. In this case too, the inequality (3.17) holds.

Now we describe the construction in simplified form. There are some facts about finite fields that the reader needs to know. More details about all the items below can be found in any standard text, for example [26, 25, 27].

1. If $q = p^a$ where p is a prime and a is an integer, then there exists a (in fact one and only one) field with precisely q elements. This field is denoted by $\mathbb{F}_q$. Note that some authors also write $GF(q)$ to stand for "Galois Field with q elements," in honor of Evariste Galois who first introduced the notion of a field with a finite number of elements.

2. The set $\mathbb{F}_q \setminus \{0\}$ consisting of all nonzero elements of $\mathbb{F}_q$ is a group under multiplication, and is denoted by $\mathbb{F}_q^*$.

3. There exists at least one (and in fact, several) elements $g \in \mathbb{F}_q^*$ that are "generators" of $\mathbb{F}_q^*$. In other words, the set of successive powers of g, namely $\{g, g^2, g^3, \ldots, g^{q-1}\}$ equals $\mathbb{F}_q^*$.

4. Consequently, if $u \in \mathbb{F}_q^*$, then there exists a unique integer $r \in [q-1]$ such that $g^r = u$. We call r the "logarithm of u with respect to g," and write $r = \log_g u$. Note that $g^{q-1} = 1$, so that if $u \neq 1$, then $\log_g u \leq q - 2$.

5. Now consider the integer q^s. Since q is a power of a prime, so is q^s. Therefore there exists a field $\mathbb{F}_{q^s}$ with q^s elements. Moreover, $\mathbb{F}_q$ is a subfield of $\mathbb{F}_{q^s}$, and $\mathbb{F}_{q^s}$ is called an "extension field" of $\mathbb{F}_q$. There exists an element $\alpha \in \mathbb{F}_{q^s} \setminus \mathbb{F}_q$ such that every element in $\mathbb{F}_{q^s}$ can be expressed as a "polynomial" $c_0 + c_1\alpha + \ldots c_{s-1}\alpha^{s-1}$, where all the "coefficients" c_i belong to $\mathbb{F}_q$. In case $s = 2$, $\mathbb{F}_{q^2}$ is called a "quadratic extension" of $\mathbb{F}_q$, in which case *every* element $\alpha \in \mathbb{F}_{q^2} \setminus \mathbb{F}_q$ has this property. However, if $s \geq 3$, the element α has to be chosen such that the above property holds.

The matrix Φ is constructed as follows: Choose a prime power $q = p^a$ and another integer s such that $q^s - 1$ is at least as large as n, the desired number of columns of Φ. Construct the Fourier matrix $\mathcal{F}_{q^s-1} \in \mathbb{C}^{q^s \times q^s}$. Then choose m rows of this Fourier matrix as follows: Choose a generator g of the multiplicative group $\mathbb{F}_{q^s}^*$, and an element $\alpha \in \mathbb{F}_{q^s} \setminus \mathbb{F}_q$ with the property described above. Let t vary over the smaller field $\mathbb{F}_q$, and define

$$M = \{\log_g(t - \alpha), t \in \mathbb{F}_q\}.$$

Note that, because $t \in \mathbb{F}_q$ and $\alpha \in \mathbb{F}_{q^s} \setminus \mathbb{F}_q$, the difference $t - \alpha$ belongs to $\mathbb{F}^*_{q^s}$. Therefore the logarithm is well-defined. Moreover, distinct values of $t \in \mathbb{F}_q$ have distinct logarithms. Therefore the set M has cardinality q. The matrix Φ consists of the rows of the Fourier matrix $\mathcal{F}_{q^s-1}$ corresponding to the indices in M.

Theorem 3.3. *With $\Phi \in \mathbb{C}^{q \times (q^s-1)}$ defined as above, we have that*

$$\mu_1(\Phi) \leq \frac{s-1}{\sqrt{q}}. \tag{3.18}$$

The next theorem, which is stated without proof, presents another method for constructing a matrix with complex entries that has optimal coherence. For a proof and background material, see [22, Proposition 5.13].

Theorem 3.4. *Let m be a prime number ≥ 5. Then there exists a matrix $B_m \in \mathbb{C}^{m \times m^2}$ such that $\mu_k(B_m) = k/\sqrt{m}$ for all $k \leq m^2 - 1$.*

Note that B_m belongs $\mathbb{C}^{m \times m^2}$, not necessarily $\mathbb{R}^{m \times m^2}$. An explicit formula for the matrix B_m is given in [22] following the above proposition, and is repeated below. For a prime number m, let $\mathbb{Z}_m$ denote the field of integers modulo m. In other words, $\mathbb{Z}_m$ consists of $\{0, 1, \ldots, m-1\}$ with addition and multiplication defined modulo m. Thus it must be remembered that all arithmetic operations in the formulae below are modulo m. Also, throughout we use $\mathbf{i}$ to denote $\sqrt{-1}$.

Let $\mathbf{x} \in \mathbb{C}^m$ denote the "Alltop" vector

$$x_j = \frac{1}{\sqrt{m}} \exp(2\pi \mathbf{i} j^3/m), j = 0, 1, \ldots, m.$$

Because each x_j has magnitude $1/\sqrt{m}$, it is clear that $\mathbf{x}$ is ℓ_2-normalized. Next define two operations T_k and M_l as follows:

$$(T_k \mathbf{x})_j = x_{j-k}, (M_l \mathbf{x})_j = z_j \cdot \exp(2\pi \mathbf{i} lj/m), 0 \leq k, l \leq m-1.$$

The reader is again reminded that expressions such as $j^3, j-k$ and lj are all to be interpreted modulo m. With these definitions, the matrix $B_m \in \mathbb{C}^{m \times m^2}$ is given by

$$B_m = [M_l T_k \mathbf{x}, 0 \leq k, l \leq m-1].$$

In other words, B_m looks like

$$B_m = [M_0 T_0 \mathbf{x} | \ldots | M_0 T_{m-1} \mathbf{x} | \ldots | M_{m-1} T_0 \mathbf{x} | \ldots | M_{m-1} T_{m-1} \mathbf{x}]. \tag{3.19}$$

It is shown in [22, pp. 121-123] through tedious but routine calculations that

$$|\langle M_l T_k \mathbf{x}, M_{l'} T_{k'} \mathbf{x} \rangle|^2 = \frac{1}{m} \text{ if } (l, k) \neq (l', k').$$

Therefore every pair of distinct columns of B_m makes an inner product of *exactly* $1/\sqrt{m}$. This means not only that $\mu_1(B_m) = 1/\sqrt{m}$, but also that $\mu_k(B_m) = k\mu_1(B_m)$.

Theorem 3.5. *Suppose n and $k \leq n$ are specified integers, and that $\delta \in (0, 1)$ is a specified real number. Choose a prime number m such that*

$$m \geq \max \left\{ \left(\frac{2k-1}{\delta} \right)^2, \sqrt{n} \right\}. \tag{3.20}$$

Let B_m be defined as in (3.19), and let A denote the submatrix of B_m consisting of the first n (in fact any n) columns of B_m. Then A satisfies the RIP of order $2k$ with constant $\delta_{2k} = \delta$.

Proof: From the claim made (without proof) following the statement of Theorem 3.4, it follows that the inner product of every pair of distinct columns of A is equal to $1/\sqrt{m}$. Therefore, for all k, we have that $\mu_k(A) = k/\sqrt{m}$. Now apply Lemma **??**. □

Now let us understand the implications of the above construction. Given integers n (size of the vector) and k (sparsity index), we can choose m so as to satisfy (3.20) and also be a prime number. Lists of prime numbers are readily available. Therefore finding a suitable m is straight-forward. Out of the two quantities on the right side of (3.20), which one is the operative constraint? If

$$\frac{2k-1}{\delta} > n^{1/4},$$

then the value of m is determined by k and not n.

Let us substitute some typical numbers. We have seen from Theorem 2.8 that the constant δ_{2k} is required to satisfy $\sqrt{2} - 1 \approx 0.414$. So if $n = 10000$, then $n^{1/4} = 10$, and $2k - 1 > (\sqrt{2} - 1)n^{1/4}$ if $k \geq 3$. So, unless k is tiny, the value of m is determined by k and not n. Now suppose $n = 10^4$, and let $k = 10$. Then

$$m \geq \left(\frac{2k-1}{\delta} \right)^2.$$

Let us choose $\delta = 0.25$, so that $1/\delta = 4$. Then we require $m \geq (19 \times 4)^2 = 5776$. The smallest prime number larger than 5776 is 5791, which more than half of n. So one cannot claim much of a "compression" in this case.

3.3 Probabilistic Methods – 1

Suppose X and Y are independent[2] zero-mean random variables with finite variance, and let V_X, V_Y denote their variances. Thus $V_X = E(X^2)$ and similarly for V_Y. If $\alpha, \beta \in \mathbb{R}$, then it is easy to verify that $Z = \alpha X + \beta Y$ also has zero mean, and that $V_Z = \alpha^2 V_X + \beta^2 V_Y$.

Now suppose m, n are specified integers, and that X is a zero-mean random variable with unit variance. One of the most common choices is to let $X \sim N(0, 1)$, that is a Gaussian random variable with zero mean and unit variance, otherwise referred to as a "normal" random variable. A normal random variable is real-valued with the probability density $(1/\sqrt{2\pi}) \exp(-x^2/2)$. Let $\phi_{11}, \ldots, \phi_{mn}$ be pairwise independent copies of X, and in a bit of sloppy notation, let the same symbols $\phi_{11}, \ldots, \phi_{mn}$ denote also *random samples* (that is, realizations) of the same random variable X. Further, let Φ denote the random $m \times n$ matrix consisting of independent copies of X, and also of mn independent samples of X. Finally, let the measurement matrix A equal $(1/\sqrt{m})\Phi$.

Let $u \in \mathbb{R}^n$ be arbitrary. Then the pairwise independence assumption guarantees that

$$E\left[\left(\sum_{j=1}^{n} \phi_{ij} u_j \right)^2 \right] = \sum_{j=1}^{n} E(\phi_{ij} u_j)^2 = \sum_{j=1}^{n} (u_j)^2 = \|u\|_2^2 \ \forall i \in [m].$$

Next, for two distinct indices i_1 and i_2 in $[m]$, the associated random variables

$$\psi_{i_1} = \sum_{j=1}^{n} \phi_{i_1 j} u_j, \psi_{i_2} = \sum_{j=1}^{n} \phi_{i_2 j} u_j$$

[2] Actually, here and in the discussion below, for the most part it is enough for the various random variables to be pairwise uncorrelated; pairwise independence is not needed.

are independent. Therefore

$$\begin{aligned} E\left(\|\Phi u\|_2^2\right) &= E\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{n}\phi_{iju_j}\right)^2\right] = \left[\sum_{i=1}^{m}E\left(\sum_{j=1}^{n}\phi_{iju_j}\right)^2\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}E[(\phi_{ij})^2]u_j^2 \\ &= \sum_{i=1}^{m}\|u\|_2^2 = m\|u\|_2^2. \end{aligned} \tag{3.21}$$

Let us fix $u \in \mathbb{R}^n$, and think of $\|\Phi u\|_2^2$ as a real-valued random variable. Then (3.21) states that the expected value of the random variable $\|\Phi u\|_2^2$ is $m\|u\|_2^2$. Because $A = (1/\sqrt{m})\Phi$, the expected value of the random variable $\|Au\|_2^2$ is $\|u\|_2^2$. Therefore, "on average," multiplication by the random matrix A preserves norms.

Next let us focus attention on the "tail probability"

$$\Pr\left\{\left|\|Au\|_2^2 - \|u\|_2^2\right| > \delta\|u\|_2^2\right\}.$$

Since $u \in \mathbb{R}^n$ is a fixed vector, it can be supposed without loss of generality that $\|u\|_2 = 1$, and we can focus attention on the quantity

$$\Pr\left\{\sup_{u\in\Sigma_k,\|u\|_2=1}\left|\|Au\|_2^2 - 1\right| > \delta\right\}. \tag{3.22}$$

Suppose we are able to show that, for a particular choice of the integer m, this probability is bounded above by ξ. Then it can be stated that, with probability $\geq 1-\xi$, the randomly generated matrix A satisfies the RIP of order k with constant δ. The remainder of this section presents various methods to find an upper bound for this tail probability in terms of the integer m and the statistical properties of the random variable X that is used to generate the measurement matrix A.

We present the main results straightaway, and defer most of the proofs until afterwards.

Definition 3.2. A real-valued random variable X is said to be (α,β)**-sub-exponential** if

$$\Pr\{|X| > t\} \leq \alpha\exp(-\beta t),\ \forall t > 0. \tag{3.23}$$

Definition 3.3. A real-valued random variable X is said to be (γ,ζ)**-sub-Gaussian** if

$$\Pr\{|X| > t\} \leq \gamma\exp(-\zeta t^2),\ \forall t > 0. \tag{3.24}$$

Definition 3.4. A real-valued random variable X is said to be c**-sub-Gaussian** if

$$E[\exp(\theta X)] \leq \exp(c\theta^2),\ \forall \theta \in \mathbb{R}. \tag{3.25}$$

Note that the function $\theta \mapsto E[\exp(\theta X)]$ is known as the **moment generating function**. This is because, if the function is well-defined for all θ in some neighborhood of $\theta = 0$, then

$$\left[\frac{d^l E[\exp(\theta X)]}{d\theta^l}\right]_{\theta=0} = E(X^l),\ \forall l \geq 0.$$

The ambiguity in the notation regarding sub-Gaussian random variables is resolved by counting the number of parameters. If two parameters are specified, then X satisfies (3.24), whereas if only one parameter was specified, then X satisfies (3.25).

There are (apparently) two distinct definitions of a sub-Gaussian random variable, one in terms of the decay rate of the tail probability $\Pr\{|X| > t\}$ and another in terms of the growth rate of the moment generating function $E[\exp(\theta X)]$. It turns out however that both definitions are equivalent.

Theorem 3.6. *Suppose X is c-sub-Gaussian. Then (i) $E(X) = 0$, and (ii) X is (γ,ζ)-sub-Gaussian with $\gamma = 2, \zeta = 1/(4c)$.*

Theorem 3.7. *Suppose X is (γ, ζ)-sub-Gaussian. Choose $\bar{\theta} < \min\{\zeta, \zeta^{1/2}\}$. Then X is c-sub-Gaussian with*

$$c = \max \left\{ \frac{\gamma}{\zeta(1 - \bar{\theta}\zeta^{-1/2})}, \frac{1}{4\bar{\theta}} + \frac{\gamma}{\zeta\bar{\theta}(1 - \bar{\theta}\zeta^{-1})} \right\}. \tag{3.26}$$

For (γ, ζ)-sub-Gaussian random variables with $\zeta \leq 1$, it is possible to simplify the bound (3.26).

Theorem 3.8. *Suppose X has zero mean, and is (γ, ζ)-sub-Gaussian with $\zeta \leq 1$. Then X is c-sub-Gaussian with*

$$c = \frac{1}{2\zeta} + \frac{4\gamma}{\zeta^2}. \tag{3.27}$$

Theorem 3.9. *Suppose X is a normal random variable. Then X is (γ, ζ)-sub-Gaussian with $\gamma = 1, \zeta = 1/2$. X is also c-sub-Gaussian with $c = 1/2$.*

If we were to substitute $\gamma = 1, \zeta = 1/2$ into (3.27), then we would get $c = 17$. Therefore the point of Theorem 3.9 is that for "true" Gaussian (or random) variables, the estimate for c given by (3.27) is extremely conservative. However, for other random variables that satisfy (3.24) with $\gamma = 1, \zeta = 1/2$, that is, the same γ, ζ as for a standard Gaussian, $c = 17$ is the best bound we can get for now. Clearly, the smaller the value of c, the tighter is the bound (3.25).

The next theorem is the main result regarding the number of measurements m needed to construct a matrix A that satisfies the RIP of order k with constant δ, with probability at least $1 - \xi$.

Theorem 3.10. *Suppose X is a random variable with zero mean, unit variance, and suppose in addition that X satisfies* (3.25) *for some constant c. Define*

$$\gamma = 2, \zeta = 1/(4c), \beta = \gamma e^{-\zeta} + e^{\zeta}, \kappa = \zeta, \tag{3.28}$$

$$\tilde{c} := \frac{\kappa^2}{4\beta + 2\kappa}. \tag{3.29}$$

Suppose an integer k and real numbers $\delta, \xi \in (0, 10$ are specified, and that $A = (1/\sqrt{m})\Phi$, where $\Phi \in \mathbb{R}^{m \times n}$ consists of independent samples of X. Then A satisfies the RIP of order k with constant δ with probability $\geq 1 - \xi$ provided

$$m \geq \frac{1}{\tilde{c}\delta^2} \left(\frac{4}{3} k \ln \frac{en}{k} + \frac{14k}{3} + \frac{4}{3} \ln \frac{2}{\xi} \right). \tag{3.30}$$

The above bound for m works whenever the generating random variable X is c-sub-Gaussian. However, in the case where X is a normal random variable, it is possible to give another bound that is often one or two orders of magnitude smaller than that given by (3.30).

Theorem 3.11. *Suppose an integer k and real numbers $\delta, \xi \in (0, 1)$ are specified, and that $A = (1/\sqrt{m})\Phi$, where $\Phi \in \mathbb{R}^{m \times n}$ consists of independent samples of a normal random variable X. Define*

$$g = 1 + \frac{1}{\sqrt{2 \ln(en/k)}}, \eta = \frac{\sqrt{1 + \delta} - 1}{g}. \tag{3.31}$$

Then A satisfies the RIP of order k with constant δ with probability $\geq 1 - \xi$ provided

$$m \geq \frac{2}{\eta^2} \left(k \ln \frac{en}{k} + \ln \frac{2}{\xi} \right). \tag{3.32}$$

Now we present the proofs of the various theorems. To assist in that process, we first prove a few theorems that are widely applicable, not just to the problem under study here.

Theorem 3.12. ***(Markov's Inequality)*** *Suppose X is a nonnegative-valued random variable with finite expectation. Then for all $a > 0$, we have that*

$$\Pr\{X > a\} \leq \frac{E(X)}{a}. \tag{3.33}$$

Proof: Let Φ_X denote the characteristic function of X. Because X is nonnegative-valued, and $E(X) < \infty$, we can write

$$\begin{aligned} E(X) &= \int_0^\infty x d\Phi_X(x) = \int_0^a x d\Phi_X(x) + \int_a^\infty x d\Phi_X(x) \\ &\geq \int_a^\infty x d\Phi_X(x) \geq a \int_a^\infty d\Phi_X(x). \end{aligned}$$

However

$$\int_a^\infty d\Phi_X(x) = \Pr\{X > a\}.$$

Therefore the above inequality can be written as

$$E(X) \geq a \Pr\{X > a\},$$

which is a rearrangement of (3.33). □

Theorem 3.13. ***(Simplified Cramèr's Theorem)*** *Suppose X is a real-valued random variable with the property that $E[\exp(\theta X)] < \infty$ for all $\theta \in \mathbb{R}$. Define the* ***moment generating function*** *$\mu(\theta)$ and the* ***logarithmic momennt generating function*** *$\lambda(\theta)$ by*

$$\mu(\theta) := E[\exp(\theta X)], \lambda(\theta) := \ln \mu(\theta). \tag{3.34}$$

For each $a > 0$, define

$$F(a) := \sup_{\theta \geq 0} \theta a - \lambda(\theta), \tag{3.35}$$

and note that $F(a)$ could be infinite for some a. Suppose $X_1, \ldots, X_l$ are independent copies of X. Then, for each $a > 0$, we have that

$$\Pr\left\{\sum_{i=1}^l X_i > la\right\} \leq \exp[-lF(a)]. \tag{3.36}$$

Proof: Let us define

$$S_l = \sum_{i=1}^l X_i.$$

Then

$$E[\exp(\theta S_l)] = E\left[\exp\left(\theta \sum_{i=1}^l X_i\right)\right] = E\left[\prod_{i=1}^l \exp(\theta X_i)\right] = [\mu(\theta)]^l < \infty, \ \forall \theta \in \mathbb{R}.$$

By applying Markov's inequality to the nonnegative random variable $\exp(\theta S_l)$ (which has finite expectation as shown above), we conclude that

$$\begin{aligned} \Pr\{S_l > la\} &= \Pr\{\exp(\theta S_l) > \exp(la)\} \\ &\leq \frac{E[\exp(\theta S_l)]}{\exp(la)} = \exp(-la)[\mu(\theta)]^l. \end{aligned}$$

Therefore

$$\log \Pr\{S_l > la\} \leq -la + l \ln \mu(\theta) = -la + l\lambda(\theta).$$

Now the point is to note that the above inequality holds for *all* $\theta \geq 0$. So we can "optimize" the inequality by choosing an appropriate θ for each a, as in (3.35). This leads to

$$\log \Pr\{S_l > la\} \leq -lF(a),$$

which is the same as (3.36). □

The above theorem is only a simplified version of Cramèr's theorem. The "true" Cramèr theorem states that the boun in (3.36) is asymptotically tight as $l \to \infty$. The interested reader can consult [16].

Theorem 3.14. ***(Bernstein's Inequality)*** *Suppose X is a real-valued zero-mean random variable that has finite moments of all orders that satisfy the bound*

$$E(|X|^k) \leq \frac{\sigma^2 k!}{2} R^{k-2}, \ \forall k \geq 2, \tag{3.37}$$

for some constants σ, R. Suppose $X_1, \ldots, X_l$ are independent copies of X. Then for all $u > 0$, we have that

$$\Pr\left\{ \left| \sum_{i=1}^{l} X_i \right| > lu \right\} \leq 2 \exp\left[-\frac{lu^2}{2(\sigma^2 + Ru)} \right]. \tag{3.38}$$

Proof: It is shown that

$$\Pr\left\{ \sum_{i=1}^{l} X_i > lu \right\} \leq \exp\left[-\frac{lu^2}{2(\sigma^2 + Ru)} \right]. \tag{3.39}$$

By replacing X by $-X$, it would then follow from (3.36) that

$$\Pr\left\{ \sum_{i=1}^{l} X_i < -lu \right\} \leq \exp\left[-\frac{lu^2}{2(\sigma^2 + Ru)} \right].$$

Combining these two inequalities would then lead to the desired conclusion (3.38).

To prove (3.39), we make use of Theorem 3.13. We begin by finding a bound for the moment generating function $\mu(\theta)$ as defined in (3.34) when the random variable X satisfies (3.37). Because $E[\exp(\theta X)]$ is well-defined for all θ, we can interchange the order of summation and integration and write

$$\mu(\theta) = E[\exp(\theta X)] = \sum_{k=0}^{\infty} \frac{\theta^k}{k!} E(X^k).$$

Note that $E(X) = 0$. Therefore

$$\begin{aligned} \mu(\theta) &= E[\exp(\theta X)] = 1 + \sum_{k=2}^{\infty} \frac{\theta^k}{k!} E(X^k) \\ &\leq 1 + \frac{1}{2} \sum_{k=2}^{\infty} \sigma^2 \theta^k R^{k-2} \\ &= 1 + \frac{\sigma^2 \theta^2}{2} \sum_{i=0}^{\infty} \theta^i R^i \\ &= 1 + \frac{2\sigma^2 \theta^2}{2(1 - \theta R)}, \end{aligned} \tag{3.40}$$

provided that $\theta < R^{-1}$ so that the power series converges. Now $\ln(\cdot)$ is a concave function. Therefore, whenever $u > 0$, we have that $\ln(1 + u) \leq u$. Using this in (3.40) leads to

$$\lambda(\theta) = \ln[\mu(\theta)] \leq \frac{2\sigma^2 \theta^2}{2(1 - \theta R)}.$$

Therefore $F(\cdot)$ defined in (3.35) satisfies

$$\begin{aligned} F(u) &= \sup_{\theta \in \mathbb{R}} \theta u - \lambda(\theta) \geq \sup_{\theta \in (0, R^{-1})} \theta u - \lambda(\theta) \\ &\geq \sup_{\theta \in (0, R^{-1})} \theta u - \frac{2\sigma^2 \theta^2}{2(1-\theta R)}. \end{aligned}$$

In particular, for a given $u > 0$, let us define

$$\bar{\theta} = \frac{u}{\sigma^2 + Ru} = \frac{1}{R} \cdot \frac{Ru}{\sigma^2 + Ru} < R^{-1}. \tag{3.41}$$

This is a valid choice for $\theta \in (0, R^{-1})$ and can be substituted into (3.10), which leads to

$$F(u) \geq \bar{\theta} u - \frac{\sigma^2 \bar{\theta}^2}{2(1 - \bar{\theta} R)}.$$

Now note that

$$1 - \bar{\theta} R = 1 - \frac{Ru}{\sigma^2 + Ru} = \frac{\sigma^2}{\sigma^2 + Ru}.$$

Therefore

$$\begin{aligned} \bar{\theta} u - \frac{\sigma^2 \bar{\theta}^2}{2(1 - \bar{\theta} R)} &= \frac{u^2}{\sigma^2 + Ru} - \frac{u^2 \sigma^2}{2(\sigma^2 + Ru)^2} \cdot \frac{\sigma^2 + Ru}{\sigma^2} \\ &= \frac{u^2}{2(\sigma^2 + Ru)}. \end{aligned}$$

We have established that

$$F(u) \geq \frac{u^2}{2(\sigma^2 + Ru)}.$$

Substituting this estimate into (3.36) shows that

$$\Pr\left\{ \sum_{i=1}^{l} X_i > lu \right\} \leq \exp\left[\frac{-lu^2}{2(\sigma^2 + Ru)} \right],$$

which is the same as (3.38). □

Proof of Theorem 3.6: Note that

$$\exp(\theta X) = \sum_{i=0}^{\infty} \theta^i X^i.$$

By assumption, (3.25) holds. To prove (i), suppose that $E(X) \neq 0$. Then for small enough $|\theta|$, we have that

$$E[\exp(\theta X)] = 1 + \theta E(X) + o(\theta),$$

where $o(\theta)$ denotes a higher order term with the property that $o(\theta)/|\theta| \to 0$ as $\theta \to 0$. Similarly,

$$\exp(c\theta^2) = 1 + c\theta^2 + o(\theta^2).$$

So if $E(X) > 0$, then for $\theta > 0$ sufficiently small we have that

$$1 + \theta E(X) > \exp(c\theta^2),$$

no matter what the constant c might be. Hence, for sufficiently small $\theta > 0$, we also have that

$$E[\exp(\theta X)] > \exp(c\theta^2).$$

If $E(X) < 0$, the above inequality holds for sufficiently small $\theta < 0$. In either case, the inequality contradicts (3.25). This shows that $E(X) = 0$.

To prove (ii), we use Markov's inequality. By assumption, we have that

$$E[\exp(\theta x)] \leq \exp(c\theta^2), \ \forall \theta \in \mathbb{R}.$$

Therefore, for all $u > 0$, it follows from Markov's inequality that

$$\begin{aligned} \Pr\{X > u\} &= \Pr\{\exp(\theta X) > \exp(\theta u)\} \\ &\leq \exp(-\theta u) E[\exp(\theta X)] \leq \exp(-\theta u + c\theta^2). \end{aligned}$$

Because the above inequality holds for *all* $\theta \in \mathbb{R}$, the exponent $-\theta u + c\theta^2$ can be minimized by choosing $\theta = u/(2c)$. This leads to

$$\Pr\{X > u\} \leq \exp[-u^2/(4c)].$$

The above argument can be repeated with X replaced by $-X$, which leads to

$$\Pr\{X < -u\} \leq \exp[-u^2/(4c)].$$

Combining these two bounds leads to

$$\Pr\{|X| > u\} \leq 2\exp[-u^2/(4c)]. \tag{3.42}$$

Therefore X is (γ, ζ)-sub-Gaussian with $\gamma = 2, \zeta = 1/(4c)$. □

Proof of Theorem 3.7: We begin by observing that it is enough to establish the following relationship: Suppose X is (γ, ζ)-sub-Gaussian. Then

$$E[\exp(\theta X)] \leq \exp(c\theta^2) \ \forall \theta \geq 0, \tag{3.43}$$

where c is defined in (3.26). The difference between (3.25) and (3.43) is that the latter relationship is required to hold only for $\theta \geq 0$. Suppose we succeed in establishing (3.43). Observe that if X satisfies (3.24) for a specific pair (γ, ζ), then $-X$ also satisfies (3.24) *for the same pair* (γ, ζ). Therefore, if we succeed in showing that (3.24) implies (3.43), and if $\theta < 0$, say $\theta = -\mu$, then we can replace X by $-X$ and conclude from (3.43) that

$$E[\exp(\theta X)] = E[\exp(-\mu X)] = E[\exp(\mu \cdot -X)] \leq \exp(c\mu^2) = \exp(c\theta^2),$$

where c is given in (3.26). Therefore, in this proof, it is assumed that $\theta \geq 0$.

For each integer k, we have that

$$\begin{aligned} E(|X|^k) &= \int_0^\infty u^k d\Phi_X(u^k) = \int_0^\infty k u^{k-1} d\Phi_X(u) \\ &= \int_0^\infty k u^{k-1} \Pr\{|X| > u\} du \leq \int_0^\infty k u^{k-1} \gamma e^{-\zeta u^2} du. \end{aligned} \tag{3.44}$$

Now let us make a change of variable $t = \zeta u^2$, so that $dt = 2\zeta u du$. Then

$$\begin{aligned} E(|X|^k) &\leq k\gamma \int_0^\infty e^{-t} \frac{t^{k/2-1}}{\zeta^{k/2-1}} \cdot \frac{dt}{2\zeta} = \frac{\gamma}{\zeta^{k/2}} \cdot \frac{k}{2} \int_0^\infty e^{-t} t^{k/2-1} dt \\ &= \frac{\gamma}{\zeta^{k/2}} \cdot \frac{k}{2} \Gamma(k/2) = \frac{\gamma}{\zeta^{k/2}} \Gamma(1 + k/2), \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function. Therefore, if $\theta < \zeta^{1/2}$, then

$$\begin{aligned} E[\exp(\theta X)] &\leq 1 + \sum_{k=2}^{\infty} \frac{\theta^k}{k!} E(|X|^k) \leq 1 + \gamma \sum_{k=2}^{\infty} \left(\frac{\theta}{\zeta^{1/2}} \right)^k \frac{\Gamma(1+k/2)}{k!} \\ &\leq 1 + \gamma \sum_{k=2}^{\infty} \left(\frac{\theta}{\zeta^{1/2}} \right)^k = 1 + \frac{\gamma\theta^2}{\zeta(1-\theta\zeta^{-1/2})}, \end{aligned} \tag{3.45}$$

where we use the facts that $E(X) = 0$, and that $\Gamma(1 + k/2) \leq k!$ whenever $k \geq 2$.

Now suppose $\bar{\theta} < \min\{\zeta, \zeta^{1/2}\}$. Then (3.43) is valid because the power series converges. Also

$$1 - \theta\zeta^{-1/2} \geq 1 - \bar{\theta}\zeta^{-1/2}, \ \forall \theta \in [0, \bar{\theta}].$$

Therefore

$$E[\exp(\theta X)] \leq 1 + \frac{\gamma}{\zeta(1-\bar{\theta}\zeta^{-1/2})}\theta^2 =: 1 + c_1\theta^2 \leq \exp(c_1\theta^2)i \ \forall \theta \in [0, \bar{\theta}],$$

where

$$c_1 = \frac{\gamma}{\zeta(1-\bar{\theta}\zeta^{-1/2})}. \tag{3.46}$$

Next we analyze the situation when $\theta > \bar{\theta}$. For all $\theta, a \geq 0$, we have the identity

$$\theta X - a\theta^2 = \frac{X^2}{4a} - \left(\frac{X}{2\sqrt{a}} - \sqrt{a}\theta \right)^2 \leq \frac{X^2}{4a},$$

or equivalently

$$\theta X \leq \frac{X^2}{4a} + a\theta^2.$$

Therefore, if

$$E[\exp(X^2/(4a))] < \infty, \tag{3.47}$$

then we can write

$$E[\exp(\theta X)] \leq e^{a\theta^2} E[\exp(X^2/(4a))]. \tag{3.48}$$

Now we make use the bounds derived in (3.44) to determine when (3.47) holds. For all $\theta \geq 0$, we have that

$$\exp(\theta X^2) = 1 + \sum_{k=1}^{\infty} \frac{\theta^k X^{2k}}{k!}.$$

Therefore, because $\bar{\theta} < \zeta^{-1}$, we can substitute from (3.44) and get

$$\begin{aligned} E[\exp(\bar{\theta} X^2)] &\leq 1 + \sum_{k=1}^{\infty} \frac{\bar{\theta}^k}{k!} E(|X|^{2k}) \\ &\leq 1 + \sum_{k=1}^{\infty} \frac{\gamma}{\zeta^k} \frac{\bar{\theta}^k}{k!} \Gamma(k+1) \\ &= 1 + \gamma \sum_{k=1}^{\infty} \left(\frac{\bar{\theta}}{\zeta} \right)^k = 1 + \frac{\gamma\bar{\theta}}{\zeta(1-\bar{\theta}\zeta^{-1})} \leq \exp(c_2), \end{aligned}$$

where

$$c_2 = \frac{\gamma\bar{\theta}}{\zeta(1-\bar{\theta}\zeta^{-1})}.$$

So we can substitute $a = 1/(4\bar{\theta})$ into (3.48), and

$$E[\exp(X^2/(4a)) = E[\exp(\bar{\theta}X^2)] \leq e^{c_2}.$$

This can be substituted into (3.48), giving

$$E[\exp(\theta X)] \leq e^{a\theta^2} \cdot e^{c_2} = \exp(a\theta^2 + c_2).$$

We would like to express the exponent in the form $c_3\theta^2$ for some suitable constant c_3 whenever $\theta \geq \bar{\theta}$. For this purpose, note that

$$c_2 = \frac{\gamma\bar{\theta}}{\zeta(1-\bar{\theta}\zeta^{-1})} \leq \frac{\gamma}{\zeta\bar{\theta}(1-\bar{\theta}\zeta^{-1})}\theta^2, \ \forall\theta \geq \bar{\theta}.$$

Hence, whenever $\theta \geq \bar{\theta}$, we have the bound

$$E[\exp(\theta X)] \leq \exp(c_3\theta^2),$$

where

$$c_3 = \frac{1}{4\bar{\theta}} + \frac{\gamma}{\zeta\bar{\theta}(1-\bar{\theta}\zeta^{-1})}. \tag{3.49}$$

Hence, if we define $c := \max\{c_1, c_3\}$ where c_1 is defined in (3.46), then it follows that

$$E[\exp(\theta X)] \leq \exp(c\theta^2), \ \forall\theta \geq 0,$$

which is the desired conclusion. □

The bound given in Theorem 3.7, specifically (3.26), is somewhat unwieldy. This is the reason for Theorem 3.8.

Proof of Theorem 3.8: Note that $\bar{\theta} < \max\{\zeta, \zeta^{1/2}\}$ in (3.26). However, if $\zeta < 1$, then the bound on $\bar{\theta}$ becomes $\bar{\theta} < \zeta$. Accordingly, let $\bar{\theta} = \zeta/2$ and let $\bar{c}_1, \bar{c}_3$ denote the resulting quantities in (3.26). Thus

$$\bar{c}_3 = \frac{1}{2\zeta} + \frac{4\gamma}{\zeta^2}a, \bar{c}_1 = \frac{\gamma}{\zeta(1-\zeta^{1/2}/2)}.$$

Now routine calculations show that $\bar{c}_1 > \bar{c}_3$, so that c in (3.26) can be taken as $\bar{c}_3$. □

Proof of Theorem 3.9: The proof of this theorem is omitted as it would take us too far afield. The interested reader is referred to [22, Proposition 7.5] and [22, Lemma 7.6] respectively.

3.4 Probabilistic Methods – 2

There is yet another type of probabilistic method for constructing matrices with RIP, which is typically used in reconstructing a vector from randomly selected components.

3.5 Case Studies

Example 3.2. In this brief example, we will compare the estimates for the number of samples required by various methods for constructing measurement matrices, both deterministic as well as probabilistic. Specifically, we will compare the bounds $m = p^2$ where p is a prime number that satisfies

$$p \geq \max\{3kr, n^{1/r}\},$$

for various integers $r \geq 3$, and the bounds in (3.30) and (3.31) respectively.

As in Example 3.1, let $n = 10,000$ and $k = 6$. The deterministic procedure of Theorem 3.1 requires $m = 3,481$ measurements. If instead we choose A to consist of independent samples of normal random variables, multiplied by $1/\sqrt{m}$, then (3.31) leads to the following estimates:

$$g \approx 1.2498, \eta \approx 0.1798, m \geq 5,785,$$

which is *more* than the number obtained using a deterministic method. If we instead use any c-sub-Gaussian random variable to generate the samples, and choose $c = 1/2$ (the same as for a normal random variable), then the resulting numbers as in (3.30) are

$$\gamma = 2, \zeta = 1/(4c) = 0.5, \beta \approx 2.8618, \kappa = 0.5, \tilde{c} \approx 0.0201,$$

and

$$m \geq 33,201,$$

which is *three times more* than the size of the unknown vector!

Example 3.3. In this example, we will study the reconstruction of a signal which is sparse in the frequency domain, from a small number of randomly selected time samples. This example is suggested by an article by Cleve Moler.

Suppose

$$x(t) = a_1 \sin f_1 t + a_2 \sin f_2 t,$$

where f_1 and f_2 are noncommensurate frequencies. Now suppose we measure $x(\cdot)$ at $n = 8,000$ samples at a frequency of $f_0 = 40$ Khz. Thus the time samples are 25 microseconds apart. Let us denote the resulting $n \times 1$ vector again as x. Then the discrete cosine transform (DCT) of x would consist of just a couple of frequencies. Thus, in principle, we should be able to reconstruct the unknown signal x from a small number of samples of x.

As before, let $\mathcal{C}_n$ denote the discrete cosine transform matrix defined in (3.15) and (3.16), and let $\mathcal{D}_n$ denote the inverse DCT matrix, which is just C_n^t. Choose a set $S \subseteq [n]$ such that $|S| \leq m$, and sample the unknown signal x at the k indices in S. By definition we know that

$$x_i = (\mathcal{D}_n)^i \tilde{x}, i \in S,$$

where $\tilde{x}$ is the DCT of x. Now let A equal the $|S| \times n$ submatrix of $\mathcal{D}_n$, consisting of the rows in the index set S, and $y := x_S \in \mathbb{R}^m$ denote the vector of samples of x. Then it is true that $y = A\tilde{x}$. We wish to reconstruct $\tilde{x}$ from y. For this purpose, we can set

$$\hat{\tilde{x}} := \operatorname*{argmin}_z \|z\|_1 \text{ s.t. } Az = y.$$

The returned vector $\hat{\tilde{x}}$ would be an approximation to the unknown DCT $\tilde{x}$. Taking the inverse DCT of $\hat{\tilde{x}}$, that is, computing $\mathcal{D}_n \hat{\tilde{x}}$, would give an approximation of the original but unknown vector x, which can be designated as $\hat{x}$.

If the matrix A satisfies an RIP of order tk with $\delta_{tk} < \sqrt{(t-1)/t}$, and if $\tilde{x}$ were to be k-sparse (with $k = 2$), then the above procedure would recover $\tilde{x}$ exactly. In the present case, if the sampling frequency of 40 Khz were to be an integer multiple of both frequencies f_1 and f_2, then the DCT $\tilde{x}$ would be sparse. However, if the two frequencies are not exact dividers of the sampling frequency, then $\tilde{x}$ would be *nearly* sparse but not exactly so.

In order for the above recovery procedure to work, the matrix A would have to satisfy an RIP. There are many papers in the literature regarding the RIP of submatrices of discrete *Fourier transform* matrices, but far fewer papers about the RIP of discrete *cosine transform* matrices. So, without any theoretical justification, let us choose $m = 500$, so that we take 500 samples. Moreover, let S consist of 500 *randomly selected* indices between 1 and 8,000. The figures below show the result of applying ℓ_1-norm minimmization

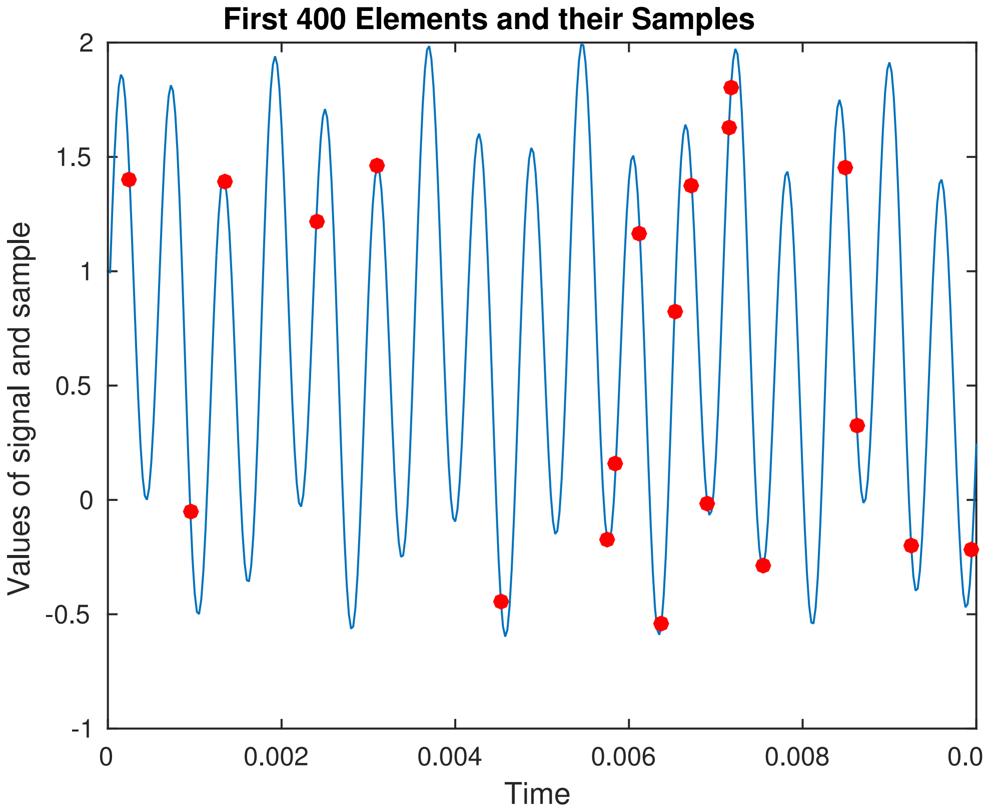

Figure 3.1: The first 400 samples and their sampled values

to reconstruct the signal, with frequencies $f_1 = 1,194$ Hz ad $f_2 = 3,387$ Hz. Figure 3.1 shows the first 400 values of x and the first 20 sampled values. Figure 3.2 shows the vector $\hat{\tilde{x}}$, which is an approximation to the true but unknown DCT $\tilde{x}$. Figure 3.3 shows the true DCT $\tilde{x}$ versus the recovered DCT $\hat{\tilde{x}}$; clearly there is virtually no difference between the two. Figures 3.4, 3.5 and 3.6 show the recovered signal $\hat{x}$ and the true signal x over various intervals. It is evident that the reconstructed signal initially starts off with some misalignment, but later samples are virtually indistinguishable from the true signal.

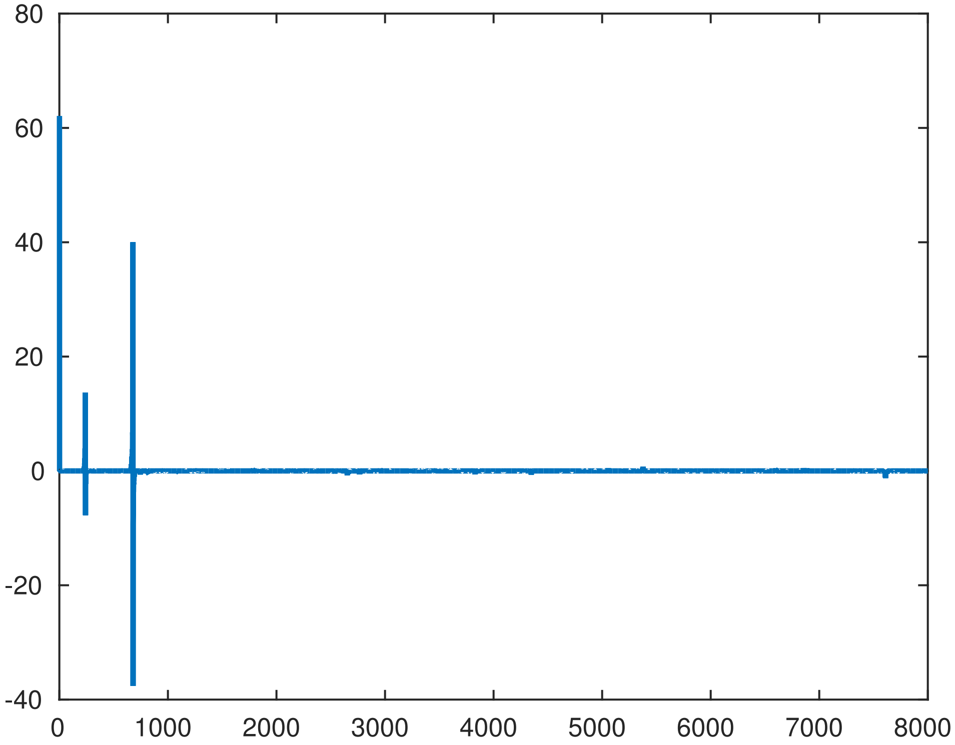

Figure 3.2: The reconstructed discrete cosine transform $\hat{\hat{x}}$.

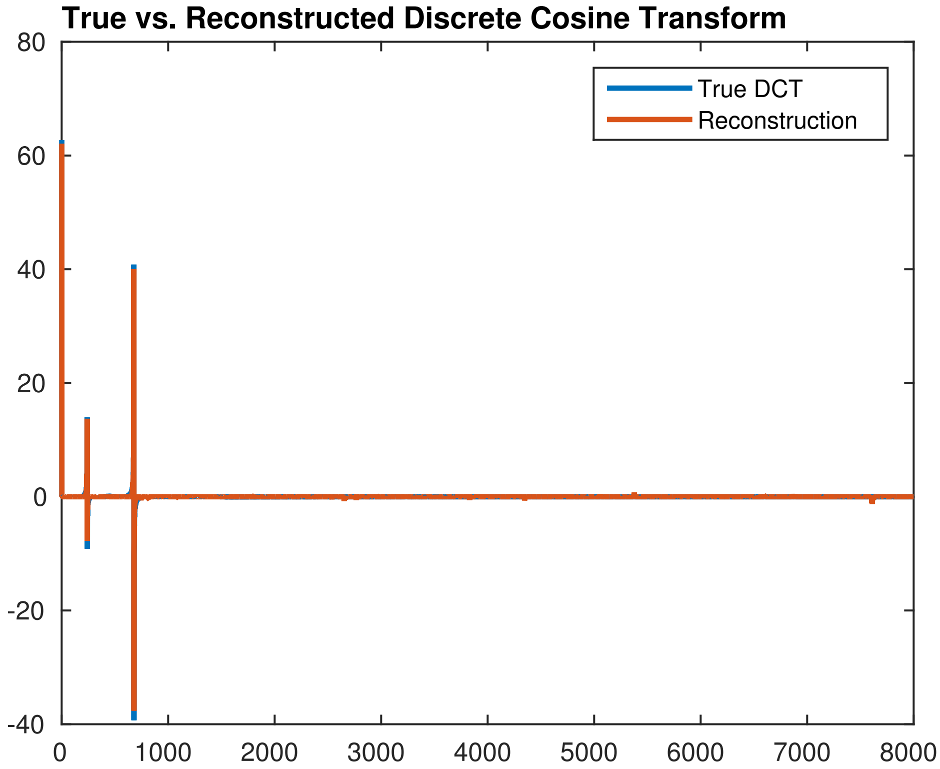

Figure 3.3: The true and reconstructed discrete cosine transforms.

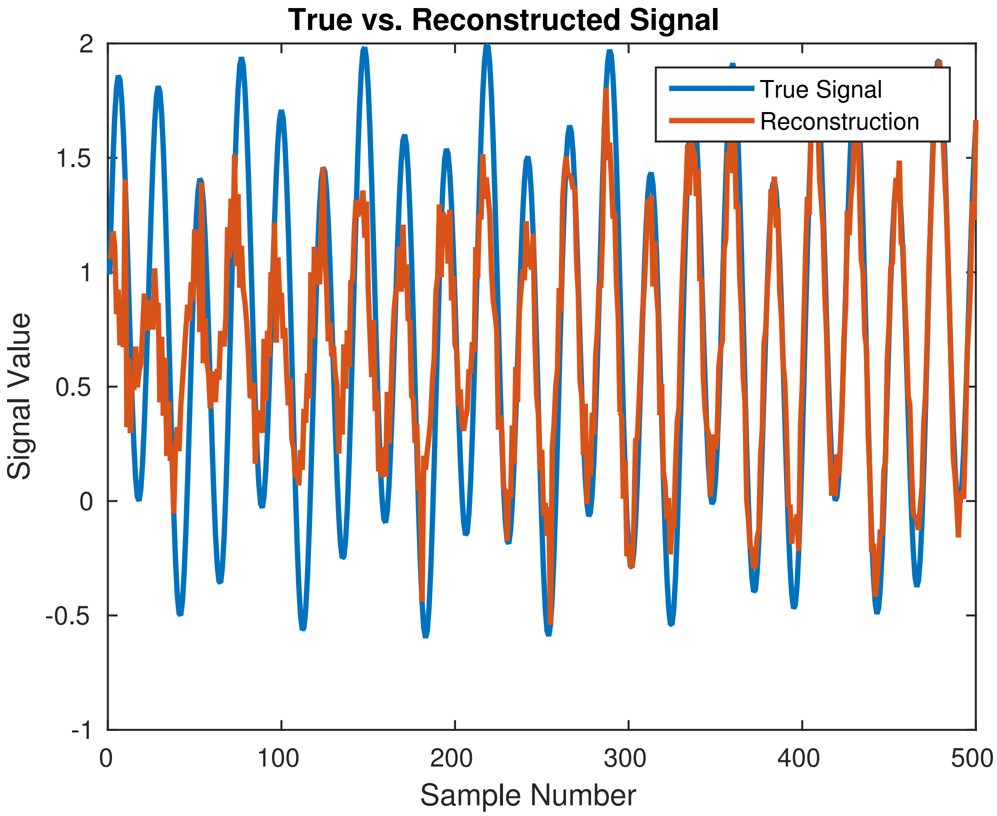

Figure 3.4: The true and reconstructed signals: Samples 1 through 500.

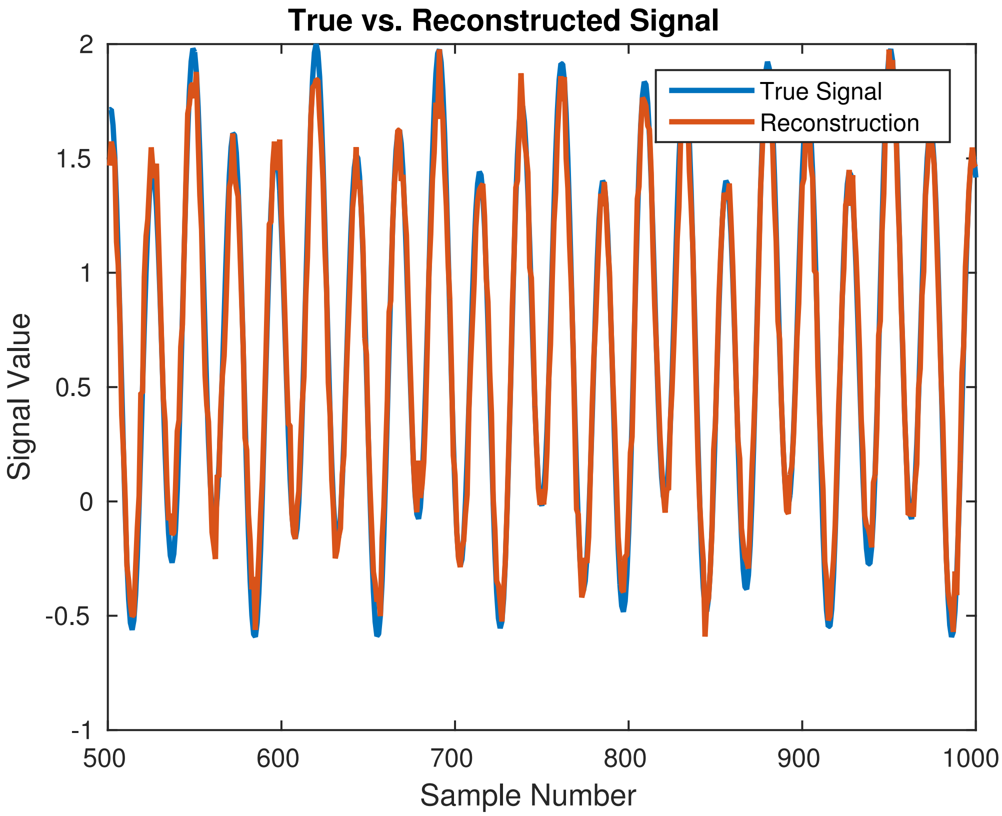

Figure 3.5: The true and reconstructed signals: Samples 501 through 1,000.

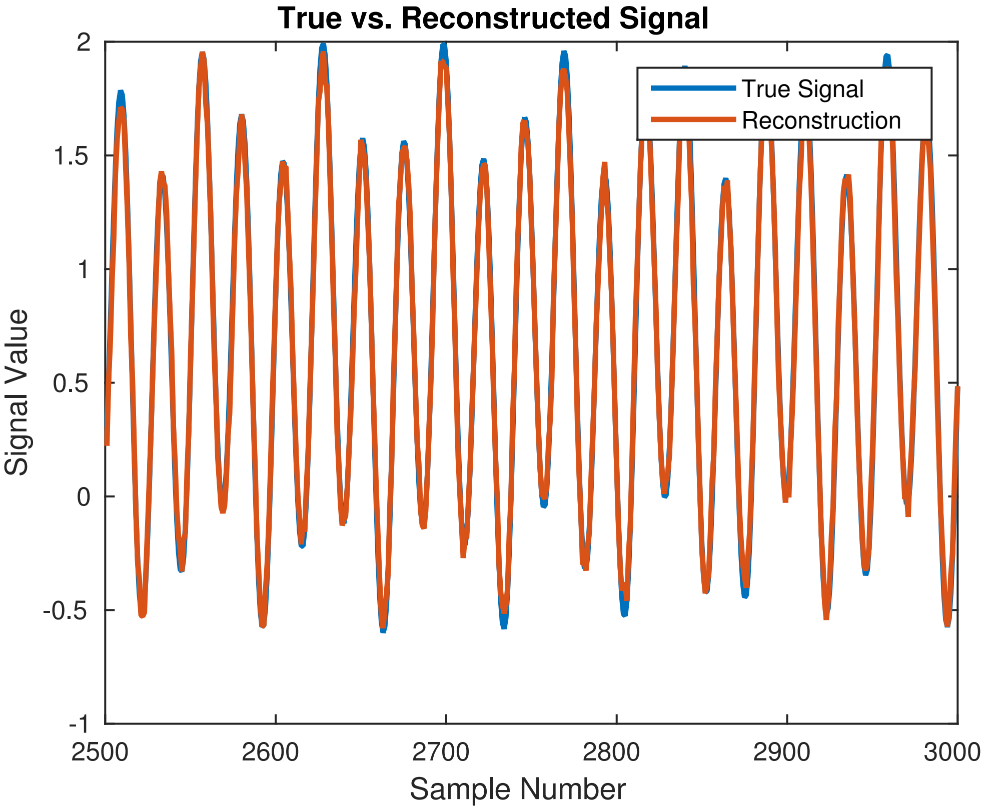

Figure 3.6: The true and reconstructed signals: Samples 2,501 through 3,000.

Chapter 4

Group Sparsity

This chapter will be based on the results of [1], and will build on earlier results from [37, 29, 32].

Chapter 5

One-Bit Compressed Sensing

Bibliography

[1] M. E. Ahsen and M. Vidyasagar. Error bounds for compressed sensing algorithms with group sparsity: A unified approach. *Applied and Computational Harmonic Analysis*, page to appear, 2016.

[2] S. A. Broughton and K. M. Bryan. *Discrete Fourier Analysis and Wavelets: Applications to Signal and Image Processing.* Wiley-Interscience, New York, 2008.

[3] T. Cai, L. Wang, and G. Xu. New bounds for restricted isometry constants. *IEEE Transactions on Information Theory*, 56(9):4388–4394, 2010.

[4] T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, 58(3):1300–1308, 2010.

[5] T. Cai, G. Xu, and J. Zhang. On recovery of sparse signal via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 55(7):3388–3397, 2009.

[6] T. Cai and A. Zhang. Shart RIP bound for sparse signal and low rank matrix recovery. *Applied and Computational Harmonic Analysis*, 35:74–93, 2013.

[7] T. Cai and A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Transactions on Information Theory*, 60(1):122–132, 2014.

[8] E. Candès. The restricted isometry property and its implications for compresed sensing. *Comptes rendus de l'Académie des Sciences, Série I*, 346:589–592, 2008.

[9] E. J. Candès and Y. Plan. Near ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[10] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications in Pure and Applied Mathematics*, 59(8):1207–1223, August 2006.

[11] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.

[12] E. J. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, December 2007.

[13] A. Cohen, Wolfgang, Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1):211–231, January 2009.

[14] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. Introduction to compressed sensing. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 1–68. Cambridge University Press, Cambridge, UK, 2012.

[15] M. E. Davies and R. Gribonval. Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$. *IEEE Transactions on Information Theory*, 55(5):2203–2214, 2009.

[16] A. Dembo and O. Zeitouni. *Large Deviation Theory and Applications.* Springer-Verlag, Berlin, 1998.

[17] R. DeVore. Deterministic construction of compressed sensing matrices. *Journal of Complexity*, 23:918–925, 2007.

[18] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[19] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, November 2007.

[20] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 48(9):2558–2567, 2002.

[21] A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 50(6):1579–1581, 2003.

[22] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Springer-Verlag, 2013.

[23] S. Gerschgorin. Über die abgrenzung der eigenwerte einer matrix. *Izvestia Akademii Nauk SSSR*, 6:749–754, 1931.

[24] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[25] R. Lidl. *Introduction to Finite Fields and Their Applications.* Cambridge University Press, 1994.

[26] R. H. McEliece. *Finite Fields for Computer Scientists and Engineers.* Kluwer Academic Publishers, Boston, 1987.

[27] G. L. Mullen and D. Panario. *Handbook of Finite Fields.* CRC Press, London, 2013.

[28] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

[29] G. Obozinski, L. Jacob, and J.-P. Vert. Group lasso with overlaps: The latest group lasso approach. *arxiv*, page 1110.0413, 2011.

[30] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. A simplified approach to recovery conditions for low-rank matrices. In *Proceedings of the International Symposium on Information Theory*, pages 2318–2322, 2011.

[31] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[32] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[33] D. Sundararajan. *The Discrete Fourier Transform: Theory, Algorithms and Applications.* World Scientific, Singapore, 2001.

[34] A. N. Tikhonov. On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39(5):195–198, 1943.

[35] R. S. Varga. *Matrix iterative Analysis (Second Revised and Expanded Edition).* Springer-Verlag, Heidelberg, 2009.

[36] G. Xu and Z. Xu. Compressed sensing matrices from Fourier matrices. *IEEE Transactions on Information Theory*, 61(1):469–478, January 2015.

[37] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.