

Chapter 2

The Geometry of Compressed Sensing

Thomas Blumensath

Abstract Most developments in compressed sensing have revolved around the exploitation of signal structures that can be expressed and understood most easily using a geometrical interpretation. This geometric point of view not only underlies many of the initial theoretical developments on which much of the theory of compressed sensing is built, but has also allowed ideas to be extended to much more general recovery problems and structures. A unifying framework is that of non-convex, low-dimensional constraint sets in which the signal to be recovered is assumed to reside. The sparse signal structure of traditional compressed sensing translates into a union of low dimensional subspaces, each subspace being spanned by a small number of the coordinate axes. The union of subspaces interpretation is readily generalised and many other recovery problems can be seen to fall into this setting. For example, instead of vector data, in many problems, data is more naturally expressed in matrix form (for example a video is often best represented in a pixel by time matrix). A powerful constraint on matrices are constraints on the matrix rank. For example, in low-rank matrix recovery, the goal is to reconstruct a low-rank matrix given only a subset of its entries. Importantly, low-rank matrices also lie in a union of subspaces structure, although now, there are infinitely many subspaces (though each of these is finite dimensional). Many other examples of union of subspaces signal models appear in applications, including sparse wavelet-tree structures (which form a subset of the general sparse model) and finite rate of innovations models, where we can have infinitely many infinite dimensional subspaces. In this chapter, I will provide an introduction to these and related geometrical concepts and will show how they can be used to (a) develop algorithms to recover signals with given structures and (b) allow theoretical results that characterise the performance of these algorithmic approaches.

T. Blumensath (✉)

ISVR Signal Processing and Control Group, University of Southampton,
Southampton, UK
e-mail: thomas.blumensath@soton.ac.uk

2.1 Introduction

How do we know something is there, if we haven't seen it, or, to use the cliché, how do we know that the falling tree still makes a sound even if there is no one to listen? This is far more than a purely philosophical question. It is at the heart of all of scientific discovery, indeed, one could say that all of science is ultimately a quest for rules that allow us to predict the unobserved. In science, this is done by observing certain aspects of nature which are then used to build models which in turn allow us to make predictions about things we have not yet seen.

Similar questions also arise in engineering. We live in a digital world where images, sounds and all kinds of other information are stored, transmitted and processed as finite collections of numbers. Whether it is your favourite TV show or the medical images acquired at your last hospital appointment, all are represented using zeros and ones on a computer. But how is this possible? Sound pressure varies continuously at your ear, so how can this continuously varying signal be described by a finite number of bits? In fact, the digital information stored on your favourite CD only describes the sound pressure measured at regular intervals. Similarly, a movie typically consists of (only) tens of images each second, yet the light intensity, originally measured by the camera, changes continuously with time. Digital movies and sound recordings are thus mere approximations of the original physical signal.

The question thus arises, “How much of the information is preserved in these approximations?” and “How do we infer what the signal was in the places we haven't seen?” that is, “How then do we interpret these approximations?” For example, our movie is only represented with a relatively small number of different images each second. Any changes that occur at a timescale that is faster than this, are not captured. In effect, to interpret the movie, we assume that such changes do not occur. Whilst this is not true in the movie example, our eyes are not able to resolve changes faster than those captured in a normal film. However, we are also all aware that this can lead to “errors”. We have all experienced the illusion of a propeller on a plane or the wheel on a car that, when changing its speed appears to change direction. This “aliasing” is due to the fact that we don't interpret the data correctly, that is, our *model* (i.e. the assumption that changes are slow) is incorrect. As we don't know what happens to the propeller or wheel between frames, our brain makes the assumption that the propeller or wheel has moved the smaller of the two possible distances between the two observations.

The moral of this story is that we constantly have to make judgements about things that “happen where we have not looked” and that we do this using assumptions or *models*. Similar judgements have to be made by any algorithm that deals with measured continuous signals and a detailed theoretical understanding of these phenomena is thus fundamental to our ability to capture, process and reconstruct continuous physical phenomena.

The Shannon Nyquist Whitaker sampling theorem [1, 2] is the classical example of such a theoretical treatment of the problem. Consider a signal $x(t)$ that changes continuously with time t . This could be, for example, the sound pressure measured

with a microphone. To represent $x(t)$ digitally, we sample it by taking equally spaced measurements $x(t_i)$ at time points t_i , where $t_i - t_{i-1} = \Delta_t$ is the constant sampling interval. Moving to bold face vector¹ notation, our representation of the original continuous signal $x(t)$ is now the vector \mathbf{x} (which has either a finite number of entries if $x(t)$ was sampled over a finite interval, or could in theory be infinitely long). \mathbf{x} is not yet a truly digital representation of $x(t)$, as each entry in the vector \mathbf{x} is a real number, which also cannot be represented exactly in digital form. Nevertheless, for the purpose of this chapter, we will ignore this additional complication and assume that the effect of the additional errors introduced by the required quantisation of real numbers are negligibly small. Instead, the leitmotif here will be the interpretation of \mathbf{x} . Which properties must $x(t)$ possess so that it is fully described by the measurements in the vector \mathbf{x} ?

We will see that there is an intimate interplay between (1) the way we measure a signal (e.g. the sampling interval in our Shannon sampling example), (2) the class of signals that we can describe exactly using the measurements \mathbf{x} and (3) the way in which we can reconstruct $x(t)$ from the measurements \mathbf{x} . For continuous signals sampled at equally spaced intervals, the relationship between these three points is precisely what is captured by the Shannon Nyquist Whitaker sampling theorem, which states that:

Theorem 1 *If a continuous signal is sampled at equally spaced intervals and if the signal is band-limited with a bandwidth of less than half the sampling rate, then the signal can be reconstructed exactly using a linear reconstruction. Furthermore, the reconstruction filter only depends on the sampling rate and the frequency band occupied by the signal.*

The signal model here assumes $x(t)$ to be band-limited and, in order to be able to interpret or reconstruct $x(t)$, the frequency support must be known. If we sample at regular intervals and use a reconstruction that assumes the incorrect frequency support, or if the signal is not band-limited, then we will not be interpreting or reconstructing the signal correctly and aliasing will occur similar to the propeller or wheel example.

In this chapter, the sampling problem will be addressed in a more general setting. In particular, more general signal models will be considered. A more general theory will bring many advantages. For example, a sampling theory that allows a more general class of signal models allows us to design particular sampling schemes that are tailored to a specific problem. This, in turn, can lead to sampling approaches capable of, for example, sampling non-bandlimited signals or, sampling at a rate well below that required by the Nyquist rate. However, this can only be achieved

¹ In this chapter, we use two, somewhat different, meanings for the term vector. On the one hand, we call any one dimensional array of real or complex numbers a vector, this is the meaning used here. Below, we will introduce a more abstract definition of vectors as elements of some mathematical space. Which of these two definitions is appropriate at any one point in this chapter should be clear from the context.

if the sampling theory provides us with the tools to model and account for known signal structures.

There are several mathematical approaches to capture and model signal structure. Our view here will be predominantly geometrical. Similar to a sphere of radius 6,371 km which is a good model to use to describe my location on the earth's surface (up to small errors that would account for the fact that the earth is not completely spherical or that I might on occasion take a plain or visit an underground cave), similar geometrical models can be used to describe constraints on signals. In general, most signals such as sounds, images and movies can be thought of as living in some signal space, where we can define the distance between two signals or can measure angles. But similar to the assumption that I am not likely to be found anywhere in space but am restricted to the earth's surface (after all, I am unlikely to spend any of my upcoming holidays on the moon), so are many types of signals only encountered in or close to a subset of the space they inhabit. For example, the assumption in Shannon's sampling theorem that signals are band-limited, translates into the geometric assumption that signals lie on a subspace (think of a subspace as the equivalent to an infinite piece of paper in our three dimensional world).

Many traditional sampling results are based on convex sets, such as subspaces. Whilst convex signal models lead to relatively simple sampling approaches, which are easily studied with current mathematical tools, non-convex models are significantly more flexible. However, the utility gained through the increased flexibility also leads to an escalation in the complexity of both the theoretical treatment of the sampling problem as well as their successful implementation. Non-convex signal models typically require non-linear reconstruction techniques, so that, for these models, an additional important aspect arises: the computational speed or complexity of signal reconstruction. In particular, many advanced signal models lead to NP-hard reconstruction problems. It thus becomes paramount to restrict sampling strategies for these signal models to a subset of linear operators that allow fast reconstruction.

The archetypal example here is compressed sensing [3–6]. Compressed sensing assumes signals to be sparse in some way. For finite dimensional signals that can be expressed as vectors, sparsity means that most of the entries in the vector that represent the signal are zero. It is important to note here that the sparse vector itself does not have to be the signal of interest. Instead, the sparse vector can equally well be a representation of a signal in some basis (wavelet and Fourier bases are popular examples). For finite dimensional signals, Fourier domain sparsity assumes a signal to be constructed from the mixture of a few sinusoids, where the frequencies of each of the sinusoids has to be taken from a fixed, finite-dimensional regularly spaced grid.

A related area that has gained more prominence recently is matrix completion [7, 8]. In the matrix completion problem, the signal of interest is a data-matrix, but, instead of measuring the data for each entry in the matrix, only a small subset of the matrix entries is filled with measurements initially. The task is then to estimate the missing entries using the measured entries only. This can again only be done if we assume the data to follow some known model. A popular model for matrix completion, which is related to the sparse model used in compressed sensing and

which is found to describe many phenomena of interest, is a low rank matrix model. In these models, the full data matrix is assumed to have a rank which is significantly smaller than the maximum rank a matrix of the same dimensions could have. A popular example is the movie recommender system, where a matrix is constructed in which each entry contains a rating of a movie by a person. For each person in the system there is thus a row in the matrix and each film has an associated column. However, people are only able to watch and rate a small fraction of all movies, so that the missing entries have to be inferred from the few ratings made. Once the missing entries have been filled in, the system can then recommend movies to people on the system that they are likely to rate highly. A common assumption in these systems is that the full data matrix is of low rank, an assumption that is justified by an argument that stipulates that a persons preference in movies is primarily driven by a small number of underlying factors.

Compressed sensing and matrix factorisation can be seen as two particular instances of a more general class of constrained inverse problems [9]. In this chapter, the main ideas that define the class of problems we discuss will be that they (1) use non-convex constraints to model the signals we will be able to reconstruct and (2) pose computationally challenging reconstruction problems so that we will require efficient reconstruction methods. As promised in the chapter title, we here take a geometrical point of view, which will allow us to study important properties of non-convex signal models and their interplay with different efficient reconstruction methods.

2.2 Geometrical Signal Models

2.2.1 A Geometrical Primer

Before continuing our study of the geometry of data recovery problems, it makes sense to define and fix several mathematical concepts and notation.

Throughout this chapter, we will talk about *signals* which will be mathematical descriptions of physical phenomena such as sounds, images or movies. From a mathematical point of view, *signals* are functions and a function is a mapping that assigns a unique real or complex number to each set of functional *parameters*. The parameters of a function are taken from the reals or from a subset of the reals. For example, sound pressure can be described as a function that assigns a unique pressure to each point in time. Similarly, an image can be understood as a function that assigns a real number (describing the image intensity) for each location in the image plain. In contrast to the sound example, where the sound parameter ran over all possible time instances, for images it is typical to restrict the domain of the image parameters to intervals of real numbers. Another important class of functions are finite length vectors. For example, a ten dimensional vector can be understood as a collection of ten real or complex numbers. Such a vector is also a function, but here, the parameters are restricted to an interval of integers (i.e. 1, 2, ..., 10).

2.2.1.1 Vector Space

The material in this section can be found in any good textbook on analysis and functional analysis. Good, however rather technical, examples are [10] and [11].

A mathematical space is a collection of mathematical objects, such as numbers or functions, together with a set of properties. Properties can include, for example, additivity of elements (so that for any two elements, there is an element of the space that is the sum of the two elements). Other properties of mathematical objects that are important for a geometrical interpretation are length or size, distance between objects and angle between objects.

In this chapters, signals will be formally described as mathematical objects that live in a *vector space*. This means that they all have a certain set of universal properties common to all vector spaces. Formally, a linear vector space \mathcal{V} over a Field \mathcal{F} (which in this chapter will either be the real numbers (\mathbb{R}) or the complex numbers (\mathbb{C})) is a selection of objects (called vectors) together with certain operations on these elements, which have the following set of properties:

1. The space has an addition operator $+$, so that for any two elements $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{V}$ the product $\mathbf{x}_1 + \mathbf{x}_2$ is also an element of the set \mathcal{V} .
2. The addition is commutative (i.e. $\mathbf{x}_1 + \mathbf{x}_2 = \mathbf{x}_2 + \mathbf{x}_1$) and associative (i.e. $(\mathbf{x}_1 + \mathbf{x}_2) + \mathbf{x}_3 = \mathbf{x}_1 + (\mathbf{x}_2 + \mathbf{x}_3)$).
3. There is a *zero* element $\mathbf{x}_0 \in \mathcal{V}$ for which $\mathbf{x} + \mathbf{x}_0 = \mathbf{x}$ holds for all $\mathbf{x} \in \mathcal{V}$. We will write the zero element as $\mathbf{0}$.
4. For all $\mathbf{x} \in \mathcal{V}$, there is an element $-\mathbf{x}$, such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.
5. The space has a scalar multiplication operator \cdot , so that for all elements $\alpha, \beta \in \mathcal{F}$ and any $\mathbf{x} \in \mathcal{V}$ the element $\alpha \cdot \mathbf{x}$ is an element in \mathcal{V} and, furthermore, $\alpha \cdot (\beta \cdot \mathbf{x}) = (\alpha\beta) \cdot \mathbf{x}$, $(\alpha + \beta) \cdot \mathbf{x} = \alpha \cdot \mathbf{x} + \beta \cdot \mathbf{x}$, $\alpha \cdot (\mathbf{x}_1 + \mathbf{x}_2) = \alpha \cdot \mathbf{x}_1 + \alpha \cdot \mathbf{x}_2$ and $1 \cdot \mathbf{x} = \mathbf{x}$.

Thus, vector spaces are collections of elements that can be added and subtracted and which can be multiplied by real or complex numbers (or more general by elements from its base field).

Banach Space

By equating real world signals with elements of a vector space, we can use vector addition and scalar multiplication to describe signal addition and scaling. The next useful concept we introduce is that of the *size* or length of a signal. Once we are able to talk about the size of a signal, then we can also talk about the *size of the difference* between two signals, which then enables us to formally define the distance or difference between two signals. The ability to talk about length and distance of signals is our first step in a geometrical interpretation of signal processing problems and is thus one of the most fundamental concepts discussed here.

The length of an element of a vector space \mathcal{V} will be measured by a *norm*. We write $\|\mathbf{x}\|$ to denote the norm of the element \mathbf{x} . A norm is a non-negative function

that assigns a real number to an element of a vector space and has the following properties:

1. The zero element $\mathbf{0}$ is the only element in the vector space that has a norm of zero, that is $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
2. The norm satisfies the triangle inequality $\|\mathbf{x}_1 + \mathbf{x}_2\| \leq \|\mathbf{x}_1\| + \|\mathbf{x}_2\|$ for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{V}$.
3. The norm increases proportionally when scaling an element in the vector space, that is, $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for all $\mathbf{x} \in \mathcal{V}$ and $\alpha \in \mathcal{F}$.

The second of these properties is one of the fundamental properties that will allow us to use some of our geometrical intuition when discussing signal properties as it links the length of the sum of two vectors to the length of each vector individually. The geometrical picture is that of a triangle, where the length of any one side of the triangle (which is the same as the sum of the two other sides) is always shorter or at most as long as the sum of the lengths of each of the other two sides. Or, using another well known geometrical property, the length between two points is the straight line.

Thus, the concept of a norm not only tells us how 'large' an element in a vector space is, it also tells us how far apart different elements in the space are. From the properties of vector spaces, we know that the element $\mathbf{x} = \mathbf{x}_1 - \mathbf{x}_2$, that is the difference between the elements \mathbf{x}_1 and \mathbf{x}_2 is itself a vector. Therefore, if we have defined a norm on the vector space, then the norm $\|\mathbf{x}_1 - \mathbf{x}_2\|$ will be defined and will measures the *distance* between these two elements.

With the definition of distance comes another property, that of convergence of sequences of elements. Assume we have a collection infinitely many elements $\{\mathbf{x}_i\}$ (which do not have to be all different). This sequence is said to be Cauchy convergent if the distance $\|\mathbf{x}_m - \mathbf{x}_n\|$ can be made arbitrary small for all $n, m > N$ if we only choose N itself large enough. In other words, if we restrict our consideration to elements that are far enough from the beginning of the sequence, then any two elements will be arbitrarily close to each other.

Cauchy convergence might seem a little bit odd at first and another form of convergence might be more intuitive to the reader new to these idea. A sequence $\{\mathbf{x}_i\}$ is said to converge to a point \mathbf{x}_{lim} , if the distance $\|\mathbf{x}_i - \mathbf{x}_{lim}\|$ converges to zero in the limit. In this form of convergence, the sequence of elements will get arbitrarily close to a certain point, which we will here call \mathbf{x}_{lim} . The difference to Cauchy convergence is that, in the definition of Cauchy convergence, which is the more general of the two properties, whilst elements in the sequence are guaranteed to stay close to each other, there might not exist a single element within our vector space, which is a limit point, that is, to which the sequence will get arbitrarily close. However, rather than this being a property of the sequence itself, this is really a property of the space from which the elements of the sequence have been picked. In a sense, spaces in which there are sequences that are Cauchy convergent but which do not have a limit point are "incomplete". Thus a space where all Cauchy sequences converge are actually called complete spaces. As convergence is such a fundamental property, it is often useful to restrict discussions to complete spaces. Complete normed vector spaces thus have their own name, they are called *Banach Spaces*. All the spaces

encountered in this chapter will be Banach spaces so that the concepts of Cauchy convergence and convergence will be identical.

Hilbert Space

A second geometrical concept which is as important as length, is that of the angle between two elements. Angles between vectors can be measured using inner products, which is a real or complex valued function of two elements of the vector space (written as $\langle \cdot, \cdot \rangle$) which satisfies the following two properties.

1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ with $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
2. $\langle \mathbf{x}_1 + \mathbf{x}_2, \mathbf{x}_3 \rangle = \langle \mathbf{x}_1, \mathbf{x}_3 \rangle + \langle \mathbf{x}_2, \mathbf{x}_3 \rangle$.
3. $\langle \lambda \mathbf{x}_1, \mathbf{x}_2 \rangle = \lambda \langle \mathbf{x}_1, \mathbf{x}_2 \rangle$, where $\lambda \in \mathbb{C}$.
4. $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = \overline{\langle \mathbf{x}_2, \mathbf{x}_1 \rangle}$, where the bar $\bar{\cdot}$ indicates the complex conjugate.

Inner products can be used to ‘induce’ a norm, that is, they can be used to define a norm as follows

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

Using the induced norm, inner products contain information on the angle between two elements. In fact, inner products combine information on angles and vector length, so that a quantity that has properties similar to the angle between two elements can be found by normalising the inner product

$$\frac{\langle \mathbf{x}_1, \mathbf{x}_2 \rangle}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|}.$$

Thus, if \mathbf{x}_1 and \mathbf{x}_2 are the same vector, then their angle will be zero and the above normalised inner product is 1. Similarly, we will say that two vectors are at right angles or *orthogonal* if their inner product is 0.

With an induced norm there is an intimate link between norms and inner products. For example, the Pythagorean theorem holds

$$\|\mathbf{x}_1 + \mathbf{x}_2\|^2 = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 \text{ if } \langle \mathbf{x}_1, \mathbf{x}_2 \rangle = 0,$$

which is a special case of the more general result that

$$\|\mathbf{x}_1 + \mathbf{x}_2\|^2 = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 + 2\langle \mathbf{x}_1, \mathbf{x}_2 \rangle.$$

In addition, the following parallelogram law also holds

$$\|\mathbf{x}_1 + \mathbf{x}_2\|^2 + \|\mathbf{x}_1 - \mathbf{x}_2\|^2 = 2\|\mathbf{x}_1\|^2 + 2\|\mathbf{x}_2\|^2$$

and so does the following inequality

$$|\langle \mathbf{x}_1, \mathbf{x}_2 \rangle| \leq \|\mathbf{x}_1\| \|\mathbf{x}_2\|.$$

A vector space that has a norm that is induced by an inner product thus has very appealing geometrical properties. Such a space is called a Hilbert space if it is furthermore complete, that is, a Hilbert space is a complete inner product space with an induced norm.

Finite and Infinite Dimensional Spaces

We live in a three dimensional world, or mathematically speaking, in a three dimensional (thus finite dimensional) Hilbert space, yet many spaces of mathematical functions are actually infinite dimensional. In infinite dimensional spaces, some of our intuition still holds, yet, care has to be taken as there are also subtle differences. In essence, an infinite dimensional space is a space in which there are infinitely many vectors that are all orthogonal to each other. Orthogonality can be measured by the inner product, in fact, the inner product of orthogonal vectors is zero. In an infinite dimensional space, there are thus infinitely many vectors which all have a zero inner product with each other.

But infinity is even more subtle than this. In fact, there are infinities of different sizes. This might come as a surprise to some, yet the typical example are the sets of integers and the sets of real numbers. There are infinitely many integers, for any integer number I name, you will always be able to find a number that is larger. Real numbers on the other hand, not only contain all integers. There are infinitely many other real numbers that lie between any two distinct real numbers. It can indeed be shown that there will be ‘more’ real numbers than there are integers. When talking about infinities it is thus helpful to distinguish the infinity that is as large as the number of integers and infinities that are larger. Sets of infinitely many elements, that have as many elements as there are integers are said to be countable. The elements in a countable set can thus be labeled using integers (that is we could count them at least in theory). Sets that cannot be labeled with integers are called uncountably infinite. We will restrict the discussion here to Hilbert spaces that are at most countably infinite.

Basis

In a similar way in which we describe locations on earth (for example, using north-south, east-west, and height), it is useful to be able to find a way to describe the ‘location’ of vectors in a vector space. This will be done using a set of basis vectors (or basis directions). An important concept here is that any such description should ideally not contain replicated information; three parameters are enough to describe any location on earth and four parameters would only replicate some of this information. A similar concept holds in general vector spaces, even in infinitely large ones.

To capture the effect of replication of information, we use the concept of linear dependency of a set of vectors. A set of vectors $\{\mathbf{x}_i\}$ is said to be linearly dependant

if there are scalars λ_i (which are not all zero) such that $\sum_i \lambda_i \mathbf{x}_i = \mathbf{0}$ or $\sum_{i \neq j} \lambda_i \mathbf{x}_i = -\lambda_j \mathbf{x}_j$. Thus, if we use the vectors \mathbf{x}_i to describe a vector \mathbf{x} as $\mathbf{x} = \sum_i \alpha_i \mathbf{x}_i$, then we can always replace one of the vectors (say \mathbf{x}_j with $\mathbf{x}_j = -\sum_{i \neq j} \lambda_i / \lambda_j \mathbf{x}_i$ so that the vector \mathbf{x} is equally well described using one less vector. On the other hand, if there is no such set of scalars λ_i such that $\sum_i \lambda_i \mathbf{x}_i = \mathbf{0}$, then we say that the set of vectors $\{\mathbf{x}_i\}$ is linearly independent.

Any set of vectors $\{\mathbf{x}_i\}$, whether linearly dependant or not, can be used to describe certain vectors \mathbf{x} as a linear combination $\mathbf{x} = \sum_i \alpha_i \mathbf{x}_i$. All those \mathbf{x} which can be written in this form for any given set $\{\mathbf{x}_i\}$ is called the linear span of the set $\{\mathbf{x}_i\}$, which is formally written as the set

$$\{\mathbf{x} = \sum_i \lambda_i \mathbf{x}_i, \text{ with } \lambda_i \in \mathcal{F}\}$$

where \mathcal{F} is the field used in the definition of the vector space (e.g. \mathcal{F} are the real or complex numbers).

A set $\{\mathbf{x}_i\}$ which is large enough to be able to describe *all* vectors in vector space and which furthermore is not too large so that its elements are linear independent is called a basis for the space.

We have already encountered the concept of orthogonality. A basis, in which any two elements are orthogonal is called an orthogonal basis. Furthermore, an orthogonal basis in which each element has unit length, is called an orthonormal basis. An important result in mathematics is the fact that every Hilbert space has an orthonormal basis. Furthermore, if the set of vectors in the basis is either finite or countably infinite, we say that the Hilbert space is separable.

We will here restrict our discussion to separable Hilbert spaces so that we can always find an at most countably infinite orthonormal basis set $\{\mathbf{x}_i\}$ that allows us to write any element of the Hilbert space as a linear combination

$$\mathbf{x} = \sum_i^{\infty} a_i \mathbf{x}_i. \quad (2.1)$$

2.2.1.2 Subspaces

A subset \mathcal{S} of a vector space is called a linear subspace if any two elements $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ have the property that their linear combination $\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2$ is also an element of the set \mathcal{S} . Here λ_1 and λ_2 are arbitrary scalars. The linear span of a set of vectors is a linear subspace.

2.2.1.3 Convex Sets

Closely related to linear subspaces are convex sets. A convex set is defined similarly to a linear subspace. A subset \mathcal{S} of a vector space is called a convex subset if any two elements $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ have the property that their linear combination $\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2$ is also an element of the set \mathcal{S} . However, the difference here is in the set of scalars allowed in the definition. Whilst in the definition of a linear subspace, λ_1 and λ_2 were allowed to be arbitrary scalars, for a set to be convex, we have the additional requirement that $\lambda_1, \lambda_2 \geq 0$ and that $\lambda_1 + \lambda_2 = 1$. It should thus be clear that a linear subspace is a convex set. A set that is not convex if there are $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ and $\lambda_1, \lambda_2 > 0$ with $\lambda_1 + \lambda_2 = 1$ for which the element $\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2$ is *not* an element of the set \mathcal{S} itself.

In a Hilbert space, in the same way in which we say that two vectors are orthogonal, we can also say that a vector \mathbf{x} is orthogonal to a subset \mathcal{S} if \mathbf{x} is orthogonal to every element in \mathcal{S} . Similarly, if we have two subsets, these can be said to be orthogonal if every vector of one subset is orthogonal to every vector of the other subset. For example, the orthogonal complement of a subset is the set of all vectors that are orthogonal to the set. The orthogonal complement of any set is a closed² convex subspace.

For any closed convex subset \mathcal{S} of a Hilbert space \mathcal{H} , it is always possible to find a *best* approximation of any vector $\mathbf{x} \in \mathcal{H}$ by an element of the closed convex subset \mathcal{S} . That is, for any $\mathbf{x} \in \mathcal{H}$ there exists a $\mathbf{x}_0 \in \mathcal{S}$ such that

$$\|\mathbf{x} - \mathbf{x}_0\| = \inf_{\tilde{\mathbf{x}} \in \mathcal{S}} \|\mathbf{x} - \tilde{\mathbf{x}}\|.$$

We will call the element \mathbf{x}_0 the projection of \mathbf{x} onto the closed convex subset \mathcal{S} .

This leads us to the important orthogonal projection theorem which states that for any closed linear subspace $\mathcal{S} \subset \mathcal{H}$ and any $\mathbf{x} \in \mathcal{H}$, we can always find a unique decomposition $\mathbf{x} = \mathbf{x}_{\mathcal{S}} + \mathbf{x}_{\mathcal{S}^\perp}$, where $\mathbf{x}_{\mathcal{S}} \in \mathcal{S}$ and where $\mathbf{x}_{\mathcal{S}^\perp}$ is orthogonal to \mathcal{S} . Furthermore, $\mathbf{x}_{\mathcal{S}} \in \mathcal{S}$ is the closest point in \mathcal{S} to \mathbf{x} .

For any closed linear subspace \mathcal{S} , let $P_{\mathcal{S}}$ be the operator that maps and $\mathbf{x} \in \mathcal{H}$ to the element $\mathbf{x}_{\mathcal{S}}$ defined in the projection theorem. The operator P is self adjoint (that is $\langle P\mathbf{x}_1, \mathbf{x}_2 \rangle = \langle \mathbf{x}_1, P\mathbf{x}_2 \rangle$ for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{H}$), $P^2 = P$ and has an operator norm $\sup_{\mathbf{x} \neq 0} \|\Phi \mathbf{x}\|/\|\mathbf{x}\| = \|P\| = 1$ whenever $P \neq 0$.

2.2.1.4 Unions of Simpler Geometrical Models

Having defined some of the basic geometric properties of Hilbert spaces, let us now return to the problem of signal modelling. Linear subspaces and closed convex sets have very appealing properties and these sets have long been used to define classes of signals which then allow us to find elements within these convex sets

² A subset $\mathcal{S} \subset \mathcal{H}$ is called closed if every sequence with elements in \mathcal{S} that converges to an element of \mathcal{H} has a limit in the subset \mathcal{S} itself.

that can act as good representatives for a particular signal. Many classical signal processing ideas have been restricted to closed convex sets, yet, recent advances in our understanding of signal geometry have allowed us to extend similar ideas to more complex signal models, models that are no longer convex. This work has primarily looked at constraint sets that are the union over several (in many cases extremely large collections of) closed convex sets. In such a signal model, we are given a number of closed convex sets and assume that any signal lies within one of these sets, however, we are not sure in which set exactly we are to look for the signal.

Let us define these unions formally. Any union of closed and convex sets is defined as

$$\mathcal{S} = \bigcup_j \mathcal{S}_j : \text{ where the } \mathcal{S}_j \text{ are closed and convex subsets,} \quad (2.2)$$

Here each \mathcal{S}_j can be any closed and convex subset of a larger Hilbert space and the union can be potentially over a countably infinite number of these sets. Of particular interest to us will be union models in which the \mathcal{S}_j are closed subspaces.

An important example of a union of subspaces model is the sparse signal model in finite dimensions. Consider the Euclidean Hilbert space of dimension N whose elements we can represent using N element vectors. In a k -sparse model, the model subset \mathcal{S} is the set of all vectors \mathbf{x} that has no more than k non-zero entries. This model is in fact a union of subspace model. To see this, consider the support of a k -sparse vector, that is, consider the pattern of the location of the non-zero elements in this vector. If we add (or subtract) a k -sparse vector that has exactly the same support (that is, whose non-zero elements are in exactly the same location), then the sum (or difference) of these two vectors will again be k -sparse and will have the same support. Thus, the set of all k -sparse vectors which have the same support is a subspace. However, for any $k < N$, there will be many different support sets. In fact there will be $\binom{N}{k}$ such sets $\binom{N}{k}$, read N choose k , is the number of different ways in which we can choose k elements from a set of N elements). Thus, the set of all k -sparse vectors (irrespective of their support) is the union of $\binom{N}{k}$ subspaces. We also see that this set is non-convex as the sum of two k -sparse vectors with different support can potentially have up to $2k$ non-zero entries. In fact, the set of the sum of two (or three) k sparse vectors will be of importance later on, and we introduce some notation to specify these sets here.

In general, if $\mathbf{x} \in \mathcal{S}$ for some union \mathcal{S} , we will write

$$\mathcal{S} + \mathcal{S} = \{\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 : \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}\} \quad (2.3)$$

and

$$\mathcal{S} + \mathcal{S} + \mathcal{S} = \{\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 : \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathcal{S}\} \quad (2.4)$$

2.2.1.5 Operators on the Elements of a Space

One last fundamental notion that will be required throughout this chapter is that of an operator. In principle, an operator takes an element of one space and *transforms* it into the element of another space. We write $\mathbf{y} = \Phi(\mathbf{x})$, where \mathbf{x} is an element of one space and \mathbf{y} is the element of another space.

A linear operator has properties similar to a matrix. In particular it is linear, that is, for any two elements \mathbf{x}_1 and \mathbf{x}_2 from one space, it does not matter if we apply the operator to the sum of the two elements or if we apply the operator to each individual element and then sum the transformed elements. That is, $\Phi(\mathbf{x}_1 + \mathbf{x}_2) = \Phi(\mathbf{x}_1) + \Phi(\mathbf{x}_2)$. For linear operators, we generally write $\Phi\mathbf{x}$ instead of $\Phi(\mathbf{x})$. The parenthesis will be used primarily to indicate non-linear operators.

For linear operators, we can define a norm on an operator as follows

$$\|\Phi\| = \sup_{\mathbf{x}: \|\mathbf{x}\| \leq 1} \|\Phi\mathbf{x}\|, \quad (2.5)$$

that is, informally speaking, the operator norm is the maximum amount by which *any* vector can be lengthened when squeezed through the operator. Note that the operator norm as defined here depends on two vector norms, the norm of $\|\mathbf{x}\|$ and the norm of $\|\Phi\mathbf{x}\|$. In general, both of these norms can be arbitrary. In the case in which both \mathbf{x} and $\Phi\mathbf{x}$ live in Hilbert spaces, then we assume that $\|\Phi\|$ is the norm defined using the Hilbert space norm.

An Operator is said to be invertible on a space \mathcal{H} (or alternatively on a subset $S \subset \mathcal{H}$), for all $\mathbf{x} \in \mathcal{H}$ (or for all $\mathbf{x} \in S$), if there exists an operator Φ^\dagger such that $\mathbf{x} = \Phi^\dagger(\Phi(\mathbf{x}))$. If Φ^\dagger is linear, then we say that Φ is linearly invertible on \mathcal{H} (or on S). An operator that is not invertible is said to be non-invertible.

If a linear operator between finite dimensional spaces is invertible, then the norm of $\|\Phi^\dagger\|$ is necessarily finite, however, in infinite dimensional spaces, it can happen that there are invertible linear operators whose norm is infinite. These operators are said to be ill-conditioned. Ill-conditioned operators would in theory allow us to recover \mathbf{x} from $\mathbf{y} = \Phi\mathbf{x}$ uniquely, however, any small perturbation of \mathbf{y} could potentially lead to an arbitrarily large change in the estimate of \mathbf{x} .

2.2.2 Examples and Sketch of Applications

So far, we have introduced an over-abundance of abstract mathematical ideas. Let us therefore take a step back here and discuss several important examples where geometrical ideas can help in the reconstruction of signals.

2.2.2.1 The Geometry of Shannon Sampling

The seminal work by Nyquist [1] and Shannon [2] is at the heart of much of traditional sampling theory. This theory deals with one instance of the signal recovery problem addressed throughout this book, although, we hardly think about it in this way any longer. The setting here is as follows, let \mathbf{x} be a function over time with a domain spanning over the real numbers. For example, this might be the sound pressure produced by your favourite band. The aim is now to measure this sound pressure. Let us do this measurement by measuring the sound pressure intensity at infinitely many equally spaced intervals in time, so that our measurement \mathbf{y} is an infinite sequence of real numbers and we again ask, how and when can we recover \mathbf{x} from \mathbf{y} . The Shannon sampling theorem answers exactly this question. In effect, if \mathbf{x} is band-limited, then there is a simple linear reconstruction method that can recover \mathbf{x} from \mathbf{y} exactly. The band-width of the signal has to be less than half the inverse of the time interval between consecutive samples for this to work. Without going into too much detail (see for example [12] for a more detailed treatment), when we say that \mathbf{x} is band-limited we mean that the Fourier transform of \mathbf{x} (call this transform \mathbf{X}) is a function whose support is restricted to a restricted frequency interval. This is our model. In fact, this is a subspace model. To see this, assume that you have two signals with the same frequency band-width. Adding these two signals (remember that the Fourier transform is linear) is the same as adding the Fourier transforms of the signals, so that the sum of any two band-limited signals is again band-limited. This is exactly our definition of a subspace. Thus, Shannon sampling uses a convex signal model and, as the model is convex, a simple reconstruction technique exists.

2.2.2.2 Sparse Signal Models in Euclidean Spaces

Instead of dealing with infinitely long sequences of numbers produced by ‘proper’ Shannon sampling, finite length approximations are the only practical approach to real problems. It is thus normal to assume that we can represent infinite dimensional signals using finite length vectors. In the same way, digitised images can be thought of as a collection of a finite number of real numbers. Let us therefore assume that our signal is well approximated using a vector in Euclidean space of dimension N .

If we were able to sample a signal using Shannon sampling ideas, then we would directly measure the elements in \mathbf{x} . However, in many situations, we are unable to make enough measurements to use Shannon theory. For example, many measurement procedures are slow (e.g. in Magnetic Resonance Imaging, a patient has to lie in the scanner for several minutes to produce a single volumetric image), pose health risks (e.g. in X-ray computed tomography, X-ray dosage has to be limited to reduce exposure to ionising radiation) or are extremely expensive (certain hyperspectral imaging devices can come at a cost of thousands of dollars for a single pixel, so that traditional million pixel cameras with these elements would be prohibitively expensive).

Thus, we would want to reduce the number of measurements further and sample at a rate significantly below that described by Shannon theory. To do this is only possible if we use a much smaller signal set as our model. Single, very low-dimensional subspaces are not versatile enough to capture the diverse information present in most signals and images (if it were, we could just use Shannon theory), instead, more complex, low-dimensional, but non-convex models have to be used. One of the most powerful sets of models are sparse models. A sparse (Euclidean) vector \mathbf{x} is a vector whose elements are zero apart from a small number of elements, which can have arbitrary magnitude. We say \mathbf{x} is k -sparse if all but k of its elements are zero. Vectors with a fixed subset of non-zero elements lie in a single subspace, but in a sparse model, we allow all possible subsets of k elements to be non-zero, so that k -sparse vectors lie in the union of $\binom{N}{k}$ different subspaces.

Instead of sparsity in the canonical basis (e.g. in an image, instead of assuming that the image has many zero pixels), a great deal of flexibility is achieved if we allow sparsity in a different basis. In our three dimensional world, the canonical basis might be a description of locations in terms of north-south, east-west and up-down, yet we are free to use a coordinate transform to represent locations in another way. In my office, it might make more sense to specify locations in terms of their distance from the window, the side walls and the floor. As my office is not exactly aligned with the north-south axis (though luckily the floor is still level), the axis in the world coordinate system are rotations of the axis in my office representation. Exactly the same principle holds in the representation of signals. For example, we are not restricted to represent an image by specifying values for each pixel (the canonical space) but could instead specify two dimensional discrete wavelet coefficients to specify spatial frequencies of the image or we might represent the image using a 2-dimensional wavelet transform. These transforms often are nothing else than a rotation of the coordinate axis. In this case, they do not change the length of vectors, just their representation. In other cases, the new coordinate system might actually have axis that are not orthogonal in the original space or even have more coordinates than the original space. In these cases, we still assume that we can find a representation of any vector in our original space in the transformed space, but the length of elements in the two spaces might now differ. The importance of these transforms from our signal recovery perspective is that many signals have sparse or approximately sparse representations in some transformed domain. For example, images are often found to be sparse in a wavelet representation. Thus, using sparsity in transform domains greatly enhances our ability to use sparse models to describe structure in real signals.

When talking about sparsity in a different domain, we assume that there is a linear mapping that maps elements \mathbf{x} of our signal space into the transformed domain. Call this mapping Ψ , so that $\mathbf{z} = \Psi \mathbf{x}$ is the representation of \mathbf{x} in the transformed domain. Importantly, we assume that there is a generalised inverse Ψ^\dagger of Ψ , such that for all $\mathbf{x} \in \mathcal{H}$, $\mathbf{x} = \Psi^\dagger \mathbf{z} = \Psi^\dagger \Psi \mathbf{x}$.

2.2.2.3 Structured Sparse Models in Euclidean Space

Sparsity can be a powerful constraint and in many applications additional structure can be brought into play, further increasing the utility of sparse models. Structured sparse models are sparse models that only allow certain subsets of sparse support sets to be present. For example, a block-sparse vector is a sparse vector in which the non-zero coefficients are contained in pre-specified blocks. For example, if $\mathbf{x} \in \mathbb{C}^N$ and assume we have J blocks that partition \mathbf{x} , that is, if $B_j \subset \{1, \dots, N\}$, $j \in \{1, \dots, J\}$ is the set of indices in block j , then we assume that the blocks (1) do not overlap (that is $B_i \cap B_j = \emptyset$) and (2) every index of \mathbf{x} is in at least one block (that is $\bigcup_{j \in J} B_j = \{1, 2, 3, \dots, N\}$). A signal that is k -block sparse is then defined as any \mathbf{x} whose support is contained in no more than $k < J$ different sets B_j , that is

$$\text{supp}(\mathbf{x}) \subset \bigcup_{\mathcal{J}} B_j : \mathcal{J} \subset \{1, 2, \dots, J\}, |\mathcal{J}| \leq k. \quad (2.6)$$

To define block-sparse signals here we imposed the restriction that the blocks do not overlap and that their union includes all elements of \mathbf{x} . In theory, we could drop these two restrictions, however, theoretical treatment of these more general models becomes much more difficult, and, in fact, this class of models would be so general that it would include all possible structured sparse models.

Another set of useful structured sparse models are tree-sparse models. Instead of partitioning the signal's support set into disjoint blocks, tree-sparse signals have non-zero coefficients that follow a tree structure in which all ancestors of a node are allowed to be non-zero whenever the node itself is non-zero. A sparse tree model is a model in which the tree is furthermore sparse, that is, the total number of non-zero elements is small. The simplest example, a one sparse tree, would only have a non-zero element at its root, whilst a two-sparse tree model would have as many possible support sets as there are children of the root, as such a model would have to include the root itself plus one of its children.

2.2.2.4 Low Rank Matrices

In many applications, data is best represented in matrix form. By specifying an appropriate inner product and norm for matrices, the Hilbert space formalism can also be applied to matrix problems so that geometrical ideas can be used to define subsets of matrices that can act as signal models. A powerful constraint here is the low-rank matrix model. The set of all M by N matrices of rank r that have the same column and row space (the space spanned by the matrix's row or column vectors) form a linear subspace, that is, we can add any two of these matrices and end up with another matrix of the same size and rank that has again the same column and row space (or, more precisely, whose row and column spaces are subspaces of the row and column spaces of the original matrices). However, a matrix with different

column or row spaces does not lie in the same subspace and adding two matrices with different column or row spaces will result in a matrix that is likely to have a different rank from that of its two components. Thus, low-rank matrices lie in a non-convex subset of the space of all matrices.

2.2.2.5 Sparsity in Continuous Signals

Our last set of examples are again taken from infinite dimensional spaces, where continuous analogues to sparsity have been developed. In Shannon sampling, the sampling rate is directly related to the bandwidth of the signal we would like to sample. In several applications, this would lead to a prohibitive sampling rate so that again, additional signal structure has to be exploited. A signal model that is in some ways similar to the sparse model in Euclidean spaces is the analogue compressed sensing model first studied in [13] for known support and in [14] for unknown support. Here a continuous and band-limited³ real valued time series $x(t)$ is assumed to have a Fourier transform $\mathcal{X}(f)$ whose support S is the union of K intervals of ‘small’ bandwidth B_K , i.e. $S \subset \bigcup_{k=1}^K [d_k, d_k + B_K]$, where the d_k are arbitrary scalars from the interval $[0, B_N - B_K]$. These signals can be understood as a continuous version of a sparse signal, but instead of having few non-zero “elements,” only a small part of the functions support (say in the Fourier domain) is non-zero. As the support of the Fourier transform of a real valued function is symmetric, we here only consider the support in the positive interval $[0, B_N]$. If $K B_K < B_N$ then $\mathcal{X}(f)$ is zero for some frequencies f in $[0, B_N]$, mirroring sparsity in a vector. If we would fix the support S , then $\mathcal{X}(f)$ and therefore $x(t)$ would lie in a subspace of the space of all square integrable functions with bandwidth B_N . However, in a model where S is not fixed and where $K B_K < B_N$, there will be infinitely many distinct sets S satisfying this definition, so that $x(t)$ will lie in the union of infinitely many infinite dimensional subspaces. The set of all signals that have energy restricted to K bands with $K B_K < B_N$ thus is a non-convex set.

Another set of powerful models are so called Finite Rate of Innovations models. Consider again a real valued function of one variable $x(t)$. Such a function is said to have a finite rate of innovation [15] if it can be written as

$$x(t) = \sum_{n \in \mathbb{Z}} \sum_{r=0}^R c_{nr} g_r \left(\frac{t - t_n}{T} \right), \quad (2.7)$$

where $T, t_n \in \mathbb{R}$ and where the $g_r(\cdot)$ are either functions (or generalised functions/distributions such as the Dirac delta function). For such signals one can define a rate of innovation as follows

³ That is, a signal whose Fourier transform $\mathcal{X}(f)$ is assumed to be zero apart from the set $S \subset [-B_N, B_N]$.

$$\rho = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} C_x \left(-\frac{\tau}{2}, \frac{\tau}{2} \right), \quad (2.8)$$

where the function $C_x(t_a, t_b)$ is a counting function that counts the number of ‘degrees of freedom’ in the interval $[t_a, t_b]$, that is, $C_x(t_a, t_b)$ counts that number of parameters c_{nr} for which the functions g are centred within the interval $[t_a, t_b]$. For a function $x(t)$ to have a finite rate of innovation, it is obviously necessary that $\rho < \infty$. Extensions of these ideas to complex valued functions of several variables are also possible and make it possible to apply similar ideas to problems in image processing.

2.3 Linear Sampling Operators, Their Properties and How They Interact with Signal Constraint Sets

Having discussed several concepts and ideas that allow us to think about signal models using geometrical ideas, we now turn to the analysis of the sampling or measurement process itself. We introduced a set of powerful constraint sets above to allow us to deal with many problems in which we are unable to sample all relevant information, either due to corruption of signals or due to constraints on resources or fundamental physical properties of our measurement system. We will now try and develop an understanding of how the measurement system itself acts on these signal models.

Assume that our sampling system is linear, so that for any signal \mathbf{x} we produce measurements $\mathbf{y} = \Phi \mathbf{x}$, where Φ is a linear sampling operator. There are two particular aspects of the sampling system Φ we should be concerned about. If we assume the signal follows a given model \mathcal{S} , then we want our measurement system to measure enough information to allow us to distinguish different signals from our model. It is thus natural to require that for any two $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ with $\mathbf{x}_1 \neq \mathbf{x}_2$ we have $\Phi \mathbf{x}_1 \neq \Phi \mathbf{x}_2$, so that any two distinct signals give distinct observations. In this case we should (at least in theory) be able to find the unique $\mathbf{x} \in \mathcal{S}$ that gave rise to an observation $\mathbf{y} = \Phi \mathbf{x}$.

The second fundamental requirement would be a certain robustness to noise. As nearly all measurements are noisy to some extent, if the measurement we have of a signal is slightly different from the measurement we would expect if there were no noise, then we would require that the signal that in a noiseless setting would give the actual measurement we observe is not too far from the true signal. More concretely, assume that we want to measure a signal \mathbf{x} , but observe the following noisy measurement $\mathbf{y} = \Phi \mathbf{x} + \mathbf{e}$ for some small noise term \mathbf{e} . Assume that there is a signal $\hat{\mathbf{x}}$ that also lies in our model and that satisfies $\mathbf{y} = \Phi \hat{\mathbf{x}} = \Phi \mathbf{x} + \mathbf{e}$. As it seems reasonable to assume that the true signal was $\hat{\mathbf{x}}$ given we observe \mathbf{y} and don’t know what \mathbf{e} is, it would not be very useful that, for small \mathbf{e} , the difference between \mathbf{x} and $\hat{\mathbf{x}}$ is large as we would then make large errors in our signal reconstruction, even under small noise perturbations.

2.3.1 A Geometrical Approach to Signal Recovery

Let us recall the signal recovery problem we would like to solve. A signal is measured and we would like to either ask specific questions about the signal or we would like to reconstruct the signal from the measurements. With our mathematical framework, both, the signal and the measurement will be represented as vectors which live in some vector space. In general, we say the signal \mathbf{x} lives in a vector space \mathcal{H} and the measurement \mathbf{y} lives in a space \mathcal{L} . For most of this chapter, \mathcal{H} and \mathcal{L} will be Hilbert spaces, that is, it will make sense to talk about distance and angle between signals (or measurements). Each measurement is a transformation of a signal \mathbf{x} into an observation \mathbf{y} . This transformation is done (mathematically speaking) by an operator $\Phi(\mathbf{x})$, which can either be linear or non-linear. For most of our discussion, we will restrict ourselves to linear operators, as these are easier to understand. However, we will also discuss how some of the ideas that hold for linear measurements can be applied to the setting where the measurements are slightly non-linear.

Nature does not follow the idealistic precision of a mathematical operator and any real measuring device will add at least some systematic or random noise to the measurements. We will thus use the following fundamental measurement equation that describes how any signal \mathbf{x} is transformed into a particular measurement

$$\mathbf{y} = \Phi(\mathbf{x}) + \mathbf{e}, \quad (2.9)$$

where \mathbf{e} is an unknown vector of measurement noise. This brings us to the fundamental problem of this book, given a measurement \mathbf{y} and knowing enough about the measurement process to be able to describe Φ , how can we recover the original signal \mathbf{x} and with what precision can we do this?

2.3.1.1 Lets Start Simple

In the simplest instance, if Φ is linear and invertible on the entire space, then we could simply estimate \mathbf{x} as

$$\hat{\mathbf{x}} = \Phi^\dagger \mathbf{y} = \Phi^\dagger \Phi \mathbf{x} + \Phi^\dagger \mathbf{e}. \quad (2.10)$$

How good this estimate is depends on how much the inverse Φ^\dagger amplifies the error \mathbf{e} . To see this, consider the difference between $\hat{\mathbf{x}}$ and \mathbf{x} .

$$\|\hat{\mathbf{x}} - \mathbf{x}\| = \|\Phi^\dagger \Phi \mathbf{x} + \Phi^\dagger \mathbf{e} - \mathbf{x}\| = \|\mathbf{x} + \Phi^\dagger \mathbf{e} - \mathbf{x}\| = \|\Phi^\dagger \mathbf{e}\|. \quad (2.11)$$

Relative to the size of \mathbf{e} , this error is thus

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{e}\|} = \frac{\|\Phi^\dagger \mathbf{e}\|}{\|\mathbf{e}\|}, \quad (2.12)$$

which, by the definition of the operator norm, cannot be larger than the operator norm of Φ^\dagger . This is related to the condition number of Φ^\dagger , which is defined as the ratio of the relative change in the size of \mathbf{e} (i.e. $\|\Phi^\dagger \mathbf{e}\|/\|\mathbf{e}\|$) to the relative change in size of $\Phi \mathbf{x}$ (i.e. $\|\Phi^\dagger \Phi \mathbf{x}\|/\|\Phi \mathbf{x}\| = \|\mathbf{x}\|/\|\Phi \mathbf{x}\|$), which is easily seen to be the same as the ratio of the operator norms $\|\Phi\|/\|\Phi^\dagger\|$. Thus, if Φ is invertible and (in an infinite dimensional setting) Φ is well-conditioned, all we need to do to recover any signal from its measurement is to calculate the inverse of the operator and apply it. To guarantee that the reconstruction error is small, we need to make sure that the operator norm of the inverse is small. The inverse itself is linked to Φ itself, so that in designing a measurement system, if we can insure that it is linearly invertible, then all we need to do is ensure that the inverse operator has small norm or that the operator has a condition number close to 1.

Before moving on to more challenging signal recovery problems, it is worth thinking about the above recovery in terms of the geometry of the signal space. Any signal that lies within a certain distance (say d) from a point \mathbf{c} is said to lie in a ball with centre \mathbf{c} and radius d . Thus, the set of all error signals that have a length of less than ϵ say, lie in an ϵ ball (with centre at zero). In the above example, where Φ was linearly invertible on the entire signal space \mathcal{H} , the norm of Φ^\dagger (i.e. $\|\Phi^\dagger\|$) together with the size of the error \mathbf{e} will then specify the radius around the point \mathbf{x} in which the estimate $\hat{\mathbf{x}}$ will lie. In the geometrical view of this chapter, we will not specify an explicit probabilistic model for the error \mathbf{e} . Instead, we assume that \mathbf{e} is of restricted size $\|\mathbf{e}\| \leq \epsilon$. There is obviously a link between a probabilistic formulation and our geometrical point of view. For example, for an independent and identically distributed Gaussian noise term \mathbf{e} , with high probability, we know that the error will very likely be smaller than several (say 3) standard deviations. Similar probabilistic arguments, where we can assume an error bound *with high probability* can be made for other noise distributions as well.

2.3.1.2 More Complex, Yet Manageable

Now the case in which Φ is linear and invertible is trivial when compared to the much more challenging task of the stable recovery of \mathbf{x} when Φ is non-invertible or ill-conditioned. If \mathbf{x} is an element in some Hilbert space, but if there are at least two $\mathbf{x}_1 \neq \mathbf{x}_2$ such that $\mathbf{y} = \Phi \mathbf{x} = \Phi \mathbf{x}_1 = \Phi \mathbf{x}_2$, then there is no way in which we can choose among the two offending \mathbf{x}_1 and \mathbf{x}_2 , given only the measurement \mathbf{y} . Typically, if Φ is linear and non-invertible, then, for each \mathbf{y} , there will be entire subspaces of elements \mathbf{x} that would give exactly the same measurement \mathbf{y} . Non-invertible linear operators have the property that there are elements $\mathbf{x}_0 \neq \mathbf{0}$ such that $\Phi \mathbf{x}_0 = \mathbf{0}$. For such an element, if we take any other \mathbf{x}_1 such that $\mathbf{y} = \Phi \mathbf{x}_1$ and add \mathbf{x}_0 to \mathbf{x}_1 we get the same observation $\mathbf{y} = \Phi \mathbf{x}_1 = \Phi(\mathbf{x}_1 + \mathbf{x}_0)$ and we are in the above situation where we can't distinguish between \mathbf{x}_1 and $\mathbf{x}_2 = \mathbf{x}_1 + \mathbf{x}_0$. Furthermore, for linear operators Φ , if $\Phi \mathbf{x}_0 = \mathbf{0}$, then $\Phi \lambda \mathbf{x}_0 = \mathbf{0}$ for all scalars λ . Thus, the set of all \mathbf{x}_0 for which $\Phi \mathbf{x}_0 = \mathbf{0}$ is a subspace. This subspace is called the *null-space* of the linear operator Φ and will be denoted as $\mathcal{N}(\Phi)$.

In the case in which Φ is non-invertible, we can therefore only recover elements from \mathcal{H} if we can restrict the search to a subset \mathcal{S} of \mathcal{H} . For this restriction to work, we require that the measurement operator Φ is invertible at least on the subset \mathcal{S} . To repeat; what we mean by this is that for any two $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$, with $\mathbf{x}_1 \neq \mathbf{x}_2$, we require that $\mathbf{y}_1 = \Phi\mathbf{x}_1 \neq \Phi\mathbf{x}_2 = \mathbf{y}_2$. Thus, if we have a signal model that restricts the class of signals we try to recover to a subset \mathcal{S} of \mathcal{H} and if Φ is invertible on the subset, then we are again able to recover $\mathbf{x} \in \mathcal{S}$, even in situations in which Φ is not invertible on all of \mathcal{H} .

The simplest constraint sets \mathcal{S} are convex sets of which subspaces are particularly nice to deal with. For a subspace \mathcal{S} it is easy to see that, if Φ is linear and invertible on \mathcal{S} , then the set $\Phi\mathcal{S} = \{\mathbf{y} = \Phi\mathbf{x} : \mathbf{x} \in \mathcal{S}\}$ is also a subspace. That is, for any two $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ the sum $\mathbf{x}_1 + \mathbf{x}_2$ is also in \mathcal{S} and so $\Phi(\mathbf{x}_1 + \mathbf{x}_2) = \Phi\mathbf{x}_1 + \Phi\mathbf{x}_2$ will be in $\Phi\mathcal{S}$. To recover \mathbf{x} from a noisy measurement $\mathbf{y} = \Phi\mathbf{x} + \epsilon$ we thus can project \mathbf{y} onto the subspace \mathcal{S} (call this projected element $\mathbf{y}_{\Phi\mathcal{S}}$ say) and then find an estimate $\hat{\mathbf{x}}$ such that $\mathbf{y}_{\Phi\mathcal{S}} = \Phi\hat{\mathbf{x}}$. In practice, as $\Phi\mathcal{S}$ is only defined implicitly, this might be a bit more involved than just described, however, conceptually, the steps of projection onto a subspace followed by the inversion on the subspace is appealing.

2.3.1.3 But Here Is the Problem

The same conceptual inversion can be carried out if \mathcal{S} is any convex subset of \mathcal{H} on which Φ is invertible. Even if Φ is no longer linear, similar ideas could be used. However, even if \mathcal{S} is convex, if Φ is non-linear, then the set $\Phi(\mathcal{S})$ might no longer be convex. Thus, finding the equivalent of a projection onto the non-convex set $\Phi(\mathcal{S})$ is now far from trivial, even if we were able to invert $\Phi(\cdot)$ on the subset \mathcal{S} . A similar situation arises when Φ is linear but the constraint set \mathcal{S} is non-convex to start with. In this case $\Phi\mathcal{S}$ is also non-convex in general and finding the closest element on $\Phi\mathcal{S}$ to an observation \mathbf{y} is non-trivial. Furthermore, the search through the set \mathcal{S} for the element that corresponds to an element in $\Phi\mathcal{S}$ is also tricky. These problems will be at the heart of this chapter.

Let us repeat the thought experiment in which we measure a signal \mathbf{x} using a measurement operator Φ and where the observation is noisy. We have $\mathbf{y} = \Phi\mathbf{x} + \epsilon$ and we want to recover \mathbf{x} from \mathbf{y} . Furthermore, the measurements are not conclusive in general that is, we are not able to distinguish all elements from the space \mathcal{H} from their measurements. Thus, we use prior knowledge and devise a model that describes a subset of elements of \mathcal{H} we expect to find. In the spirit of this chapter, this model comes in the form of a geometrical constraint set \mathcal{S} in which we assume \mathbf{x} to lie. Now, if our measurements have been designed appropriately for our model, then Φ will be invertible on \mathcal{S} , and our theoretical approach to reconstruct \mathbf{x} from $\Phi\mathbf{x} + \epsilon$ in general would be

1. Find a point in $\Phi\mathcal{S}$ that is closest to the observation \mathbf{y} . Call this point $\mathbf{y}_{\mathcal{S}}$.
2. As Φ is invertible on \mathcal{S} , find the point $\hat{\mathbf{x}} \in \mathcal{S}$ for which $\mathbf{y}_{\mathcal{S}} = \Phi\hat{\mathbf{x}}$.

For general sets \mathcal{S} in general Hilbert spaces, there is no guarantee that there actually is a unique point in $\Phi\mathcal{S}$ that is closer to \mathbf{y} than all other points. In this case, we would have to select arbitrarily from among the ‘closest’ points. For general sets \mathcal{S} , another problem that arises is that there might not even be a point that is closer to a given $\mathbf{y} \in \mathcal{H}$ than all the other points in \mathcal{S} . For example, let $\Phi\mathcal{S}$ be the set $\{\mathbf{y} = 1/n, \text{ where } n \text{ is a positive integer}\}$. If we were to observe any non-positive number (including zero), then there actually is no element in $\Phi\mathcal{S}$ that is the closest element. That is, for which ever element we choose from \mathcal{S} (say we choose element $1/N$) there is always an infinite number of other elements which are closer to all non-positive numbers. In this case, we would have to be contempt in step (1) of our recovery scheme with the selection of a point in $\Phi\mathcal{S}$ that is nearly as close to \mathbf{y} as possible.

The closest we can get to any one point \mathbf{y} with any element in \mathcal{S} is given by the infimum

$$\inf_{\mathbf{x} \in \mathcal{S}} \|\mathbf{y} - \Phi\mathbf{x}\|. \quad (2.13)$$

We have to take the infimum here instead of the minimum, as there might actually not be an element \mathbf{x} that reaches this minimal distance. From the definition of the infimum and baring in mind that $\inf_{\mathbf{x} \in \mathcal{S}} \|\mathbf{y} - \Phi\mathbf{x}\|^2 < \infty$, we can derive the following lemma

Lemma 1 *Let \mathcal{S} be a nonempty closed subset of a Hilbert space \mathcal{H} . Let Φ be an operator from \mathcal{H} into a Hilbert space \mathcal{L} , then for all $\delta > 0$ and $\mathbf{y} \in \mathcal{L}$, there exist an element $\tilde{\mathbf{x}} \in \mathcal{S}$ for which*

$$\|\mathbf{y} - \Phi\tilde{\mathbf{x}}\| \leq \inf_{\mathbf{x} \in \mathcal{S}} \|\mathbf{y} - \Phi\mathbf{x}\| + \delta. \quad (2.14)$$

All this lemma is saying is that we can actually find an element in $\Phi\mathcal{S}$ that is up to an arbitrarily small distance as close to \mathbf{y} as any other element in $\Phi\mathcal{S}$. Thus, we can talk about a relaxed form of projection, where, instead of finding the closest point in a set, we are contempt with a nearly closest point.

Thus consider the following mapping that for each \mathbf{y} and a fixed and arbitrarily small δ returns a set of elements

$$m_{\mathcal{S}}^{\delta}(\mathbf{y}) = \{\tilde{\mathbf{y}} : \tilde{\mathbf{y}} \in \mathcal{S} \text{ and } \|\mathbf{y} - \tilde{\mathbf{y}}\| \leq \inf_{\mathbf{x} \in \mathcal{S}} \|\mathbf{y} - \Phi\mathbf{x}\| + \delta\}. \quad (2.15)$$

By the above lemma, the sets $m_{\mathcal{S}}^{\delta}(\mathbf{y})$ are non-empty for all $\delta > 0$. An operator that for each \mathbf{y} returns a single element from the set $m_{\mathcal{S}}^{\delta}(\mathbf{y})$ will be said to be an δ -projection.

Thus, for each \mathbf{y} , we can find the δ -best $\mathbf{y}_{\mathcal{S}} \in \Phi\mathcal{S}$ and then search through \mathcal{S} to find the unique $\hat{\mathbf{x}}$ such that $\mathbf{y}_{\mathcal{S}} = \Phi\hat{\mathbf{x}}$.

2.3.1.4 It Only Works If...

How far will $\hat{\mathbf{x}}$ be from \mathbf{x} ? To answer this question we need to introduce a further property of the operator Φ , namely a property that describes how much Φ ‘stretches’ or ‘shrinks’ elements. For example, if we have a vector \mathbf{x} of length $\|\mathbf{x}\|$, once we have mapped this vector into the space \mathcal{L} , how does the length change? If Φ is linear, then we say that Φ is bounded if $\|\Phi\mathbf{x}\| \leq c\|\mathbf{x}\|$ holds for all $\mathbf{x} \in \mathcal{H}$ and for some fixed c , so that bounded linear operators can never ‘stretch’ vectors by more than the operator norm (which is finite for bounded operators). But how much can \mathbf{x} be ‘shrunk’? Remember that we are interested in problems in which Φ is ill-conditioned and non-invertible. For these problems, we have necessarily the tight lower bound $0 \leq \|\Phi\mathbf{x}\|$, that is, vectors in the null-space of Φ are mapped to zero vectors whilst for ill-conditioned Φ , vectors are potentially shrunk to arbitrarily small length. But this is exactly why we introduced the constraint set \mathcal{S} . Thus, instead of asking what happens to the length of all vectors in \mathcal{H} , we instead would like to know what happens to those vectors that live in our constraint set. Furthermore, as will become clear later, we are actually mainly interested in the difference between vectors, thus, we ask, what happens to the length of the difference of any two vectors \mathbf{x}_1 and \mathbf{x}_2 that lie in the subset \mathcal{S} . What is the maximum these differences are stretched and by how much might they be shrunk? More formally, we want to find the largest real number α and the smallest real number β such that

$$\alpha\|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|\Phi(\mathbf{x}_1 - \mathbf{x}_2)\| \leq \beta\|\mathbf{x}_1 - \mathbf{x}_2\| \quad (2.16)$$

holds for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$. We call the above inequality the bi-Lipschitz condition, with α and β being the bi-Lipschitz constants.

For once, if Φ is linear and if $\alpha > 0$, then Φ will actually be invertible on \mathcal{S} , that is, assume that $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ are different vectors, i.e. $\|\mathbf{x}_1 - \mathbf{x}_2\| > 0$, so that the lower bound in the bi-Lipschitz condition is non-zero. By the bi-Lipschitz condition this then implies that $\|\Phi(\mathbf{x}_1 - \mathbf{x}_2)\|$ will also be non-zero, which in turn requires that $\|\Phi\mathbf{x}_1 \neq \Phi\mathbf{x}_2\|$ so that Φ is *one to one* on \mathcal{S} (that is, Φ maps distinct points in \mathcal{S} into distinct points in \mathcal{L}).

However, a non-zero bound with $\alpha > 0$ actually tells us more. If we use our theoretical reconstruction technique, that is, we project \mathbf{y} onto $\Phi\mathcal{S}$ (assuming for now that this projection exists, though a similar argument can be made for ϵ -projections) and then find the corresponding $\mathbf{x} \in \mathcal{S}$. Say $\mathbf{y}_{\mathcal{S}}$ is the projection and $\tilde{\mathbf{x}}$ is the corresponding element in \mathcal{S} so that $\Phi\tilde{\mathbf{x}} = \mathbf{y}_{\mathcal{S}}$. How far will \mathbf{x} be from $\tilde{\mathbf{x}}$? We have

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \frac{1}{\alpha}\|\Phi\mathbf{x} - \Phi\tilde{\mathbf{x}}\| = \frac{1}{\alpha}\|\mathbf{y} - \mathbf{e} - \mathbf{y}_{\mathcal{S}}\| \leq \frac{1}{\alpha}\|\mathbf{y} - \mathbf{y}_{\mathcal{S}}\| + \frac{1}{\alpha}\|\mathbf{e}\| \leq \frac{2}{\alpha}\|\mathbf{e}\|,$$

where the second to last inequality is the triangle inequality (which is one of the properties of a norm) and where the last inequality is due to the fact that $\mathbf{y}_{\mathcal{S}}$ is the closest element in $\Phi\mathcal{S}$ to \mathbf{y} and is thus closer to \mathbf{y} than $\Phi\mathbf{x}$ itself. Thus $\|\Phi\mathbf{x} - \mathbf{y}\| = \|\mathbf{e}\| \geq \|\mathbf{y} - \mathbf{y}_{\mathcal{S}}\|$.

We thus have the following Lemma.

Lemma 2 *For any $\mathbf{x} \in \mathcal{S}$, let $\mathbf{y} = \Phi\mathbf{x} + \mathbf{e}$, where Φ satisfies the bi-Lipschitz condition with $\alpha > 0$ and let $\mathbf{y}_{\Phi\mathcal{S}}$ be the closest element in $\Phi\mathcal{S}$ to \mathbf{y} , then the error between \mathbf{x} and $\tilde{\mathbf{x}} \in \mathcal{S}$ uniquely defined by $\mathbf{y}_{\Phi\mathcal{S}} = \Phi\tilde{\mathbf{x}}$ satisfies*

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \frac{2}{\alpha} \|\mathbf{e}\|. \quad (2.17)$$

Therefore, if $\mathbf{x} \in \mathcal{S}$ and if Φ is linear and satisfies the bi-Lipschitz condition, then our theoretical reconstruction technique will recover a signal $\tilde{\mathbf{x}}$ that is no more than $\frac{2}{\alpha} \|\mathbf{e}\|$ away from the true signal \mathbf{x} . This is good news, we have just shown that, at least in theory, we should be able to recover any $\mathbf{x} \in \mathcal{S}$ as long as Φ is bi-Lipschitz on \mathcal{S} . The worst case accuracy of our recovered signal will then only depend on the amount of measurement noise \mathbf{e} and on the inverse of the lower bi-Lipschitz constant α .

The same argument would hold for non-linear Φ if $\alpha\|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)\|$, so that a non-linear operator Φ with this condition also guarantees that the theoretical inverse is stable, that is, if we find that element $\mathbf{y}_{\mathcal{S}}$ in $\Phi\mathcal{S}$ closest to \mathbf{y} , then the $\tilde{\mathbf{x}}$ that satisfies $\mathbf{y}_{\mathcal{S}} = \Phi(\tilde{\mathbf{x}})$ will be close to \mathbf{x} .

2.3.1.5 All Models Are Wrong

\mathcal{S} is a model for our signal and we assumed above that $\mathbf{x} \in \mathcal{S}$. However, as all models are wrong (at least in general), any errors in the model have to be taken into account in our discussion. Let us therefore consider what happens if \mathbf{x} does not lie exactly in \mathcal{S} , but only ‘near by’. To deal with this case in our framework, we will consider the projection of \mathbf{x} onto \mathcal{S} . Again, as \mathcal{S} can be a general non-convex set, this ‘projection’ is not guaranteed to exist and is definitely not required to be unique. The first problem can be dealt with in a similar way in which we dealt with the projection onto $\Phi\mathcal{S}$. We can either find an δ optimal point or, restrict discussions to sets \mathcal{S} that allow us to find a closest point in \mathcal{S} to all points $\mathbf{x} \in \mathcal{H}$. To simplify notation, we restrict ourselves here to the second case and assume there is such a closest point. However, this might not be unique. If there are more than one point that is closest to a point \mathbf{x} we will thus assume that we choose one of these. We call this point $\mathbf{x}_{\mathcal{S}}$, so that $\mathbf{x}_{\mathcal{S}} \in \mathcal{S}$ and $\|\mathbf{x} - \mathbf{x}_{\mathcal{S}}\| \leq \inf_{\bar{\mathbf{x}} \in \mathcal{S}} \|\mathbf{x} - \bar{\mathbf{x}}\|$. What about error $\|\mathbf{x} - \mathbf{x}_{\mathcal{S}}\|$? To recover \mathbf{x} from $\mathbf{y} = \Phi(\mathbf{x}) + \mathbf{e}$ we follow the same steps as before, we ‘project’ \mathbf{y} onto $\Phi(\mathcal{S})$ and then find the corresponding $\tilde{\mathbf{x}} \in \mathcal{S}$. How far is this estimate now from \mathbf{x} ? Again, we require the stability condition $\alpha\|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)\|$ to hold, so that (this time using the non-linear notation)

$$\begin{aligned} \|\mathbf{x} - \tilde{\mathbf{x}}\| &= \|\mathbf{x} - \mathbf{x}_{\mathcal{S}} + \mathbf{x}_{\mathcal{S}} - \tilde{\mathbf{x}}\| \\ &\leq \|\mathbf{x} - \mathbf{x}_{\mathcal{S}}\| + \|\mathbf{x}_{\mathcal{S}} - \tilde{\mathbf{x}}\| \end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{\alpha} \|\Phi(\mathbf{x}_S) - \Phi(\tilde{\mathbf{x}})\| + \|\mathbf{x} - \mathbf{x}_S\| \\
&= \frac{1}{\alpha} \|\mathbf{y} - \tilde{\mathbf{e}} - \mathbf{y}_S\| + \|\mathbf{x} - \mathbf{x}_S\| \\
&\leq \frac{1}{\alpha} \|\mathbf{y} - \mathbf{y}_S\| + \frac{1}{\alpha} \|\tilde{\mathbf{e}}\| + \|\mathbf{x} - \mathbf{x}_S\| \\
&\leq \frac{1}{\alpha} \|\mathbf{e}\| + \frac{1}{\alpha} \|\tilde{\mathbf{e}}\| + \|\mathbf{x} - \mathbf{x}_S\|, \\
&\leq \frac{2}{\alpha} \|\mathbf{e}\| + \frac{1}{\alpha} \|\Phi(\mathbf{x}) - \Phi(\mathbf{x}_S)\| + \|\mathbf{x} - \mathbf{x}_S\|.
\end{aligned}$$

where $\tilde{\mathbf{e}} = \mathbf{e} + \Phi(\mathbf{x}) - \Phi(\mathbf{x}_S)$ and where the first inequality is again the triangle inequality. Thus, if our model is wrong, then our recovery lemma reads (spot the two small differences to the previous version, (1) \mathbf{x} is no longer required to lie in S and (2) the distances of \mathbf{x} from \mathbf{x}_S and of $\Phi(\mathbf{x})$ from $\Phi(\mathbf{x}_S)$ now join the error bound)

Lemma 3 *For any \mathbf{x} , let $\mathbf{y} = \Phi\mathbf{x} + \mathbf{e}$, where Φ satisfies the bi-Lipschitz condition with $\alpha > 0$ and let $\mathbf{y}_{\Phi S}$ be the closest element in ΦS to \mathbf{y} , then the error between \mathbf{x} and $\tilde{\mathbf{x}} \in S$ uniquely defined by $\mathbf{y}_{\Phi S} = \Phi\tilde{\mathbf{x}}$ satisfies*

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \frac{2}{\alpha} \|\mathbf{e}\| + \frac{1}{\alpha} \|\Phi(\mathbf{x}) - \Phi(\mathbf{x}_S)\| + \|\mathbf{x} - \mathbf{x}_S\|. \quad (2.18)$$

Thus, even if \mathbf{x} is no longer within our model, we can still use the model S to recover \mathbf{x} . All we loose in the accuracy of our reconstruction is then the additional error terms $\mathbf{x} - \mathbf{x}_S$ and $\Phi(\mathbf{x}) - \Phi(\mathbf{x}_S)$. Thus, if \mathbf{x} is close to S , then we can still recover \mathbf{x} with high accuracy.

We have thus demonstrated that it is possible to recover elements from \mathcal{H} which are close to elements in S from noisy observations $\mathbf{y} = \Phi(\mathbf{x}) + \mathbf{e}$ whenever $\alpha\|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)\|$ holds for all $\mathbf{x}_1, \mathbf{x}_2 \in S$. However, our approach to do this recovery required two steps, (1) find an element \mathbf{y}_S in ΦS closest to \mathbf{y} and (2) find that $\tilde{\mathbf{x}} \in S$ such that $\Phi\tilde{\mathbf{x}} = \mathbf{y}_S$. For many complex models S , both of these steps are far from trivial. For several sets S that are of interest in many applications, we will thus study more practical methods to recover \mathbf{x} . Crucially, not only are these approaches computationally much more efficient than the approach described above, they will also be shown to have a similar worst case recovery error.

2.4 Geometry of Convex Relaxation

The first efficient approach we will discuss that can be used to recover data in certain data-recovery problems under non-convex constraints uses convexification of the constraint set. This is the traditional approach used in compressed sensing and its operation relies on some beautiful geometrical reasoning. Convexification based

ideas have been developed predominantly for sparse problems, where there is a natural and powerful convex version of the constraint. Consider a real vector \mathbf{x} and let $\|\mathbf{x}\|_0$ be the number of non-zero entries in the vector \mathbf{x} . If we want to optimise with the constraint that $\|\mathbf{x}\|_0$ is smaller than some specified integer, then we have a non-convex constraint. Similarly, if we would like to optimise \mathbf{x} so that $\|\mathbf{x}\|_0$ is as small as possible, subject to some other constraint (for example $\mathbf{y} = \Phi\mathbf{x}$), then we are dealing with a non-convex cost function. To simplify these problems, we can replace the non-convex function $\|\mathbf{x}\|_0$ with the norm $\|\mathbf{x}\|_1$, i.e. with the ℓ_1 vector norm, which directly leads to convex problems that are much easier to solve numerically.

The question now is, under which conditions are the solutions to problems that use $\|\mathbf{x}\|_0$ equivalent or similar to the solutions solved with their convex version based on the norm $\|\mathbf{x}\|_1$? To study this problem, we will look at the geometry of the constraint set $\|\mathbf{x}\|_1 \leq 1$.

2.4.1 The Null-Space and Its Properties

Our treatment of the topic here is inspired by the work in [16, 18]. Consider the compressed sensing problem: minimise $\|\mathbf{x}\|_0$ such that $\mathbf{y} = \Phi\mathbf{x}$ and its convex counterpart: minimise $\|\mathbf{x}\|_1$ such that $\mathbf{y} = \Phi\mathbf{x}$. Let $\hat{\mathbf{x}}$ be the solution to the second one of these problems and let \mathbf{x}_k be the best k -term approximation of the vector \mathbf{x} , that is \mathbf{x}_k satisfies $\|\mathbf{x}_k - \mathbf{x}\| = \min_{\tilde{\mathbf{x}}: \|\tilde{\mathbf{x}}\|_0 = k} \|\tilde{\mathbf{x}} - \mathbf{x}\|$.

The null-space of Φ will play a fundamental role in this section. Let \mathbf{h} be a vector in this null-space, that is, we have $\Phi\mathbf{h} = \mathbf{0}$. We will also use the following measure that characterises how well vectors in this null-space align with the co-ordinate axis. The null-space property of Φ is defined as follows. Let C_k be the largest constant such that

$$C_k \sum_{i \in \mathcal{K}} |\mathbf{h}_i| \leq \sum_{i \notin \mathcal{K}} |\mathbf{h}_i|, \quad (2.19)$$

holds for all vectors \mathbf{h} in the null-space of Φ and for all index sets \mathcal{K} of size k or less. Importantly, if the above condition holds for all subsets of \mathcal{K} elements of the vector \mathbf{h} , then it must also hold for the subset of the k largest elements. We can therefore write the above condition as

$$C_k \|\mathbf{h}_k\| \leq \|\mathbf{h} - \mathbf{h}_k\|, \quad (2.20)$$

where \mathbf{h}_k is again the vector with the largest (in magnitude) k elements of \mathbf{h} and zeros elsewhere. This condition is known as the null-space property of Φ and if it holds for $C_k \leq 1$, then we say that Φ satisfies the null-space property of order k .

2.4.2 The Null-Space Property for Signal Recovery

The null-space property directly implies a bound on the quality of the solution $\hat{\mathbf{x}}$ to the convex optimisation problem: minimise $\|\mathbf{x}\|_1$ such that $\mathbf{y} = \Phi \mathbf{x}$.

To see this, let \mathbf{x} be any vector such that $\mathbf{y} = \Phi \mathbf{x}$ and let $\hat{\mathbf{x}}$ be the minimum of the optimisation problem so that $\|\hat{\mathbf{x}}\|_1 \leq \|\mathbf{x}\|_1$ and $\mathbf{y} = \Phi \hat{\mathbf{x}} = \Phi \mathbf{x}$. We want to bound the length of the error $\hat{\mathbf{x}} - \mathbf{x}$. To do this, we first note that the vector $\mathbf{h} = \hat{\mathbf{x}} - \mathbf{x}$ lies in the null-space of Φ . To see this we use the fact that $\mathbf{y} = \Phi \hat{\mathbf{x}} = \Phi \mathbf{x}$, so that $\mathbf{0} = \Phi \hat{\mathbf{x}} - \Phi \mathbf{x} = \Phi(\hat{\mathbf{x}} - \mathbf{x})$.

Note that the ℓ_1 norm has the property that for any vector \mathbf{x} , we have $\|\mathbf{x}\|_1 = \|\mathbf{x}_k\|_1 + \|\mathbf{x} - \mathbf{x}_k\|_1$. Furthermore, note that the null-space property implies that for all \mathbf{h} that lie in the null-space of Φ

$$\begin{aligned} (C-1)(\|\mathbf{h}_k\|_1 + \|\mathbf{h} - \mathbf{h}_k\|_1) &= C\|\mathbf{h}_k\|_1 - \|\mathbf{h}_k\|_1 + C\|\mathbf{h} - \mathbf{h}_k\|_1 - \|\mathbf{h} - \mathbf{h}_k\|_1 \\ &\leq \|\mathbf{h} - \mathbf{h}_k\|_1 - \|\mathbf{h}_k\|_1 + C\|\mathbf{h} - \mathbf{h}_k\|_1 - C\|\mathbf{h}_k\|_1 \\ &= C+1(\|\mathbf{h} - \mathbf{h}_k\|_1 - \|\mathbf{h}_k\|_1), \end{aligned} \quad (2.21)$$

which we will use in the following form

$$\|\mathbf{h}_k\|_1 + \|\mathbf{h} - \mathbf{h}_k\|_1 \leq \frac{C+1}{C-1} (\|\mathbf{h} - \mathbf{h}_k\|_1 - \|\mathbf{h}_k\|_1) \quad (2.22)$$

Using these two inequalities, we can then decompose and bound the ℓ_1 norm of the error $\mathbf{x} - \hat{\mathbf{x}}$.

$$\begin{aligned} \|\mathbf{x} - \hat{\mathbf{x}}\|_1 &= \|\mathbf{h}\|_1 = \|\mathbf{h}_k\|_1 + \|\mathbf{h} - \mathbf{h}_k\|_1 \\ &\leq \frac{C+1}{C-1} (\|\mathbf{h} - \mathbf{h}_k\|_1 - \|\mathbf{h}_k\|_1) \\ &= \frac{C+1}{C-1} (\|\mathbf{h} - \mathbf{h}_k\|_1 - \|\mathbf{h}_k\|_1 - \|\mathbf{x} - \mathbf{x}_k\|_1 + \|\mathbf{x} - \mathbf{x}_k\|_1) \\ &\leq \frac{C+1}{C-1} (\|(\mathbf{h} - \mathbf{h}_k) + (\mathbf{x} - \mathbf{x}_k)\|_1 - \|\mathbf{h}_k\|_1 + \|\mathbf{x} - \mathbf{x}_k\|_1) \\ &= \frac{C+1}{C-1} (\|(\mathbf{h} - \mathbf{h}_k) + (\mathbf{x} - \mathbf{x}_k)\|_1 - \|\mathbf{h}_k\|_1 + \|\mathbf{x}_k\|_1 - \|\mathbf{x}_k\|_1 + \|\mathbf{x} - \mathbf{x}_k\|_1) \\ &\leq \frac{C+1}{C-1} (\|(\mathbf{h} - \mathbf{h}_k) + (\mathbf{x} - \mathbf{x}_k)\|_1 + \|\mathbf{h}_k + \mathbf{x}_k\|_1 - \|\mathbf{x}_k\|_1 + \|\mathbf{x} - \mathbf{x}_k\|_1) \\ &= \frac{C+1}{C-1} (\|\mathbf{x} + \mathbf{h}\|_1 - \|\mathbf{x}_k\|_1 + \|\mathbf{x} - \mathbf{x}_k\|_1) \\ &= \frac{C+1}{C-1} (\|\hat{\mathbf{x}}\|_1 - \|\mathbf{x}_k\|_1 + \|\mathbf{x} - \mathbf{x}_k\|_1) \\ &\leq \frac{C+1}{C-1} (\|\mathbf{x}\|_1 - \|\mathbf{x}_k\|_1 + \|\mathbf{x} - \mathbf{x}_k\|_1) \end{aligned}$$

$$\begin{aligned}
&= \frac{C+1}{C-1} (\|\mathbf{x} - \mathbf{x}_k\|_1 + \|\mathbf{x} - \mathbf{x}_k\|_1) \\
&= 2 \frac{C+1}{C-1} \|\mathbf{x} - \mathbf{x}_k\|_1
\end{aligned} \tag{2.23}$$

Let us walk through this chain of equalities and inequalities at a more pedestrian speed. The first equality just re-states that the error $\mathbf{x} - \hat{\mathbf{x}}$ lies in the null-space of Φ . The second equality is then the first property above, whilst the first inequality is the second property in (2.22). The next equality simply adds and subtracts $\|\mathbf{x} - \mathbf{x}_k\|_1$, whilst in the following line we use the triangle inequality

$$\begin{aligned}
\|(\mathbf{h} - \mathbf{h}_k) + (\mathbf{x} - \mathbf{x}_k)\|_1 &= \|(\mathbf{h} - \mathbf{h}_k) + (\mathbf{x} - \mathbf{x}_k) + (\mathbf{x}_k + \mathbf{h}_k) - (\mathbf{x}_k + \mathbf{h}_k)\|_1 \\
&\leq \|(\mathbf{h} - \mathbf{h}_k) + (\mathbf{x} - \mathbf{x}_k) + (\mathbf{x}_k + \mathbf{h}_k)\|_1 + \|(\mathbf{x}_k + \mathbf{h}_k)\|_1.
\end{aligned}$$

We again add and subtract the same number, before making a second use of the triangle inequality. We then use the fact that the two vectors $(\mathbf{h} - \mathbf{h}_k) + (\mathbf{x} - \mathbf{x}_k)$ and $\mathbf{h}_k + \mathbf{x}_k$ have different support, so that we can again use property one. The next equality just uses the definition of $\hat{\mathbf{x}} = \mathbf{x} + \mathbf{h}$, whilst the last inequality uses the fact that $\|\hat{\mathbf{x}}\|_1 \leq \|\mathbf{x}\|_1$ (remember, $\hat{\mathbf{x}}$ minimises the ℓ_1 norm among all \mathbf{x} that satisfy $\mathbf{y} = \Phi\mathbf{x}$). We finish the argument by a final application of property one.

Interestingly, the requirement that the null-space property holds is not only sufficient for the above bound to hold (as we have just shown) but is also necessary in the following sense. If the null-space property is violated, then there exists a measurement matrix with this null-space so that the above bound is violated for some k [16]. Note however that this does not imply that the bound is violated necessarily for any particular measurement matrix Φ even if it has a null-space that violates the condition.

Note also that the result here is slightly different from that of the “ideal” algorithm of the previous section and is also different from the bounds we derive in the next section. Firstly, the null-spaced based results are not able to account for measurement errors. Secondly, the bound here is in terms of the ℓ_1 norm of the error $\mathbf{x} - \mathbf{x}_k$, that is, it tells us how well we can approximate vectors whose $N - k$ smallest coefficients have a small ℓ_1 norm. A theory based on ideas similar to the bi-Lipschitz condition on Φ can also be derived. This is done for example in [5, 17]. For example, in [17] we have the following result which is more similar to that in Lemma 3.

Theorem 2 *For any \mathbf{x} , assume Φ satisfies the bi-Lipschitz property*

$$(1 - \gamma)\|\mathbf{x}_1 + \mathbf{x}_2\|^2 \leq \|\Phi(\mathbf{x}_1 + \mathbf{x}_2)\|^2 \leq (1 + \gamma)\|\mathbf{x}_1 + \mathbf{x}_2\|^2, \tag{2.24}$$

where $\gamma < \sqrt{2} - 1$. Given observations $\mathbf{y} = \Phi\mathbf{x} + \mathbf{e}$, the minimiser of the problem $\min_{\tilde{\mathbf{x}}} \|\tilde{\mathbf{x}}\|_1$ subject to the constraint that $\|\mathbf{y} - \Phi\tilde{\mathbf{x}}\| \leq \|\mathbf{e}\|$ recovers an estimate $\hat{\mathbf{x}}$ that satisfies

$$\|\mathbf{x} - \hat{\mathbf{x}}\| \leq_0 C\|\tilde{\mathbf{e}}\| + C_1\|\mathbf{x}_k - \mathbf{x}\|, \tag{2.25}$$

where $\tilde{\mathbf{e}} = \Phi(\mathbf{x} - \mathbf{x}_k) + \mathbf{e}$ and where C_o and C_1 are constants depending on γ .

Instead of proving this result here (the interested reader is redirected to [17]), we instead return to the null-space property and study the geometrical implications this property has for the recovery of sparse vectors in somewhat more detail.

2.4.3 Random Null-Spaces and the Grassman Angle

To build a measurement system that would allow us to use ℓ_1 recovery with the tight error bounds derived in (2.23), we thus need to ensure that the measurement system satisfies the null-space property. One particularly powerful approach to construct measurement systems is through random construction methods and it can be shown that these systems often satisfy the required null-space properties. As the null-space property is fundamentally geometrical in nature, geometrical ideas can also be used to study and understand these construction techniques.

Instead of the careful construction of a matrix whose null-space satisfies the null-space property, it is significantly simpler to randomly choose a null-space and then construct a matrix that has the same null-space. In fact, this random construction is one of the only few known construction method that can build matrices that on the one hand satisfy the null-space property and on the other hand, are optimal in terms of the number of measurements. However, we must note that, if we use a random construction, then our desired property will only hold with high probability and is not absolutely guaranteed.

We will assume that the null-space is chosen randomly in such a way that its distribution is rotation invariant. With this we mean that, if \mathbf{B} is a basis for a null-space of dimension N and if \mathbf{U} is an orthonormal rotation matrix, then any rotation invariant distribution $p(\mathbf{B})$ must satisfy $p(\mathbf{B}) = p(\mathbf{UB})$. For example, if we choose the entries of the matrix $\Phi \in \mathbb{R}^{M \times N}$ to be drawn independently from a zero-mean unit variance normal distribution, and if $M < N$, then the distribution of the null-space of Φ will have this property.

The null-space property of a matrix Φ is related to the following property (see [16]).

Lemma 4 *Let \mathcal{K} be a subset of k of the indices of a vector in \mathbb{R}^N . Then, the null-space property*

$$C \|\mathbf{h}\|_1 \leq \|\mathbf{h} - \mathbf{h}_k\|_1, \quad (2.26)$$

for all \mathbf{h} in the null-space is equivalent to the property that all vectors \mathbf{x} supported on \mathcal{K} satisfy

$$\|\mathbf{x} + \mathbf{h}_k\|_1 + \left\| \frac{\mathbf{h} - \mathbf{h}_k}{C} \right\|_1 \geq \|\mathbf{x}\|_1 \quad (2.27)$$

for all \mathbf{h} in the null-space.

We here use the notation \mathbf{h}_k to refer to a version of the vector \mathbf{h} in which all entries are set to 0 apart from those elements with indices in the set \mathcal{K} .

To derive a lower bound on the probability under which a randomly sampled subspace satisfies the null-space property, we can therefore derive an upper bound on the probability with which the above condition fails. That is, what is the probability that for any k -sparse vector \mathbf{x} the condition in (2.27) will fail?

To answer this question, we first note that we can restrict our attention to vectors \mathbf{x} that satisfy $\|\mathbf{x}\|_1 = 1$. This is because if (2.27) holds or fails for any \mathbf{x} , then it will also hold or fail for $c\mathbf{x}$ for any c .

Let us now look at the probability that a randomly chosen null-space violates (2.27) for a particular \mathbf{x} with a given support set \mathcal{K} and a particular sign pattern. We will call this probability $P_{\mathcal{K}}$. To understand the geometric properties of $P_{\mathcal{K}}$ let us consider all vectors \mathbf{x} which satisfy $\|\mathbf{x}\|_1 = 1$ and which have a support \mathcal{K} with $|\mathcal{K}| = k$.

As we assume $\|\mathbf{x}\|_1 = 1$, the condition in (2.27) is related to the following geometrical object.

$$WB = \{\hat{\mathbf{x}} \in \mathbb{R}^N : \|\hat{\mathbf{x}}_k\|_1 + \|\frac{\hat{\mathbf{x}} - \hat{\mathbf{x}}_k}{C}\|_1 \leq 1\}. \quad (2.28)$$

We call this cross-polytope the weighted ℓ_1 ball. A sketch of WB , \mathbf{x} and \mathbf{h} is given in Fig. 2.1. The probability $P_{\mathcal{K}}$ is thus the probability that there exist a vector $\mathbf{h} \neq \mathbf{0}$ in the null-space of Φ so that for at least one k -sparse vector \mathbf{x} with $\|\mathbf{x}\|_1 = 1$ and support \mathcal{K} , where $\text{sign}(\mathbf{x}_k)$ is fixed, we have

$$\|\mathbf{x} + \mathbf{h}_k\|_1 + \|\frac{\mathbf{h} - \mathbf{h}_k}{C}\|_1 < \|\mathbf{x}\|_1 = 1. \quad (2.29)$$

Note that all the k -sparse \mathbf{x} we consider here (those with $\|\mathbf{x}\|_1 = 1$) lie on the surface of an ℓ_1 ball. Furthermore, because \mathbf{x} is assumed to be k -sparse, \mathbf{x} lies on a k -dimensional face. To get a geometrical intuition for this, think of a diamond (build by sticking two equal pyramids that have a square base and equal-lateral triangle sides together at the square surface). Such a diamond is a cross-polytope in three dimensions. Each of its eight triangular sides is a two dimensional face. Furthermore, the diamond has eight ridges, which are, in high-dimensional geometry language, one-dimensional faces. Finally, the six sharp corners are called zero-dimensional faces or, alternatively, vertices. To further build our geometrical intuition, assume that our co-ordinate axis are aligned with the diamond edges (or vertices). If these vertices lie at exactly the points $[100]$, $[-100]$, $[010]$, $[0-10]$, $[001]$ and $[00-1]$, then the diamond is the unit ℓ_1 ball in three dimensions. Importantly, note also that any 2-sparse vector with unit ℓ_1 norm will lie on one of the ridges (or two-dimensional faces). Now, once we fix the support set \mathcal{K} , then in our two dimensional example, x_k will lie in one of the three planes that align exactly with four of the eight ridges of the diamond. Exactly which four ridges depends on the support set \mathcal{K} . The weighted ℓ_1 ball in this three dimensional example would then be a stretched diamond in

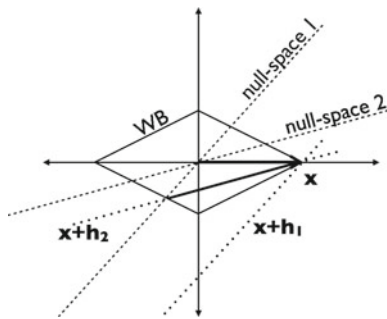


Fig. 2.1 Low dimensional sketch of the vectors, subspaces and ℓ_1 ball involved in the discussion of this section. For this simple two dimensional example, the null-spaces are chosen to be one-dimensional. A 1-sparse vectors \mathbf{x} can be recovered if the null-space is null-space 1, as there is no null-space vector \mathbf{h}_1 such that $\mathbf{x} + \mathbf{h}_1$ lies within the cross-polytope WB (see dotted line labeled $\mathbf{x} + \mathbf{h}_1$). If, on the other hand, the null-space is null-space 2, then there exist null-space vectors \mathbf{h}_2 such that $\mathbf{x} + \mathbf{h}_2$ lies within WB (see the solid section of the dotted line labeled $\mathbf{x} + \mathbf{h}_2$) and \mathbf{x} is not recoverable

which the two vertices that do not lie on the two dimensional subspace defined by the support set \mathcal{K} are further away from the coordinate centre. Furthermore, we only consider one of the four possible sign patterns, which means that \mathbf{x}_k is assumed to lie on only one of the four ridges.

The same principle holds in higher dimensions. Consider any particular k -sparse \mathbf{x} that lies in the interior of a k -face of the “stretched” ℓ_1 ball (the k -face itself lies in the plane where no “stretching” has occurred). With interior of the face we here mean that we assume that \mathbf{x} lies exactly within the k -face but not in any one of the $k-1$ -faces, i.e. \mathbf{x} has exactly k non-zero entries and not fewer. In our three dimensional pyramid for $k=2$ this would mean that \mathbf{x} lies on a ridge, but not exactly at a corner.

Let us now consider the probability that a randomly drawn subspace has vectors that satisfy

$$\|\mathbf{x} + \mathbf{h}_k\|_1 + \left\| \frac{\mathbf{h} - \mathbf{h}_k}{C} \right\|_1 < \|\mathbf{x}\|_1 = 1 \quad (2.30)$$

for at least one such \mathbf{x} , which as stated above is exactly the probability with which such a null-space would violate the null-space property for one of the possible sign patterns.

In our three dimensional example, the stretched cross-polytope is a stretched diamond and any 2-sparse vector \mathbf{x} would be assumed to lie on one of the non-stretched ridges. Now if we were to draw a direction \mathbf{d} at random (our null-space) and consider the affine subspace $\mathbf{x} + \mathbf{d}$, then what is the probability that this subspace does not go through the stretched diamond? Whilst in this three dimensional example with a two sparse vector, a one-dimensional subspace is really the only interesting scenario, in high dimensions, there is substantially more space and there is actually space to “attach” significantly higher dimensional subspaces onto low-dimensional faces of our diamond without the subspace actually cutting into the diamond itself.

Let us first try and think about the probability that a one-dimensional subspace does not intersect our stretched diamond if we attach it to a particular ridge. This probability is equivalent to a randomly drawn vector lying within a particular cone. To see this, take your diamond and shift it so that the point \mathbf{x} lies at the centre of our coordinate system. In our three-dimensional example, all that we really care about in terms of the intersection of our subspace and the diamond is the intersection with the two *two*-faces that intersect at the ridge on which \mathbf{x} lies. Now after shifting our diamond, our condition is violated as soon as a randomly drawn vector intersect any one of these two faces. And this happens with exactly the same probability with which a randomly drawn vector would lie within the cone generated by these two faces. Our problem is thus the same as one of specifying the probabilities with which randomly drawn subspaces lie within a cone specified by the faces of our cross-polytope that intersect with the face on which \mathbf{x} lies.

Luckily, this is a problem that has been studied before. In fact, the probability that a randomly chosen low-dimensional subspace intersects with a skewed cross-polytope is equal to a geometric property known as the complementary Grassmann angle [19]. There even is a ready made formula available to calculate the complementary Grassman angle for any $(k - 1)$ dimensional face [20].

$$P_{|\mathcal{K}|} = 2 \sum_i \sum_{FACE_i} \beta(FACE_k, FACE_i) \gamma(FACE_i, WB). \quad (2.31)$$

The first sum is over all non-negative integers i and the second sum is over all $(M+1+2i)$ dimensional faces $FACE_i$ of the skewed cross-polytope. Here $FACE_k$ is the k dimensional face on which we assume that \mathbf{x} lies whilst WB is the entire cross-polytope itself. Both functions $\beta(\cdot, \cdot)$ and $\gamma(\cdot, \cdot)$ are functions of two faces of the cross-polytope (note that the entire polytope also counts as a face).

$\beta(FACE_1, FACE_2)$ is known as the internal angle. The internal angle is a geometrical property of the two faces. The angle is calculated by considering the following cone C . For all \mathbf{x} in $FACE_1$, shift the polytope so that $\mathbf{x} = \mathbf{0}$ and let $C(\mathbf{x})$ be the cone of all vectors that leave \mathbf{x} and intersect the face $FACE_2$. C is then the intersection of all cones $C = \bigcup_{\mathbf{x} \in FACE_1} C(\mathbf{x})$. The internal angle is now the proportion of the unit sphere of the same dimension as the cone, that is covered by the cone. The internal angle is zero if the two faces do not intersect and is unity if the faces are identical.

$\gamma(FACE_1, FACE_2)$ on the other hand is known as the external angle. The external angle is defined in a similar way, however, the cone is constructed differently by considering all outward normals to the hyperplanes that support the two faces. The external angle is again zero if the two faces do not intersect and is unity if the faces are identical.

The main effort now is the derivation for expressions that quantify these angles [16], but instead of going through this lengthy derivation here, or in fact, stating the expressions themselves, let us instead consider how these can be used to bound the probability we are interested in.

Now the above probability was for a given support set and a given sparsity pattern. However, we require the condition to hold for all support sets. To derive such a bound, let us first count the number of different support sets. For each support set, there are $2^{|\mathcal{K}|}$ different sign patterns. Furthermore, there are $\binom{n}{k}$ different support sets with k non-zero elements. We can therefore use a so called union bound to bound the probability of failure. A union bound uses the following simple probabilistic fact. If A and B are two events, then the probability that A or B holds (write $P(A \cup B)$) is always smaller or equal the probability that A holds ($P(A)$) added to the probability that B holds ($P(B)$). Thus

$$P(A \cup B) \leq P(A) + P(B). \quad (2.32)$$

If we apply this principle to the probabilities that (2.27) is violated for one of the support sets and one of the sign patterns, then we can bound this probability as

$$P(\mathbf{Failure}) \leq \binom{n}{k} 2^k P_{\mathcal{K}}. \quad (2.33)$$

We thus see that the probability is bounded as a function of k and $P_{\mathcal{K}}$. $P_{\mathcal{K}}$ itself depends on the dimensions of the problem M and N as well as k . It also depends on C (as C specifies the amount of stretching in our weighted ℓ_1 ball). The main message is that, if we require a level of robustness (as defined by C and k) and want to observe a vector of length N , then we need to choose the number of observations large enough, so that the probability $\binom{n}{k} 2^k P_{\mathcal{K}}$ is sufficiently small. In this case, a randomly chosen $N - M$ dimensional subspace will (with a probability bounded by $\binom{n}{k} 2^k P_{\mathcal{K}}$) allow us to reconstruct our vector within the required precision. Unfortunately, closed form expressions are not available for the probabilistic bound derived here, however, numerical methods can be used to evaluate the required Grassmann angles for any required combination of C , M , N and k [16].

2.5 Geometry of Iterative Projection Algorithms

There are two main approaches to the solution of signal recovery problems under non-convex constraints. The first approach, discussed in the previous section, replaced the non-convex problem with a convex one, thus greatly simplifying it. In this section we look at an alternative, greedy methods. Greedy methods are iterative schemes that replace a non-convex optimisation problem with a sequence of simpler problems. The moniker ‘greedy’ here indicates that these methods ‘greedily’ grab a signal from the non-convex constraint set to satisfy these local optimisation constraints. Whilst there are many greedy algorithms, we here discuss a conceptually simple, yet extremely powerful approach that has similar performance guarantees to the convex relaxation based approaches discussed above, yet can also be used with many non-convex constraints for which there are no simple convex relaxations.

2.5.1 The Iterative Hard Thresholding Algorithm

The Iterative Hard Thresholding algorithm [21, 22], also known as the Iterative Projection or Projected Landweber Algorithm, is an iterative method that, as the name suggests, thresholds or projects an estimate iteratively. To see how this method works, let us consider again the optimisation problem we are trying to solve.

$$\min_{\mathbf{x}} \|\mathbf{y} - \Phi \mathbf{x}\|^2 : \mathbf{x} \in \mathcal{S}, \quad (2.34)$$

where \mathcal{S} is a possibly non-convex constraint set.

Without any constraint, the simplest approach to tackle the above problem would be to use gradient optimisation (assuming that the gradient of $\|\mathbf{y} - \Phi \mathbf{x}\|^2$ exists). If \mathbf{g} is the negative gradient of $\|\mathbf{y} - \Phi \mathbf{x}\|^2$ (or the Gâteaux derivative if \mathbf{x} is a more general function), then this optimisation would update an estimate \mathbf{x}^n using the iteration

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \Omega \mathbf{g}, \quad (2.35)$$

where Ω is either a scalar step size, or more generally, a linear map to precondition and stabilise the problem. For example, Ω could be the inverse of $\Phi \Phi^T$ [23, 24] or the Hessian operator as in Newton's method. However, if Φ is non-invertible or ill-conditioned, then this optimisation will not lead to a unique and stable solution, which was the reason why the constraint set \mathcal{S} was introduced in the first place. Thus, to utilise the constraint, we simply enforce the requirement that \mathbf{x}^{n+1} lies in \mathcal{S} . To do this, the estimate $\mathbf{a} = \mathbf{x}^n + \mu \mathbf{g}$ has to be mapped to an element in \mathcal{S} and, to keep the potential increase in our cost $\|\mathbf{y} - \Phi \mathbf{x}\|^2$ this mapping entails to a minimum, this mapping should not take us too far away from \mathbf{a} itself. Thus, we would ideally like to find a point in \mathcal{S} that is as close to \mathbf{a} as possible. If we are able to calculate such a projection for the non-convex set \mathcal{S} , then we can use the Iterative Hard Thresholding algorithm.

$$\mathbf{x}^{n+1} = P_{\mathcal{S}}(\mathbf{x}^n + \Omega \mathbf{g}), \quad (2.36)$$

where $P_{\mathcal{S}}$ is this projection mapping.

This procedure might remind the reader of the approach we have discussed above, in which the reconstruction was done via a projection of \mathbf{y} onto the closest element in $\Phi \mathcal{S}$. In principle, the projection $P_{\mathcal{S}}$ is defined in a similar way to the projection onto the set $\Phi \mathcal{S}$. Thus, if we are able to efficiently calculate the projection onto $\Phi \mathcal{S}$, then there would be no need to use the more complex Iterative Hard Thresholding algorithm. The point however is that, for many constraint sets \mathcal{S} used in practice, calculating the projection $P_{\mathcal{S}}$ is significantly more efficient than to try and project onto the set $\Phi \mathcal{S}$. Several examples are given next.

2.5.2 Projections onto Non-convex Sets

Let us start by formalising again what we will mean when talking about projections onto the set $\mathcal{S} \subset \mathcal{H}$. A projection operator $P_{\mathcal{S}}$ will be any map that, for a given $\mathbf{x} \in \mathcal{H}$ returns a unique element $\mathbf{x}_{\mathcal{S}} \in \mathcal{S}$ such that

$$\|\mathbf{x} - \mathbf{x}_{\mathcal{S}}\| = \inf_{\tilde{\mathbf{x}} \in \mathcal{S}} \|\mathbf{x} - \tilde{\mathbf{x}}\|. \quad (2.37)$$

Again, in certain circumstances, there might not be any $\mathbf{x} \in \mathcal{S}$ for which this property holds with equality. In those cases, one can again relax the requirement on the projection and talk of ϵ projections as those mappings $P_{\mathcal{S}}$ that, for a given $\mathbf{x} \in \mathcal{H}$ returns a unique element $\mathbf{x}_{\mathcal{S}} \in \mathcal{S}$ such that

$$\|\mathbf{x} - \mathbf{x}_{\mathcal{S}}\| \leq \inf_{\tilde{\mathbf{x}} \in \mathcal{S}} \|\mathbf{x} - \tilde{\mathbf{x}}\| + \epsilon. \quad (2.38)$$

2.5.2.1 Sparsity

In Euclidean space, a sparse vector \mathbf{x} is an element of \mathbb{R}^N or \mathbb{C}^N for which $x_i = 0$ for “many” of the indices $i \in [1, 2, \dots, N]$. A popular constraint set is then the set \mathcal{S}_k of all vectors \mathbf{x} in \mathbb{R}^N (or in \mathbb{C}^N) that have no more than $k < N$ non-zero entries. As discussed above, this is a non-convex set and for general Φ finding the projection onto $\Phi\mathcal{S}_k$ is far from trivial, in fact this is a combinatorial search problem in general and we would have to look at each of the k -sparse subspaces in \mathcal{S}_k in turn. However, projecting a vector \mathbf{x} onto \mathcal{S}_k itself is trivial, all one has to do is to identify the k largest (in magnitude) components $|x_i|$ and setting all other components to zero.⁴

2.5.2.2 Block-Sparsity

Like many other structured sparsity constraints, block-sparsity is not easy to deal with directly in the observation domain, that is, it is difficult to project onto $\Phi\mathcal{S}$. Yet again, projection onto \mathcal{S} itself is trivial and is done in a similar way to the sparse case. The only difference now is that we have to calculate the length of \mathbf{x} when restricted to each of the blocks. For example, if the individual blocks are labeled with indices j and if \mathbf{x}_j is the sub-vectors of \mathbf{x} containing only those elements in block j , then we calculate the length of each \mathbf{x}_j and set all blocks to zero apart from those elements that are in the k largest blocks.

⁴ We assume here that we use the norm $\sqrt{\sum x_i^2}$, though other Euclidean norms are treated with equal ease.

2.5.2.3 Tree-Sparsity

Tree sparse models are the other main structured sparsity model of interest. For a given Euclidean vector \mathbf{x} and a pre-defined tree structure, finding the closest sparse vector that respects the tree structure is somewhat more complicated than the projection in the previous two examples. Luckily, there exist fast (yet in the worst case only approximate) algorithms that can be used. In particular, the *condensing sort and select algorithm* (CSSA) is relatively fast as it only requires a computational effort that, in many instances, is of the order of $\mathcal{O}(\mathcal{N} \log \mathcal{N})$ [25].

2.5.2.4 Sparsity in Other Bases

In the three examples above we have used constraint sets in which the signal model assumed sparsity in the canonical basis, that is, we thought about vectors as collections of N numbers and sparsity simply meant we were only allowed a few non-zero numbers. To do this, we have implicitly assumed that we write the vector as a collection of N real or complex numbers, that is, we assumed that we write our vectors in the traditional linear algebra notation as

$$\mathbf{x} = [x_1 x_2 x_3 \cdots x_N]^T. \quad (2.39)$$

Such a notation only specifies a vector if it is made with respect to some basis. Remember, it is best to think of a vector as a point in spaces, say the location of a particular flat in an apartment block, whose location you could specify as 3 floors up, corridor to the left and the third flat on the right, which could be written as [3 3 1]. However, other coordinate systems are possible and would lead to a different set of three numbers. The same is, as said before, also possible for our signal representation. If our signals \mathbf{x} is a vector in Euclidean space, then we can write it as

$$\mathbf{x} = \sum_i a_i \mathbf{x}_i, \quad (2.40)$$

where the a_i are the numbers that specify the location and where the \mathbf{x}_i are a particular basis. For example, for a sampled time series, we typically use what is called the canonical basis, where each basis vector is used to specify the signal intensity at each of the sample time-points. But, if we think about signals as points in space, then we are free to choose more convenient coordinate axis. This is particularly useful as sparsity is a property of collections of numbers (which are often informally called vectors, as we here freely commit too, but which, as we stressed above, are not to be confused with the definition of a vector as a point in space) and not of a point in space. A point in space can only be sparse if we define the appropriate basis in which its representation is sparse and different signals are sparse in different bases. For example, many natural sounds are made up of a small number of harmonic components so that sounds are often fairly sparse using Fourier or other sinusoidal

bases. Images, on the other hand, are often found to be sparse in representations based on wavelet bases.

It is easy to compute the projection of any signal onto any one of the basis vectors. This is done simply as

$$\langle \mathbf{x}, \mathbf{x}_i \rangle / \|\mathbf{x}_i\|, \quad (2.41)$$

where \mathbf{x}_i is the basis vector we project onto. If all basis vectors are orthogonal, then we can also use this approach to project onto the linear subspace spanned by a subset of the basis vectors. Importantly, for orthogonal bases, the optimal choice of the coefficient for one basis vector does not depend on the choice of the other basis vectors. This nice property does no longer hold if the basis is not orthogonal, however. Thus, finding a sparse approximation in any orthogonal basis is simple and can be computed by finding the representation of the signal in the basis, followed by a simple thresholding where only the elements are kept that have the largest magnitude. However, this simple approach is no longer possible in general when the basis is not orthogonal and the non-orthogonality will have to be taken into account.

2.5.2.5 Low Rank Matrices

As discussed above, data that comes in matrix form also allows the specification of powerful non-convex constraints. For matrices that are known to have a low rank we require a projection onto low rank matrices. Again, these projections are easy to calculate. The best approximation of a matrix with a matrix of rank k can be calculated using the Singular Value Decomposition of the matrix followed by thresholding of the singular values, such that only the largest k singular values are retained [26].

2.5.3 Convergence and Stable Recovery

The main question we should ask at this point is, “How good is the Iterative Hard Thresholding algorithm?” that is, if we are given an observation \mathbf{y} , where $\mathbf{y} = \Phi \mathbf{x} + \mathbf{e}$ and if we run the algorithm for a number of iterations, how close will our estimate $\hat{\mathbf{x}}$ be to the true, unknown signal \mathbf{x} ?

An answer to this question is provided by the following theorem taken from [9].

Theorem 3 *Assume an arbitrary signal \mathbf{x} in some Hilbert space \mathcal{H} . Assume you are given an observation \mathbf{y} and a measurement operator Φ and you assume that $\mathbf{y} = \Phi \mathbf{x} + \mathbf{e}$ where \mathbf{e} is an unknown error term. You furthermore know, from prior experience, that \mathbf{x} lies close to a non-convex constraint-set \mathcal{S} . Then, if Φ satisfies the bi-Lipschitz condition on \mathcal{S} with constants $\beta/\alpha < 1.5$, then the Iterative Hard Thresholding algorithm run with a step size μ that satisfies $\beta \leq \frac{1}{\mu} < 1.5\alpha$ and run for*

$$n^* = \left\lceil 2 \frac{\log(\delta \frac{\|\tilde{\mathbf{e}}\|}{\|P_{\mathcal{S}}(\mathbf{x})\|})}{\log(2/(\mu\alpha) - 2)} \right\rceil \quad (2.42)$$

iterations, will calculate a solution $\hat{\mathbf{x}}$ satisfying

$$\|\mathbf{x} - \hat{\mathbf{x}}\| \leq (\sqrt{c} + \delta) \|\tilde{\mathbf{e}}\| + \|P_{\mathcal{S}}(\mathbf{x}) - \mathbf{x}\| \quad (2.43)$$

where $c \leq \frac{4}{3\alpha-2\mu}$, $\tilde{\mathbf{e}} = \Phi(\mathbf{x} - P_{\mathcal{S}}(\mathbf{x})) + \mathbf{e}$ and $\delta > 0$ is arbitrary.

There are several interesting observations to be made here. Let us start by looking at the number of iterations required by the theorem.

$$n^* = \left\lceil 2 \frac{\log(\delta \frac{\|\tilde{\mathbf{e}}\|}{\|P_{\mathcal{S}}(\mathbf{x})\|})}{\log(2/(\mu\alpha) - 2)} \right\rceil, \quad (2.44)$$

which depends on the ratio $\delta \frac{\|\tilde{\mathbf{e}}\|}{\|P_{\mathcal{S}}(\mathbf{x})\|}$, which is a form of signal to noise ratio, however, the signal component here is $P_{\mathcal{S}}(\mathbf{x})$, that is, the projection of the true signal onto the closest element in \mathcal{S} . Similarly, the error term $\tilde{\mathbf{e}} = \Phi(\mathbf{x} - P_{\mathcal{S}}(\mathbf{x})) + \mathbf{e}$, does not only account for the observation noise \mathbf{e} , but also for the distance between the true signal and the model $\mathbf{x} - P_{\mathcal{S}}(\mathbf{x})$. The flexibility in the choice of δ in the theorem allows us furthermore to trade the number of iterations with approximation accuracy. Importantly, δ influences the error bound linearly (halving δ will decrease the error bound dependance on $\tilde{\mathbf{e}}$ with a constant proportion) but it feeds into the required iteration count within the logarithm, so that a linear change in the approximation error only requires a logarithmic increase in computation time.⁵

Let us also look closer at the approximation error itself. This is made up of two error terms, $\|\tilde{\mathbf{e}}\|$ and $\|P_{\mathcal{S}}(\mathbf{x}) - \mathbf{x}\|$. The second one of these terms $\|P_{\mathcal{S}}(\mathbf{x}) - \mathbf{x}\|$, is the distance between the true signal \mathbf{x} and its best approximation with an element from \mathcal{S} . Clearly, all our estimates are from the set \mathcal{S} , so we will be unable to get an approximation that is better than $\|P_{\mathcal{S}}(\mathbf{x}) - \mathbf{x}\|$. The second terms, $\tilde{\mathbf{e}} = \Phi(\mathbf{x} - P_{\mathcal{S}}(\mathbf{x})) + \mathbf{e}$, is made up of two error contributions, the observation noise \mathbf{e} and the error $(\mathbf{x} - P_{\mathcal{S}}(\mathbf{x}))$ again, but this time, after being mapped into the observation space. The fraction of this error we actually have to suffer depends on the number of iterations we use (through δ) and the constant $c \leq \frac{4}{3\alpha-2\mu}$, which is bounded by μ and α . As μ ultimately depends on β , the constant c thus depends on the bi-Lipschitz properties of Φ on \mathcal{S} .

⁵ Note that for $\delta < \frac{\|P_{\mathcal{S}}(\mathbf{x})\|}{\|\tilde{\mathbf{e}}\|}$ and for μ and α as in the theorem, both, the numerator and the denominator in the iteration count are negative numbers, so that a decrease in delta leads to an increase in the required number of iterations. If we were to choose δ such that the numerator becomes positive, we would get a *negative* number of iterations. This has to be interpreted as meaning that we actually don't need to run the algorithm at all, as the associated estimate error is already achieved by the estimate $\hat{\mathbf{x}} = \mathbf{x}^0 = \mathbf{0}$.

2.5.4 The Proof

Proof We now show how the above theorem can be derived using the geometrical ideas developed throughout this chapter. The derivation here follows that in [9]. Our aim is to bound the distance between the true signal \mathbf{x} and its estimate $\hat{\mathbf{x}}^n$ after iteration n . To do this, we start with the trusted triangle inequality to split this vector into two components, the error between \mathbf{x} and $\mathbf{x}_S = P_S(\mathbf{x})$ and the error between $\mathbf{x}_S = P_S(\mathbf{x})$ and $\hat{\mathbf{x}}^n$. This gives the bound

$$\|\mathbf{x} - \hat{\mathbf{x}}^n\|_2 \leq \|\mathbf{x}_S - \hat{\mathbf{x}}^n\|_2 + \|\mathbf{x}_S - \mathbf{x}\|_2. \quad (2.45)$$

We see that the term $\|\mathbf{x}_S - \mathbf{x}\|_2$ is already the last term in our error bound in the theorem and, as discussed before, we can't expect to do better than this, so we are done with this term and instead concentrate on the first term, the length of $\mathbf{x}_S - \hat{\mathbf{x}}^n$ which we will bound further. Our aim here will be to bound $\|\mathbf{x}_S - \hat{\mathbf{x}}^n\|_2$ in terms of the length of the error in the previous iteration plus some extra error terms independent of $\hat{\mathbf{x}}$.

Note that both \mathbf{x}_S and $\hat{\mathbf{x}}^n$ lie within the set \mathcal{S} , so that we can use the bi-Lipschitz condition for both of these vectors, in particular, we have

$$\|\mathbf{x}_S - \hat{\mathbf{x}}^n\|_2^2 \leq \frac{1}{\alpha} \|\Phi(\mathbf{x}_S - \hat{\mathbf{x}}^n)\|_2^2. \quad (2.46)$$

If we now use the definition $\tilde{\mathbf{e}} = \Phi(\mathbf{x} - \mathbf{x}_S) + \mathbf{e}$, we see that $\Phi\mathbf{x}_S - \Phi\hat{\mathbf{x}}^n = \Phi\mathbf{x}_S - \Phi\hat{\mathbf{x}}^n + \Phi\mathbf{x} - \Phi\mathbf{x}_S + \mathbf{e} - \Phi(\mathbf{x} - \mathbf{x}_S) + \mathbf{e} = (\mathbf{y} - \Phi\hat{\mathbf{x}}^n) - \tilde{\mathbf{e}}$. We can thus express the square of the length of $\Phi(\mathbf{x} - \mathbf{x}_S) + \mathbf{e}$ as the sum of the square of the length of $\mathbf{y} - \Phi\hat{\mathbf{x}}^n$ and $\tilde{\mathbf{e}}$.

$$\begin{aligned} \|\Phi(\mathbf{x}_S - \hat{\mathbf{x}}^n)\|_2^2 &= \|\mathbf{y} - \Phi\hat{\mathbf{x}}^n\|_2^2 + \|\tilde{\mathbf{e}}\|_2^2 - 2\langle \tilde{\mathbf{e}}, (\mathbf{y} - \Phi\hat{\mathbf{x}}^n) \rangle \\ &\leq \|\mathbf{y} - \Phi\hat{\mathbf{x}}^n\|_2^2 + \|\tilde{\mathbf{e}}\|_2^2 + \|\tilde{\mathbf{e}}\|_2^2 + \|\mathbf{y} - \Phi\hat{\mathbf{x}}^n\|_2^2 \\ &= 2\|\mathbf{y} - \Phi\hat{\mathbf{x}}^n\|_2^2 + 2\|\tilde{\mathbf{e}}\|_2^2, \end{aligned} \quad (2.47)$$

with the last inequality derived through the inequalities

$$\begin{aligned} -2\langle \tilde{\mathbf{e}}, (\mathbf{y} - \Phi\hat{\mathbf{x}}^n) \rangle &= -\|\tilde{\mathbf{e}} + (\mathbf{y} - \Phi\hat{\mathbf{x}}^n)\|_2^2 + \|\tilde{\mathbf{e}}\|_2^2 + \|\mathbf{y} - \Phi\hat{\mathbf{x}}^n\|_2^2 \\ &\leq \|\tilde{\mathbf{e}}\|_2^2 + \|\mathbf{y} - \Phi\hat{\mathbf{x}}^n\|_2^2. \end{aligned} \quad (2.48)$$

We are now ready to bound the first term in (2.47). This is done using the abbreviation $\mathbf{g}^{n-1} = 2\Phi^T(\mathbf{y} - \Phi\hat{\mathbf{x}}^{n-1})$ and the inequality

$$\|\mathbf{y} - \Phi\hat{\mathbf{x}}^n\|_2^2 \leq (\mu^{-1} - \alpha)\|\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}\|_2^2 + \|\tilde{\mathbf{e}}\|_2^2 + (\beta - \mu^{-1})\|\hat{\mathbf{x}}^n - \hat{\mathbf{x}}^{n-1}\|_2^2, \quad (2.49)$$

which is due to the inequality

$$\begin{aligned}
& \|\mathbf{y} - \Phi \hat{\mathbf{x}}^n\|_2^2 - \|\mathbf{y} - \Phi \hat{\mathbf{x}}^{n-1}\|_2^2 \\
& \leq -\langle (\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}), \mathbf{g}^{n-1} \rangle + \mu^{-1} \|\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}\|_2^2 + (\beta - \mu^{-1}) \|\hat{\mathbf{x}}^n - \hat{\mathbf{x}}^{n-1}\|_2^2 \\
& \leq -\langle (\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}), \mathbf{g}^{n-1} \rangle + \|\Phi(\mathbf{x}_S - \hat{\mathbf{x}}^{n-1})\|_2^2 \\
& \quad + (\mu^{-1} - \alpha) \|\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}\|_2^2 + (\beta - \mu^{-1}) \|\hat{\mathbf{x}}^n - \hat{\mathbf{x}}^{n-1}\|_2^2 \\
& = \|\tilde{\mathbf{e}}\|_2^2 - \|\mathbf{y} - \Phi \hat{\mathbf{x}}^{n-1}\|_2^2 + (\mu^{-1} - \alpha) \|\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}\|_2^2 + (\beta - \mu^{-1}) \|\hat{\mathbf{x}}^n - \hat{\mathbf{x}}^{n-1}\|_2^2.
\end{aligned} \tag{2.50}$$

Here, the second inequality is due to the non-symmetric RIP whilst the first inequality follows from the lemma [9]

Lemma 5 *If $\hat{\mathbf{x}}^n = H_k(\hat{\mathbf{x}}^{n-1} + \mu \Phi^T(\mathbf{y} - \Phi \hat{\mathbf{x}}^{n-1}))$, then*

$$\begin{aligned}
& \|\mathbf{y} - \Phi \hat{\mathbf{x}}^n\|_2^2 - \|\mathbf{y} - \Phi \hat{\mathbf{x}}^{n-1}\|_2^2 \\
& \leq -\langle (\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}), \mathbf{g}^{n-1} \rangle + \mu^{-1} \|\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}\|_2^2 + (\beta - \mu^{-1}) \|\hat{\mathbf{x}}^n - \hat{\mathbf{x}}^{n-1}\|_2^2
\end{aligned} \tag{2.51}$$

We can now combine the inequalities (2.45), (2.46) and (2.49). If $\beta \leq \mu^{-1}$, then we have

$$\|\mathbf{x}_S - \hat{\mathbf{x}}^n\|_2^2 \leq 2 \left(\frac{1}{\mu\alpha} - 1 \right) \|\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}\|_2^2 + \frac{4}{\alpha} \|\tilde{\mathbf{e}}\|_2^2. \tag{2.52}$$

This is exactly the bound we were looking for as now the error between \mathbf{x}_S and the current estimate is smaller than a fraction of the difference between \mathbf{x}_S and the previous estimate (plus some additional noise term). Because we also have the restriction that $2(\frac{1}{\mu\alpha} - 1) < 1$, so that if we replace $\|\mathbf{x}_S - \hat{\mathbf{x}}^{n-1}\|_2^2$ with the bound in terms of $\|\mathbf{x}_S - \hat{\mathbf{x}}^{n-2}\|_2^2$ and then $\|\mathbf{x}_S - \hat{\mathbf{x}}^{n-2}\|_2^2$ with the bound in terms of $\|\mathbf{x}_S - \hat{\mathbf{x}}^{n-3}\|_2^2$ and so on until we end up with a bound in terms of $\|\mathbf{x}_S - \hat{\mathbf{x}}^0\|_2^2$, where we assume that $\hat{\mathbf{x}}^0 = \mathbf{0}$, then we have

$$\|\mathbf{x}_S - \hat{\mathbf{x}}^n\|_2^2 \leq \left(2 \left(\frac{1}{\mu\alpha} - 1 \right) \right)^n \|\mathbf{x}_S\|_2^2 + c \|\tilde{\mathbf{e}}\|_2^2, \tag{2.53}$$

with $c \leq \frac{4}{3\alpha - 2\mu^{-1}}$. These arguments then lead to the claim in the theorem. To see this, we first bound the distance of \mathbf{x} from our estimate at iteration n

$$\begin{aligned}
\|\mathbf{x} - \hat{\mathbf{x}}^n\|_2 & \leq \sqrt{\left(\frac{2}{\mu\alpha} - 2 \right)^n \|\mathbf{x}_S\|_2^2 + c \|\tilde{\mathbf{e}}\|_2^2} + \|\mathbf{x}_S - \mathbf{x}\|_2 \\
& \leq \left(\frac{2}{\mu\alpha} - 2 \right)^{n/2} \|\mathbf{x}_S\|_2 + c^{0.5} \|\tilde{\mathbf{e}}\|_2 + \|\mathbf{x}_S - \mathbf{x}\|_2,
\end{aligned} \tag{2.54}$$

which shows that after $n^* = \left\lceil 2 \frac{\log(\|\tilde{\mathbf{e}}\|_2 / \|\mathbf{x}_S\|_2)}{\log(2/(\mu\alpha) - 2)} \right\rceil$ iterations we have

$$\|\mathbf{x} - \mathbf{x}^{n*}\|_2 \leq (c^{0.5} + 1)\|\tilde{\mathbf{e}}\|_2 + \|\mathbf{x}_S - \mathbf{x}\|_2. \quad (2.55)$$

2.6 Extensions to Non-linear Observation Models

We are here interested in the development of a better understanding of what happens to the compressed sensing recovery problem when a signal is measured with some non-linear system. In particular, the hope is that, if the system is not too non-linear, then recovery should still be possible under similar assumptions to those made in linear compressed sensing. Considering non-linear measurements is not only of academic interest but has important implications for many real-world sampling systems, where measurement system can often not be designed to be perfectly linear. Assume therefore that our measurements are described by a nonlinear mapping $\Phi(\cdot)$ that maps elements of the normed vector spaces \mathcal{H} into the normed vector spaces \mathcal{B} . The observation model is therefore

$$\mathbf{y} = \Phi(\mathbf{x}) + \mathbf{e}, \quad (2.56)$$

where $\mathbf{e} \in \mathcal{B}$ is an unknown but bounded error term.

In order to keep our development as general as possible, we will allow the error between \mathbf{y} and $\Phi(\mathbf{x})$ to be measured with some general norm, that is, whilst we assume that \mathbf{x} is an element from some Hilbert spaces \mathcal{H} , \mathbf{y} will be allowed to lie in a more general Banach space \mathcal{B} with norm $\|\cdot\|_{\mathcal{B}}$. Whilst we have not yet derived a full understanding of this recovery problem, some progress has been made. For example, we could show that the Iterative Hard Thresholding algorithm can also solve quite general non-convex optimisation problems under general Union of Subspaces non-convex constraints, given that a condition similar to the bi-Lipschitz property holds [28].

2.6.1 The Iterative Hard Thresholding Algorithm for Nonlinear Optimisation Under Non-convex Constraints

We start by treating the problem in a quite general framework where we want to optimise a non-convex function $f(\mathbf{x})$ under the constraint that \mathbf{x} lies in a union of subspaces \mathcal{S} . This optimisation will be done using the Iterative Hard Thresholding method and to do this, we will need to specify an update direction. For example, we could assume that $f(\mathbf{x})$ is Fréchet differentiable with respect to \mathbf{x} . The Fréchet derivative is an extension of differentiation to function spaces and is defined as follows. A function is Fréchet differentiable if for each \mathbf{x}_1 there exist a linear functional $D_{\mathbf{x}_1}(\cdot)$ such that

$$\lim_{\mathbf{h} \rightarrow 0} \frac{f(\mathbf{x}_1 + \mathbf{h}) - f(\mathbf{x}_1) - D_{\mathbf{x}_1}(\mathbf{h})}{\|\mathbf{h}\|} = 0. \quad (2.57)$$

So not to have to deal with an abstract linear functional, we will use the Riesz representation theorem [29] which tells us that for each linear functional, we can find an equivalent inner product representation. Thus, we can always find a function ∇ so that we can write the functional $D_{\mathbf{x}_1}(\cdot)$ as an inner product

$$D_{\mathbf{x}_1}(\cdot) = \langle \nabla(\mathbf{x}_1), \cdot \rangle. \quad (2.58)$$

$\nabla(\mathbf{x}_1) \in \mathcal{H}$ is now an element of our function space.

In situations in which the space \mathcal{H} is Euclidean, the Fréchet derivative is the differential of $f(\mathbf{x})$ at \mathbf{x}_1 , in which case $\nabla(\mathbf{x}_1)$ is the gradient and $\langle \cdot, \cdot \rangle$ the Euclidean inner product. To simplify the discussion, we will therefore abuse terminology and call $\nabla(\mathbf{x}_1)$ the gradient even in more general Hilbert space settings.

Once we have specified the update direction $\nabla(\mathbf{x})$, we are in a good position to define an algorithmic strategy to optimise $f(\mathbf{x})$. In particular, the Iterative Hard Thresholding algorithm for non-linear optimisation problems can now be written as

$$\mathbf{x}^{n+1} = P_{\mathcal{S}}(\mathbf{x}^n - (\mu/2)\nabla(\mathbf{x}^n)), \quad (2.59)$$

where $\mathbf{x}^0 = \mathbf{0}$ and μ is a step size parameter chosen to satisfy the condition in theorem below.

2.6.2 Some Theoretic Considerations

Unfortunately, we can not expect the method to work for all constraint sets and for all problems. To specify those problems that can be recovered, we use the following generalisation of the bi-Lipschitz property called the *Restricted Strong Convexity Property* (RSGP) which, to our knowledge, was first introduced in [30]. The *Restricted Strong Convexity Constants* α and β are the largest respectively smallest constants for which

$$\alpha \leq \frac{f(\mathbf{x}_1) - f(\mathbf{x}_2) - \text{Re}\langle \nabla(\mathbf{x}_2), (\mathbf{x}_1 - \mathbf{x}_2) \rangle}{\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \leq \beta, \quad (2.60)$$

holds for all $\mathbf{x}_1, \mathbf{x}_2$ for which $\mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{S} + \mathcal{S}$, where the set $\mathcal{S} + \mathcal{S} = \{\mathbf{x} = \mathbf{x}_a + \mathbf{x}_b : \mathbf{x}_a, \mathbf{x}_b \in \mathcal{S}\}$.

Note that the bi-Lipschitz property is recovered if $f(\mathbf{x}) = \|\mathbf{y} - \Phi\mathbf{x}\|_2^2$, where Φ is linear. Also note that the main result in the next section requires the *Restricted Strong Convexity Property* to hold for all vectors \mathbf{x}_1 and \mathbf{x}_2 , such that $\mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{S} + \mathcal{S} + \mathcal{S}$, where the set $\mathcal{S} + \mathcal{S} + \mathcal{S} = \{\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 : \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathcal{S}\}$.

The performance result now mirrors that derived for the linear compressed sensing setting and states that for $f(\mathbf{x})$ which satisfy the Restricted Strong Convexity Property, the iterative hard thresholding algorithm can be used to find a vector $\mathbf{x} \in \mathcal{S}$ that is close to the true minimiser of $f(\mathbf{x})$. The formal theorem reads as follows.

Theorem 4 Let \mathcal{S} be a union of subspaces. Given the optimisation problem $f(\mathbf{x})$, where $f(\mathbf{x})$ is a positive function that satisfies the Restricted Strict Convexity Property

$$\alpha \leq \frac{f(\mathbf{x}_1)f - f(\mathbf{x}_2) - \text{Re}\langle \nabla f(\mathbf{x}_2), (\mathbf{x}_1 - \mathbf{x}_2) \rangle}{\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \leq \beta, \quad (2.61)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{H}$ for which $\mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{S} + \mathcal{S} + \mathcal{S}$ with constants $\beta \leq \frac{1}{\mu} \leq \frac{4}{3}\alpha$, then, after

$$n^* = 2 \frac{\ln \left(\delta \frac{f(\mathbf{x}_S)}{\|\mathbf{x}_S\|} \right)}{\ln 4(1 - \mu\alpha)}, \quad (2.62)$$

iterations, the Iterative Hard Thresholding Algorithm calculates a solution \mathbf{x}^{n^*} that satisfies

$$\|\mathbf{x}^{n^*} - \mathbf{x}\| \leq \left(2\sqrt{\frac{\mu}{1-c}} + \delta \right) f(\mathbf{x}_S) + \|\mathbf{x} - \mathbf{x}_S\| + \sqrt{\frac{2}{1-c}}\epsilon, \quad (2.63)$$

where $\mathbf{x}_S = \arg\min_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x})$.

2.6.3 Proof of Theorem 4

Proof The proof, first presented in [28], is based around a subspace Γ , which is defined as the sum of no more than three subspaces of \mathcal{S} , such that $\mathbf{x}_S, \mathbf{x}^n, \mathbf{x}^{n+1} \in \Gamma$. We also define P_Γ to be the orthogonal projection onto the subspace Γ and use the short hand notation $\mathbf{a}_\Gamma^n = P_\Gamma \mathbf{a}^n$ and $P_\Gamma \nabla f(\mathbf{x}^n) = \nabla_\Gamma f(\mathbf{x}^n)$.

As in [28], we start by establishing a few basic equalities, which follow from the fact that for all orthogonal projections P , we have $\langle P\mathbf{x}_1, P\mathbf{x}_2 \rangle = \langle \mathbf{x}_1, P\mathbf{x}_2 \rangle$. As both \mathbf{x}_S and \mathbf{x}^n lie in Γ we have

$$\begin{aligned} \text{Re}\langle \nabla_\Gamma f(\mathbf{x}^n), (\mathbf{x}_S - \mathbf{x}^n) \rangle &= \text{Re}\langle P_\Gamma \nabla f(\mathbf{x}^n), (\mathbf{x}_S - \mathbf{x}^n) \rangle \\ &= \text{Re}\langle \nabla f(\mathbf{x}^n), P_\Gamma (\mathbf{x}_S - \mathbf{x}^n) \rangle \\ &= \text{Re}\langle \nabla f(\mathbf{x}^n), (\mathbf{x}_S - \mathbf{x}^n) \rangle \end{aligned} \quad (2.64)$$

and

$$\begin{aligned} \|\nabla_\Gamma f(\mathbf{x}^n)\|^2 &= \langle \nabla_\Gamma f(\mathbf{x}^n), \nabla_\Gamma f(\mathbf{x}^n) \rangle = \langle P_\Gamma \nabla f(\mathbf{x}^n), P_\Gamma \nabla f(\mathbf{x}^n) \rangle \\ &= \langle \nabla f(\mathbf{x}^n), P_\Gamma^* P_\Gamma \nabla f(\mathbf{x}^n) \rangle \\ &= \langle \nabla f(\mathbf{x}^n), \nabla_\Gamma f(\mathbf{x}^n) \rangle, \end{aligned} \quad (2.65)$$

We will also make use of the following lemma.

Lemma 6 *Under the assumptions of the theorem,*

$$\|\frac{\mu}{2}\nabla_{\Gamma}(\mathbf{x}^n)\|^2 - \mu f(\mathbf{x}^n) \leq 0. \quad (2.66)$$

Proof The lemma can be established as follows. Using the *Restricted Strict Convexity Property* we have

$$\begin{aligned} \|\frac{\mu}{2}\nabla_{\Gamma}(\mathbf{x}^n)\|^2 &= -\frac{\mu}{2}\text{Re}\langle \nabla(\mathbf{x}^n), -\frac{\mu}{2}\nabla_{\Gamma}(\mathbf{x}^n) \rangle \\ &\leq \frac{\mu}{2}\beta\|\frac{\mu}{2}\nabla_{\Gamma}(\mathbf{x}^n)\|^2 + \frac{\mu}{2}f(\mathbf{x}^n) - \frac{\mu}{2}f(\mathbf{x}^n - \frac{\mu}{2}\nabla_{\Gamma}(\mathbf{x}^n)) \\ &\leq \frac{\mu}{2}\beta\|\frac{\mu}{2}\nabla_{\Gamma}(\mathbf{x}^n)\|^2 + \frac{\mu}{2}f(\mathbf{x}^n). \end{aligned} \quad (2.67)$$

Thus

$$(2 - \mu\beta)\|\frac{\mu}{2}\nabla_{\Gamma}(\mathbf{x}^n)\|^2 \leq \mu f(\mathbf{x}^n), \quad (2.68)$$

which is the desired result as $\mu\beta \leq 1$ by assumption.

The main point of the theorem is to bound the distance between the current estimate \mathbf{x}^{n+1} and the optimal estimate $\mathbf{x}_{\mathcal{S}}$. To derive this bound, let us write $\mathbf{a}_{\Gamma}^n = \mathbf{x}_{\Gamma}^n - \mu/2\nabla_{\Gamma}(\mathbf{x}^n)$. We then note that \mathbf{x}^{n+1} is, up to ϵ the closest element in \mathcal{S} to \mathbf{a}_{Γ}^n , so that

$$\begin{aligned} \|\mathbf{x}^{n+1} - \mathbf{x}_{\mathcal{S}}\|^2 &\leq \left(\|\mathbf{x}^{n+1} - \mathbf{a}_{\Gamma}^n\| + \|\mathbf{a}_{\Gamma}^n - \mathbf{x}_{\mathcal{S}}\| \right)^2 \\ &\leq 4\|\mathbf{a}_{\Gamma}^n - \mathbf{x}_{\mathcal{S}}\|^2 + 2\epsilon \\ &= 4\|\mathbf{x}^n - (\mu/2)\nabla_{\Gamma}(\mathbf{x}^n) - \mathbf{x}_{\mathcal{S}}\|^2 + 2\epsilon \\ &= 4\|(\mu/2)\nabla_{\Gamma}(\mathbf{x}^n) + (\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n)\|^2 + 2\epsilon \\ &= \mu^2\|\nabla_{\Gamma}(\mathbf{x}^n)\|^2 + 4\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n\|^2 + 4\mu\text{Re}\langle \nabla_{\Gamma}(\mathbf{x}^n), (\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n) \rangle + 2\epsilon \\ &= \mu^2\|\nabla_{\Gamma}(\mathbf{x}^n)\|^2 + 4\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n\|^2 + 4\mu\text{Re}\langle \nabla(\mathbf{x}^n), (\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n) \rangle + 2\epsilon \\ &\leq 4\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n\|^2 + \mu^2\|\nabla_{\Gamma}(\mathbf{x}^n)\|^2 \\ &\quad + 4\mu[-\alpha\|\mathbf{x}^n - \mathbf{x}_{\mathcal{S}}\|^2 + f(\mathbf{x}_{\mathcal{S}}) - f(\mathbf{x}^n)] + 2\epsilon \\ &= 4(1 - \mu\alpha)\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n\|^2 + 4\mu f(\mathbf{x}_{\mathcal{S}}) + 2\epsilon \\ &\quad + 4[\|(\mu/2)\nabla_{\Gamma}(\mathbf{x}^n)\|^2 - \mu f(\mathbf{x}^n)] \\ &\leq 4(1 - \mu\alpha)\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n\|^2 + 4\mu f(\mathbf{x}_{\mathcal{S}}) + 2\epsilon. \end{aligned} \quad (2.69)$$

Here, the second to last inequality is the RSCP and the last inequality is due to lemma 6.

We could thus bound the difference between the current estimate and $\mathbf{x}_{\mathcal{S}}$ in terms of the previous estimate and $\mathbf{x}_{\mathcal{S}}$ plus some error term.

$$\|\mathbf{x}^{n+1} - \mathbf{x}_S\|^2 \leq 4(1 - \mu\alpha)\|\mathbf{x}_S - \mathbf{x}^n\|^2 + 4\mu f(\mathbf{x}_S) + 2\epsilon. \quad (2.70)$$

If we define the constant $c = 4(1 - \mu\alpha)$ and iterate the above expression (i.e. use the same bound to bound the last error with the one before that and so on), then we see that

$$\|\mathbf{x}^n - \mathbf{x}_S\|^2 \leq c^n \|\mathbf{x}_S\|^2 + \frac{4\mu}{1-c} f(\mathbf{x}_S) + \frac{2}{1-c} \epsilon, \quad (2.71)$$

where the constant $\frac{1}{1-c}$ in front of the error term is a bound of the geometric series $\sum_n c^n$ due to the iterative procedure. Importantly, if $\frac{1}{\mu} < \frac{4}{3}\alpha$ we have $c = 4(1 - \mu\alpha) < 1$, so that c^n decreases with n . Taking the square root on both sides and noting that for positive a and b , $\sqrt{a^2 + b^2} \leq a + b$, we then have

$$\|\mathbf{x}^n - \mathbf{x}_S\| \leq c^{n/2} \|\mathbf{x}_S\| + 2\sqrt{\frac{\mu}{1-c}} f(\mathbf{x}_S) + \sqrt{\frac{2}{1-c}} \epsilon. \quad (2.72)$$

The theorem now follows using the triangle inequality

$$\begin{aligned} \|\mathbf{x}^n - \mathbf{x}\| &\leq \|\mathbf{x}^n - \mathbf{x}_S\| + \|\mathbf{x} - \mathbf{x}_S\| \\ &\leq c^{n/2} \|\mathbf{x}_S\| + 2\sqrt{\frac{\mu}{1-c}} f(\mathbf{x}_S) + \sqrt{\frac{2}{1-c}} \epsilon \\ &\quad + \|\mathbf{x} - \mathbf{x}_S\| \end{aligned} \quad (2.73)$$

and the iteration count is determined by setting

$$c^{n/2} \|\mathbf{x}_S\| \leq \delta(\mathbf{x}_S). \quad (2.74)$$

so that after

$$n = 2 \frac{\ln \left(\delta \frac{f(\mathbf{x}_S)}{\|\mathbf{x}_S\|} \right)}{\ln c}, \quad (2.75)$$

iterations

$$\|\mathbf{x}^n - \mathbf{x}\| \leq \left(2\sqrt{\frac{\mu}{1-c}} + \delta \right) f(\mathbf{x}_S) + \|\mathbf{x} - \mathbf{x}_S\| + \sqrt{\frac{2}{1-c}} \epsilon. \quad (2.76)$$

2.6.4 An Important Caveat

Whilst this is an important result that shows how the Iterative Hard Thresholding algorithm can be used for many non-linear optimisation problems, it does not directly translate into a simple application to Compressed Sensing under non-linear observations. It seems tempting to use $f(\mathbf{x}) = \|\mathbf{y} - \Phi(\mathbf{x})\|_B^2$, where $\|\cdot\|_B$ is some Banach

space norm and where $\Phi(\cdot)$ is some non-linear function. If this $f(\mathbf{x})$ would satisfy the Restricted Strict Convexity property, then we could clearly use the algorithm to solve the non-linear compressed sensing problem in which we are given noisy observations

$$\mathbf{y} = \Phi(\mathbf{x}) + \mathbf{e}. \quad (2.77)$$

Unfortunately, whilst properties such as the restricted strict convexity property hold for certain non-linear functions such as those encountered in certain logistic regression problems [31], it is far from clear under which conditions on $f(\mathbf{x}) = \|\mathbf{y} - \Phi(\mathbf{x})\|_B^2$ similar properties would hold.

Indeed, the following lemma shows that such a condition cannot be fulfilled in general for Hilbert spaces.

Lemma 7 *Assume \mathcal{B} is a Hilbert space and assume $f(\mathbf{x})$ is convex on $\mathcal{S} + \mathcal{S}$ for all \mathbf{y} (i.e. it Satisfies the Restricted Strict Convexity Property), then Φ is affine on all subspaces of $\mathcal{S} + \mathcal{S}$.*

Proof The proof is by contradiction. Assume Φ is not affine on any subspace of $\mathcal{S} + \mathcal{S}$. Thus, there is a subspace $\mathcal{S} = \mathcal{S}_i + \mathcal{S}_j$, and $\mathbf{x}_n \in \mathcal{S}$, such that for $\mathbf{x} = \sum_n \lambda_n \mathbf{x}_n$, where $\sum_n \lambda_n = 1$ and $0 \leq \lambda_n$, we have $\sum_n \Phi(\mathbf{x}_n) - \Phi(\mathbf{x}) \neq \mathbf{0}$. Now by assumption of strong convexity on \mathcal{S} , we have (using $\mathbf{y}_n = \Phi(\mathbf{x}_n)$ and $-\bar{\mathbf{y}} = \mathbf{x}$)

$$\begin{aligned} \mathbf{0} &\leq \sum_n \lambda_n \|\mathbf{y} - \Phi(\mathbf{x}_n)\|^2 - \|\mathbf{y} - \Phi(\mathbf{x})\|^2 = \sum_n \lambda_n \|\mathbf{y} - \mathbf{y}_n\|^2 - \|\mathbf{y} - \bar{\mathbf{y}}\|^2 \\ &= 2\langle \mathbf{y}, \bar{\mathbf{y}} - \sum_n \lambda_n \mathbf{y}_n \rangle + \sum_n \lambda_n \|\mathbf{y}_n\|^2 - \|\bar{\mathbf{y}}\|^2. \end{aligned} \quad (2.78)$$

where the inequality is due to the assumption of convexity. But the above inequality cannot hold for all \mathbf{y} (it fails for example for a multiple of $-(\bar{\mathbf{y}} - \sum_n \lambda_n \mathbf{y}_n)$). Thus Φ needs to be affine on the linear subsets of $\mathcal{S} + \mathcal{S}$.

2.6.5 An Alternative Approach

Fortunately, the above result does not prevent the existence of Φ for which $\|\mathbf{y} - \Phi(\mathbf{x})\|_B^2$ has the Restricted Strict Convexity Property for at least some \mathbf{y} . Alternatively, one could also envisage an approach where the linearisation error is dealt with by considering a local linear approximation to $\Phi(\mathbf{x})$ of the form $\Phi(\mathbf{x}) = \Phi_{\mathbf{x}^*} \mathbf{x} + g_{\mathbf{x}^*}(\mathbf{x})$, where $\Phi_{\mathbf{x}^*}$ is linear and satisfies a form of the linear bi-Lipschitz condition. In this case, one would need to bound the error $g_{\mathbf{x}^*}(\mathbf{x})$. If this can indeed be done, then similar recovery results to those available in the linear case would seem feasible also for non-linear problems.

For example, we have [32]

Theorem 5 Assume that $\mathbf{y} = \Phi(\mathbf{x}) + \mathbf{e}$ and that $\Phi_{\mathbf{x}^*}$ is a linearisation of $\Phi(\cdot)$ at \mathbf{x}^* (i.e. the Jacobian of $\Phi(\mathbf{x})$ evaluated at \mathbf{x}^*) so that the Iterative Hard Thresholding algorithm uses the iteration $\mathbf{x}^{n+1} = P_{\mathcal{S}}(\mathbf{x}^n + \Phi_{\mathbf{x}^n}^*(\mathbf{y} - \Phi(\mathbf{x}^n)))$. Assume that $\Phi_{\mathbf{x}^*}$ satisfies RIP

$$\alpha \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \leq \alpha \|\Phi_{\mathbf{x}^*}(\mathbf{x}_1 - \mathbf{x}_2)\|_2^2 \leq \beta \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \quad (2.79)$$

for all $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}^* \in \mathcal{S}$. Define $\epsilon_{\mathcal{S}} = \sup_{\mathbf{x} \in \mathcal{S}} \|\mathbf{y} - \Phi_{\mathbf{x}} \mathbf{x}_{\mathcal{S}}\|_2$ and let $\mathbf{e}_{\mathcal{S}}^n = \mathbf{y} - \Phi_{\mathbf{x}^n} \mathbf{x}_{\mathcal{S}}$, then after

$$k^* = \left\lceil 2 \frac{\ln(\delta \frac{\|\mathbf{e}_{\mathcal{S}}\|}{\|\mathbf{x}_{\mathcal{S}}\|})}{\ln(2/(\mu\alpha) - 2)} \right\rceil \quad (2.80)$$

iterations we have

$$\|\mathbf{x} - \mathbf{x}^{k^*}\| \leq (c^{0.5} + \delta) \|\mathbf{e}_{\mathcal{S}}\| + \|\mathbf{x}_{\mathcal{S}} - \mathbf{x}\| + \sqrt{\frac{c\epsilon}{2\mu}}, \quad (2.81)$$

Proof The proof is similar to the linear case, with a few minor variations. In particular, we introduce the error term $\mathbf{e}_{\mathcal{S}}^n = \mathbf{y} - \Phi(\mathbf{x}^n) - \Phi_{\mathbf{x}^n}(\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n)$ to bound $\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^{n+1}\|^2$ using the expression

$$\begin{aligned} & \|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^{n+1}\|^2 \\ & \leq \frac{1}{\alpha} \|\Phi_{\mathbf{x}^n}(\mathbf{x}_{\mathcal{S}} - \mathbf{x}^{n+1})\|^2 \\ & = \frac{1}{\alpha} \|\mathbf{y} - \Phi(\mathbf{x}^n) - \Phi_{\mathbf{x}^n}(\mathbf{x}^{n+1} - \mathbf{x}^n) - (\mathbf{y} - \Phi(\mathbf{x}^n) - \Phi_{\mathbf{x}^n}(\mathbf{x}_{\mathcal{S}} - \mathbf{x}^n))\|^2 \\ & = \frac{1}{\alpha} \left(\|\mathbf{y} - \Phi(\mathbf{x}^n) - \Phi_{\mathbf{x}^n}(\mathbf{x}^{n+1} - \mathbf{x}^n)\|^2 + \|\mathbf{e}_{\mathcal{S}}^n\|^2 \right. \\ & \quad \left. - 2\langle \mathbf{e}_{\mathcal{S}}^n, (\mathbf{y} - \Phi(\mathbf{x}^n) - \Phi_{\mathbf{x}^n}(\mathbf{x}^{n+1} - \mathbf{x}^n)) \rangle \right) \\ & \leq \frac{2}{\alpha} \|\mathbf{y} - \Phi(\mathbf{x}^n) - \Phi_{\mathbf{x}^n}(\mathbf{x}^{n+1} - \mathbf{x}^n)\|^2 + \frac{2}{\alpha} \|\mathbf{e}_{\mathcal{S}}^n\|^2. \end{aligned} \quad (2.82)$$

Again using similar ideas to those of the linear proof, we use $\mathbf{g} = 2\Phi_{\mathbf{x}^n}^*(\mathbf{y} - \Phi(\mathbf{x}^n))$ and expand

$$\begin{aligned} & \|\mathbf{y} - \Phi(\mathbf{x}^n) - \Phi_{\mathbf{x}^n}(\mathbf{x}^{n+1} - \mathbf{x}^n)\|^2 - \|\mathbf{y} - \Phi(\mathbf{x}^n)\|^2 \\ & = -\langle (\mathbf{x}^{n+1} - \mathbf{x}^n), \mathbf{g} \rangle + \|\Phi_{\mathbf{x}^n}(\mathbf{x}^{n+1} - \mathbf{x}^n)\|^2 \\ & \leq -\frac{2}{\mu} \langle (\mathbf{x}^{n+1} - \mathbf{x}^n), \frac{\mu}{2} \mathbf{g} \rangle + \frac{1}{\mu} \|\mathbf{x}^{n+1} - \mathbf{x}^n\|^2 \\ & = \frac{1}{\mu} \left[\|\mathbf{x}^{n+1} - \mathbf{x}^n - \frac{\mu}{2} \mathbf{g}\|^2 - \frac{\mu}{2} \|\mathbf{g}\|^2 \right] \\ & \leq \frac{1}{\mu} \left[\inf_{\mathbf{x} \in \mathcal{S}} \|\mathbf{x} - \mathbf{x}^n - \frac{\mu}{2} \mathbf{g}\|^2 + \epsilon - \frac{\mu}{2} \|\mathbf{g}\|^2 \right] \end{aligned}$$

$$\begin{aligned}
&= \inf_{\mathbf{x} \in \mathcal{S}} \left[-\langle (\mathbf{x} - \mathbf{x}^n), \mathbf{g} \rangle + \frac{1}{\mu} \|\mathbf{x} - \mathbf{x}^n\|^2 + \frac{\epsilon}{\mu} \right] \\
&\leq -\langle (\mathbf{x}_S - \mathbf{x}^n), \mathbf{g} \rangle + \frac{1}{\mu} \|\mathbf{x}_S - \mathbf{x}^n\|^2 + \frac{\epsilon}{\mu} \\
&= -2\langle (\mathbf{x}_S - \mathbf{x}^n), \Phi_{\mathbf{x}^n}^*(\mathbf{y} - \Phi(\mathbf{x}^n)) \rangle + \frac{1}{\mu} \|\mathbf{x}_S - \mathbf{x}^n\|^2 + \frac{\epsilon}{\mu} \\
&= -2\langle (\mathbf{x}_S - \mathbf{x}^n), \Phi_{\mathbf{x}^n}^*(\mathbf{y} - \Phi(\mathbf{x}^n)) \rangle + \alpha \|\mathbf{x}_S - \mathbf{x}^n\|^2 \\
&\quad + \left(\frac{1}{\mu} - \alpha\right) \|\mathbf{x}_S - \mathbf{x}^n\|^2 + \frac{\epsilon}{\mu} \\
&\leq -2\langle (\mathbf{x}_S - \mathbf{x}^n), \Phi_{\mathbf{x}^n}^*(\mathbf{y} - \Phi(\mathbf{x}^n)) \rangle + \|\Phi_{\mathbf{x}^n}(\mathbf{x}_S - \mathbf{x}^n)\|^2 \\
&\quad + \left(\frac{1}{\mu} - \alpha\right) \|\mathbf{x}_S - \mathbf{x}^n\|^2 + \frac{\epsilon}{\mu} \\
&= \|\mathbf{y} - \Phi(\mathbf{x}^n) - \Phi_{\mathbf{x}^n}(\mathbf{x}_S - \mathbf{x}^n)\|^2 - \|\mathbf{y} - \Phi(\mathbf{x}^n)\|^2 \\
&\quad + \left(\frac{1}{\mu} - \alpha\right) \|\mathbf{x}_S - \mathbf{x}^n\|^2 + \frac{\epsilon}{\mu} \\
&= \|\mathbf{e}_S^n\|^2 - \|\mathbf{y} - \Phi(\mathbf{x}^n)\|^2 + \left(\frac{1}{\mu} - \alpha\right) \|\mathbf{x}_S - \mathbf{x}^n\|^2 + \frac{\epsilon}{\mu}, \tag{2.83}
\end{aligned}$$

where the first inequality is due to the bi-Lipschitz property and the choice of $\beta \leq \frac{1}{\mu}$ and the second inequality is the definition of $\mathbf{x}^{n+1} = P_S^c(\mathbf{x}^n + \frac{\mu}{2}\mathbf{g})$. The third inequality is due to the fact that $\mathbf{x}_S \in \mathcal{S}$ whilst the last inequality is the bi-Lipschitz property again.

This gives the bound

$$\|\mathbf{y} - \Phi(\mathbf{x}^n) - \Phi_{\mathbf{x}^n}(\mathbf{x}^{n+1} - \mathbf{x}^n)\|^2 \leq \left(\frac{1}{\mu} - \alpha\right) \|\mathbf{x}_S - \mathbf{x}^n\|^2 + \|\mathbf{e}_S^n\|^2 + \frac{\epsilon}{\mu}, \tag{2.84}$$

so that

$$\|\mathbf{x}_S - \mathbf{x}^{n+1}\|^2 \leq 2 \left(\frac{1}{\mu\alpha} - 1 \right) \|\mathbf{x}_S - \mathbf{x}^n\|^2 + \frac{4}{\alpha} \|\mathbf{e}_S^n\|^2 + \frac{2\epsilon}{\mu\alpha}. \tag{2.85}$$

This again expresses the distance of \mathbf{x}^{n+1} from \mathbf{x}_S in terms of the distance of the estimate \mathbf{x}^n calculated in the previous iteration.

The condition of the theorem $(2(\frac{1}{\mu\alpha} - 1) < 1)$ again allows us to iterate this expression so that

$$\|\mathbf{x}_S - \mathbf{x}^k\|^2 \leq \left(2 \left(\frac{1}{\mu\alpha} - 1 \right) \right)^k \|\mathbf{x}_S\|^2 + c\epsilon_S + \frac{c\epsilon}{2\mu}, \tag{2.86}$$

where $c \leq \frac{4}{3\alpha - 2\frac{1}{\mu}}$.

In conclusion, using the square root of (2.86), we have thus shown that

$$\begin{aligned}
\|\mathbf{x} - \mathbf{x}^k\| &\leq \sqrt{\hat{c}^k \|\mathbf{x}_S\|^2 + c \|\mathbf{e}_S\|^2 + \frac{c\epsilon}{2\mu}} + \|\mathbf{x}_S - \mathbf{x}\| \\
&\leq \hat{c}^{k/2} \|\mathbf{x}_S\| + c^{0.5} \|\mathbf{e}_S\| + \sqrt{\frac{c\epsilon}{2\mu}} + \|\mathbf{x}_S - \mathbf{x}\|,
\end{aligned}$$

where $\hat{c} = \frac{2}{\mu\alpha} - 2$. The theorem directly follows from this.

2.7 Conclusions

The use of geometrical ideas in signal processing can often lead to new insights and solutions. This is particularly true in the field of sampling. Sampling, the transition from the continuous world of physical phenomena to the discretised world of concrete computation, fundamentally relies on approximations and these, in turn, must be based on prior assumptions that incorporate models of the physical world. Geometric descriptions of these models have over the years proven exceedingly useful, culminating in the recent ascend of compressed sensing. Here, geometric considerations have led to significant advances in signal reconstruction and interpretation, particularly in settings, where complex prior constraints are to be imposed.

In this chapter, we have provided an introductory tour of some of the underlying mathematical concepts that make up modern geometry, focussing especially on those aspects relevant to modern sampling theory. Building on this mathematical framework, several aspects of sampling, and in particular compressed sensing, have been explored. For example, we have shown how geometric ideas can be used to extend the sparse signal models traditionally used in compressed sensing to much more general union of subspaces models, which are much more widely applicable. Geometric interpretations were further shown to be fundamental in the development and understanding of algorithmic signal reconstruction strategies that try to solve optimisation problems that are constraint by these models. But not only do these ideas allow us to construct efficient algorithms, geometric insights are also likely to play a major role in future developments, such as those discussed here in the context of non-linear sampling.

Acknowledgments This work was supported in part by the UKs Engineering and Physical Science Research Council grants EP/J005444/1 and D000246/1 and a Research Fellowship from the School of Mathematics at the University of Southampton.

References

1. Nyquist H (1928) Certain topics in telegraph transmission theory. Trans AIEE 47:617–644
2. Shannon CA, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana

3. Donoho DL (2006) For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Commun Pure Appl Math* 59(6):797–829
4. Candès E, Romberg J (2006) Quantitative robust uncertainty principles and optimally sparse decompositions. *Found Comput Math* 6(2):227–254
5. Candès E, Romberg J, Tao T (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans Inform Theory* 52(2):489–509
6. Candès E, Romberg J, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. *Commun Pure Appl Math* 59(8):1207–1223
7. Abernethy J, Bach F, Evgeniou T, Vert J-P (2006) Low-rank matrix factorization with attributes. [arxiv:0611124v1](https://arxiv.org/abs/0611124)
8. Recht B, Fazel M, Parrilo PA (2009) Guaranteed minimum-rank solution of linear matrix equations via nuclear norm minimization. *Found Comput Math* 9:717–772
9. Blumensath T (2010) Sampling and reconstructing signals from a union of linear subspaces. *IEEE Trans Inf Theory* 57(7):4660–4671
10. Rudin W (1976) Principles of mathematical analysis, 3rd edn. McGraw-Hill Higher Education, New York
11. Conway JB (1990) A course in functional analysis. Graduate texts in mathematics, 2nd edn. Springer, Berlin
12. Unser M (2000) Sampling-50 years after Shannon. *Proc IEEE* 88(4):569–587
13. Landau HJ (1967) Necessary density conditions for sampling and interpolation of certain entire functions. *Acta Math* 117:37–52
14. Mishali M, Eldar YC (2009) Blind multi-band signal reconstruction: compressed sensing for analog signals. *IEEE Trans Signal Process* 57(3):993–1009
15. Vetterli M, Marziliano P, Blu T (2002) Sampling signals with finite rate of innovation. *IEEE Trans Signal Process* 50(6):1417–1428
16. Xu W, Hassibi B (to appear) Compressive sensing over the grassmann manifold: a unified geometric framework. *IEEE Trans Inf Theory*
17. Cands EJ (2006) The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences, Serie I*(346):589–592
18. Donoho DL (2006) High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput Geom* 35(4):617–652
19. Gruenbaum B (1968) Grassmann angles of convex polytopes. *Acta Math* 121:293–302
20. Gruenbaum B (2003) Convex polytopes. Graduate texts in mathematics, vol 221, 2nd edn. Springer-Verlag, New York
21. Blumensath T, Davies M (2008) Iterative thresholding for sparse approximations. *J Fourier Anal Appl* 14(5):629–654
22. Blumensath T, Davies M (2009) Iterative hard thresholding for compressed sensing. *Appl Comput Harmon Anal* 27(3):265–274
23. Qui K, Dogandzic A (2010) ECME thresholding methods for sparse signal reconstruction. [arXiv:1004.4880v3](https://arxiv.org/abs/1004.4880v3)
24. Cevher V, (2011) On accelerated hard thresholding methods for sparse approximation. EPFL Technical Report, February 17, 2011
25. Baraniuk RG (1999) Optimal tree approximation with wavelets. *Wavelet Appl Sig Image Process VII* 3813:196–207
26. Goldfarb D, Ma S (2010) Convergence of fixed point continuation algorithms for matrix rank minimization. [arXiv:09063499v3](https://arxiv.org/abs/09063499v3)
27. Needell D, Tropp JA (2008) CoSaMP: iterative signal recovery from incomplete and inaccurate samples. *Appl Comput Harmon Anal* 26(3):301–321
28. Blumensath T (2010) Compressed sensing with nonlinear observations. Technical report. <http://eprints.soton.ac.uk/164753>
29. Rudin W (1966) Real and complex analysis, McGraw-Hill, New York
30. Negahban S, Ravikumar P, Wainwright MJ, Yu B (2009) A unified framework for the analysis of regularized M-estimators. *Advances in neural information processing systems*, Vancouver, Canada

31. Bahmani S, Raj B, Boufounos P (2012) Greedy sparsity-constrained optimization. arXiv:1203.5483v1
32. Blumensath T (2012) Compressed sensing with nonlinear observations and related nonlinear optimisation problems. arXiv:1205.1650v1

Compressed Sensing & Sparse Filtering

Carmi, A.Y.; Mihaylova, L.S.; Godsill, S.J. (Eds.)

2014, XII, 502 p. 135 illus., Hardcover

ISBN: 978-3-642-38397-7