University of the Witwatersrand

Doctoral Thesis

A Dynamical Trajectory-Based Method for Sparse Recovery

Author:
Matthews M. Sejeso

Supervisor:
Prof. Montaz Ali

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

to the

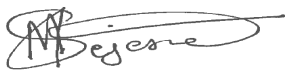
Faculty of Science
School of Computer Science and Applied Mathematics

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

October 2022

Declaration of authorship

I, *Matthews M.* Sejeso, hereby declare that this thesis titled, '*A Dynamical Trajectory-Based Method for Sparse Recovery*' and the work presented in it are my own. Any work done by others or by myself previously has been acknowledged and referenced accordingly. This thesis is submitted for the degree of *Doctor of Philosophy* in Computational and Applied Mathematics at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any other degree or examination at any other institute.



Matthews M. Sejeso

18 October 2022

University of the Witwatersrand

Johannesburg

UNIVERSITY OF THE WITWATERSRAND

Abstract

Faculty of Science

School of Computer Science and Applied Mathematics

Doctor of Philosophy

A Dynamical Trajectory-Based Method for Sparse Recovery

by Matthews M. Sejeso

Many applications in emerging technologies call for efficient sensing systems to acquire and process high-resolution signals. This task is impractical for the traditional sampling scheme, the Nyquist-Shannon sampling theory. Compressed sensing was developed as the new signal acquisition scheme to address this issue. The compressed sensing theory asserts that linear encoded measurements can be used to simultaneous acquire and compress the signal. The technique requires much less computational resources than traditional sampling schemes.

In compressed sensing, an optimization problem comprised of a data fidelity term and a nonlinear sparsity enforcing term are used to recover sparse vectors from a few linear measurements. In most applications, the problems are large-scale and characterized by high-dimensional decision variables. They also require the real-time processing of data. Specialised optimization algorithms have been used to solve sparse recovery problems in compressed sensing. However, these algorithms tend to be slow and computationally intensive to perform real-time recovery. On the contrary, continuous-time dynamical systems have recently gained attention as efficient optimization problem solvers. They have the potential to yield significantly speed and power improvements over their discrete counterpart.

The focus of this thesis is to understand the type of continuous-time dynamical systems that can be used to solve sparse recovery problems and analyse their performance mathematically. It is essential to understand the dynamical system's behaviour before it can be used to solve optimization problems. First, we present a general dynamical system modelled by the subgradient of nonsmooth objective function coupled with the sparse promoting activation function. Convergence analysis of this gradient-like differential inclusion is done using the recently developed nonsmooth Łojasiewicz inequality. The trajectories of the dynamical system are shown to have finite lengths and are globally convergences to equilibrium points.

The equilibrium points of the dynamical system correspond to the critical point of the sparse optimization problem. An estimate of convergence rate, which depends on the Łojasiewicz exponent, is obtained

Second, the Bregman integrated dynamical system for solving the ℓ_1-minimization problem is presented. The dynamical system integrates the Bregman distance in the design, resulting in an improved convergence rate. The proposed dynamical system fits well within the framework differential inclusion presented and analysed early. Thus the Bregman integrated dynamical system is well suited to solve the ℓ_1-minimizer problem. We show that the proposed dynamical system takes an efficient path towards the optimal solution and recovers the expected support set of the sparse solution. The Bregman integrated dynamical system yields the exponential convergence rate, which significantly improves the convergence of the previously proposed dynamical system of Locally Competitive Algorithm.

Computational results are presented to support the developed theory and the good performance of the proposed dynamical system. Several comparative experiments on sparse recovery problems demonstrate that the proposed dynamical system approach is efficient and effective.

To every person, living or dead, who has contributed knowledge to the advancement of Science and Humanity.

Acknowledgements

I want to pass sincere gratitude to my supervisor Prof. Montaz Ali for his guidance and support throughout the research work of this thesis.

Special thanks to my family and friends for their unconditional support and encouragement throughout this research work.

I want to thank the School of Computer Science and Applied Mathematics, Faculty of Science, and the University of Witwatersrand for financial support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Many situations in a wide range of science and engineering applications are modelled as inverse problems. Inverse problems seek to infer information about the features of a specific system from observed measurements. In applications of interest, it suffices to model the problem as a linear relationship between the features and the observable quantities. Moreover, the observable quantities are much less than the unknown features, with the knowledge that the unknown features are sparse. The mathematical model becomes an underdetermined linear system. Recently, finding a sparse solution to an underdetermined linear system has been intensively studied. Mainly because solving such a system constitutes a critical step in the emerging methodology in digital signal processing - called *compressed sensing*.

Compressed sensing is a signal processing technique for efficiently acquiring and processing signals [1, 2]. Compressed sensing outperforms the traditional Shannon-Nyquist technique by pushing the computational burden to the processing stage. Recovering signals from their compressed sensing measurements can be a computationally expensive problem. The principle of reconstruction is based on nonsmooth convex optimization theory. The classic approach to this problem is the sparse recovery problem, which involves finding sparse solutions to an underdetermined linear algebraic system.

Specialised optimization algorithms have been used to solve sparse recovery problems in compressed sensing. However, these algorithms tend to be slow and computationally intensive

to perform real-time recovery. Responsiveness of any real-time system must be sufficiently small. On the contrary, continuous-time dynamical systems have recently gained attention as efficient optimization problem solvers. They have the potential to yield significantly speed and power improvements over their discrete counterpart. This thesis studies the type of continuous-time dynamical systems that solves sparse recovery problems. We design and provide the mathematical analysis of a fast dynamic system to solve sparse recovery problems.

1.2 Background

The theoretical foundation of acquiring signals was pioneered by Nyquist [3] and Shannon [4] on sampling continuous-time band-limited signals. A signal is band-limited if the amplitude of its spectrum goes to zero for all frequencies above some threshold. Nyquist and Shannon showed that analog signals could be exactly recovered from a set of uniformly spaced samples taken at the *Nyquist rate*. The Nyquist rate is a sampling rate equal to twice the highest frequency present in the signal. As a result of the Nyquist-Shannon sampling theory, most signal processing has shifted from analog to digital. Although, the success of digitization resulted in sensing systems generating a large amount of data. The digitization of signals has enabled the invention of more robust, flexible, and cheaper sensing systems to acquire and process signals.

Unfortunately, the resulting Nyquist rate has severe repercussions in many applications in various emerging technologies. The Nyquist rate is so high that too many samples are needed. Despite advancements in computational power, the acquisition and processing of signals in many applications such as imaging, medical imaging and biology continue to pose tremendous challenges in terms of computational resources (transmission, storage, e.t.c). Additionally, prominent signal structures such as sparsity are not fully utilized. Nyquist-Shannon theory would heavily oversample sparse signals since much of the signals have little to no vital information.

Compression often achieves the conventional way of mitigating computational resource challenges involved in signal processing. Compression aims to find the most concise representation of a signal with modest distortion. A typical example of an image compression technique

is *transform coding*. Transform coding relies on finding a basis that provides sparse or compressible signal representation. Sparse representation means a signal has few nonzero components compared to its length. Compressible representation means a signal can be well-approximated by a sparse signal (precise definitions of sparse and compressible representation are reserved until Section 3.1 of Chapter 3). Sparse and compressible signals can be accurately represented by preserving signal components with the largest magnitudes. This process is known as a *sparse approximation* and plays a fundamental role in transforming coding schemes. A sensing system usually samples a signal at the Nyquist rate. Before transmission, the samples are compressed, discarding samples of minimal value, thus minimizing storage and ultimately computational cost.

Leveraging the concept of transform coding, *compressed sensing* has emerged as a new framework for signal acquisition and sensor design. Compressed sensing significantly reduces sampling and computation costs for sensing signals with a sparse or compressible representation. The Nyquist-Shannon sampling theory states that a minimum number of samples is required to capture an arbitrary band-limited signal perfectly. However, when a signal is sparse on a known basis, it can vastly reduce the number of measurements needed to be stored. Thus, when sensing sparse signals, we might be able to do better than the traditional Nyquist-Shannon sampling. This is the fundamental idea behind compressed sensing. Rather than first sampling at a high rate and then compressing the sampled signal, it would be best to find ways to sense the signal in a compressed form directly. Compressed sensing grew out of the work of Candiès [1] and Donoho [2]. They showed that a finite-dimensional signal admitting a sparse or compressible representation can be recovered from a small set of linear, nonadaptive measurements.

1.3 Problem statement

Many applications in emerging technologies call for efficient sensing systems to acquire and process high-resolution signals. This task is impractical for the traditional sampling scheme, the Nyquist-Shannon sampling theory. Compressed sensing was developed as the new signal acquisition scheme to address this issue. The compressed sensing theory asserts that linear encoded measurements can be used to simultaneous acquire and compress the signal. The

technique requires much less computational resources for sensing/acquisition of signals than the traditional sampling schemes.

At the heart of compressed sensing theory lies the sparse recovery problem. The sparse recovery problem seeks to find a solution to an undetermined linear system of equations requiring that the solution is sparse. The problem can be solved using the brute force approach by checking all possible solutions and selecting one with the smallest number of nonzero components. This task is computationally intractable. Candiès [1], and Donoho [2] showed that with an appropriate choice of a sparse promoting function, the sparse recovery problem can be posed as the convex optimization program. Thus, algorithms for solving the resulting problem are computationally tractable. However, the choice of sparse prompting function often results in a nonsmooth optimization problem. It is mathematically challenging to deal with a nonsmooth optimization problem.

Developing efficient discrete-time algorithms to solve the sparse recovery problem has driven many research efforts. Various algorithms have been proposed and can generally be grouped into three categories: the convex relaxation methods [5–7], iterative thresholding methods [8, 9], and greed methods [10–12]. Despite numerous discrete-time algorithms, none are currently efficient enough to achieve real-time processing as required in some applications. Real-world applications are invariably large-scale, posing severe computational challenges. The speed and power efficiency of digital computers reach a bottleneck when the size of the data becomes enormous.

Hence, there is a need for research to address the challenges mentioned above. The objective of the current research work is to develop an effective and efficient algorithm for solving nonsmooth optimization resulting from the formulation of a sparse recovery problem. The new algorithm must be practical; it should scale well as the problem size increase.

Recently, there has been much research effort investigating the ability of dynamical systems as an efficient tool to solve optimization problems. A continuous-time dynamical system approach is adopted in this research work to answer the need for a fast algorithm that solves the sparse recovery problem.

1.4 Motivation

In this section, we motivate the use of the continuous-time dynamical system to solve the sparse recovery problem. The advantages of this approach are twofold:

First, dynamical systems have a rich and well-developed theory. Thus, the use of dynamical systems in optimization provides good insights into the design and analysis of algorithms. This approach has recently gained attention in the optimization research community. In [13], Su *et.al* derive a second-order ordinary differential equation which is the limit of Nestrove's accelerated gradient methods [14, 15]. The resulting differential equation allows a better understanding of the behaviour of Nesterov's scheme. We limit our research work to first-order dynamical systems derived from the Hopfield network [16–18]. The Hopfield approach provides an intuitive way of constructing the dynamical system. Furthermore, the approach is very efficient in solving quadratic programs. We will see later that sparse recovery can be formulated as a quadratic program.

Second, for the usefulness of dynamical systems to real-world applications, we consider the following compressed sensing problems:

- **Compressive Imaging.** Compressed sensing has far-reaching implications for compressive imaging systems and cameras. Rice university invented the single-pixel camera that can operate efficiently across a broader spectral range than the conventional camera, see [19–21]. The camera is based on a single photon detector adaptable to images at wavelengths that were impossible with conventional cameras. The advantage of a single-pixel camera over its conventional counterpart is that a tiny number of measurements are required to create an image. The resulting image does not suffer from the usual aberration and focusing problems associated with the lens. Real-time processing of data is paramount to realise a practical single-pixel camera fully.

- **Medical Imaging.** The early success of compressed sensing comes from medical imaging, particularly Magnetic Resonance Imaging (MRI) [22–24]. Generally, MRI is a costly and time-consuming process because its data collection process depends on sensitive physical and physiological constraints. However, compressive sensing-based techniques have reduced the computational cost of collecting data while retaining the

image quality, essential for accurate diagnosis. Furthermore, the images are usually large, motivating the need for efficient algorithms to achieve the real-time process of the data.

- **Microarrays Sequencing.** Microarray is used in biology to identify specimens. Fluorescent tags are used to determine where samples bind. Most samples have only a few active parts. Thus, using ideas from compressive sensing, it is possible to take fewer measurements and accurately infer which specimens are present. To fully realise the potential of this approach, an efficient algorithm that can perform real-time processing of data is sought after. This approach is helpful in practice; see [25, 26].

These applications are characterized by their large size and require real-time data processing. Current discrete-time algorithms do not achieve these requirements. Thus, there is a need to use continuous-time methods. The continuous-time dynamical system has the potential to perform real-time data processing, primarily when implemented on dedicated hardware.

Hence, the theoretical benefit of a continuous-time dynamical system is that it provides an easy means for the design and analysis of optimization algorithms. In the case of real-time processing, having dedicated hardware to implement the algorithm will provide added benefits. Thus, there is a need to study continuous-time dynamical systems to solve sparse recovery problems.

1.5 Contributions

Before using dynamical systems to solve optimization problems, it is crucial to assess their performance guarantees. This thesis aims to study and understand the types of dynamical systems that efficiently solve sparse recovery problems. The contributions of this thesis are twofold.

- First, we consider a general dynamical system for solving sparse recovery problems. Theoretical tools for analysing the dynamical system are developed. In particular, new results are presented for the convergence study of a class of differential inclusion. A differential inclusion is constructed from the subdifferential of the nonsmooth objective

function. The differential inclusion is coupled with an activation function that promotes sparsity on the solution. The sparsity promoting activation function does not fit the standard activation functions such as the sigmoid function or saturation function. The activation function is identical to zero in some specific interval. Furthermore, the activation function is not bounded, making the mathematical analysis difficult.

We have used the recently developed Lojaseiwicz inequality for a nonsmooth function. Using the inequality, we have proved that the trajectory of the differential inclusion converges to equilibrium. The equilibrium point corresponds to the local minimizer of the nonsmooth objective function. Convergence is guaranteed even when the optimal points are not isolated. Furthermore, we have estimated the convergence speed in terms of the Lojaseiwicz exponent.

- Second, we consider a particular case of sparse recovery problem, the ℓ_1-minimization problem. The ℓ_1 minimization problem is the most famous optimization program for sparse recovery in compressed sensing and comes with strong performance guarantees. We construct a dynamical system that integrates the $\ell_1 - minimization$ problem with the Bregman distance. We call it the *Bregman integrated dynamical system.*

  We have provided the mathematical analysis of the solution produced by the Bregman integrated dynamical system. The analysis shows that the Bregman integrated dynamical system takes an efficient trajectory toward the optimal solution of the ℓ_1-minimization problem. Thus the approach reaches a more accurate solution much faster than the previously proposed dynamical system called Locally Competitive Algorithm [27].

This thesis improves the convergence rate and the accuracy of a continuous-time dynamical system to recover static signals. The existing analysis has focused on discrete-time algorithms to recover sparse vectors. Iterative threshold algorithms are widely used due to their balanced use of the greed method's speed and the convex method's performance guarantees. The best-known results of iterative thresholding methods are linear convergence rate. The proposed Bregman integrated dynamical system provides an exponential rate of convergence. It generally scales well with problem size compared to the discrete-time algorithms. Also, the continuous-time dynamical system can be easily adapted to recover time-varying signals.

The contributions of this thesis are presented in the following manuscripts prepared for publication:

1. Convergence analysis of subgradient dynamical system for solving sparse recovery problems.

2. Bregman integrated dynamical system for solving ℓ_1-minimization problems.

1.6 Organization of the thesis

The thesis is organized in the following manner. Chapter 2 gives mathematical preliminaries needed for mathematical analysis in the research work. In Chapter 3, we review the standard compressed sensing theory, highlighting some of the essential theoretical results. We also review the previous work on using the dynamical system to solve optimization problems, in particular, the research based on the Hopfield approach. In Chapter 4, we prove the convergence analysis of a differential inclusion using the Lojaseiwicz inequality for nonsmooth function. The results are general for sparse recovery problems. In Chapter 5, we construct and provide mathematical analysis of the proposed Bregman integrated dynamical system for solving ℓ_1-minimization problem. In Chapter 6, we demonstrate the computational performance of the proposed dynamical system. Finally, we give a summary of the current research work and discuss possible future research options in Chapter 7.

Chapter 2

Mathematical Preliminaries

This chapter gives the mathematical preliminaries needed to develop and analyse mathematical concepts in this research work. A signal can be viewed as a vector living in a certain vector space. In Section 2.1, we review the normed vector space. The convex nonsmooth objective functions are at the heart of optimization problems considered in this research. In Section 2.2, a review convex analysis and subgradient of nonsmooth functions is presented. The use of the dynamical system to solve optimization problems is the main concern of the current research work. In Section 2.3, we review dynamical system analysis. In Section 2.4, we present Gronwall's lemma, which provides a way of solving some simple linear dynamical systems. In Section 2.5, we present the Łojasiewicz inequality for nonsmooth function. The inequality plays an important role in the convergence analysis of dynamical systems considered in this research work.

2.1 Normed vector space

A vector space is a set that is closed under finite vector addition and scalar multiplications. Any addition of two vectors and scalar multiplication with any vector, the result belongs to the vector space. Vector space allows us to apply intuition and tools from geometry to describe and compare signals of interest. We can compute quantities such length of the signal, the distance between two signals and so forth. Throughout this research work, we consider the n-dimensional Euclidean space, denoted by $\mathbb{R}^n$. Moreover, we are concerned

with *normed vector spaces*, that is, vector spaces endowed with a norm. A norm is defined as follows.

Definition 2.1 (Norm)**.** *A function* $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+$ *is said to be a norm if it satisfies the following condition:*

1. $\|\mathbf{x}\| \geq 0$ *for all* $\mathbf{x} \in \mathbb{R}$ *and* $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$.
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ *for all* $\alpha \in \mathbb{R}$ *and* $\mathbf{x} \in \mathbb{R}^n$
3. $\|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ *for all* $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$.

These properties are known as: 1 positive definiteness, 2 homogeneity, and 3 triangle inequality.

When dealing with vectors in $\mathbb{R}^n$, often ℓ_p-norm becomes the norm of choice, as they are easy to handle. For $p \in [1, \infty]$ the ℓ_p-norms are defined as

$$\|\mathbf{x}\|_p = \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, & p \in [1, \infty); \\ \max_i |x_i|, & p = \infty. \end{cases} \tag{2.1}$$

In Euclidean space, we can also consider the standard *inner product* in $\mathbb{R}^n$, defined as

$$\langle \mathbf{x}, \mathbf{z} \rangle = \mathbf{x}^T \mathbf{z} = \sum_{i=1}^n x_i z_i.$$

The norm induced by the inner product corresponds to the ℓ_2-norm: $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

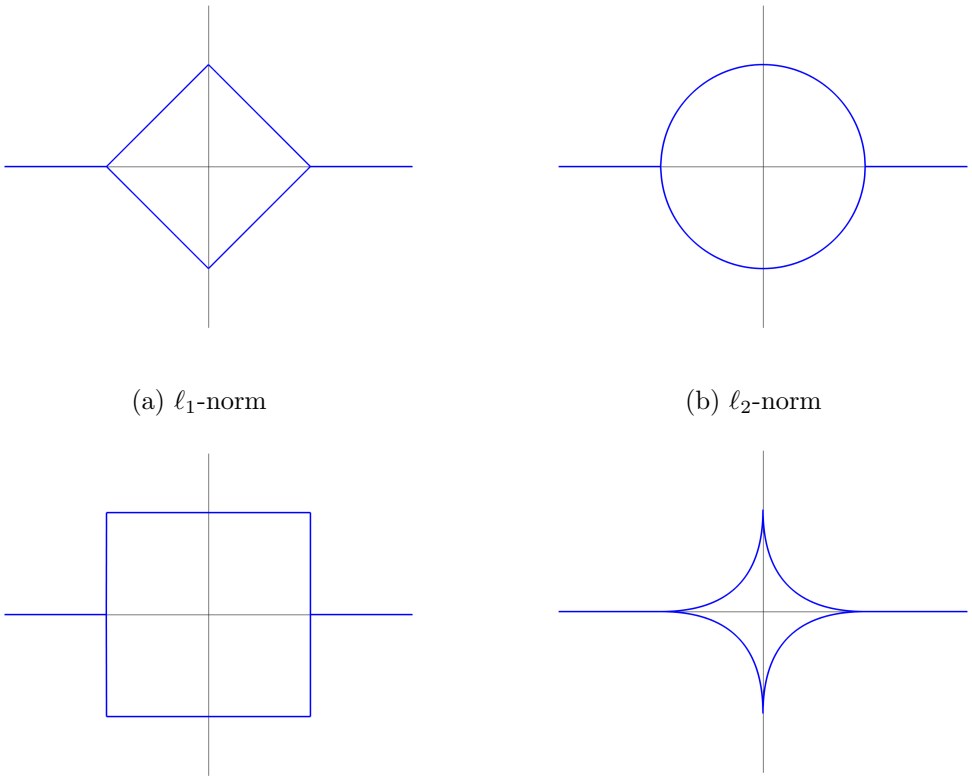
In some contexts it is useful to extend the notion of ℓ_p-norms to the case where $p \in (0, 1)$. However, the ℓ_p-norms defined in (2.1) for $p \in (0, 1)$ cease to be a proper norm as it fails to satisfy the triangle inequality. Instead, we get quasinorms for $p \in (0, 1)$. In the case where $p = 0$, we define the the ℓ_0-norm of $\mathbf{x}$ of the vector $\mathbf{x} \in \mathbb{R}^n$ as the number of nonzero components

$$\|\mathbf{x}\|_0 := \lim_{p \to 0} \|\mathbf{x}\|_p^p = |\text{supp}(\mathbf{x})|, \tag{2.2}$$

where $\text{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$ is the support of $\mathbf{x}$ and $|\text{supp}(\mathbf{x})|$ denotes the cardinality of $\text{supp}(\mathbf{x})$. We will make frequent use of ℓ_0-norm to measure the sparsity of vectors. Note, we

abuse notation here; the ℓ_0-norm is neither a norm nor quasinorm as it fails to satisfy the homogeneity condition.

The ℓ_p-norms (and quasinorms) have different properties for different values of p. To illustrate this, consider Figure 2.1. The plots show the unit sphere, defined as $\{\mathbf{x} : \|\mathbf{x}\|_p = 1\}$, induced by each of these norms in $\mathbb{R}^2$. We observe that for the well-defined norms $p \in [1, \infty]$, any line segment joining two points in the sphere belongs entirely in the sphere. In contrast, this is not the case for quasinorms $p \in (0, 1)$ due to the violation of the triangle inequality. This observation leads to convexity, which we define in the next section.



(a) ℓ_1-norm (b) ℓ_2-norm

(c) ℓ_∞-norm (d) $\ell_{\frac{1}{2}}$-quasinorm

FIGURE 2.1: Unit spheres in $\mathbb{R}^2$ for the ℓ_p-norms with $p = 1, 2, \infty$, and for the ℓ_p-quasinorm with $p = \frac{1}{2}$.

2.2 Convex analysis and subdifferentials

The concept of convexity is fundamental in optimization theory. There is a rich theory regarding the analysis of convex function, especially when the objective function is smooth.

However, we are dealing with nonsmooth functions; thus, we need tools to analyze convex nonsmooth optimization problems. We review the properties of convex functions and some essential tools for nonsmooth analysis.

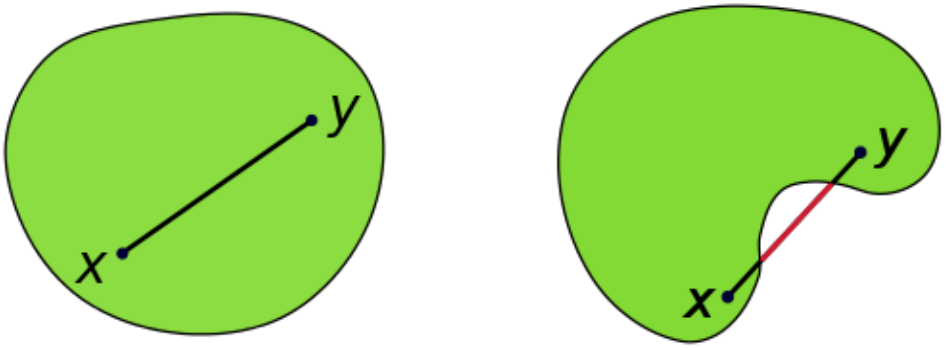
2.2.1 Convexity

If $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$, the vector $\lambda\mathbf{x} + (1 - \lambda)\mathbf{z}$ is called a *convex combination* of $\mathbf{x}$ and $\mathbf{z}$. A set Ω is convex if every convex combination of any two elements of Ω belongs to Ω. A formal mathematical definition of a convex set is given below.

Definition 2.2 (Convex set)**.** *A set $\Omega \subseteq \mathbb{R}^n$ is said to be a convex set if for any $\mathbf{x}, \mathbf{z} \in \Omega$, the following condition holds*

$$\lambda\mathbf{x} + (1 - \lambda)\mathbf{z} \in \Omega, \quad \textit{for all} \quad \lambda \in [0, 1].$$

In $\mathbb{R}^2$ a convex set can be easily visualized. If $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$, then the set $\{\lambda\mathbf{x}+(1-\lambda)\mathbf{z} : \lambda \in [0, 1]\}$ is the line segment form $\mathbf{z}$ to $\mathbf{x}$. Thus, a set $\Omega \subseteq \mathbb{R}^2$ is convex if and only if the line segment joining any two points in Ω belongs entirely in Ω. The set in Figure 2.2b is not convex, because the line segment connecting $\mathbf{x}$ and $\mathbf{z}$ does not lie entirely in the set. On the other hand, the set in Figure 2.2a is convex. This interpretation can be conceptualized to high dimensions.



(a) Convex set (b) Nonconvex set

FIGURE 2.2: An illustration of convex set and nonconvex set in $\mathbb{R}^2$.

Given a nonconvex set Ω, we define a *convex hull* as the smallest convex set containing Ω. The convex hull of a set Ω is denoted as $co\{\Omega\}$.

Definition 2.3 (Convex function)**.** *A function $F : \mathbb{R}^n \to \mathbb{R}$ defined on a convex set $\Omega \subseteq \mathbb{R}^n$, is said to be convex if for any two points $\mathbf{x}, \mathbf{y} \in \Omega$ the following inequality is holds*

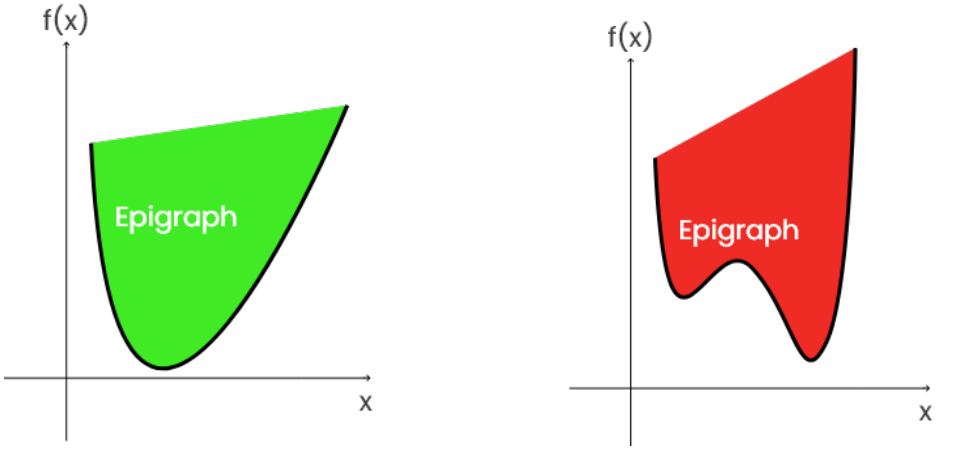
$$F(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda F(\mathbf{x}) + (1-\lambda)F(\mathbf{y}), \quad \textit{for all} \quad \lambda \in [0,1].$$

Figure 2.3 gives a pictorial representation of convex and nonconvex functions. A very convenient and equivalent definition of a convex function is in terms of its *epigraph*. A convex function is completely characterised by its epigraph.

Definition 2.4 (Epigraph of a function)**.** *The epigraph of function $F : \mathbb{R}^n \to \mathbb{R}$ denoted by $epiF$ is defined by the set*

$$epiF = \{(\mathbf{x}, \mu) : \mathbf{x} \in \Omega, \mu \geq F(\mathbf{x})\}.$$

In Figure 2.3 an epigraph is shown by the shaded area. Hence, a function F is convex if and only if its epigraph $epiF$ is a convex set.



(a) Nonconvex function (b) Nonconvex function

FIGURE 2.3: An illustration of convex set and nonconvex functions.

Simple instances of convex functions are norms. In particular, ℓ_p-norms form a convex function for $p \in [1, \infty]$ and a nonconvex function for $p \in (0, 1)$. As we will see later, both

the ℓ_1-norm and ℓ_0-norm are of particular interest in the development of compressed sensing theory.

The following properties of convex functions are essential in developing and analysing optimization algorithms.

Definition 2.5 (Properties of convex function)**.** *Let $F : \Omega \to \mathbb{R}$ be a convex function . We then say that*

- *F is* proper*, if $F(\mathbf{x}) < \infty$ for some $\mathbf{x} \in \Omega$, and $F(\mathbf{x}) > -\infty$ for all $\mathbf{x} \in \mathbb{R}$.*
- *F is* lower semicontinuous *at $\mathbf{x}$ if for any sequence $\{\mathbf{x}^i\}_{i=1}^{\infty} \subset \mathbb{R}^n$, with $\mathbf{x}^i \to \mathbf{x}$ holds*

$$F(\mathbf{x}) \leq \liminf_{i \to \infty} F(\mathbf{x}^i).$$

  Moreover, F is lower semicontinuous *if it is lower semicontinuous at every $\mathbf{x} \in \Omega$.*

- *F is* closed *if and only if epi$F$ is a closed set.*

These properties are important for optimization problems, as evident by the following proposition, which characterizes the existence of infimum for a function satisfying the above properties.

Proposition 2.6 (Existence of infimum)**.** *Let $F : \Omega \to \mathbb{R}$ be a proper and lower semicontinuous, and $\Omega \subset \mathbb{R}^n$ be closed and bounded. Then there exist $\mathbf{x}^* \in \Omega$ such that*

$$F(\mathbf{x}^*) = \inf_{\mathbf{x} \in \Omega} F(\mathbf{x}),$$

and this value is finite.

2.2.2 Clarke subdifferential

Here, we recall the concepts of generalized derivatives in the sense of Clarke subdifferential. The discussions here follow the work of Clarke [28]. The aim is to characterize the optimal solution to a nonsmooth optimization problem. We also discuss calculus rules associated with the Clarke subdifferential.

Definition 2.7 (Lipschitz continues function)**.** *The function $F : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous at $\mathbf{x} \in \mathbb{R}^n$ if there exist $\delta > 0$ and a constant $L > 0$ (called a Lipschitz constant) such that*

$$|F(\mathbf{x}) - F(\mathbf{z})| \leq L\|\mathbf{x} - \mathbf{z}\|_2 \quad \forall \mathbf{x}, \mathbf{z} \in \mathcal{N}_\delta(\mathbf{x}). \tag{2.3}$$

We refer to the set $\mathcal{N}_\delta(\mathbf{x})$ as the Lipschitz neighbourhood of $\mathbf{x}$. A function F is Lipschitz continuous if it is locally Lipschitz continuous everywhere.

A Lipschitz continuous function is not necessarily differentiable. However, Rademacher's theorem states that if a function F is Lipschitz continuous, then F is differentiable almost everywhere [29]. With this knowledge, we now quantify the derivative of an arbitrary Lipschitz continuous nonsmooth function. There are several definitions of the standard notion of directional derivative for a Lipschitz function.

Definition 2.8 (One-sided directional derivative)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}$, the one-sided directional derivative of F at $\mathbf{x} \in \mathbb{R}^n$ in the direction $\mathbf{d} \in \mathbb{R}^n$ is*

$$F'(\mathbf{x}; \mathbf{d}) := \lim_{\alpha \to 0} \frac{F(\mathbf{x} + \alpha\mathbf{d}) - F(\mathbf{x})}{\alpha}, \quad \alpha \geq 0. \tag{2.4}$$

Some nonsmooth functions may fail to admit one-sided derivatives. The definition of one-sided directional derivative may be relaxed to the following notion of generalized directional derivative.

Definition 2.9 (Generalized directional derivative)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}$, the generalized directional derivative of F at $\mathbf{x} \in \mathbb{R}^n$ in the direction $\mathbf{d} \in \mathbb{R}^n$ is*

$$F^\circ(\mathbf{x}; \mathbf{d}) := \limsup_{\mathbf{y} \to \mathbf{x}, \alpha \to 0} \frac{F(\mathbf{y} + \alpha\mathbf{d}) - F(\mathbf{y})}{\alpha}. \tag{2.5}$$

We can now define the subgradient of nonsmooth functions.

Definition 2.10 (Subgradient)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}$, the subdifferential of F at $\mathbf{x} \in \mathbb{R}^n$ is*

$$\partial F(\mathbf{x}) := \{\xi \in \mathbb{R}^n : \langle \xi, \mathbf{d} \rangle \leq F^\circ(\mathbf{x}; \mathbf{d}), \forall \mathbf{d} \in \mathbb{R}^n\}. \tag{2.6}$$

For a locally Lipschitz continuous function, the differential ∂F is well-defined, nonempty and convex. The Rademacher's theorem guarantees that F is differentiable all most everywhere. Taking this into account, (2.6) can be simplified to define the *Clarke subdifferential* of F at $\mathbf{x}$ as

$$\partial F(\mathbf{x}) = co\left\{\lim_{i\to\infty} \nabla F(\mathbf{x}^i) : \mathbf{x}^i \to \mathbf{x}\right\}.$$

In the case where F is a smooth function, the Clarke subdifferential $\partial F(\mathbf{x})$ reduce to a singleton that is equal to the classic gradient of F, that is $\partial F(\mathbf{x}) = \{\nabla F(\mathbf{x})\}$. Equipped with subdifferential, we can characterise the optimal solution of nonsmooth optimization problem using the Fermat principle

Proposition 2.11 (Fermat principle)**.** *If $F : \mathbb{R}^n \to \mathbb{R}$ has a local minimum at a point $\mathbf{x}^*$, then the inclusion $\mathbf{0} \in \partial F(\mathbf{x}^*)$ holds.*

The following notion of regularity plays a curial role in this research work.

Definition 2.12 (Regular function)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz continuous function, F is regular at $\mathbf{x} \in \mathbb{R}^n$ if $F'(\mathbf{x}; \mathbf{d})$ exists and $F'(\mathbf{x}; \mathbf{d}) = F^\circ(\mathbf{x}; \mathbf{d})$ for all the directions $\mathbf{d} \in \mathbb{R}^n$.*

In this research work, we deal with composite function of the form $F(\mathbf{x}(t))$. If F is regular, we can use chain rule to compute the time derivative of $F(\mathbf{x}(t))$ as stated in the following lemma.

Lemma 2.13 (Chain Rule)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz continuous and regular on $\mathbb{R}^n$ and $\mathbf{x} : \mathbb{R}_+ \to \mathbb{R}^n$ be differentiable. The composite function $F(\mathbf{x}(t))$ is also Lipschitz continuous and regular on $\mathbb{R}^n$, and the derivative $\dot{F}(\mathbf{x}(t))$ exits for almost all $t \geq 0$,*

$$\dot{F}(\mathbf{x}(t)) = \xi^T \dot{\mathbf{x}}, \quad \forall \xi \in \partial F(\mathbf{x}(t)). \tag{2.7}$$

The theorem states that any element ξ in the subdifferential ∂F can be used to compute the time derivative of $F(\mathbf{x}(t))$. This result is important in the study of trajectories of dynamic system.

2.3 Dynamic system analysis

A study of dynamical systems is at the core of this thesis. Before using dynamical systems to solve optimization problems, we must quantify their behaviour. The main goal is to determine the behaviour of the solution with respect to time. A solution of a dynamical system must settle to an appropriate equilibrium point as time evolves. Then the dynamical system is suitable for solving optimization problems. An investigation of stability and convergence is needed. We present some definitions and tools for determining the dynamic system's behaviour.

2.3.1 Stability and convergence

Consider the general dynamical system:

$$\dot{\mathbf{x}}(t) = F(\mathbf{x}(t)) \tag{2.8}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the system state at time t and the function $F : \mathbb{R}^n \to \mathbb{R}^n$ determine the dynamics of the system. Assume that $F(\mathbf{x}(t))$ is continuous, it is natural to define solutions as continuously differentiable functions satisfying the equation (2.8) in all points of some time interval. Any continuous function $\mathbf{x} : \mathbb{R}_+ \to \mathbb{R}^n$ is called a curve in $\mathbb{R}^n$.

Definition 2.14 (Absolute continuous curve)**.** *The curve $\mathbf{x} : \mathbb{R}_+ \to \mathbb{R}^n$ is absolutely continuous if there exits a map $\mathbf{z} : \mathbb{R}_+ \to \mathbb{R}^n$ that is integrable on any compact interval and satisfies*

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t \mathbf{z}(\eta) d\eta, \quad \forall t \geq 0.$$

Moreover, if this is the case, then the equality $\mathbf{z}(t) = \dot{\mathbf{x}}(t)$ holds for all $t \geq 0$.

Henceforth, we call absolutely continuous curves *trajectories*. We will often use the observation that if $F : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous and $\mathbf{x}(t)$ is a trajectory, then the composition $F(\mathbf{x}(t))$ is absolutely continuous.

Consider the system (2.8), we call the point $\mathbf{x}^* \in \mathbb{R}^n$ an equilibrium point if $F(\mathbf{x}^*) = \mathbf{0}$. The linear time-invariant systems have only one isolated equilibrium point. Nonlinear systems

can have more than one isolated equilibrium point. Therefore, we can only analyse local stability around the equilibrium points when dealing with nonlinear systems. There are different definitions of stability in the literature. Here, we adopt the Lyapunov stability.

Definition 2.15. *The system (2.8) is said to be:*

1. ***Locally Lyapunov stable*** *around the equilibrium point* $\mathbf{x}^*$*, if for any* $\varepsilon > 0$*, there exists* $\delta > 0$ *such that, if* $\|\mathbf{x}(0) - \mathbf{x}^*\| < \delta$*, then* $\|\mathbf{x}(t) - \mathbf{x}^*\| < \varepsilon$ *for all* $t > 0$*;*
2. ***Locally asymptotically stable*** *around the equilibrium point* $\mathbf{x}^*$*, if there exits* $\delta > 0$ *such that, if* $\|\mathbf{x}(0) - \mathbf{x}^*\| < \delta$*, then* $\lim_{t\to\infty} \|\mathbf{x}(t) - \mathbf{x}^*\| = 0$*.*
3. ***Locally finite stable*** *around the equilibrium point* $\mathbf{x}^*$*, if there exits* $\delta > 0$ *and* $T > 0$ *such that, if* $\|\mathbf{x}(t) - \mathbf{x}^*\| < \delta$*, then* $\|\mathbf{x}(t) - \mathbf{x}^*\| = 0$ *for all* $t > T$*.*

Lyapunov stability describes the behaviour of the trajectories locally around an equilibrium point. It guarantees that trajectory $\mathbf{x}(t)$ starting from the neighbourhood of the equilibrium point $\mathbf{x}^*$ will remain inside the neighbourhood. This type of stability does not guarantee that trajectories approach an equilibrium point as time evolves. However, *asymptotic stability* guarantees that the trajectory of the system converges to $\mathbf{x}^*$ as time goes to infinity. The strongest definition is the *finite-time stability*. Finite-time stability implies that the trajectory $\mathbf{x}(t)$ converges to an exact equilibrium point $\mathbf{x}^*$ after finite time T.

The local stability can be extended to *global stability* by relaxing the neighbourhood requirement around the equilibrium point (i.e $\|\mathbf{x}(0) - \mathbf{x}^*\| < \delta$). Global stability allow convergence from any starting point $\mathbf{x}(0) \in \mathbb{R}^n$. If the system is globally convergent to $\mathbf{x}^*$, it also implies that $\mathbf{x}^*$ is a unique equilibrium point. Suppose the trajectories can only be shown to approach a set of equilibrium points. In that case, the dynamical system is said to be *quasi-convergent*.

In addition to stability, it is essential to know how fast trajectories converge for real-time applications. Of interest in this thesis is the notion of exponential rate of convergence. The dynamical system is called exponentially convergent to an equilibrium point $\mathbf{x}^*$ if there exits a constant $c > 0$ such that for any initial point $\mathbf{x}(0)$, there exists a constant $\kappa > 0$ for which the trajectory $\mathbf{x}(t)$ of (2.8) with $\mathbf{x}(0) = \mathbf{x}^0$ satisfy

$$\|\mathbf{x}(t) - \mathbf{x}^*\| \leq \kappa e^{-ct}, \quad \forall t \geq 0.$$

The constant c is referred to as the convergence speed of the system. When a dynamic system is exponentially convergent, the distance to the fixed point decays rapidly.

2.3.2 Lyapunov's function method

Lyapunov's function method makes the mathematical analysis of stability and convergence of the system easy [30]. The key to this method resides in finding a positive definite function that represents a notion of energy of the dynamical system (2.8). The equilibrium points are stable if the energy function is nonincreasing along the system's trajectories.

Proposition 2.16 (Lyapunov function method)**.** *Suppose* $\mathbf{x}^*$ *is the equilibrium point of the dynamical system (2.8), and denote* $\bar{\mathbf{x}}(t) = \mathbf{x}(t) - \mathbf{x}^*$*. If there exits a Lypapunov function* $E(\bar{\mathbf{x}})$ *which is continuous and positive definite for all* $\bar{\mathbf{x}} \neq \mathbf{0}$ *and* $E(\mathbf{0}) = 0$*, then the system (2.8) is said to be:*

1. *Locally Lyapunov stable around the equilibrium point* $\mathbf{x}^*$*, if*

$$\dot{E}(\bar{\mathbf{x}}) \leq 0, \quad \text{for all } \bar{\mathbf{x}} \neq \mathbf{0};$$

2. *Locally asymptotically stable with rate* κ *around the equilibrium point* $\mathbf{x}^*$ *if*

$$\dot{E}(\bar{\mathbf{x}}) \leq -\kappa E(\bar{\mathbf{x}}), \quad \text{for all } \bar{\mathbf{x}} \neq \mathbf{0},$$

   with $\kappa > 0$*;*

3. *Locally finite-time stability around the equilibrium point* $\mathbf{x}^*$*, if*

$$\dot{E}(\bar{\mathbf{x}}) \leq -\kappa E^{\alpha}(\bar{\mathbf{x}}), \quad \text{for all } \bar{\mathbf{x}} \neq \mathbf{0},$$

   with $\kappa > 0$ *and* $\alpha \in (0, 1)$*.*

Without the need to solve the dynamical system, Lyapunov function method can be used to analyse the behaviour of the system.

2.4 Gronwall's Lemma

Gronwall's Lemma plays a very crucial role in estimating the solution of the dynamical system [31]. The Lemma allows one to bound a function that is known to satisfy certain differential inequality. There are many variations of the Lemma. The following variation of Gronwall's Lemma will be used repeatedly in this search work.

Lemma 2.17 (Gronwall's Lemma)**.** *Let $a \in \mathbb{R}$. If $x : \mathbb{R}_+ \to \mathbb{R}$ is absolutely continuous and satisfies*

$$\frac{dx(t)}{dt} \leq -ax(t) + F(x(t), t), \qquad x(0) = x^0. \tag{2.9}$$

where $F(\mathbf{x}(t), t)$ is the nonautonomous form of $F(\mathbf{x}(t))$ in (2.8). Then the following holds for all $t \geq 0$:

$$x(t) \leq e^{-at}x_0 + e^{-at}\int_0^t e^{a\eta}F(x(\eta), \eta)d\eta. \tag{2.10}$$

Proof. The following derivation holds for all $t \geq 0$:

$$\frac{d}{dt}\left(e^{at}x(t)\right) = ae^{at}x(t) + e^{at}\frac{d}{dt}x(t) \tag{2.11a}$$

$$\leq ae^{at}x(t) + e^{at}\left(-ax(t) + F(x(t), t)\right) \tag{2.11b}$$

$$= e^{at}F(x(t), t), \tag{2.11c}$$

where (2.11b) results from (2.9). Now, integrating both sides from 0 to t and using positivity of the integral yields

$$e^{at}x(t) - x(0) \leq \int_0^t e^{a\eta}F(x(\eta), \eta)d\eta.$$

It follows that

$$x(t) \leq e^{-at}x_0 + e^{-at}\int_0^t e^{a\eta}F(x(\eta), \eta)d\eta.$$

□

As an extension of Gronwall's Lemma, the following result applies to a linear system of differential equations with a constant matrix.

Lemma 2.18 (Extension of Gronwall's Lemma)**.** *Let* $\mathbf{x} : \mathbb{R}_+ \to \mathbb{R}^n$ *be an absolutely continuous curve,* $\mathbf{A}$ *be a matrix* $\mathbb{R}^{n\times n}$ *and* $\mathbf{b} : \mathbb{R}_+ \to \mathbb{R}^n$ *be absolute continuous. Consider the following differential equation*

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}(t), \tag{2.12}$$

with the initial condition at t_k *being* $\mathbf{x}(t_k)$*. Then, for all* $t > t_k$ *the solution of the differential equation (2.12) is*

$$\mathbf{x}(t) = e^{\mathbf{A}(t-t_k)}\mathbf{x}(t_k) + e^{\mathbf{A}t}\int_{t_k}^{t} e^{-\mathbf{A}\eta}\mathbf{b}(\eta)d\eta. \tag{2.13}$$

Moreover, if $\mathbf{b}(s) = \mathbf{b}$ *is a constant vector in* $\mathbb{R}^n$*, the solution can be written explicitly as*

$$\mathbf{x}(t) = e^{\mathbf{A}(t-t_k)}\mathbf{x}(t_k) + \left(\mathbf{I} - e^{\mathbf{A}(t-t_k)}\right)\mathbf{A}^{-1}\mathbf{b}. \tag{2.14}$$

In the expression (2.14), $\left(\mathbf{I} - e^{\mathbf{A}t}\right)\mathbf{A}^{-1}$ is well-defined. To illustrate this fact, the matrix $\mathbf{A}$ can be expressed as $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$, where $\mathbf{\Lambda}$ is a diagonal matrix formed by eigenvalues λ_i of $\mathbf{A}$ and $\mathbf{P}$ is formed by the eigenvectors of $\mathbf{A}$. Using this, the decomposition yields

$$\begin{aligned}\left(\mathbf{I} - e^{\mathbf{A}t}\right)\mathbf{A}^{-1} &= \mathbf{P}\left(\mathbf{I} - e^{\mathbf{\Lambda}t}\right)\mathbf{\Lambda}^{-1}\mathbf{P}^{-1} \\ &= \mathbf{P}\begin{bmatrix}(1-e^{\lambda_1 t})\lambda_1^{-1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & (1-e^{\lambda_n t})\lambda_n^{-1}\end{bmatrix}\mathbf{P}^{-1}\end{aligned}$$

The following Taylor expansion as λ_i goes to zero can be used to show that the diagonal elements are well-defined even when $\lambda_i = 0$:

$$\lambda_i^{-1}\left(1 - e^{\lambda_i t}\right) = \lambda_i^{-1}\left(-\lambda_i t + \mathcal{O}(\lambda_i^2)\right) = -t + \mathcal{O}(\lambda_i).$$

Hence, by continuity, $\lambda_i^{-1}\left(1-e^{\lambda_i t}\right)=-t$ when $\lambda_i=0$. As a results, the matrix $\left(\mathbf{I}-e^{\mathbf{A}t}\right)\mathbf{A}^{-1}$ is well defined.

2.5 Subanalicity and Łojasiewicz inequality

Lyapunov's function method is often used to prove the convergence of dynamical systems. Consider the proof of convergence to a set of isolated equilibrium points. Suppose there exists a set of connected equilibrium points. In this case, the trajectories are only guaranteed to evolve towards this set. However, there is no certainty that they will converge towards one unique point in the set. In other words, the trajectories are not prevented from growing unbounded or oscillating indefinitely as they approach the solution set. A new technique based on the Łojasiewicz gradient inequality [32] was developed to overcome this limitation.

The Łojasiewicz gradient inequality relies on the geometric properties of a function. It relates the difference of a function value near a certain point to the norm of its gradient. Formally, it states that for a real-analytic function $F:\mathbb{R}^n\to\mathbb{R}$ and for all $\mathbf{x}^*\in\mathbb{R}^n$, there exit $\theta\in(0,1]$, $c>0$ and $r>0$ such that the function F satisfies

$$|F(\mathbf{x})-F(\mathbf{x}^*)|^\theta\leq c\|\nabla F(\mathbf{x})\|_2,\qquad \text{for all } \mathbf{x} \text{ such that } \|\mathbf{x}-\mathbf{x}^*\|_2<r.$$

Using this inequality, Łojaseiwicz showed that the trajectories of gradient flow $\dot{\mathbf{x}}(t)=-\nabla F(\mathbf{x}(t))$ have finite length. Thus ensuring convergence to a singleton even when the equilibrium points are not isolated [32].

Recently, an extension of the Łojasiewicz inequality was developed for nonsmooth functions [33]. The gradient in the original formulation is replaced by nonsmooth slope, which represents the smallest norm of any vector in the subdifferential $\partial F(\mathbf{x})$.

Theorem 2.19 (Nonsmooth Łojasiewicz inequality)**.** *Suppose that a function $F:\mathbb{R}^n\to\mathbb{R}$ is subanalytic and continuous on $\mathbb{R}^n$. Then, for any $\mathbf{x}^*\in\mathbb{R}^n$, there exists $\theta\in[0,1]$, $c>0$, and $r>0$ such that*

$$|F(\mathbf{x})-F(\mathbf{x}^*)|^\theta\leq cm\left(\partial F(\mathbf{x})\right),\qquad \textit{for all } \mathbf{x} \textit{ such that } \|\mathbf{x}-\mathbf{x}^*\|_2<r.$$

where the nonsmooth slope of F *at* $\mathbf{x} \in \mathbb{R}^n$ *is defined as*

$$m\left(\partial F(\mathbf{x})\right) = \inf\left\{\|\xi\|_2 : \xi \in \partial F(\mathbf{x})\right\}. \tag{2.15}$$

The nonsmooth Łojasiewicz inequality requires the function F to be *subanalytic*. This property does not require the function to be differentiable, but involves geometric properties of the graph, such as algebraic manipulations of sets defined by real-analytic equations and inequalities. We adopt the definition of subanalytic from [33]. A set $\mathcal{A} \in \mathbb{R}^n$ is said to be semianalytic if each point $\mathbf{x} \in \mathbb{R}^n$ admits a neighbourhood $\mathcal{N}$ for which

$$\mathcal{A} \cap \mathcal{N} = \bigcup_{i=1}^{p} \bigcap_{j=1}^{q} \left\{\mathbf{x} \in \mathcal{N} : f_{ij}(\mathbf{x}) = 0, g_{ij}(\mathbf{x}) > 0\right\},$$

where $f_{i,j}, g_{i,j} : \mathcal{N} \to \mathbb{R}$ are real analytic functions for all $1 \leq i \leq p$, $1 \leq j \leq q$ and p and q are some integers. A set $\mathcal{B}$ is said to be subanalytic if it is locally the projection of a semianalytic set, that is, each point $\mathbf{x} \in \mathbb{R}^n$ admits a neighbourhood $\mathcal{N}$ such that $\mathcal{B} \cap \mathcal{N} = \{\mathbf{x} \in \mathbb{R}^n : (\mathbf{x}, \mathbf{y}) \in \mathcal{A}\}$, where $\mathcal{A}$ is a bounded semianalytic subset of $\mathbb{R}^n \times \mathbb{R}^m$ from $m \geq 1$. Finally, a function $F : \mathbb{R}^n \to \mathbb{R}$ is said to be subanalytic if its graph, $graf(F) = \{(\mathbf{x}, \mathbf{y}) : \mathbf{y} = F(\mathbf{x})\}$, is a subanalytic subset of $\mathbb{R}^n \times \mathbb{R}$.

Chapter 3

Background

This chapter reviews the standard compressed sensing theory and the dynamical system approach used for solving optimization problems. Sparsity plays a crucial role in compressed sensing. Thus, in Section 3.1 we review sparse representation and sparse recovery problem. The challenges of compressed sensing are twofold: the design of measurement matrix and efficient sparse recovery algorithms. In Section 3.2, we review the properties of the measurement matrix needed to quantify the matrix as appropriate for compressed sensing use. In Section 3.3, we review some basic algorithmic approaches to compressed sensing. In Section 3.4, we review the Hopfield approach used to derive the dynamical systems for solving optimization problems. In Section 3.5, we review previous work done relating to the analysis of dynamical systems associated with the Hopfield approach.

3.1 Sparse representation

The concepts of sparse and compressible representation of signals was introduced earlier. The concepts are now formalised mathematically.

3.1.1 Sparse vectors

A set $\{\psi_i\}_{i=1}^n$ is called basis for $\mathbb{R}^n$ if the vectors in the set span $\mathbb{R}^n$ and are linearly independent. This means that each vector in the space has a unique representation as linear

combination of these basis vectors. Specifically, for any vector $\mathbf{x} \in \mathbb{R}^n$, there exist an unique set of coefficients $\{c_i\}_{i=1}^n$ such that

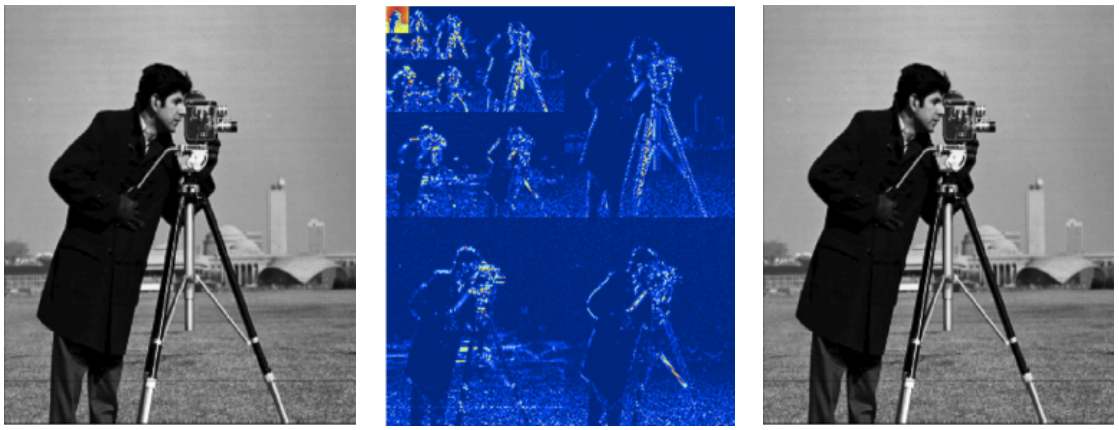
$$\mathbf{x} = \sum_{i=1}^{n} c_i \psi_i. \tag{3.1}$$

Note that if we let $\mathbf{\Psi} \in \mathbb{R}^{n \times n}$ denote the $n \times n$ matrix of real values whose columns are formed by ψ_i and let vector $\mathbf{c} \in \mathbb{R}^n$ denote the vector whose components are c_i, then we can represent the above relation more compactly as $\mathbf{x} = \mathbf{\Psi c}$.

A vector $\mathbf{x} \in \mathbb{R}^n$, is said to be s-sparse in the basis $\mathbf{\Psi}$ if its associated coefficient vector $\mathbf{c} \in \mathbb{R}^n$ has at most s components and s is much smaller than the length of the vector $\mathbf{x}$. (that is, $s << n$). We will refer to $\mathbf{x}$ as being s-sparse, with the understanding that we can express $\mathbf{x}$ as $\mathbf{x} = \mathbf{\Psi c}$ where $\|\mathbf{c}\|_0 \leq s$. We define the set of all s-sparse vectors as

$$\Sigma_s := \{\mathbf{x} : \|\mathbf{x}\|_0 \leq s\}.$$

Sparsity has been exploited heavily in image processing tasks. Most natural images are characterized by larger smooth or textured regions and relatively few sharp edges; see Figure 3.1a. Signals with this structure are very sparse when represented using a multiscale wavelet transform. Figure 3.1b is the multiscale wavelet transform of the image in Figure 3.1a. The multiscale wavelet transform consists of recursively dividing the image into low- and high-frequency components. The lower frequency components provide a coarse-scale approximation of the image. While the higher frequency components fill in the details and resolves edges. Most components are tiny, represented by dark pixels, and only a few components are high, represented by light pixels. This clearly, indicates the sparse representation of the original image on the wavelet basis. We obtain a good approximation of the original image by setting components below a certain threshold to zero, thus, leaving high-frequency components. Figure 3.1c is a reconstruction of the original image by keeping only 10% of randomly selected high-frequency components of the multiscale wavelet transform.



(a) Original image (b) Wavelet representation (c) Reconstruction

FIGURE 3.1: Sparse representaion of an image using a multiscale wavelet transform.

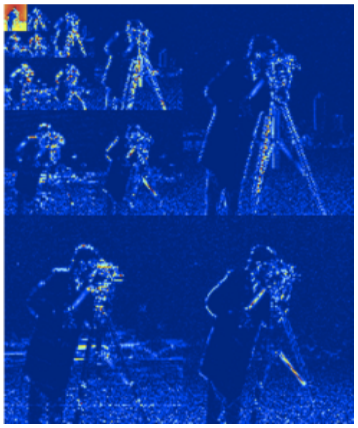
3.1.2 Compressible vectors

In practice, signals are often not exactly sparse. Instead, they are compressible; see Figure 3.2. For example, consider the natural image represented on a multiscale wavelet basis, Figure 3.2a. Most components are very small − close to zero, and few are very high. Figure 3.2b is a histogram of the magnitude of wavelet components sorted in descending order. The components of $\mathbf{x}$ obey the power-law decay. There exist constants $C, q > 0$ such that $|x_i| \leq Ci^{-q}$, for all i. The large the value of q, the faster the magnitude of the signal decay and the more compressible the signal. Because the magnitude of the components decays rapidly, compressible signals can be represented accurately by sparse vector. We can obtain a good approximation by retaining s largest components and setting all other components to zero. The remaining large components can represent the original signal with hardly noticeable distortion. Figure 3.2c is a reconstruction of the original image from only 10% of the largest multiscale wavelet components.

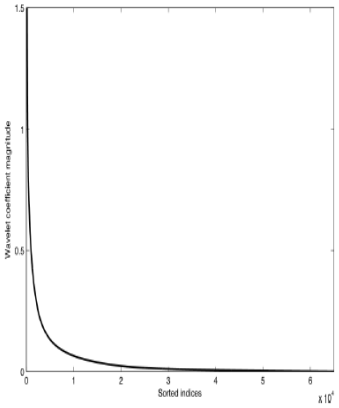
Alternatively, the compressibility of a signal can be quantified by calculating the error incurred by the best s-term approximation as

$$\sigma_s(\mathbf{x})_p = \min_{\mathbf{z} \in \Sigma_s} \|\mathbf{x} - \mathbf{z}\|_p. \tag{3.2}$$

If $\mathbf{x} \in \Sigma_s$, then clearly $\sigma_s(\mathbf{x})_p = 0$ for any ℓ_p-norm. Moreover, the thresholding strategy by keeping s largest components results in the optimal best s-term approximation as measured by (3.2) for all the ℓ_p-norm.



(a) Wavelet representation



(b) Wavelet magnitude

(c) Reconstruction

FIGURE 3.2: Compressible representaion of an image using a multiscale wavelet transform.

3.1.3 Sparse recovery

We define the standard sparse recovery problem for compressed sensing. We start with the acquisition of the signal.

Signal acquisition

We consider the mechanism in which the information is obtained. Instead of acquiring the full set of signal samples of length n, we consider the measurement system that acquire m linear measurements. The measurement y_i, is obtained by correlating the signal $\mathbf{x} \in \mathbb{R}^n$ with a set of sensing vectors $\{\phi_i\}_1^m$, that is,

$$y_i = \langle \mathbf{x}, \phi_i \rangle \quad \text{for} \quad i = 1, \ldots, m \quad \text{or equivalently} \quad \mathbf{y} = \mathbf{\Phi x}. \tag{3.3}$$

Here $\mathbf{y} \in \mathbb{R}^m$ is the measurement vector, and $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ is called the measurement matrix with the vectors ϕ_i^T as its rows. For standard compressed sensing, we assume that the measurements are non-adaptive, meaning that the rows of $\mathbf{\Phi}$ are fixed in advance and do not depend on the previously acquired measurements.

The major challenge associated with compressed sensing is in the case where m is typically much smaller than n. One needs to solve an underdetermined system of linear equations in order to recover the original signal $\mathbf{x}$ from a few measurements $\mathbf{y}$. The basic linear algebraic theory asserts that solving such a solution will produce infinitely many solutions. It is necessary to impose the constraints on the candidate solutions to identify the desirable

solution. The constraint of interest in this regard is sparsity. Thus the problem state as: find sparse vector $\mathbf{x}$ such that $\mathbf{y} = \mathbf{\Phi x}$.

There are two main theoretical questions in compressed sensing. First, how should we design the sensing matrix $\mathbf{\Phi}$ to ensure that it preserves the information in the signal $\mathbf{x}$? Second, how can we recover the original signal $\mathbf{x}$ from measurements $\mathbf{y}$? When the signal is sparse or compressible, it is possible to design a measurement matrix that ensures the recovery of the signal accurately and efficiently using a variety of practical algorithms.

Signal recovery

Based on the knowledge of the measurement matrix $\mathbf{\Phi} \in \mathbb{R}^n$ and measurement vector $\mathbf{y} \in \mathbb{R}^m$ such that equation (3.3) is satisfied, the task of compressed sensing is to find a sparse vector $\mathbf{x} \in \mathbb{R}^n$. Using the knowledge that $\mathbf{x}$ is sparse, a unique solution $\mathbf{x}^*$ can be obtained by posing the the problem as ℓ_0-minimization problem of the form:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{\Phi x}. \tag{3.4}$$

The ℓ_0-norm counts the number on nonzero components of $\mathbf{x}$. Thus, this formulation is equivalent to finding all possible solutions to the problem. The search for a solution of (3.4) by trying all possible combinations is computationally intractable even for a medium-sized problem. The ℓ_0-minimization problem has been declared NP-hard. An alternative has been proposed in the literature, and it can obtain a solution similar to the ℓ_0-norm minimization in near polynomial time.

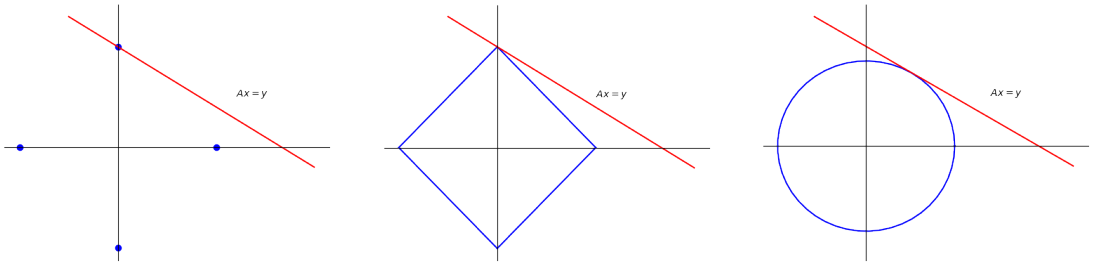
Candès *et.al* [1] and Donoho [2] provided the mathematical foundation to overcome the drawback of ℓ_0-minimization problem. Their profound idea was to pose the problem as an ℓ_1-norm minimisation problem of the form:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{\Phi x}. \tag{3.5}$$

The ℓ_1-norm of $\mathbf{x} \in \mathbb{R}^n$ is defined as $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$. The ℓ_1-norm enforce sparsity on the solution. The formulation (3.5) is a convex optimization programm and it is usually called *Basis Pursuit* in literature [6]. Algorithms for solving problem (3.5) are computationally

tractable. The main challenge becomes the nondifferentiability of the objective function due to the absolute value function.

It is possible to used a differentiable norm, such as ℓ_2-norm. However, such norm does not promote sparsity - an important property of interest, see Figure 2.1. The sparse solution in $\mathbb{R}^2$ must lie of on one of the coordinate axes. We observe that for ℓ_2-minimization the error is spread out evenly among the two components. While for ℓ_p-minimization, where $p = 0, 1$ the error is more unevenly distributed and tends to be sparse.



(a) ℓ_0-minimization (b) ℓ_1-minimization (c) ℓ_2-minimization

FIGURE 3.3: A two dimensional depiction of the contour of ℓ_p-norm for $p = 0, 1, 2$ along with the constriants $\mathbf{y} = \mathbf{\Phi x}$.

3.2 Characteristics of measurement matrices

The sparsity is a necessary but not a sufficient for finding a unique sparse solution of (3.5). Sufficient conditions that guarantee successful recovery of sparse solution from an underdetermined linear algebraic system are associated with the design of the measurement matrix $\mathbf{\Phi}$. Rather than presenting a straightforward design procedure, we instead consider some properties that the measurement matrix must have to guarantee the recovery of a unique sparse solution.

3.2.1 Null space conditions

When designing the measurement matrix, the natural place to begin is by considering the null space of $\mathbf{\Phi}$.

Definition 3.1 (Null space)**.** *Consider the matrix* $\mathbf{\Phi}^{m\times n}$ *where* $m < n$*. The null space of the matrix* $\mathbf{\Phi}$ *(denoted by* $\mathcal{N}(\mathbf{\Phi})$*) is defined by*

$$\mathcal{N}(\mathbf{\Phi}) = \{\mathbf{x} : \mathbf{\Phi x} = \mathbf{0}\}$$

Using this definition, we establish the first simple condition on $\mathbf{\Phi}$ necessary to recover sparse vector from underdetermined linear measurements. In order to recover all s-sparse vectors form the linear measurements (3.3), we must have $\mathbf{\Phi x} \neq \mathbf{\Phi z}$ for any pair of vectors $\mathbf{x}, \mathbf{z} \in \Sigma_s$. Otherwise, it will be impossible to distinguish $\mathbf{x}$ from $\mathbf{z}$ based only on the measurements $\mathbf{y}$. Note that if $\mathbf{\Phi x} = \mathbf{\Phi z}$ then $\mathbf{\Phi}(\mathbf{x} - \mathbf{z}) = \mathbf{0}$ with $\mathbf{x} - \mathbf{z} \in \Sigma_{2s}$. It follows that $\mathbf{\Phi}$ uniquely represents all vectors $\mathbf{x} \in \Sigma_s$ if and only if $\mathcal{N}(\mathbf{\Phi})$ contains no vectors in Σ_{2s}. While there are many equivalent ways of characterizing this property the most common is known as the *spark* [34].

Definition 3.2 (Spark)**.** *The spark of a matrix,* $\mathbf{\Phi}$ *(denoted by* $spark(\mathbf{\Phi})$*) is the smallest subset of columns of* $\mathbf{\Phi}$ *that are linearly dependent.*

The spark sets a fundamental limit of the sparsity of vectors that can be recovered uniquely from linear measurements.

Theorem 3.3. *For any vector* $\mathbf{y} \in \mathbb{R}^m$*, there exists at most one vector* $\mathbf{x} \in \Sigma_s$ *such that* $\mathbf{y} = \mathbf{\Phi x}$ *if and only if* $spark(\mathbf{\Phi}) > 2s$*.*

When dealing with exactly sparse vectors, the spark provides a complete characterization of when sparse recovery is possible. However, when dealing with approximately sparse vectors, we must consider more restrictive conditions on the null space of the measurement matrics $\mathbf{\Phi}$ [35]. We must ensure that the null space of $\mathbf{\Phi}$ does not contain vectors that are too compressible in addition to sparse vectors. The following notation will aid in the definition of null space property. Suppose that $\mathcal{S} \subset \{1, \ldots, n\}$ is a subset of indices and let $\mathcal{S}_c$ be the complementary set of $\mathcal{S}$. By $\mathbf{x}_\mathcal{S}$ we typically mean a vector of length n obtained by setting the components of $\mathbf{x}$ indexed by $\mathcal{S}_c$ to zero, that is

$$\mathbf{x}_\mathcal{S} = \begin{cases} x_i & \text{if } i \in \mathcal{S} \\ 0 & \text{otherwise .} \end{cases}$$

Similarly, by $\mathbf{\Phi}_{\mathcal{S}}$ we typically mean the $m \times n$ matrix obtained by setting the columns of $\mathbf{\Phi}$ indexed by $\mathcal{S}_c$ to zero. The notations $\mathbf{x}_{|s}$ sets all components of $\mathbf{x}$ to zero excepts s largest components.

Definition 3.4 (Null Space Property)**.** *A matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ satisfies the null space property of order s if there exists a constant $C > 0$ such that,*

$$\|\mathbf{x}_{\mathcal{S}}\|_2 \leq \frac{C}{\sqrt{s}} \|\mathbf{x}_{\mathcal{S}_c}\|_1 \tag{3.6}$$

holds for all $\mathbf{x} \in \mathcal{N}(\mathbf{\Phi})$ and for all $\mathcal{S}$ such that $|\mathcal{S}| \leq s$.

The null space property quantifies the notion that vectors in the null space of $\mathbf{\Phi}$ should not be too concentrated on a small subset of indices. For example, if a vector $\mathbf{x}$ is exactly s-sparse, then there exits an $\mathcal{S}$ such that $\|\mathbf{x}_{\mathcal{S}_c}\|_1 = 0$ and hence (3.6) implies that $\mathbf{x}_{\mathcal{S}} = \mathbf{0}$ as well. Thus, if $\mathbf{\Phi}$ satisfies the null space property then the only s-sparse vector in $\mathcal{N}(\mathbf{\Phi})$ is $\mathbf{x} = \mathbf{0}$.

The null space property provides a necessary and sufficient condition for the following recovery guarantee for sparse vectors.

Theorem 3.5. *Every s-sparse vector $\mathbf{x}^*$ is a unique minimizer of the basis pursuit problem*

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \qquad \textit{such that} \qquad \mathbf{y} = \mathbf{\Phi}\mathbf{x},$$

with $\mathbf{y} = \mathbf{\Phi}\mathbf{x}^*$ if and only if $\mathbf{\Phi}$ satisfies the null space property of order s.

The theorem presents an invaluable results regarding the necessary and sufficient conditions for a matrix $\mathbf{\Phi}$ to recover sparse vector from underdetermined linear measurements. However, the theorem makes no statement regarding the existence of such matrix. As it turns out, constructing deterministic matrices which directly satisfy the null space property constitutes a highly nontrivial problem. In fact, even verifying whether a given matrix satisfies the null space property was eventually shown to be an NP-hard decision problem [36]. Fortunately, it can be shown that matrices satisfying the null space property still exits in abundance in the from of random matrices. While it is possible to directly establish the existence of such matrices probabilistically, it has become common practice in commpressed sensing to

consider an alternative property of measurements matrices to establish recovery guarantees. The property is the restricted isometry property (RIP) [37].

3.2.2 Restricted isometry property

The null space property provides necessary and sufficient conditions for establishing recovery guarantees of approximately sparse vectors. These guarantees do not account for noise. Candès and Tau [37] introduced the *restricted isometry property* and showed in [38] that it allows robust recovery of approximately sparse vectors in the presence of measurement noise. The restricted isometry property is defined as follows:

Definition 3.6. *(Restricted Isometry Property) The matrix $\mathbf{\Phi}$ satisfies the restricted isometry property of order s if there exists a constant $\delta \in (0,1)$, such that any s-sparse vector $\mathbf{x} \in \mathbb{R}^n$, the following condition holds:*

$$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2. \tag{3.7}$$

If this is the case, the matrix $\mathbf{\Phi}$ is also said to satisfy the RIP with parameters (s,δ).

When the measurement matrix $\mathbf{\Phi}$ satisfies the restricted isometry property of order $2s$, then the information about s-sparse vectors is approximately preserved by taking $\mathbf{\Phi}\mathbf{x}$. It also implies that no two s-sparse vectors can be mapped to the exact measurement vector $\mathbf{y}$ through $\mathbf{\Phi}$.

In the following, we consider noisy measurements such that $\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{e}$, where the additive noise term $\mathbf{e} \in \mathbb{R}^m$ is assumed to be bounded as $\|\mathbf{e}\|_2 \leq \varepsilon$. Under the assumption of the restricted isometry property, the following results are established regarding stable and robust recovery of approximately sprase vector from noise measurements [39].

Theorem 3.7. *Let $\mathbf{\Phi} \in \mathbb{R}^{m\times n}$ be a matrix satisfying the restricted isometry property of order $2s$ with restricted isometry constant δ_{2s}. For $\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \varepsilon$, let $\mathbf{x}^*$ be the solution of the quadratically-constrained basis pursuit problem*

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_1 \quad \textit{subject to} \quad \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2 \leq \varepsilon.$$

Then for $|\mathcal{S}| \leq s$ we have error bounds

$$\|\mathbf{x} - \mathbf{x}^*\|_1 \leq C_1 \sigma_s(\mathbf{x})_1 + C_2 \sqrt{s}\varepsilon,$$

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \frac{C_1}{\sqrt{s}} \sigma_s(\mathbf{x})_1 + C_2 \varepsilon,$$

where C_1 and C_2 are constants depending on δ_{2s} and $\sigma_s(\mathbf{x})_p$ is defined by (3.2).

These results are both stable with respect to sparsity defect and robust against additive noise as the error bounds only depend on the model mismatch quantified by the best s-term approximation error of $\mathbf{x}$ as well as on the extrinsic noise level ε. In the case of exact s-sparsity of $\mathbf{x}$, and the absence of measurements noise, the theorem immediately implies perfect recovery.

In addition to being used to establish recovery results, the restricted isometry property yields several bounds on the eigenvalues of certain submatrices of $\mathbf{\Phi}^T\mathbf{\Phi}$ and will be useful in the analysis of dynamical systems for sparse recovery. The following properties relating to the restricted isometry property are required in this research work to prove the main result. The properties are adopted from [12].

Lemma 3.8. *Suppose the matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ satisfy the restricted isometry property with parameters (s, δ). Let $\mathcal{S}$ be a set of s indices or fewer. Then for all $\mathbf{x} \in \mathbb{R}^n$ supported on $\mathcal{S}$ and for all $\mathbf{y} \in \mathbb{R}^m$, the following holds:*

1. $\left\|\mathbf{\Phi}_{\mathcal{S}}^T \mathbf{y}\right\|_2 \leq \sqrt{1+\delta}\|\mathbf{y}\|_2$,
2. $(1-\delta)\|\mathbf{x}\|_2 \leq \|\mathbf{\Phi}_{\mathcal{S}}^T \mathbf{\Phi}_{\mathcal{S}} \mathbf{x}\|_2 \leq (1+\delta)\|\mathbf{x}\|_2$,
3. $\frac{1}{1+\delta}\|\mathbf{x}\|_2 \leq \left\|\left(\mathbf{\Phi}_{\mathcal{S}}^T \mathbf{\Phi}_{\mathcal{S}}\right)^{-1} \mathbf{x}\right\|_2 \leq \frac{1}{1-\delta}\|\mathbf{x}\|_2$,
4. $\frac{1}{\sqrt{1+\delta}}\|\mathbf{y}\|_2 \leq \left\|\left(\mathbf{\Phi}_{\mathcal{S}}^T \mathbf{\Phi}_{\mathcal{S}}\right)^{-1} \mathbf{\Phi}_{\mathcal{S}}^T \mathbf{y}\right\| \leq \frac{1}{\sqrt{1-\delta}}\|\mathbf{y}\|_2$.

Proof. The restricted isometry property implies that the singular values of $\mathbf{\Phi}_{\mathcal{S}}$ lie between $\sqrt{1-\delta}$ and $\sqrt{1+\delta}$. Thus, the first inequality readily follows from this fact.

Consider the reduced form of singular value decomposition of $\mathbf{\Phi}_{\mathcal{S}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ is $m \times |\mathcal{S}|$ matrix of orthogonal columns, $\mathbf{\Sigma}$ is $|\mathcal{S}| \times |\mathcal{S}|$ diagonal matrix whose entries are

singular values of $\mathbf{\Phi}_{\mathcal{S}}$, and $\mathbf{V}$ is $|\mathcal{S}| \times n$ unitary matrix. It is easy to show that singular values of $\mathbf{\Phi}_{\mathcal{S}}^T \mathbf{\Phi}_{\mathcal{S}} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$ lie between $1-\delta$ and $1+\delta$. Thus, the prove of the second inequalities follows from this fact.

The above fact also implies that singular values of $\left(\mathbf{\Phi}_{\mathcal{S}}^T \mathbf{\Phi}_{\mathcal{S}}\right)^{-1} = \mathbf{V}\left(\mathbf{\Sigma}^2\right)^{-1}\mathbf{V}^T$ lie between $(1+\delta)^{-1}$ and $(1-\delta)^{-1}$. Thus, this fact entails the prove of third inequalities.

Finally, the singular values of $\left(\mathbf{\Phi}_{\mathcal{S}}^T \mathbf{\Phi}_{\mathcal{S}}\right)^{-1}\mathbf{\Phi}_{\mathcal{S}}^T = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$ lie between the $(\sqrt{1+\delta})^{-1}$ and $(1-\delta)^{-1}$. Thus, the prove of the fourth inequalities follows from this fact.

□

The following provides slightly more complicated consequences of the restricted isometry property that involves two subsets of indices which are not necessarily disjoint, [12]

Lemma 3.9. *Suppose the matrix $\mathbf{\Phi} \in \mathbb{R}^{m\times n}$ satisfy the restricted isometry property with parameters $(s+q, \delta)$. Let $\mathcal{S}_1$ contains less than q indices and $\mathcal{S}_2$ contain less that s indices. Then for every $\mathbf{x} \in \mathbb{R}^n$ supported on $\mathcal{S}_1 \cup \mathcal{S}_2$, the following holds:*

1. $\left\|\mathbf{\Phi}_{\mathcal{S}_1}^T \mathbf{\Phi}_{\mathcal{S}_1^c \cup \mathcal{S}_2}\mathbf{x}\right\|_2 \leq \delta\|\mathbf{x}\|_2$
2. $\left\|\left(\mathbf{I}_{\mathcal{S}_1} - \mathbf{\Phi}_{\mathcal{S}_1}^T \mathbf{\Phi}_{\mathcal{S}_1\cup\mathcal{S}_2}\right)\mathbf{x}\right\|_2 \leq \delta\|\mathbf{x}\|_2$

Proof. Since $\mathcal{S}_1 \cup \mathcal{S}_2$ contains less than $s+q$ indices, the restricted isomery property implies that the eigenvalues of the product matrix $\mathbf{\Phi}_{\mathcal{S}_1\cup\mathcal{S}_2}^T \mathbf{\Phi}_{\mathcal{S}_1\cup\mathcal{S}_2}$ lies between $(1-\delta)$ and $1+\delta$. Let $\lambda(\cdot)$ denote the eigenvalues of the a given matrix. The eigenvalues of the following matrix

$$\mathbf{G}_{\mathcal{S}_1\cup\mathcal{S}_2} = \mathbf{I}_{\mathcal{S}_1\cup\mathcal{S}_2} - \mathbf{\Phi}_{\mathcal{S}_1\cup\mathcal{S}_2}^T \mathbf{\Phi}_{\mathcal{S}_1\cup\mathcal{S}_2}$$

can be deduced as:

$$\lambda(\mathbf{G}_{\mathcal{S}_1\cup\mathcal{S}_2}) \leq \max\{1-(1-\delta), -1+(1+\delta)\} = \delta$$

and

$$\lambda(\mathbf{G}_{\mathcal{S}_1\cup\mathcal{S}_2}) \geq \min\{1-(1+\delta), -1+(1-\delta)\} = -\delta$$

The matrix $\mathbf{\Phi}_{\mathcal{S}_1}^T \mathbf{\Phi}_{\mathcal{S}_1^c \cup \mathcal{S}_2}$ and $\left(\mathbf{I}_{\mathcal{S}_1} - \mathbf{\Phi}_{\mathcal{S}_1}^T \mathbf{\Phi}_{\mathcal{S}_1 \cup \mathcal{S}_2}\right)$ are submatrix of $\mathbf{G}_{\mathcal{S}_1 \cup \mathcal{S}_2}$, in the sense that

$$\mathbf{\Phi}_{\mathcal{S}_1}^T \mathbf{\Phi}_{\mathcal{S}_1^c \cup \mathcal{S}_2} = \prod_{\mathcal{S}_1} \mathbf{G}_{\mathcal{S}_1 \cup \mathcal{S}_2}$$

and

$$\left(\mathbf{I}_{\mathcal{S}_1} - \mathbf{\Phi}_{\mathcal{S}_1}^T \mathbf{\Phi}_{\mathcal{S}_1 \cup \mathcal{S}_2}\right) = \prod_{\mathcal{S}_1} \mathbf{G}_{\mathcal{S}_1 \cup \mathcal{S}_2}$$

where $\prod_{\mathcal{S}_i}$ is the projection onto set of indices $\mathcal{S}_i$. The operator norm of the projection $\prod_i$ is 1. Thus, the operator norms of the two matrices $\mathbf{\Phi}_{\mathcal{S}_1}^T \mathbf{\Phi}_{\mathcal{S}_1^c \cap \mathcal{S}_2}$ and $\left(\mathbf{I}_{\mathcal{S}_1} - \mathbf{\Phi}_{\mathcal{S}_1}^T \mathbf{\Phi}_{\mathcal{S}_1 \cup \mathcal{S}_2}\right)$ are bounded by the operator norm of the larger matrix $\mathbf{G}_{\mathcal{S}_1 \cup \mathcal{S}_2}$, which is its largest eigenvalue and it is equal to δ. Hence, prove of the two inequalities follows. □

3.2.3 Mutual coherence

Although both null space and restricted isometry property allow for the derivation of strong results in terms of stability and robustness of general recovery algorithms, verifying that a general matrix $\mathbf{\Phi}$ satisfies any of these properties in practice remain an NP-hard problem [36]. In many cases, it is preferable to use properties of a measurement matrix $\mathbf{\Phi}$ that is easily computable to provide more concrete recovery guarantees. The *coherence* of a matrix provide such property [34].

Definition 3.10 (Mutual Coherence)**.** *The coherence (denoted $\mu(\mathbf{\Phi})$) of the matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ is defined as*

$$\mu(\mathbf{\Phi}) = \max_{1 \leq i \neq j \leq n} \frac{|\langle \phi_i, \phi_j \rangle|}{\|\phi_i\|_2 \|\phi_j\|_2}$$

where ϕ_i is the i-th column of the matrix $\mathbf{\Phi}$.

Coherence measure the absolute largest correlation between any two columns of $\mathbf{\Phi}$. It is possible to show from linear algebra that the range of coherence is bounded by $\mu(\mathbf{\Phi}) \in \left[\sqrt{\frac{n-m}{m(n-1)}}, 1\right]$. If two columns of $\mathbf{\Phi}$ are correlated then coherence is large. Otherwise, coherence is small and the columns are said to be incoherent. Compressed sensing is mainly concerned with low coherence between the columns of $\mathbf{\Phi}$.

Unfortunately, coherence-based analyses are rather pessimist in terms of the number of measurements required to establish stable and recovery guarantees. We will use restricted isometry property for the analysis in this work.

3.2.4 Measurements bounds

Some classes of matrices are known to satisfy the restricted isometry property with high probability [1, 38]. In particular, if $\mathbf{\Phi} \in \mathbb{R}^{m\times n}$ is a random matrix whose columns ϕ_i are independent subgaussian random vectors with $\|\phi_i\|_2 = 1$, then for any sparsity level $1 < s < n$ and $\delta \in (0, 1)$, the matrix $\mathbf{\Phi}$ satisfies the RIP with parameters (s, δ) with high probability, provided

$$m \geq C\frac{s}{\delta^2}\log(n/s), \tag{3.8}$$

where log denotes the natural logarithmic and inequality hold up to some scaling factor C [38]. In practice, it is unknown how to determine the restricted isometry property constant for any given matrix in polynomial time. However, with algebraic manipulation the inequality (3.8) can be rearranged to get the following estimate:

$$\delta \approx C\sqrt{\frac{s\log(n/s)}{m}}, \tag{3.9}$$

for some constant C. This estimate is often used to estimate the number of measurements necessary for a discrete sparse optimization algorithms.

3.3 Algorithmic approaches for sparse recovery

The second fundamental aspect of compressed sensing is the design of efficient and effective algorithms. Given the noisy compressible measurements $\mathbf{y} = \mathbf{\Phi x} + \mathbf{e}$ of a signal $\mathbf{x}$, a core problem in compressed sensing is to recovery a sparse signal $\mathbf{x}$ from a set of measurements $\mathbf{y}$. Considerable efforts have been made towards developing algorithms that perform fast, accurate and stable recovery of $\mathbf{x}$ from $\mathbf{y}$. A good compressed sensing measurement matrix

typically satisfies certain conditions, such as restricted isometry property (RIP). Practical algorithms exploit this fact in various ways to reduce the number of measurements, enable faster recovery and ensure robustness to noise.

The development of sparse recovery algorithms are guided by various criteria. Some important ones are: minimal number of measurements, robustness to noisy measurements, speed, and performance guarantees. Various algorithms satisfying some of the above requirements have been proposed in the literature. There are multiple criteria by which sparse recovery algorithms are divided. The most generic classification usually considers three distinct categories: convex optimization-based methods, greedy methods and iterative thresholding methods. We review some examples and basic algorithmic properties of each mentioned category.

3.3.1 Convex optimization-based methods

This approach poses the compressed sensing problem as a convex optimization program, the ℓ_1-minimization - which can be then solved using techniques from linear programming. While the optimization problem can be solved using general-purpose convex optimization software, various algorithmic techniques exist to solve the problem in the context of compressed sensing. There are multiple equivalent formulations of the problem.

As usual, we model the measurements process of a perfectly sparse or compressible vector $\mathbf{x} \in \mathbb{R}^n$ via the measurement $\mathbf{y} = \mathbf{\Phi x} + \mathbf{e}$ where $\mathbf{e} \in \mathbb{R}^m$ is the additive noise and bounded as $\|\mathbf{e}\|_2 \leq \varepsilon$ with $\varepsilon \geq 0$. If the upper bound of the error term $\mathbf{e}$ is known, we naturally consider the quadratically constrained basis pursuit problem as discussed before

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{x}\|_1 \quad \text{subject to } \|\mathbf{y} - \mathbf{\Phi x}\|_2^2 \leq \varepsilon. \tag{3.10}$$

For perfect measurements, without noise, $\varepsilon = 0$, and this problem immediately reduces to the original basis pursuit problem (3.5).

We characterized the recovery behaviour of this problem when we were discussing the restricted isometry property. If $\mathbf{x} \in \mathbb{R}^n$ is approximately sparse, we obtain the following characterization for minimizers $\mathbf{x}^*$ of problem (3.10): if $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ satisfies the restricted isomerty property

of order $2s$, then we have

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \frac{C_1}{\sqrt{k}}\sigma_s(\mathbf{x})_1 + C_2\varepsilon, \tag{3.11}$$

where C_1 and C_2 are constants depending on δ_{2s}. This results implies prefect recovery in the case where the measurements are strictly s-sparse vectors in a noise free case.

The following unconstrained problem is a prevalent variant of the quadratically constrained basis pursuit program

$$\min_{\mathbf{x}\in\mathbb{R}^{\mathbf{n}}} \mu\|\mathbf{x}\|_1 + \|\mathbf{\Phi x} - \mathbf{y}\|_2^2 \tag{3.12}$$

where $\mu > 0$ is the regularization parameter, and the problem is referred to as *basis pursuit denoising* (BBDN). The BPDN problem is exciting in situations where no reasonable estimation of noise level ε is available. In this case, we can use the regularization parameter μ to control the trade-off between sparsity and data fidelity.

Another important formulation which was originally proposed in the context of sparse model selection in statistics is the *least-absolute shrinkage selection operator* (LASSO):

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{\Phi x} - \mathbf{y}\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}\|_1 \leq \varepsilon_2, \tag{3.13}$$

for some parameter $\varepsilon_2 > 0$. Since the ℓ_1-norm operates as a sparsity enforcing function, this formulation might be of special interest in the case where the sparsity level can be well estimated rather than the noise level.

The last approach we discuss is the program known as the *Dantzig selector*:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject ot} \quad \|\mathbf{\Phi}^T(\mathbf{\Phi x} - \mathbf{y})\|_\infty \leq \varepsilon_3, \tag{3.14}$$

for some parameter $\varepsilon_3 > 0$. The key idea here is to impose a maximum tolerance on the worst case correlation between the residual $\mathbf{r} = \mathbf{\Phi x} - \mathbf{y}$ and the columns of measurement

matrix $\mathbf{\Phi}$. In the extreme case the parameter $\varepsilon_3 = 0$, the Dantzig selector reduces to the original basis pursuit problem.

Conveniently, despite the difference in formulations used, the problem BPDN (3.12), LASSO (3.13) and Dantzig slector (3.14) all share the some recovery guarantee from equation (3.11) up to the nonlinear transformation of the parameters μ, ε_1, and ε_2 [39]. While the Dantzig selector is the odd one out, similar guarantees can still be derived [7].

The above mentioned problems can be solved using a variety of related methods such as alternating direction method of multipliers (ADMM), forward-backward splitting, Douglass-Rachford splitting, homotopy methods or fixed point methods, we refer the reader to a survey [40] for more algorithms.

3.3.2 Greedy methods

While convex optimization techniques are potent methods for computing sparse representations, various greedy methods also solve such problems. Greedy methods rely on an iterative approximation of the signal support and coefficients by following a metaheuristic of choosing the best immediate local solution. In these methods, an important feature is associated with active columns of the measurement matrix. At the optimal solution, active columns correspond with nonzero components of the sparse signal $\mathbf{x}$ and the measurements $\mathbf{y}$ are merely a weighted linear combination of active columns of the measurement matrix. In each iteration, the support set is updated by selecting one or more columns of the measurement matrix that are highly correlated in some way with the measurements. Once the support set is known, usually, a least squared subproblem restricted to the support set is solved to approximate the components of $\mathbf{x}$.

In literature, there are two main categorizations of greedy methods. In the first categorization, methods are either *serial* and *parallel*. In the serial implementation, only one column is selected at each iteration. While in parallel implementation, multiple columns can be selected simultaneously at each iteration. In the second categorization, methods are either *irreversible* or reversible. In the irreversible implementation, once a component is added to the support set, it remains there until the algorithm terminates. In contrast, the reversible

implementation allows modification of the support set as the algorithm progresses. Early greedy methods tend to implement a serial-irreversible strategy, while contemporary algorithms tend to implement a parallel-reversible strategy. In the latter approach, greedy methods can actually be shown to have performance guarantees that match those obtained for convex relaxation methods.

Orthogonal Matching Pursuit.

While technically a successor to the lesser used matching pursuit algorithm [10], orthogonal matching pursuit (OMP) remains one of the most popular greedy algorithm due to the fact that it is one of the methods with the lowest computational complexity. Consider the Algorithm 3.1, at the k-th iteration, OMP updates its estimated support set one column at a time by identifying the column ϕ_i that exhibits the strongest correlation with the residual $\mathbf{r}^k = \mathbf{\Phi}\mathbf{x}^k - \mathbf{y}$ as measured by the inner product $|\langle \phi_i, \mathbf{r}^k \rangle|$.

Algorithm 3.1 Orthogonal Matching Pursuit

1: **Input:** $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$, and $\mathbf{y} \in \mathbb{R}^m$
2: **Initialization:** $\mathbf{x}^0 = \mathbf{0}$, $\mathbf{r}^0 = \mathbf{y}$, and $\mathcal{S}^0$.
3: **for** $k+ = 1$ till stopping criterion is met **do**
4: $\quad j^{k+1} = \arg\max_j |(\mathbf{\Phi}^T \mathbf{r}^k)_j|$ { Selection of column}
5: $\quad \mathcal{S}^{k+1} = \mathcal{S}^k \cup \{j^{k+1}\}$ { Support extension }
6: $\quad \mathbf{x}^{k+1} = \arg\min_{x_{\mathcal{S}^{k+1}}} \|\mathbf{\Phi}_{\mathcal{S}^{k+1}} \mathbf{x}_{\mathcal{S}^{k+1}} - \mathbf{y}\|_2^2$ { Least squares projection}
7: $\quad \mathbf{r}^{k+1} = \mathbf{\Phi}\mathbf{x}^{k+1} - \mathbf{y}$ { Calculation of residual}
8: **end for**
9: **return** $\mathbf{x}^k$.

The column selection step in each OMP iteration can be interpreted as identifying the component of $\mathbf{x}^k$ with respect to which the function $f(\mathbf{x}^k) = \frac{1}{2}\|\mathbf{\Phi}\mathbf{x}^k - \mathbf{y}\|_2^2$ varies the most. This is due to the fact that the gradient of f at $\mathbf{x}^k$ reads $\nabla f(\mathbf{x}^k) = \mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{x}^k - \mathbf{y}) = \mathbf{\Phi}^T\mathbf{r}^k$. The update step from $\mathbf{x}^k$ to $\mathbf{x}^{k+1}$ corresponds to a projection of $\mathbf{y}$ on the subspace spanned by the the columns of $\mathbf{\Phi}$ indexed by the current support set $\mathcal{S}^{k+1}$.

In general, OMP does not require an estimate of the sparsity level of the vector one aims to recover. The algorithm terminates as soon as the same column of measurement matrix $\mathbf{\Phi}$ is selected twice in subsequent iterations. Other stopping conditions include the relative

change of estimates $\mathbf{x}^k$ between iterations and tolerance criteria of data fidelity measures with respect to the residual $\mathbf{r}^k$. Considering OMP updates the support set one index at the time per iteration, OMP require at least s iterations for find s-sparse candidate vector. If the sparsity level is known prior, it can be used as stopping condition to set the allowed number of iterations.

While theoretical guarantees in the noise free and exactly sparse case exist for OMP, robust and stable recovery guarantees are not as well developed as one might expect given the maturity of the theory and the popularity of OMP in general. One of the earliest recovery guarantees for OMP was the coherence based condition $(2s-1)\mu(\mathbf{\Phi}) < 1$ which allows OMP to recover any s-sparse vector from noiseless linear measurements in s iterations [41]. Currently, one of the best known sufficient conditions for exact s-sparse recovery in the noise free setting, in terms of restricted isomerty property of order s, requires $\delta < 1/\sqrt{s+1}$ [42]

Compressive Sampling Matching Pursuit.

The compressive sampling matching pursuit (CoSaMP) algorithm shares a lot of similarity with OMP algorithm [12]. However, it have better recovery guarantees. The procedure is summarised in Algorithm 3.2. Given the current estimation of the solution $\mathbf{x}^k$, CoSaMP proceeds by first identifying the $2s$ columns of measurement matrix $\mathbf{\Phi}$ which best correlate with the residual $\mathbf{r}^k = \mathbf{\Phi}\mathbf{x}^k - \mathbf{y}$ at the k-th iteration. The algorithm then continue to solve a least squares problem with respect to columns submatrix defined by the support of $\mathbf{x}^k$ and the $2s$ columns indices selected in the previous step. Since the algorithm ultimately aims to obtain strictly s-sparse solution, the next estimate of the solution $\mathbf{x}^{k+1}$ is found via a hard thresholding of the least squares update. Solving the least squares problem over a column index set of size at most $3s$ effectively allows CoSaMP to adaptively correct previous choices of the support set of its estimate solution. This is one of the main drawbacks of the OMP algorithm, which will never remove a previously selected column.

Algorithm 3.2 Compressive Sampling Matching Pursuit

1: **Input:** $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$, $\mathbf{y} \in \mathbb{R}^m$, and s.
2: **Initialization:** $\mathbf{x}^0 = \mathbf{0}$, $\mathbf{r}^0 = \mathbf{y}$, and $\mathcal{S}^0$.
3: **for** $k+ = 1$ till stopping criterion is met **do**
4: $\quad \mathbf{z} = (\mathbf{\Phi}^T \mathbf{r}^k)_{|2s}$
5: $\quad \mathcal{S}^{k+1} = \text{supp}(\mathbf{x}^k) \cup \text{supp}(\mathbf{z})$ { Support merge }
6: $\quad \mathbf{v}^{k+1} = \arg\min_{x_{\mathcal{S}^{k+1}}} \|\mathbf{\Phi}_{\mathcal{S}^{k+1}} \mathbf{x}_{\mathcal{S}^{k+1}} - \mathbf{y}\|_2^2$ { Least squares projection}
7: $\quad \mathbf{x}^{k+1} = \mathbf{v}^{k+1}_{|s}$ { Hard threshold}
8: $\quad \mathbf{r}^{k+1} = \mathbf{\Phi}\mathbf{x}^{k+1} - \mathbf{y}$ { Calculation of residual}
9: **end for**
10: **return** $\mathbf{x}^k$.

3.3.3 Thresholding algorithms

Convex optimization based recovery procedures provide the strongest recovery guarantees. However, they become less practical as the problem size increase. The thresholding algorithm category presents an attractive compromise between solid theoretical guarantees and highly efficient and predictable running times.

Thresholding algorithms are generally divided into so-called *hard* and *soft* thresholding algorithms. In the following, we review the most popular representatives from each class, namely iterative hard thresholding (IHT) and hard thresholding pursuit (HTP) for the former, and iterative soft thresholding algorithm (ISTA) and the fast iterative soft thresholding algorithm (FISTA) for the latter. Other popular thresholding based algorithms includes subspace pursuit [11], NESTA [43], SpaRSA [44] and many more.

Hard Thresholding.

At the heart of any hard threshold algorithm lies the hard threholding operator $H_s : \mathbb{R}^n \to \Sigma_s$ defined as

$$H_s(\mathbf{x}) = \arg\min_{\mathbf{z} \in \Sigma_s} \|\mathbf{x} - \mathbf{z}\|_p ,$$

where $p \geq 1$ which projects an arbitrary n dimensional vector on the set of s-sparse vector. The value of H_s is constructed by identifying the index set $\mathcal{S}$ of size $|\mathcal{S}| = s$ which supports the largest magnitude of $\mathbf{x}$ and setting any other values to zero. In other words, the vector $H_s(\mathbf{x})$ achieves the best s-term approximation error $\sigma_s(\mathbf{x})_p$ for any $p \geq 1$. For convenience, we also define the set valued operator L_s, defined by the support set of the best s-term approximation of $\mathbf{x} \in \mathbb{R}^n$, that is $L_s(\mathbf{x}) = \text{supp}(H_s(\mathbf{x}))$, which results the support set of s largest components of $\mathbf{x}$.

Using the above definition, we describe the first hard thresholding algorithm namely the *iteratiative hard thresholding* (IHT). The key idea of IHT is to reduce the smooth function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{\Phi x} - \mathbf{y}\|_2^2$ with gradient $\nabla f(\mathbf{x}) = \mathbf{\Phi}^T(\mathbf{\Phi x} - \mathbf{y})$ at every iteration by mean of gradient descent update before pruning the solution to the set of s-sparse vector by mean of hard thresholding function operator. A full description of the algorithm is summarised in Algorithm 3.3

Algorithm 3.3 Iterative Hard Thresholding

1: **Input:** $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$, and $\mathbf{y} \in \mathbb{R}^m$
2: **Initialization:** $\mathbf{x}^0 = \mathbf{0}$.
3: **for** $k+ = 1$ till stopping criterion is met **do**
4: $\quad \mathbf{z}^{k+1} = \mathbf{x}^k - \mathbf{\Phi}^T\left(\mathbf{\Phi x}^k - \mathbf{y}\right)$ { Gradient descent step}
5: $\quad \mathbf{x}^{k+1} = H_s(\mathbf{z}^{k+1})$ {Hard thresholding }
6: **end for**
7: **return** $\mathbf{x}^k$.

The following results from [39], demonstrate performance guarantees of IHT in terms of robustness to sparsity defect and stability with respect to measurement noise. Consider and arbitrary vector $\mathbf{x} \in \mathbb{R}^n$ which is measured according to the model $\mathbf{y} = \mathbf{\Phi x} + \mathbf{e}$. If the measurement matrix satisfies RIP of order $6s$ with constant $\delta < 1/\sqrt{3}$, Algorithm 3.3 produce a sequence of iterates $\{\mathbf{x}^k\}_{k\geq 0}$ that satisfies

$$\|\mathbf{x}^k - \mathbf{x}^*\|_2 \leq 2\rho^k\|\mathbf{x}^*\|_2 + C_1 s^{-1/2}\sigma_s(\mathbf{x}^*)_1 + C_1\|\mathbf{e}\|_2,$$

where C_1, $C_2 > 0$ and $\rho \in (0, 1)$ are constants which only depend on RIP constant δ. For $k \to \infty$, this sequence converges to a cluster of points $\mathbf{x}^*$ satisfying

$$\|\mathbf{x}^k - \mathbf{x}^*\|_2 \leq C_1 s^{-1/2} \sigma_s(\mathbf{x}^*)_1 + C_2 \|\mathbf{e}\|_2.$$

If the vector we wish to recovery is in reality supported on an index set $\mathbf{S}$ of size s, and the measurements are not distubed by noise, $\mathbf{e} = 0$, one has $\sigma_s(\mathbf{x})_1 = 0$, this imply perfect recovery.

The second hard thresholding algorithm is *hard thresholding pursuit* (HTP). The fundamental difference between IHT and HTP is the fact that HTP merely uses hard thresholding gradient descent updates to estimate the support set of $\mathbf{x}^*$. In particular, it propagates least-squares solutions of $\mathbf{y} = \mathbf{\Phi x}$ with respect to a submatrix of $\mathbf{\Phi}$ obtained by pursuing the active support set of coefficients in each iteration based on the operator $L_s = \text{supp}(H_s)$. A description of the full algorithm is summarised in Algorithm 3.4

Algorithm 3.4 Hard Thresholding Pursuit

Input: $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$, and $\mathbf{y} \in \mathbb{R}^m$

Initialization: $\mathbf{x}^0 = \mathbf{0}$.

for $k+ = 1$ till stopping criterion is met **do**

$\mathbf{z}^{k+1} = \mathbf{x}^k - \mathbf{\Phi}^T \left(\mathbf{\Phi x}^k - \mathbf{y}\right)$ { Gradient descent step}

$\mathcal{S}^{k+1} = L_s(\mathbf{z}^{k+1})$ { Support identification}

$\mathbf{x}^{k+1} = H_s(\mathbf{z}^{k+1})$ {Hard thresholding }

$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}_{\mathcal{S}^{k+1}}} \|\mathbf{\Phi}_{\mathcal{S}^{k+1}} \mathbf{x}_{\mathcal{S}^{k+1}} - \mathbf{y}\|_2^2$ { Least squares projection}

end for

return $\mathbf{x}^k$.

Surprisingly, the stability and robustness analysis are identical for IHT and HTP barring the change of parameters (C_1, C_2, ρ) for HTP. Most importantly, this change results in a faster rate of convergence for the HTP algorithm [39].

Soft Threshoding.

The soft thresholding methods promote sparsity by incorporating an ℓ_1-norm in their objective functions, and applying the proximal gradient algorithm [45] or it variant. In particular, soft

thresholding algorithms solves the unconstrained optimization problem of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}), \tag{3.15}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function and $g : \mathbb{R}^n \to \mathbb{R}$ is nonsmooth convex function. If g were smooth, this problem could be solved by standard optimization tools such as conjugate gradient methods or Newton's method. However, in order to promote sparsity we often choose $g(\mathbf{x}) = \mu\|\mathbf{x}\|_1$, thus standard approach is not applicable due to nonsmoothness of the function.

In the proximal gradient method, the smooth part, f, of the objective function (3.15) is replaced by it's second-order Taylor approximation. That is, we consider an iterative procedure of the form

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ g(\mathbf{x}) + \hat{f}(\mathbf{x}, \mathbf{x}^k) \right\},$$

where for some parameter $\alpha > 0$, the second-order approximation of f is given by

$$\hat{f}\left(\mathbf{x}, \mathbf{x}^k\right) = f(x) + \left\langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \right\rangle + \frac{1}{2\alpha} \left\| \mathbf{x} - \mathbf{x}^k \right\|_2^2. \tag{3.16}$$

Hence, it is easy to verify that the expression for $\mathbf{x}^{k+1}$ can be rewritten as

$$\begin{aligned} \mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}} g(\mathbf{x}) + f(\mathbf{x}^k) + \left\langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \right\rangle + \frac{1}{2\alpha} \left\| \mathbf{x} - \mathbf{x}^k \right\|_2^2, \\ &= \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\alpha} \left\| \mathbf{x} - \left(\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k) \right) \right\|_2^2. \end{aligned} \tag{3.17}$$

While this formulation might give the expression that we merely trade one difficult optimization problem for another, it turns out that the result in equation (3.17) corresponds to the so-called proximal operator [45]

$$\text{prox}_{\alpha g}(\mathbf{x}^k) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\alpha} \left\| \mathbf{x} - \mathbf{z}^k \right\|_2^2, \tag{3.18}$$

applied to the gradient descent update $\mathbf{z}^k = \mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k)$. Conveniently, the proximal operator (3.18) has a close-form solution for a variety of nonsmooth functions g. In particular,

for $g(x) = \mu\|\mathbf{x}\|$, it is easy to check via subdifferential calculus over its individual components that $\text{prox}_{\alpha\mu\|\cdot\|_1}(\mathbf{x}) = S_{\alpha\mu}(\mathbf{x})$ where

$$S_\mu(x_i) = \begin{cases} \text{sign}(x_i)(|x_i| - \mu), & \text{if } |x_i| > \mu, \\ 0, & \text{if } |x_i| \leq \mu, \end{cases}$$

is the so-called *shrinkage operator* (or soft-thresholding operator) that is applied component wise to $\mathbf{x}$. Overall, equation (3.17) reduce to the following iteration

$$\mathbf{x}^{k+1} = S_{\alpha\mu}(\mathbf{x}^k - \alpha_k \mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{x} - \mathbf{y})), \tag{3.19}$$

if we apply this method to the Basis Pursuit Denoising problem (3.12). In this particular formulation, the parameter α acts as a stepsize which we may choose via backtracking line-search, while $\mu > 0$ can be used to control the trade-off between sparsity of the solution $\mathbf{x}^*$ and the data fidelity term $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{\Phi}\mathbf{x} - \mathbf{y}\|_2^2$.

This algorithm requires $\mathcal{O}(1/\varepsilon)$ iterations to achive $\left|f(\mathbf{x}^*) - f(\mathbf{x}^k)\right| \leq \varepsilon$ accuracy, implying convergence rate of $\mathcal{O}(1/k)$ [9]. According to a celebrated results by Nesterov [15], the best achievable convergence rate in the class of nonsmooth first-order methods is $\mathcal{O}(1/k^2)$. This rate is achievable by Nesterov's acceleration method, resulting in the well-known *fast iterative soft thresholding algorithm*(FISTA) due to Beck and Teboulle when applied to the iterative soft thresholding algorithm [9]. The key idea of FISTA is to add a momentum like term depending on the the last two iterations to avoid erratic changes in the search direction. One updates the iterates according to

$$\mathbf{z}^{k+1} = \mathbf{x}^k + \frac{k-2}{k+1}\left(\mathbf{x}^k - \mathbf{x}^{k-1}\right),$$

$$\mathbf{x}^{k+1} = S_{\alpha_k\mu}(\mathbf{z}^{k+1} - \alpha_k \mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{x} - \mathbf{y})),$$

where $\alpha_k > 0$ is the stepsize at iteration k.

3.4 Sparse recovery via dynamical system

In this section, we present the existing research on sparse recovery using dynamical systems. We start with the connection between the dynamical systems and optimization problems.

A sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ generated by an iterative algorithm may be considered as a discrete dynamical system. The step length often establishes the connection between the dynamics system and numerical optimization. Taking the step length to be very small, the solution path of the iterative algorithm converges to a curve modelled by an ordinary differential equation (ODE). For example, in the case of gradient descent method applied to an unconstrained optimization problem $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$, one can inspect that

$$\lim_{\alpha\to 0}\left\{\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha\nabla f(\mathbf{x}^k)\right\} \approx \dot{\mathbf{x}}(t) = -\nabla f(\mathbf{x}(t)), \tag{3.20}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function, $\dot{\mathbf{x}} = \frac{d\mathbf{x}}{dt}$ denotes the time parameter derivative. The time parameter in the ODE is related to the step length via $t = k\alpha$, where k is the iteration index. There is a long history relating dynamical systems to optimization, see Courant [46], Schropp [47], and Helmke and Moore [48]. The motivation to this line of research is due to well-established analysis tools in continuous dynamical systems described by differential equations.

This research associates the sparse recovery problem with a dynamical system. The aim is to derive an efficient iterative algorithm from the resulting dynamical system. The dynamical system approach in optimization has many advantages. It provides deep insights into the expected behaviour of the method. In addition, sometimes the techniques used in the continuous case can be adapted to obtain results for the discrete algorithm. On the other hand, a continuous dynamical system satisfying excellent theoretical properties may suggest a new iterative scheme.

3.4.1 Hopfield network

A network is a collection of connected units acting together to perform a specific task. Hopfield [16] demonstrated that a highly interconnected network of nonlinear analogue

processors could efficiently perform computational tasks. These computational tasks were initially related to storing and retrieving embedded memories modelled using a network of binary-state processors and a stochastic updating algorithm. Later, Hopfield [17] expanded the idea by proposing a continuous version. Hopfield proved that convergence of the dynamics associated with the network minimizes a Lyapunov function. Over time, it became clear that Hopfield networks can minimize any function provided the network parameters are appropriately set. We present two Hopfield networks: the original discrete model which is stochastic [16] and the continuous model which is deterministic [17].

Discrete Hopfield network

Given an optimization problem on n variables, the corresponding Hopfield network [16] made up of a fully interconnected system of n computational nodes, see Figure 3.4. Each network node i is characterized by the internal state variable denoted by u_i, and the output denoted by x_i. The internal state variable is used to describe the dynamical behaviour of the node. The output describes the stimuli of the node, which will be communicated to all other connected nodes. The original Hopfield network used binary-state node output. The output takes the value of 0 for the inhabitation of the node and 1 to signal excitation of the node. Note that the output also corresponds to the original optimization variable. Thus, the original Hopfield network was proposed to solve discrete problems. In particular, it was proposed for the computational task of associative memory.

The dynamical behaviour of each node is derived as follows. Each node's inputs come from two sources, the external input I_i associated with the environment, and the input due to the weighted sum of stimulation from all other connected nodes in the network excluding itself. The total input of node i is then

$$\text{Input of node } i = \sum_{j \neq i} W_{ij} x_j + I_i, \tag{3.21}$$

where W_{ij} describe the synaptic interconnection strength between node i and j. The computations in the network involve the change of state of nodes with time. The motion of the nodes in the state space describes the computation that a set of nodes is performing. A network must describe how the state evolves in time, and the original Hopfield network describes this in terms of stochastic evolution. Each node samples its input at random. At

time instance t_k, the time behaviour of the network is described using (3.21), by the following dynamical model:

$$\begin{cases} u_i(t_{k+1}) & = \sum_{j \neq i} W_{ij} x_j(t_k) + I_i \\ x_i(t_{k+1}) & = B_\mu(u_i(t_{k+1})). \end{cases} \tag{3.22}$$

Here B_μ is the binary activation function which changes the value of the output or leaves it fixed according to the threshold rule with the threshold μ

$$B_\mu(u_i(t_{k+1})) = \begin{cases} 1 & \text{if } u_i(t_{k+1}) > \mu, \\ 0 & \text{if } u_i(t_{k+1}) \leq \mu. \end{cases} \tag{3.23}$$

The activation function stipulates that a node will be excited if it receives sufficient stimulation from its connecting nodes, and the corresponding output will be one. Otherwise, the node will be inhibited if the stimulation is insufficient, and the corresponding output will be zero.

Convergence flow to a stable state is the essential feature of the Hopfield network. Hopfield showed that if the weight matrix W is symmetric with zero diagonal entries (that is $W_{ij} = W_{ji}$ and $W_{ii} = 0$), the dynamical system associated with the network is guaranteed to converge to a stable equilibrium state. In particular, the proof of this property follows from the construction of an appropriate energy function that is always nonincreasing by any state change produced by the dynamical system. Consider the energy function

$$E = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} W_{ij} x_i x_j - \sum_{i=1}^{n} I_i x_i + \sum_{i=1}^{n} \mu x_i. \tag{3.24}$$

We consider the change in one node at a time and taking node i, we derive the change in energy $\Delta E = E(t_{k+1}) - E(t_k)$ as follows:

$$\Delta E = -\left[\sum_{j \neq i} W_{ij} x_j + I_i - \mu \right] \Delta x_i, \tag{3.25}$$

where $\Delta x_i = x_i(t_{k+1}) - x_i(t_k)$.

According the the dynamical system (3.22), Δx_i is positive only when the bracket is positive, similar for the negative case. Thus, any change in E under the dynamical system (3.22) is

negative. The energy function E is bounded, so the evolution of the dynamical system must lead to a stable state. The stable state of the network correspond to the local minimum of the energy function (3.24).

The continuous deterministic Hopfield Network [17] was recently adopted to solve sparse optimization problems by Rozell *et.al* [27]. The full description is give in the next section.

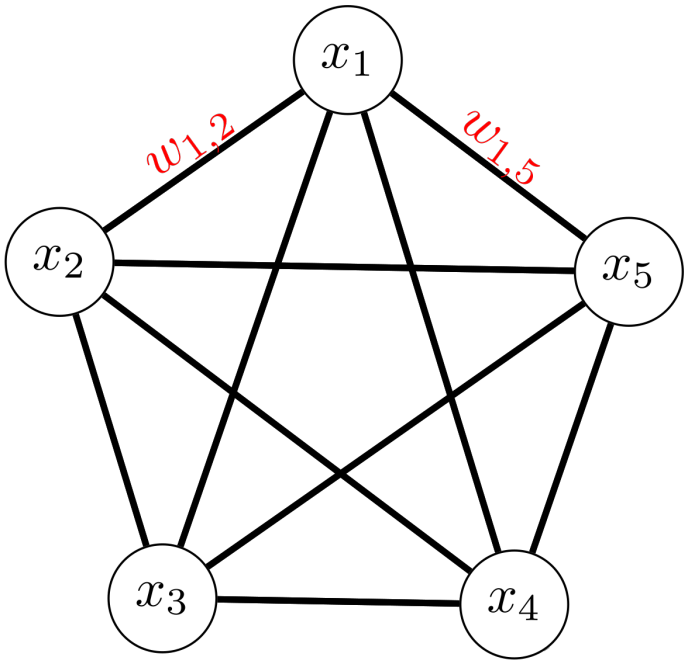


FIGURE 3.4: The schematic representation of Hopfield network.

.

3.4.2 Locally competitive algorithm

The Local Competitive Algorithm (LCA) proposed initially by Rozell *et al.* [27] is a continuous-time dynamical system designed to solve sparse recovery problems. LCA is an instance of Hopfield network that operates on a network of nodes, where each node competes with neighbouring nodes for a chance to represent the signal. The first-order ordinary differential equation models the evolution of nodes. The steady-state of the differential equation represents a solution to the optimization problem. The LCA is a specific instance of the Hopfield network that has a long history of being used to solve optimization problems [16].

The LCA network is a collection of n nodes representing an optimization problem of n variables. The i-th node of the LCA network (see Figure 3.5) is associated with ϕ_i, the

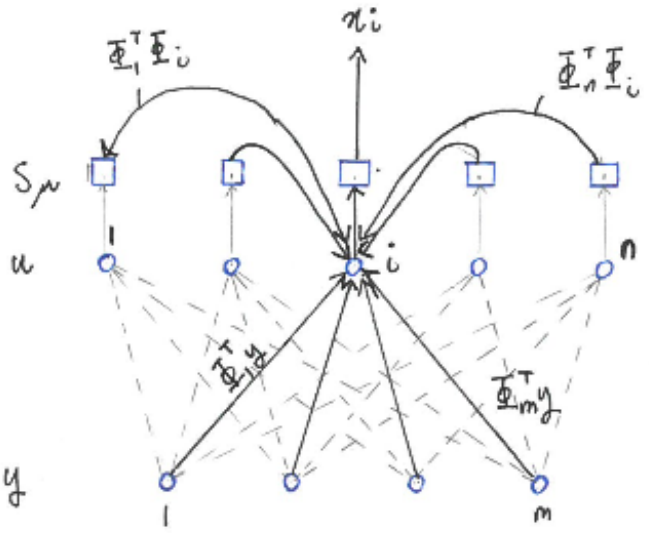


FIGURE 3.5: The schematic representation of LCA network, a Hopfield type network for solving sparse approximation problems

.

i-th column of the measurement matrix $\mathbf{\Phi}$. Without loss of generality, we assume that columns of $\mathbf{\Phi}$ are normalized to have a unit norm. This node is described at a given time t by an internal *state variable* $u_i(t)$. The state variable produces the *output* $x_i(t)$ through the activation function $S_\mu(u_i)$. The goal is to design a dynamical system associated with the LCA network such that few outputs have non-zero values while minimizing the energy function that requires both faithful and efficient representation of the signal:

$$F(\mathbf{x}, t) = \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^{n} \phi_j x_j \right\|_2^2 + \mu \sum_{i=1}^{n} R_i(x_i) \tag{3.26}$$

where $R_i : \mathbb{R} \to \mathbb{R}$ is referred to as sparsity penalty function and is chosen to enforce sparsity requirement on the solution. Equation (3.26) resembles equation (3.12) for $R_i = |x_i|$. Next we derive the LCA dynamical system associated with the LCA network.

Following the gradient flow, we take the derivative of the energy function (3.26) with respect to the individual node activity, thus the gradient descent:

$$-\frac{\partial F(\mathbf{x}, t)}{\partial x_i} = \langle \phi_i, \mathbf{y} \rangle - \sum_{j \neq i}^{n} \langle \phi_i, \phi_j \rangle x_j - x_i - \mu \frac{\partial R_i(x_i)}{\partial x_i}. \tag{3.27}$$

We have used the assumption that the columns of the measurement matrix have unit norm, that is $\langle \phi_i, \phi_i \rangle = 1$. Regarding the network, see Figure (3.5), the right hand side of (3.27)

can be interpreted as follows: Given the input vector $\mathbf{y}$, individual nodes are driven by the excitation input force proportional to $\langle \phi_i, \mathbf{y} \rangle$. This indicates the connection strength between the input $\mathbf{y}$ and the columns of Φ. The stronger the similarity between the input and the column ϕ_i, the larger the driving excitation force. This causes certain nodes to be active much quicker than others. The lateral connections between nodes characterize the competition that allows high active nodes to suppress less active nodes. The active nodes inhibit other nodes with an inhibition signal propositional to both their activity level and the similarity of the node's respective fields. Specifically, magnitude of inhibition from the active node j to any other node i is propositional to $\langle \phi_i, \phi_j \rangle x_j$. Self-inhibition is not allowed. We define the $f_\mu : \mathbb{R} \to \mathbb{R}$ to characterize the self-inhibition:

$$f_\mu(x_i) = x_i + \mu \frac{\partial R(x_i)}{\partial x_i}. \tag{3.28}$$

Self-inhibition imposes sparsity by penalizing the nodes own output, which is different from the sparsity inducing input from other laterally connected nodes. Hence, the partial derivative of energy function (3.27) becomes:

$$-\frac{\partial F(\mathbf{x}, t)}{\partial x_i} = \langle \phi_i, \mathbf{y} \rangle - \sum_{j \neq i}^{n} \langle \phi_i, \phi_j \rangle x_j - f_\mu(x_i). \tag{3.29}$$

At this point we could update $\mathbf{x}(t)$ using the gradient descent following (3.20) to produce a sparse solution from the given input. However, a more energy efficient solution will be to have the nodes maintain some internal state. Only a few nodes will be active since inhibition gives strong nodes a chance to prevent weak nodes from becoming active. A node is set to produce output when it exceeds some threshold. Following this logic, Rozel *et al.* [27] define an internal state variable u_i that represents potential for node i at time t. When a node's potential climbs above some threshold, it communicates in the form of an activation $x_i(t)$. We make the rate of change of internal state to be proportional to the gradient descent: $\dot{u}(t) \propto -\frac{\partial F(t)}{\partial x_i}$. Thus, the state dynamics is governed by the differential equation:

$$\dot{u}_i(t) = \frac{1}{\tau} \left(\langle \phi_i, \mathbf{y} \rangle - f_\mu(x_i(t) - \sum_{j \neq i}^{n} \langle \phi_i, \phi_j \rangle x_j(t) \right), \tag{3.30}$$

where the parameter τ represents the time constant of the system implementing the differential equation and it is characterized by the physical properties of the system. Since it does not affect the mathematical analysis of the LCA, it is assumed to be $\tau = 1$ except when its influence on convergence speed is made explicit.

In order to have a complete description of the model, we need to describe a relationship between the state variable u_i and the output x_i. The function $f_\mu(x_i)$ is described as self-inhibition that promote sparsity on the output. We assign the internal state, $u_i(t)$ to this function:

$$u_i(t) = f_\mu(x_i(t)). \tag{3.31}$$

Hence, we can then invert the function to get our node output activity:

$$x_i(t) = f_\mu^{-1}(u_i(t)) := S_\mu(u_i(t)), \tag{3.32}$$

where S_μ is the activation function and it will be derived shortly. Hence, this gives the LCA nodes update equation:

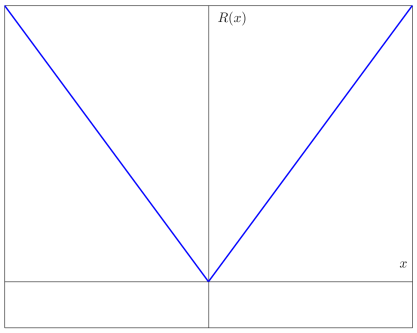
$$\begin{cases} \dot{u}_i(t) &= \langle \phi_i, \mathbf{y} \rangle - u_i(t) - \sum_{j \neq i}^{n} \langle \phi_i, \phi_j \rangle \, x_j(t), \\ x_i(t) &= S_\mu(u_i(t)). \end{cases} \tag{3.33}$$

The nodes become active when they exceed some threshold defined by $x_i(t) = S_\mu(u_i(t))$ - the thresholding function. The thresholding function must be monotonically increasing to grauntee gradient descent of the energy function (3.26) [27]. It can take various forms determined by the nature of the sparsity penalty function R_i. Our interest is on the ℓ_1-norm, $R_i(x_i) = |x_i|$, which gives the *soft-thresholding* function:

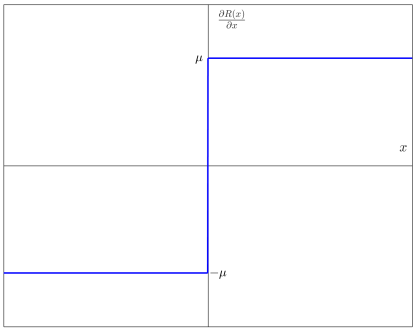
$$S_\mu(u_i) = \begin{cases} 0 & \text{for } |u_i| \leq \mu. \\ u_i - \mu \cdot \text{sign}(u_i) & \text{for } |u_i| > \mu. \end{cases} \tag{3.34}$$

The parameter, $\mu > 0$ determines the threshold that the node must exceed in order to be active. An illustration of how the threshodling function (3.34) is obtained from ℓ_1-norm

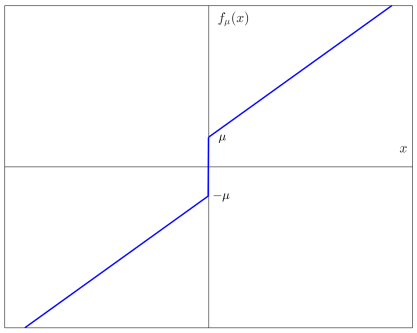
penalty function is shown in Figure 3.6.



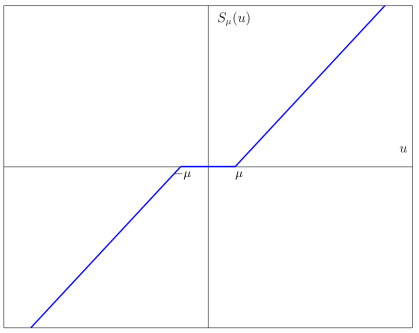
(a) The spareness penalty: $R(x_i) = |x_i|$.



(b) Subdifferential of the penalty: $\frac{\partial R(x_i)}{\partial x_i}$



(c) Leak integrator: $f_\mu(x_i) = x_i + \frac{\partial R(x_i)}{\partial x_i}$



(d) Soft-thresholding function: $S_\mu(u_i)$

FIGURE 3.6: Derivation of the LCA soft-thresholding function

.

In compact vector form, the evolution of the state variable with respect to time is governed by a set of coupled nonlinear ordinary differential equations of the form:

$$\begin{cases} \tau\dot{\mathbf{u}}(t) & = -\mathbf{u}(t) - (\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\mathbf{x}(t) + \mathbf{\Phi}^T\mathbf{y}, \\ \mathbf{x}(t) & = S_\mu(\mathbf{u}(t)). \end{cases} \tag{3.35}$$

The sparse penalty function is separable, thus the computations of the thresholding function are done element-wise.

3.5 Related work

Using Lyapunov's direct method, Hopfield showed that the dynamical system derived from the Hopfield network converges to a stable equilibrium point that corresponds to the minimum

of the energy function [17]. Later, these ideas naturally led Hopfield to consider the reverse problem. Starting from an objective function to be minimized, Hopfield showed how to construct the network and its associated dynamical system such that it performs the desired computation [18]. Hopfield applied this technique to the travelling salesman problem in [18] and to linear programming in [49]. These were the pioneering steps in using a dynamical system to solve optimization problems, and they paved the way for many extensions. In particular, the locally competitive algorithm [27] descends from this lineage of dynamical systems designed for specific optimization. In this work, we propose a new dynamical system based on a locally competitive algorithm, incorporating several new ideas to accelerate the algorithm.

Unfortunately, not all optimization programs have the necessary properties for Lyapunov's method to apply. Specifically, Hopfield's paper on liner programming restricts the matrix to be symmetric with zeros on the diagonal. The activation function is nondecreasing everywhere. The activation and objection functions must be smooth and accept a derivative everywhere. The need for dynamical systems to solve more complex optimization programs has led researchers to analyze neural networks extending the classic Hopfield network.

To remove the symmetry condition on the interconnection matrix, the authors of [50] proved global asymptotic convergence when the interconnection matrix is lower triangular. In [51], the interconnection matrix can be non-symmetric but must have symmetric and positive semidefinite submatrices. The results in [52] remove the symmetric assumption altogether. However, these results require the activation function to be bounded and strictly increasing. This specification does not suit the case of sparse recovery, whose activation is unbounded and contains a thresholding region where the outputs are precisely zero over some interval.

Although the LCA dynamical system is a Hopfield-type, its objective function does not satisfy the smoothness requirement of the traditional Lyapunov approach. Several articles have considered nonsmooth objective function to extend the dynamical system to more general optimisation classes. Their analysis relies upon the notion of subgradients developed by Clarke [28] and on the theory of differential inclusion as studied by Filippov [53].

Chapter 4

Subgradient Dynamical System for Sparse Recovery

This chapter presents new convergence results of the dynamical system modelled by subgradients of nonsmooth objective function coupled with sparse promoting activation function. The convergence analysis of this gradient-like differential inclusion is done using the recently developed Łojasiewicz gradient inequality for nonsmooth functions. The optimization problem solved using the dynamical system and the related work have been presented in Section 4.1. In Section 4.2, we derive the subgradient dynamical system for sparse recovery. In Section 4.3, we present the existence and uniqueness of the solution of the dynamical system. In Section 4.4, we show that the dynamical system has finite-length trajectories using Łojasiewicz inequality. Thus, the dynamical system converges to an unique solution, even when solutions are not isolated. In Section 4.5, we present an estimate of the convergence rate in terms of Łojasiewic exponents. Finally, we summarise the work done in this chapter in Section 4.6.

4.1 Introduction

Consider the following general nonsmooth convex optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}, \tag{4.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function whose gradient is Lipschitz continuous, and $g : \mathbb{R}^n \to \mathbb{R}$ is a continuous convex function which is possibly nonsmooth. All the functions are assumed to be proper and lower-semicontinuous. Problem formulation (4.1) covers many problems arising in inverse problem, image processing, statistics and machine learning. In this thesis, our interest is on sparse recovery problem arising in compressed sensing. Here, f measure data discrepancy and g enforces sparsity on the solution.

We construct the dynamical system for solving (4.1) using the intuition from the Hopfield network. Overall, the dynamical system is modelled by the subgradient of the nonsmooth objective function - resulting in a differential inclusion. The differential inclusion is coupled with the spare promoting activation function. This approach for solving optimization problems is effective and particularly attractive in applications where it is crucial to obtain the optimal solution in real-time. Before using the dynamical system to solve optimization, it is essential to understand its behaviour. Using the new method based on nonsmooth Łojasiewicz inequality [33], we establish some results on the convergence of trajectories of the system. The results are more substantial than those obtained using Lyapunov's function method.

The contributions of this chapter are the results related to the convergence analysis of the differential inclusion (4.2). First, we show using the Łojasiewicz inequality that each trajectory of the differential inclusion has a finite length. Thus, each trajectory converges to a singleton. Hence, the system is convergent. This property is true when the system has infinitely many nonisolated equilibrium points, such as sparse recovery problems. Using Lyapunov's function method, the system can only be shown to be quasi-convergent. Quasi-convergence is not enough as the system may oscillate near the equilibrium point.

Second, we address the issue of the convergence rate of trajectories generated by the differential inclusion. Knowing how fast the trajectories converge to the equilibrium point is crucial for the excellent design of a dynamic system to solve optimization problems in real-time. Thus, we have provided the estimate of convergence rate in terms of the Łojasiewicz exponents.

4.2 The dynamical system model

The dynamical system for solving (4.1) is constructed using the intuition from the Hopfield network. The system consists of n computational nodes corresponding to n optimization variables. The computational nodes are fully inter-connected. Each node is characterized by time-dependent internal state variable $u_i(t)$ and the output variable $x_i(t)$. The dynamics of the system are described by the internal state $\mathbf{u}(t)$. The time derivative of $\mathbf{u}(t)$ is generated by negative subgradients of the objective function, $\partial F : \mathbb{R}^n \to \mathbb{R}^n$. The output $\mathbf{x}(t)$ is generated by a nonlinear activation function $S_\mu : \mathbb{R}^n \to \mathbb{R}^n$ from the internal variable, $\mathbf{u}(t)$. The dynamical system is a gradient descent-like differential inclusion described mathematically as

$$\begin{cases} \dot{\mathbf{u}}(t) \in -\partial F\left(\mathbf{x}(t)\right) \\ \mathbf{x}(t) = S_\mu(\mathbf{u}(t)). \end{cases} \tag{4.2}$$

We will present the construction and properties of the general activation function S_μ associated with the sparse recovery problems later. First, for a mathematical analysis of the dynamical system (4.2), we make the following standing assumption about the objective function.

Assumption 4.1. *The objective function F satisfy the following conditions:*

1. *F is locally Lipschitz continuous,*
2. *F is regular in $\mathbb{R}^n$,*
3. *F is subanalytic in $\mathbb{R}^n$.*

The Lipschitz continuity and regularity of the objective function are assumed to ensure the existence of subdifferentials. The subanalyticity of the objective function is assumed to allow the easy convergence analysis using nonsmooth Łojasiewicz inequality.

The primary motivation for the study of differential inclusion (4.2) is the need for an efficient solver of sparse recovery problems. The function, g, is utilized to enforce a specific structure

on the solution, and it determines the choice of the activation function in (4.2). Our interest is on sparsity promoting functions. We assume a separable function, that is generally defined as $g(\mathbf{x}) = \mu \sum_{i=1}^{n} R_i(x_i)$, where μ is the regularization parameter and R_i is the regularization function chosen to enforce specific solution structure. The following example outlines various choices of regularisation function R_i and their associated activation functions.

Example 4.1. *The examples of different choices of regularization functions and their corresponding activation functions used to enforce a desired structure on the solution are as follows:*

1. *The ℓ_1-norm, $\|\mathbf{x}\|_1 = \sum_i |x_i|$, is usually a preferred choice to enforce sparsity on the solution. The corresponding activation function is the well-known* soft-thresholding *function*

$$R_i(x_i) = |x_i| \quad \Longleftrightarrow \quad x_i = S_\mu(u_i) = \begin{cases} 0, & |u_i| \leq \mu \\ u_i - \mu \ sign(u_i), & |u_i| > \mu. \end{cases} \tag{4.3}$$

2. *The ℓ_0-norm, $\|\mathbf{x}\|_0 = |\{i : x_i \neq 0\}|$, where $|\cdot|$ is the cardinality of the given set. The corresponding activation function is the* hard-thersholding *function, H_μ,*

$$R_i(x_i) = I(x_i \neq 0) \quad \Longleftrightarrow \quad x_i = H_\mu(u_i) = \begin{cases} 0, & |u_i| \leq \mu \\ u_i, & |u_i| > \mu. \end{cases} \tag{4.4}$$

   where $I(x_i \neq 0)$ is the standard indicator function.

3. *The ℓ_2-norm, $\|\mathbf{x}\|_2 = \sum_i x_i^2$, gives raise to the Tikhonov regularization. The corresponding activation function is known as the Tikhonov-thresholding, T_μ,*

$$R_i(x_i) = x_i^2 \quad \Leftrightarrow \quad x_i = T_\mu(u_i) = \frac{u_i}{1 + 2\mu}. \tag{4.5}$$

From the observations of the above examples, specifically from equation (4.3) and (4.4), we can construct a general activation function for sparse recovery problems. Notice that the activation function must satisfy several requirements to be eligible for solving optimization problems. First, Rozell *et.al* [27] established that the activation function must be nondecreasing everywhere. This property is necessary for the objective function, F, to decrease along

the trajectories of the dynamical system (4.2). Thus, the objective function qualifies as a candidate Lyapunov function for the dynamical system.

Second, the activation function must be continuous to ensure that the continuity assumption on the objective function, F, is still satisfied. This requirement prevents scenarios where the objective function is decreasing for all time but returns to a high value at discontinuity points and may never reach a stable minimum. Based on the observation in conditions (4.3) and (4.4), together with the above mentioned requirements we construct the general activation function for sparse recovery with the followings assumption [54].

Assumption 4.2. *The activation function is locally Lipschitz continuous, odd and non decreasing. Specifically, there exists $\mu > 0$ such that*

$$x(t) = S_\mu(u(t)) = \begin{cases} 0 & |u(t)| \leq \mu \\ h(u) & |u(t)| > \mu \end{cases} \tag{4.6}$$

where the function $h : \mathbb{R} \to \mathbb{R}$ is a real valued function defined on the domain $\mathcal{D} = (-\infty, -\mu) \cup (+\mu, \infty)$, continuous on $\mathcal{D}$, differentiable in the interior of $\mathcal{D}$, and satisfies the following properties;

$$h(-u_i) = -h(u_i), \qquad \forall u_i \in \mathcal{D}, \text{ and } h(\mu) = 0 \tag{4.7}$$

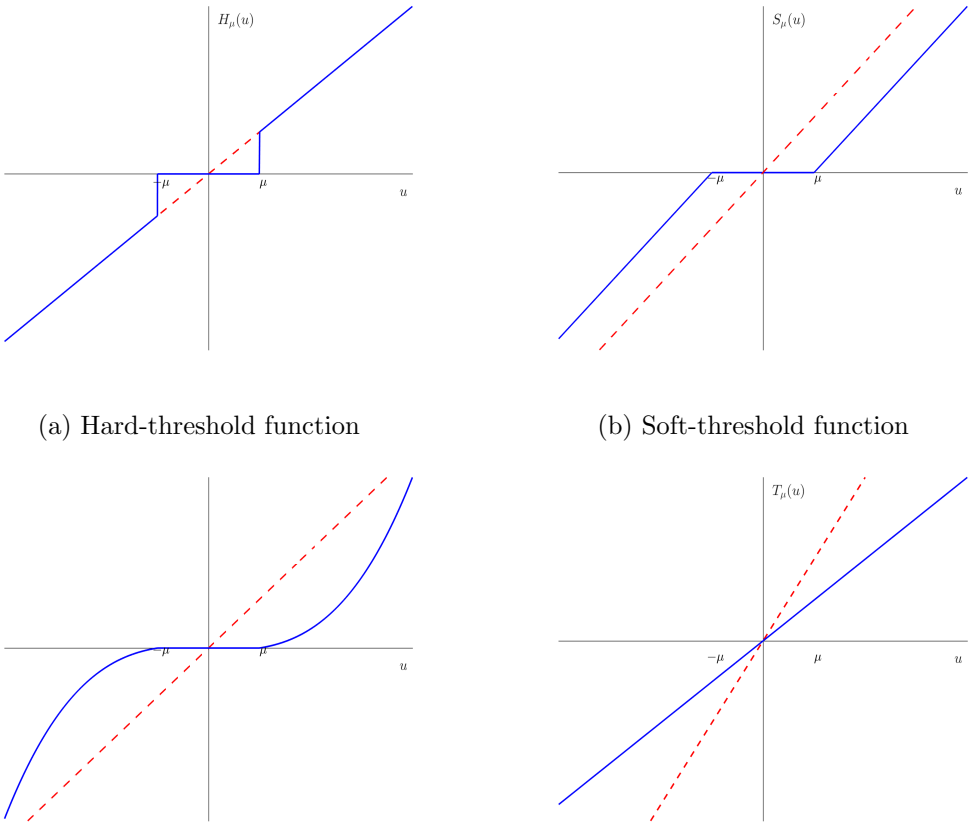
$$h'(u_i) > 0, \qquad \forall u_i \in \mathcal{D} \tag{4.8}$$

$$h(u_i) \leq u_i, \qquad \forall u_i \in \mathcal{D} \text{ such that } u_i \geq 0 \tag{4.9}$$

$$h(u_i) \text{ is subanalytic on } \mathcal{D}. \tag{4.10}$$

A generic activation function satisfying conditions (4.7) - (4.10) of Assumption 4.2 is shown in Figure 4.1. More generally, the activation function $S_\mu(\mathbf{u}(t))$, satisfying Assumption 4.2, corresponds to a large class of sparse promoting functions that are often used in practice [55]. For any $\mu > 0$, it is guaranteed that $S_\mu(u) > 0$ for all $u > \mu$. In the case where $\mu > 0$, the activation function is exactly zero on the interval $[-\mu, +\mu]$. Intuitively, many components

with small amplitude are forced to be zero, thus promoting a sparse output. The case where $\mu = 0$ is less interesting for sparse recovery problems, as it does not yield sparse output.



(a) Hard-threshold function

(b) Soft-threshold function

(c) Generic-threshold function

(d) Tikhonov-threshold function

FIGURE 4.1: Examples of activations functions satisfying conditions (4.7)-(4.10)

Model (4.2) generally encompass the Hopfield networks [16, 17], which includes the dynamical system associated with Locally Competitive Algorithm (LCA) [27],

$$\begin{cases} \dot{\mathbf{u}}(t) &= -\mathbf{u}(t) - (\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\mathbf{x}(t) + \mathbf{\Phi}^T\mathbf{y}, \\ \mathbf{x}(t) &= S_\mu(\mathbf{u}(t)). \end{cases} \tag{4.11}$$

A Lyapunov function can be designed for the above dynamic system. An appropriate Lyapunov function for (4.11) is the typical objective function used to solve sparse approximation problems:

$$F(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \mu\sum_{i=1}^{n} R_i(x_i), \tag{4.12}$$

where $R_i(x_i) = |x_i|$. The first term is the least-squared error of the approximation, while the second term $g(\mathbf{x}) = \sum_{i=1}^{n} R_i(x_i)$ is a cost penalty on the solution chosen to encourage sparsity - we will refer to it as the sparsity penalty function. The parameter μ controls the trade off between the two terms. The most prominent sparse approximation problem is the ℓ_1-minimization, for which the problem (4.12) is known as *Basis Pursuit Denoising*. It is widely used to solve sparse recovery problems for compressed sensing due to its unconstrained nature. The following lemma demonstrates that the dynamical system (4.11) is modelled by (4.2).

Lemma 4.3. *If the activation function $S_\mu : \mathbb{R} \to \mathbb{R}$ in (4.11) and the sparse penalty function R_i satisfy, for all $x_i \in \mathbb{R}$,*

$$u_i - x_i = u_i - S_\mu(u_i) \in \mu \partial R_i(x_i), \tag{4.13}$$

then the trajectories of the dynamic system (4.11) satisfy the differential inclusion (4.2).

Proof. The energy function (4.12) is locally Lipschitz and regular on $\mathbb{R}^n$. Thus, by applying the rules of calculus for subgradients, we get

$$\partial F(\mathbf{x}(t)) = -\mathbf{\Phi}^T\mathbf{y} + \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{x}(t) + \mu \sum_{i}^{n} \partial R_i(x_i(t)).$$

The dynamical system (4.11) together with condition (4.13) yield

$$\begin{aligned} -\dot{\mathbf{u}}(t) &= \mathbf{u}(t) - \mathbf{x}(t) + \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{x}(t) - \mathbf{\Phi}^T\mathbf{y} \\ &\in \mu \sum_{i}^{n} \partial R_i(\mathbf{x}(t)) + \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{x}(t) - \mathbf{\Phi}^T\mathbf{y} \\ &\in \partial F(\mathbf{x}(t). \end{aligned}$$

□

Note that condition (4.13) provides an important aspect of evaluating the subgradient associated with the sparsity penalty function. The condition plays a very crucial role in analysing dynamical systems for solving sparse recovery problems.

4.2.1 Properties of the dynamical model

Based on the activation function (4.6), the computational nodes of the network associated with the dynamical system (4.2) can be split into active nodes and inactive nodes. The active nodes form an *active set* $\mathcal{A}(t)$ which contains indices such that

$$i \in \mathcal{A}(t) \quad \Leftrightarrow \quad |u_i(t)| > \mu \text{ and } |x_i(t)| > 0. \tag{4.14}$$

On the contrary, state variables that satisfy $|u_i(t)| \leq \mu$ generate outputs $x_i(t) = 0$ and the associated nodes are called inactive nodes. Their indices are collected to form the inactive set $\mathcal{A}_c(t)$. The active set and inactive set are time-dependent since the magnitude of the internal state variables, $u_i(t)$ can cross the threshold μ at any time. For readability purposes, we remove the explicit dependence on time in the notation and write the sets as $\mathcal{A}$ and $\mathcal{A}_c$. The following set of lemmas will be useful in proving the convergence results. The discussion follows the works of [54, 56]

Lemma 4.4. *Suppose the objective function $F(\mathbf{x}(t))$ satisfies Assumption 4.1 and the activation function satisfies Assumption 4.2. Then the time derivative of the objective function $F(\mathbf{x}(t))$ satisfies the following equalities*

$$\dot{F}(\mathbf{x}(t)) = -\sum_{i \in \mathcal{A}} h'(u_i)|\dot{u}_i(t)|^2, \tag{4.15}$$

$$\dot{F}(\mathbf{x}(t)) = -\sum_{i \in \mathcal{A}} \frac{1}{h'(u_i)}|\dot{x}_i(t)|^2, \tag{4.16}$$

for all $t \geq 0$. Moreover, $\dot{F}(\mathbf{x}(t)) = 0$ for all $i \in \mathcal{A}_c$.

Proof. Since the objective function is Lipschitz and regular by Assumption 4.1, we can use any element in $\partial F(\mathbf{x}(t))$ to calculate the time derivative of $F(\mathbf{x}(t))$ along the trajectories of the dynamical system (4.2). We split the nodes into the active and inactive nodes. In the case of active nodes, $x_i = h(u_i(t))$. Using the chain rule we get $\dot{x}_i(t) = h'(u_i(t))\dot{u}_i(t)$. Thus,

the results follow as

$$\begin{aligned}\dot{F}(\mathbf{x}(t)) &= \xi^T \dot{\mathbf{x}}(t) \qquad \forall \xi \in \partial F(\mathbf{x}(t)) \\ &= -\dot{\mathbf{u}}(t)^T \dot{\mathbf{x}}(t) \\ &= -\sum_{i=1}^{n} \dot{u}_i(t)\dot{x}_i(t) \\ &= -\sum_{i=1}^{n} h'(u_i(t))|\dot{u}_i(t)|^2 \\ &= -\sum_{i=1}^{n} \frac{1}{h'(u_i(t))}|\dot{x}_i(t)|^2.\end{aligned}$$

In the case of inactive nodes, $x_i = 0$. Thus, it follows that $\dot{F}(\mathbf{x}(t)) = 0$ for all $i \in \mathcal{A}_c$. □

The activation function is non-decreasing as alluded by conditions (4.8) of Assumption 4.2. Hence, Lemma 4.4 demonstrates that the objective function is strictly decreasing on non-stationary trajectories of the differential inclusion (4.2). These are positive and necessary results, indicating that it is possible to use the differential inclusion (4.2) as a tool to solve optimization problems. However, the results are insufficient as the conditions do not guarantee that the objective function will decrease until the optimal solution is reached.

The following Lemma gives some properties of the sparse penalty function.

Lemma 4.5. *Consider the objective function of the form* $F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})$*, where the sparsity penalty function is separable,* $g(\mathbf{x}) = \mu \sum_{i=1}^{n} R_i(x_i)$*. Without loss of generality, assume that* $R_i(0) = 0$*. Then, condition (4.13) and Assumption 4.2 yield the following properties:*

$$R_i(x_i) \geq 0 \quad \textit{and} \quad R_i(x_i) = R_i(-x_i) \quad \forall x_i \in \mathbb{R} \tag{4.17}$$

$$sign(u_i) = sign(x_i) \qquad \forall u_i \in \mathcal{D} \tag{4.18}$$

$$|x_i|^2 \leq x_i u_i \leq |u_i|^2 \qquad \forall u_i \in \mathbb{R}. \tag{4.19}$$

Proof. Since $h : \mathbb{R} \to \mathbb{R}$ is continuous and strictly increasing on $\mathcal{D}$, the inverse $h^{-1} : \mathbb{R} \to \mathbb{R}$ is well-defined and strictly increasing on $h(\mathcal{D})$. In addition, property (4.7) of $h(\cdot)$ implies that $h^{-1}(\cdot)$ satisfies $h^{-1}(-x_i) = -h^{-1}(x_i)$ and $h^{-1}(\mu) = 0$. Thus, using these relations and equation (4.13) we get the following expression

$$\partial R_i(x_i) = u_i - x_i = u_i - h(u_i) = h^{-1}(x_i) - x_i. \tag{4.20}$$

This quantity is positive by (4.9). This proves that $R_i(x_i) \geq R_i(0) = 0$ for all $x_i > 0$. Moreover, for all $x_i > 0$, the following holds:

$$\begin{aligned} R_i(-x_i) &= \int_0^{-x_i} \left(h^{-1}(\eta) - \eta\right) d\eta \\ &= \int_0^{x_i} \left(h^{-1}(-\eta) + \eta\right) (-d\eta) \\ &= \int_0^{x_i} \left(h^{-1}(\eta) - \eta\right) d\eta = R_i(x_i). \end{aligned}$$

So $R_i(-x_i) = R_i(x_i)$ for all $x_i \in \mathbb{R}$ and thus (4.17) holds.

Now, for $u_i > \mu$, it follows from (4.9) that $0 < x_i = h(u_i) \leq u_i$. Multiplying by negative, we get $-u_i \leq -h(u_i) = h(-u_i) = -x_i < 0$. Thus, this proves property (4.18), that is, the state variable, u_i, and the corresponding output, x_i, always have the same sign.

Hence, it follows that $x_i u_i = |x_i||u_i|$, and $|x_i| \leq |u_i|$ for all $u_i \in \mathcal{D}$. The last inequality can be extended to $\mathbb{R}$, since $|u_i| \leq \mu$, $x_i = 0$. Finally, for all $u_i \in \mathbb{R}$, we obtain $|x_i|^2 \leq |x_i||u_i| = x_i u_i \leq |u_i|^2$, which proves property (4.19). □

Note that we could choose any value for $R_i(0)$. In all cases, the objective function F will have a lower bound $\mu n R_i(0)$. A lower bound on F is all that is required in the proof of the main results. Taking $R(0) = 0$ simplifies the lower bound to $F(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. Using these properties, the following lemma states that the objective function is also upper-bounded for all time, and so are the state and output variables.

Lemma 4.6. *Suppose that the objective function $F(\mathbf{x}(t))$ satisfies Assumption 4.1 and the activation function satisfies Assumption 4.2. Then $F(\mathbf{x}(t)) \leq F(\mathbf{x}(0))$ for all $t \geq 0$. In*

addition, the state variable $\mathbf{u}(t)$ *and output* $\mathbf{x}(t)$ *of the system (4.2), are bounded for all* $t \geq 0$.

Proof. It follows from Lemma 4.4 and property (4.8) that $\dot{F}(\mathbf{x}(t)) \leq 0$ for all $t \geq 0$. As a consequence, $F(\mathbf{x}(t))$ is non-increasing for all $t > 0$, which implies

$$F(\mathbf{x}(t)) - F(\mathbf{x}(0)) = \int_0^t \dot{F}(\mathbf{x}(\eta))d\eta. \tag{4.21}$$

Since $0 < t$ and $\dot{F}(\mathbf{x}(t)) \leq 0$ for all $\eta \in (0, t)$, we see that $F(\mathbf{x}(t)) \leq F(\mathbf{x}(0))$ for all $t \geq 0$.

Hence, $F(\mathbf{x}(t)) \leq F(\mathbf{x}(0))$ for all $t \geq 0$ implies that all level set of the function F are bounded. Using this fact together with $\dot{F}(\mathbf{x}(t)) \leq 0$ for all $t \geq 0$, the Lyapunov boundedness theorem asserts that all trajectories of the system (4.2) are bounded, that is, the exist $M > 0$ such that $\|\mathbf{x}\|_2 < M$. Since $\mathbf{x}(t)$ is bounded, then $\mathbf{u}(t)$ must also be bounded. □

4.3 Existence of solution and equilibrium

Before solving the dynamical system, it is important to establish if the system has any solution. The following theorem guarantees at least one solution of the dynamical system. Before stating the theorem, we define what is meant by the solution.

Definition 4.7 (Solution of dynamical system (4.2))**.** *The pair* $\mathbf{u}(t)$ *and* $\mathbf{x}(t)$ *constitute a solution of the dynamical system (4.2) if both* $\mathbf{u}(t)$ *and* $\mathbf{x}(t)$ *are absolutely continuous function for all* $t \geq 0$*, which satisfy the differential inclusion (4.2) with* $\mathbf{u}(0) = \mathbf{u}^0$ *and* $\mathbf{x}(0) = S_\mu(\mathbf{u}^0)$.

Theorem 4.8 (Existence of solution)**.** *Suppose the objective function* F *satisfies Assumption 4.1 and the activation function satisfies Assumption 4.2. Then, the dynamic system (4.2) has at least one solution* $\mathbf{u}(t)$*, initiated at* $\mathbf{u}(0) = \mathbf{u}^0$*, for any* $\mathbf{u}^0 \in \mathbb{R}^n$ *with* $\mathbf{x}(0) = S_\mu(\mathbf{u}^0)$ *and defined on* $[0, +\infty)$.

Proof. Since F satisfies Assumption 4.1 and the activation function satisfies 4.2, the subgradients $\partial F(\mathbf{x}(t))$ with $\mathbf{x}(t) = S_\mu(\mathbf{u}(t))$ is nonempty, convex, closed and locally bounded for $\mathbf{u}(t) \in \mathbb{R}^n$. According to [57], there exists at least one solution of the differential inclusion (4.2) with initial condition $\mathbf{u}(0) = \mathbf{u}^0$ on the time interval $[0, T)$ which satisfies that either

$\lim_{t\to T} \|\mathbf{u}(t)\|_2 = +\infty$ or $T = +\infty$. Since $\mathbf{u}(t)$ is bounded by Lemma 4.6, then the maximum existing time interval is $[0, \infty)$. □

Furthermore, we can conclude that the dynamical system (4.2) has at least one equilibrium point. Clearly, a point $\mathbf{u}^* \in \mathbb{R}^n$ is an equilibrium point of (4.2) if and only if $\mathbf{u}^*$ satisfies the algebraic inclusion $\dot{\mathbf{u}}^*(t) = \mathbf{0} \in \partial F(\mathbf{x}^*)$, where $\mathbf{x}^* = S_\mu(\mathbf{u}^*)$. This is exactly the Fermat principle that characterizes the optimal solution of nonsmooth optimization problems. Thus the equilibrium point of the dynamical system (4.2) coincide with the critical points of the nonsmooth objective function F.

4.4 Convergence analysis

Before using a dynamical system to solve optimization problems, it is important to establish that the system converges to a stable equilibrium point. We state and prove a theorem that guarantees that the trajectories have finite length and they converge to the stable equilibrium point. The proof will follow the works in [33, 58] by using the nonsmooth Łojasiewics inequality. The results are presented for a general nonsmooth objective function $F(\mathbf{x}(t))$ and sparse promoting activation function $S_\mu(\mathbf{u}(t))$. Note that the activation function is not bounded, which is fundamentally different from previous works of [58, 59].

4.4.1 The output convergence

The output variable $\mathbf{x}(t)$ is related to the state variable $\mathbf{u}(t)$ through the activation function as $\mathbf{x}(t) = S_\mu(\mathbf{u}(t))$. We discuss the proof of the output convergence to a stable equilibrium point. The state convergence follows easily from the results. The following theorem states that output variable $\mathbf{x}(t)$ converges in finite length and is globally convergent to a stable equilibrium point.

Theorem 4.9 (Output convergence)**.** *Suppose the differential inclusion (4.2) has an objective function $F(\mathbf{x}(t))$ satisfying Assumption 4.1 and an activation function $S_\mu(\mathbf{u}(t))$ satisfying Assumption 4.2, then the trajectories of the output $\mathbf{x}(t)$ has a finite length on $[0, +\infty)$ and*

the trajectories are globally convergent, that is there exits $\mathbf{x}^* \in \mathbb{R}^n$ *such that*

$$\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{x}^*, \qquad \textit{for all } \mathbf{x}(0) \in \mathbb{R}^n. \tag{4.22}$$

Moreover, $\mathbf{x}^*$ *is a critical point of the objective function* F*.*

Proof. By Lemma 4.6, $F(\mathbf{x}(t))$ is continuous and bounded, and by Lemma 4.4, $F(\mathbf{x}(t))$ is nonincreasing for all $t \geq 0$. Thus, the objective function $F(\mathbf{x}(t))$ converges to a constant limit F^* for $t \to \infty$.

By Lemma 4.6, the output $\mathbf{x}(t)$ is bounded for all $t \geq 0$. Applying the Bolzano-Weierstrass theorem, there exits a sequence of times $\{t_k\}_{k\in\mathbb{N}}$ such that the sequence $\{\mathbf{x}(t_k)\}_{k\in\mathbb{N}}$ converges as $k \to \infty$. We define $\mathbf{x}^*$ as the limiting point of this converging sequence. Thus, we need to show that the output $\mathbf{x}(t)$ converges to $\mathbf{x}^*$. Note that due to continuity of the objective function with respect to time, the limit of the sequence $\{F(\mathbf{x}(t_k))\}_{k\in\mathbb{N}}$ converges to $F(\mathbf{x}^*) = F^*$.

To show $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{x}^*$, we apply the Lojasiewicz inequality (Theorem 2.19) for nonsmooth functions at the point $\mathbf{x}^*$. There exits $\theta \in [0, 1)$, $c > 0$, and $r > 0$ such that

$$|F(x(t)) - F^*|^\theta \leq c\, m\left(\partial F(\mathbf{x}(t))\right) \qquad \text{for all } t > t_p \text{ such that } \|\mathbf{x}(t) - \mathbf{x}^*\|_2 < r, \tag{4.23}$$

where $m\left(\partial F(\mathbf{x}(t))\right) = \inf\{\|\xi\|_2 : \xi \in \partial F(\mathbf{x}(t))\}$. The existence of time t_p is guaranteed since the sequence of time $\{t_k\}_{k\in\mathbb{N}}$ is increasing and goes to infinity.

We need the following results to complete the proof. By Lemma 4.6, the state variable $\mathbf{u}(t)$ is bounded, and from the property (4.8) of Assumption 4.2, $h'(u_i(t))$ is strictly positive and continuous for all $i \in \mathcal{A}$ and $h'(u_i(t))$ is bounded from below and above. Thus we can find $0 < L < U < +\infty$ such that

$$L \leq h_i'(u_i(t)) \leq U. \tag{4.24}$$

Using the fact that for the active set $\mathcal{A}$, the output $x_i(t) = h(u_i(t))$ and the lower bound (4.24) we get the following inequality

$$|\dot{x}_i(t)| = |h'(u_i(t)) \cdot \dot{u}_i(t)| \geq L|\dot{u}_i|, \qquad \text{for all } i \in \mathcal{A}. \tag{4.25}$$

Hence, it follows that

$$\begin{aligned}
\|\dot{\mathbf{x}}(t)\|_2 &= \|\dot{\mathbf{x}}_{\mathcal{A}}(t)\|_2 && \text{(4.26a)}\\
&\geq L\|\dot{\mathbf{u}}_{\mathcal{A}}(t)\|_2 && \text{(4.26b)}\\
&\geq Lm\left(\partial F(\mathbf{x}_{\mathcal{A}}(t))\right) && \text{(4.26c)}\\
&= Lm\left(\partial F(\mathbf{x}(t))\right) && \text{(4.26d)}\\
\Rightarrow \|\dot{\mathbf{x}}(t)\|_2 &\geq Lm\left(\partial F(\mathbf{x}(t))\right), && \text{(4.26e)}
\end{aligned}$$

where (4.26b) results from inequality (4.25) and (4.26c) results from the differential inclusion (4.2). Furthermore, from Lemma 4.4 and the upper bound (4.24) we get the following

$$\dot{F}(\mathbf{x}(t)) = -\sum_{i\in\mathcal{A}} \frac{1}{h'(u_i(t))}|\dot{x}_i(t)|^2 \leq -\frac{1}{U}\|\dot{\mathbf{x}}(t)\|_2^2. \tag{4.27}$$

Hence, putting inequalities (4.26e) and (4.27) together, and using the Lojasiewicz inequality (4.23), we get

$$\begin{aligned}
\dot{F}(\mathbf{x}(t)) &\leq -\frac{1}{U}\|\dot{\mathbf{x}}(t)\|_2^2 && \text{(4.28a)}\\
&\leq -\frac{L}{U}\|\dot{\mathbf{x}}\|_2 m\left(\partial F(\mathbf{x}(t))\right) && \text{(4.28b)}\\
\Rightarrow \dot{F}(\mathbf{x}(t)) &\leq -\frac{L}{cU}\|\dot{\mathbf{x}}\|_2 \left|F(x(t)) - F^*\right|^{\theta}, && \text{(4.28c)}
\end{aligned}$$

where (4.28b) results from (4.26e), and (4.28c) results from the Lojasiewicz inequality (4.23). Hence, after rearranging (4.28c) we get the following inequality

$$\|\dot{\mathbf{x}}(t)\|_2 \leq -\frac{cU}{L}\frac{\dot{F}(\mathbf{x}(t))}{\left(F(\mathbf{x}(t)) - F^*\right)^{\theta}}. \tag{4.29}$$

Since the Łojasiewicz inequality (4.23) is satisfied. Then there exits t_1 and t_2 such that $\|\mathbf{x}(t) - \mathbf{x}^*\|_2 < r$ for all $t_2 > t > t_1 \geq t_p$ and we get the following bound

$$\int_{t_1}^{t_2} \|\dot{\mathbf{x}}(t)\|_2 dt \leq -\frac{cU}{L} \int_{t_1}^{t_2} \frac{\dot{F}(\mathbf{x}(\eta))}{(F(\mathbf{x}(\eta)) - F^*)^\theta} d\eta, \tag{4.30a}$$

$$= -\frac{cU}{L} \int_{F(\mathbf{x}(t_1))}^{F(\mathbf{x}(t_2))} \frac{dF}{(F - F^*)^\theta}, \tag{4.30b}$$

$$= -\frac{cU}{L(1-\theta)} \left[(F(\mathbf{x}(t_2)) - F^*)^{1-\theta} - (F(\mathbf{x}(t_1)) - F^*)^{1-\theta} \right], \tag{4.30c}$$

$$= \frac{cU}{L(1-\theta)} \left[(F(\mathbf{x}(t_1)) - F^*)^{1-\theta} - (F(\mathbf{x}(t_2)) - F^*)^{1-\theta} \right], \tag{4.30d}$$

$$\leq \frac{cU}{L(1-\theta)} \left[F(\mathbf{x}(t_1)) - F^* \right]^{1-\theta}, \tag{4.30e}$$

where (4.30a) is the integral of (4.29) between t_1 and t_2, followed by algebraic manipulation.

To proceed, note that since we have $\lim_{k\to\infty} F(\mathbf{x}(t_k)) = F^*$ and $\lim_{k\to\infty} \mathbf{x}(t_k) = \mathbf{x}^*$, then there exits t_r such that for $t_r > t_p$ we have

$$F(\mathbf{x}(t_r)) - F^* \leq \left(\frac{rL(1-\theta)}{4cU} \right)^{\frac{1}{1-\theta}} \quad \text{and} \quad \|\mathbf{x}(t_r) - \mathbf{x}^*\| < \frac{r}{4}. \tag{4.31}$$

Note that the constant is the bracket can be chosen any how, we define it this way to aid our analysis.

Now, we argue by contradiction to show that we have $\|\mathbf{x}(t) - \mathbf{x}^*\|_2 < r$ for all $t > t_r$. This implies that for all $t > t_r$ the output trajectory remains within a ball of radius r around the equilibrium point $\mathbf{x}^*$. We prove by contradiction that the output trajectory never leaves this ball. Suppose the output trajectory leaves the ball at some time t_e, that is, for $t_e > t_r$ we have $\|\mathbf{x}(t_e) - \mathbf{x}^*\|_2 = r$, while $\|\mathbf{x}(t) - \mathbf{x}^*\|_2 < r$ for $t_e > t > t_r > t_p$. Then, we can find an

upper bound

$$\|\mathbf{x}(t_e) - \mathbf{x}(t_r)\|_2 = \left\| \int_{t_r}^{t_e} \dot{\mathbf{x}}(t)dt \right\|_2 \leq \int_{t_r}^{t_e} \|\dot{\mathbf{x}}(t)\|_2 dt \tag{4.32a}$$

$$\leq \frac{cU}{L(1-\theta)} \left[(F(\mathbf{x}(t_r)) - F^*)^{1-\theta} \right] \tag{4.32b}$$

$$\leq \frac{cU}{L(1-\theta)} \left(\frac{rL(1-\theta)}{4cU} \right) \tag{4.32c}$$

$$\leq \frac{r}{4}, \tag{4.32d}$$

where (4.32b) results from (4.30e) and the inequality (4.32c) results from (4.31). Finally, we see that

$$r = \|\mathbf{x}(t_e) - \mathbf{x}^*\|_2 \leq \|\mathbf{x}(t_e) - \mathbf{x}(t_r)\|_2 + \|\mathbf{x}(t_r) - \mathbf{x}^*\|_2, \tag{4.33a}$$

$$\leq \frac{r}{4} + \frac{r}{4}, \tag{4.33b}$$

$$= \frac{r}{2}, \tag{4.33c}$$

where (4.33b) results from (4.32d) and (4.31). Thus, we get a contradiction. Therefore, the output trajectory never leaves the ball of radius r around the equilibrium point $\mathbf{x}^*$.

Consequently, $\|\mathbf{x}(t) - \mathbf{x}^*\|_2 < r$ for all $t \geq t_r$, we obtain from (4.30e)

$$\int_{t_r}^{+\infty} \|\dot{\mathbf{x}}(t)\|_2 dt \leq \frac{cU}{L(1-\theta)} \left(F(\mathbf{x}(t_r)) - F^*\right)^{(1-\theta)} < \frac{r}{4}. \tag{4.34}$$

Therefore, $\mathbf{x}(t)$ has finite length on $[0, +\infty)$ and, since r can be chosen arbitrarily small, it follows that $\lim_{t\to+\infty} \mathbf{x}(t) = \mathbf{x}^*$, and thus the output converges to a singleton.

□

4.4.2 The state convergence

Theorem 4.10 (State variable convergence)**.** *Suppose the differential inclusion (4.2) has an objective function satisfying Assumption 4.1 and an activation function satisfying Assumption*

4.2, then the state variable trajectory has a finite length on $[0, +\infty)$ *and is globally convergent, that is there exits* $\mathbf{u}^* \in \mathbb{R}^n$ *such that*

$$\lim_{t\to\infty} \mathbf{u}(t) = \mathbf{u}^*, \qquad \forall \mathbf{u}(0) \in \mathbb{R}^n. \tag{4.35}$$

Moreover, $\mathbf{x}^* = S_\mu(\mathbf{u}^*)$.

Proof. The proof is similar to that of output trajectories. We first have to establish the Lyapunov function in terms of $\mathbf{u}(t)$, them apply the nonsmooth Lojaseiwicz inequality.

□

4.5 Convergence rate analysis

It is essential to know how much progress is made towards the optimal solution in optimization. We state the estimates of convergence rate in terms of Łojasiewicz exponents.

Theorem 4.11. *Suppose the differential inclusion (4.2) has an objective function satisfying Assumption 4.1 and an activation function satisfying Assumption 4.2. The output trajectory* $\mathbf{x}(t)$ *converges to some critical point* $\mathbf{x}^* \in \mathbb{R}^n$ *of the objective function* F. *Let* $\theta \in (0, 1]$ *be a Lojasiewicz exponent near* $\mathbf{x}^*$. *Then, there exits* $c_1, c_2 > 0$, $T_1, T_2, T_3 > 0$ *and* $k_1, k_2 > 0$ *such that the following estimates hold:*

1. *If* $\theta \in (1/2, 1)$, *then* $\|\mathbf{x}(t) - \mathbf{x}^*\| \leq (c_1 t + k_1)^{\frac{1-\theta}{2\theta-1}}$ *holds for all* $t \geq T_1$.
2. *If* $\theta = \frac{1}{2}$, *then* $\|\mathbf{x}(t) - \mathbf{x}^*\| \leq c_2 exp(-k_1 t)$ *holds for all* $t \geq T_2$.
3. *If* $\theta \in [0, 1/2)$, *then* $\mathbf{x}(t)$ *converges in finite time, that is* $\mathbf{x}(t) = \mathbf{x}^*$ *for all* $t \geq T_3$.

Proof. Since the output trajectory $\mathbf{x}(t)$ converges to some critical point $\mathbf{x}^* \in \mathbb{R}^n$ of the objective function F, there exits $t_p > 0$ such that the Lojasiewicz inequality (4.23) holds for every $t \geq t_p$. We consider the tail length function $\sigma : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$\sigma(t) = \int_t^{+\infty} \|\dot{\mathbf{x}}(\eta)\|_2 d\eta, \qquad \forall t > 0. \tag{4.36}$$

Since $\mathbf{x}(t)$ is an absolutely continuous function, it follows that for $T > t$ we have

$$\mathbf{x}(t) = \mathbf{x}(T) - \int_t^T \dot{\mathbf{x}}(\eta) d\eta.$$

Now introducing the $\mathbf{x}^*$ on both side of the above equation and taking the norm we have

$$\|\mathbf{x}(t) - \mathbf{x}^*\|_2 = \left\| \mathbf{x}(T) - \mathbf{x}^* - \int_t^T \dot{\mathbf{x}}(\eta) d\eta \right\|_2, \tag{4.37a}$$

$$\leq \|\mathbf{x}(T) - \mathbf{x}^*\|_2 + \int_t^T \|\dot{\mathbf{x}}(\eta) d\eta\|_2. \tag{4.37b}$$

Now, we let $T \to +\infty$ and the inequality (4.37b) becomes

$$\|\mathbf{x}(t) - \mathbf{x}^*\|_2 \leq \int_t^{+\infty} \|\dot{\mathbf{x}}(\eta)\|_2 d\eta = \sigma(t), \tag{4.38}$$

which is the definition of σ, see equation (4.36).

Taking the derivative (4.38), we get $\dot{\sigma}(t) = -\|\dot{\mathbf{x}}(t)\|_2$, note that $\|\dot{\mathbf{x}}(\infty)\|_2 = \|\dot{\mathbf{x}}^*\|_2 = 0$. For $t > t_r$ and on the basis of inequality (4.34) and the Łojasiewics inequality (4.23), we obtain the following

$$\sigma(t) = \int_t^{+\infty} \|\dot{\mathbf{x}}(\eta)\|_2 d\eta, \tag{4.39a}$$

$$\leq \frac{cU}{L(1-\theta)} \left(F(\mathbf{x}(t)) - F^*\right)^{(1-\theta)}, \tag{4.39b}$$

$$\leq \frac{c^{\frac{\theta+1}{\theta}} U}{L(1-\theta)} m(\partial F(\mathbf{x}(t)))^{\frac{1-\theta}{\theta}}, \tag{4.39c}$$

$$\leq \frac{c^{\frac{\theta+1}{\theta}} U}{(1-\theta)} \|\dot{\mathbf{x}}(t)\|_2^{\frac{1-\theta}{\theta}}, \tag{4.39d}$$

$$= \frac{c^{\frac{\theta+1}{\theta}} U}{(1-\theta)} \left[-\dot{\sigma}(t)\right]^{\frac{1-\theta}{\theta}}, \tag{4.39e}$$

where (4.39b) follows from (4.34), the inequality (4.39c) follows from the Łojasiewicz inequality (4.23) and (4.39d) follows from (4.26e). Hence, rearranging the inequality (4.39e), $\sigma(t)$ is an

absolute continuous function that satisfies the following differential inequality:

$$\dot{\sigma}(t) \leq -C\left[\sigma(t)\right]^{\frac{\theta}{1-\theta}}$$

where $C = c^{\frac{1+\theta}{1-\theta}}U/(1-\theta)^{\frac{\theta}{1-\theta}}$ is a positive constant. To obtain the estimate of convergence rate it suffices to solve the following differential equation:

$$\begin{cases} \dot{z}(t) = -C\left[z(t)\right]^{\frac{\theta}{1-\theta}} & \forall t > t_r, \\ z(0) = \sigma(t_r). \end{cases} \tag{4.40}$$

Consequently, the convergence estimates follows from (4.38) and due to the fact that $\sigma(t) \leq z(t)$ for all $t > t_r$. Considering each case separately we obtain:

- Assume that $\theta \in (1/2, 1)$ and note that $\frac{1-\theta}{1-2\theta} < 0$, then from differential equation (4.40) we obtain

  $$\frac{d}{dt}\left(z(t)^{\frac{1-2\theta}{1-\theta}}\right) = C\frac{2\theta-1}{1-\theta} \qquad \forall t > t_r.$$

  By integration, there exits $c_1 > 0$, $k_1 > 0$ and $T_1 = t_r$ such that

  $$z(t) = (c_1 t + k_1)^{-\frac{1-\theta}{2\theta-1}} \qquad \forall t > T_1.$$

  Thus, the conclusion of statement 1 follows from (4.38).

- If $\theta = \frac{1}{2}$, then from differential equation (4.40) we obtain

  $$\dot{z}(t) = -Cz(t) \qquad \forall t > t_r.$$

  By multiplying with the integrating factor $\exp(-Ct)$ and integrating from t_r to t, it follows that there exits $c_2 > 0$, $k_2 > 0$ and $T_2 = t_r$ such that

  $$z(t) = c_2 \exp(-k_2 t) \qquad \forall t > T_2.$$

  Hence, the conclusion of item 2 follows immediately from (4.38). Thus the output trajectory $\mathbf{x}(t)$ converges exponentially towards the point $\mathbf{x}^*$.

- Finally, suppose that $\theta \in [0, 1/2)$ and note that $\frac{1-\theta}{1-2\theta} > 0$, then from differential equation (4.40) we obtain

$$\frac{d}{dt}\left(z(t)^{\frac{1-2\theta}{1-\theta}}\right) = -C\frac{1-2\theta}{1-\theta} \qquad \forall t > t_r.$$

By integrating this equation, there exits $c_3 > 0$ and $k_3 > 0$ such that

$$z(t)^{\frac{1-2\theta}{1-\theta}} = -c_3 t + k_3 \qquad \forall t > t_r.$$

Thus, there exits $T_3 \geq 0$ such that

$$z(t) \leq 0 \qquad \forall t > T_3.$$

Hence, since $\sigma(t) \leq z(t)$, it follows that $\|\mathbf{x}(t) - \mathbf{x}^*\| = 0$, which proves statement 3. The output trajectory converges to $\mathbf{x}^*$ in finite finite time.

□

Note that it is generally difficult to find the Łojasiewic exponents for any given function.

4.6 Chapter summary

This chapter addressed the convergence of trajectories of differential inclusion aimed at solving the sparse recovery problems. The main result is that the trajectories of the differential inclusion are either exponentially convergent or finite-time convergent towards a singleton. These convergence properties are independent of the nature of the set of equilibrium points. Hence, they hold even when the differential inclusion has infinitely many nonisolated equilibrium points. The proof exploits the recently proposed method for convergence, based on the use of Łojasiewicz inequality for nonsmooth functions. The results are of interest for the efficient computation of solutions to sparse recovery problems.

Chapter 5

Bregman Integrated Dynamical System

This chapter presents the Bregman integrated dynamical system for solving the sparse recovery problem. We analyse the effectiveness and efficiency of the proposed dynamical system for solving the ℓ_1-minimization problem. We start by presenting the challenges of ℓ_1-minimization problem and highlighting the contributions of this work in Section 5.1. In Section 5.2, we present the Bregman integrated dynamical system. In Section 5.3, we present the properties of the Bregman integrated dynamical system. In particular, we show that the trajectories of the system follow an efficient path towards the equilibrium points and that the trajectories are always bounded relative to the equilibrium point. In Section 5.4, we show that the optimal support set bounds the active set of the solution. In Section 5.5, we present convergence rate of the proposed dynamical system. Finally, we summarise the work done in this chapter in Section 5.6.

5.1 Introduction

Our main objective is to solve the sparse recovery problem efficiently. Given, the measurements $\mathbf{y} = \mathbf{\Phi x} \in \mathbb{R}^m$ of $\mathbf{x} \in \mathbb{R}^n$, the problem is generally modelled by the objective function of the

form

$$F(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \mu \sum_{i=1}^{n} R_i(x_i), \tag{5.1}$$

where the first term measures data discrepancy and $R_i : \mathbb{R} \to \mathbb{R}$, is chosen to enforce a specific structure on the solution. The solution structure of interest is sparsity. Thus, we use the ℓ_1-norm penalty function, which is the sum of the absolute value of the components of $\mathbf{x}$. Therefore, the regularization function is $R_i(x_i) = |x_i|$. In this case, we call the problem the ℓ_1-minimization problem. The main difficulty for the analysis is that the objective function is not differential everywhere. The dynamical system approach can be used to solve the problem (5.1) efficiently.

The design of the dynamical system, in this chapter, is based on the Bregman proximal mapping. This approach has the ability to accelerate the convergence rate of the trajectories towards the optimal solution, achieving a better convergence rate compared to the dynamical system proposed for Locally Completive Algorithm (proposed in [27]). The proposed dynamical system is termed the Bregman integrated dynamical system. It is important to provide the mathematical analysis that qualifies the dynamical system to be suitable for solving optimization problems.

The contributions of this chapter are the design and analysis of the Bregman integrated dynamical system. Specifically, we show the following: First, the Bregman integrated dynamical system is well suited for solving sparse recovery problems. The trajectories of the Bregman dynamical system converge to an equilibrium point from any starting point. Moreover, the equilibrium points correspond to the optimal points of ℓ_1-minimization problem.

Second, we show that under certain conditions based on restricted isometry property, the size of the active set is guaranteed to remain bounded throughout convergence. We show that under certain conditions, the active set of the output trajectories of the Bregman dynamical system is bounded by the optimal support set. For practical propose, this result may be strong in that the active set may contain more components than the optimal support set due to numerical implementations. In that case, we show that the active set of the output trajectories is bounded by a constant. This constant is a small fraction of the size of the optimal support set.

Thirdly, we present an estimate of the convergence rate of the dynamical system. In particular, we show that the proposed dynamical system has an exponential convergence rate.

5.2 Description of the dynamical model

The proposed dynamical system is derived from the Bregman proximal mapping. Here, the objective function (5.1) is integrated with the Bregman distance. The Bregman integrated dynamical system is then constructed using the intuition of the Hopfield network.

5.2.1 Bregman proximal mapping

The Bregman distance [60] is an important concept in convex analysis and optimization methods which we present below:

Definition 5.1 (Bregman distance)**.** *Suppose* $v : \mathbb{R}^n \to \mathbb{R}$ *is a convex function and* $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$*. The Bregman distance associated with* v *is defined by*

$$D_v(\mathbf{x}, \mathbf{z}) = v(\mathbf{x}) - v(\mathbf{z}) - \langle \nabla v(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle . \tag{5.2}$$

It naturally measures the proximity between the two points ($\mathbf{x}$, $\mathbf{z}$). The Bregman distance has several properties that make it an efficient tool in optimisation. In particular, for a general convex function v, the Bregman distance is non-negative, $D_v(\mathbf{x}, \mathbf{z}) \geq 0$ and $D_v(\mathbf{x}, \mathbf{z}) = 0$ if and only if $\mathbf{x} = \mathbf{z}$, which implies a basic distance-like property. However, D_v is not symmetric in general, unless v is the energy function like

$$v(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2. \tag{5.3}$$

We adopt the energy function in this chapter. There are other functions used in literature [61, 62], but they are not suited for the sparse recovery problem.

Instead of solving (5.1) as it is, the focus of this work is on a more general and flexible methods based on Bregman proximal mapping defined for some parameter γ:

$$\text{prox}_{\gamma F}^{v}(\mathbf{z}) = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min}\, F(\mathbf{x}) + \gamma D_v(\mathbf{x}, \mathbf{z}). \tag{5.4}$$

The rationale for the relevance and usefulness of combining (5.1) and Bregman distance comes from the algorithmic context. The need to improve and extend the convergence properties of optimization algorithms. Adding the Bregman term in the formulation can significantly improve the convergence rate of an algorithm.

5.2.2 Bregman integrated dynamical system

We define the proximal function $F_p(\mathbf{x}) := F(\mathbf{x}) + \gamma D_v(\mathbf{x}, \mathbf{z})$ for some predefined point $\mathbf{z}$ and parameter γ. The function $F(\mathbf{x})$ is the sparse recovery objective function (5.1) with $R_i = |x_i|$ and v, associated with the Bregman distance, is the energy function (5.3). Thus, we have

$$F_p(\mathbf{x}) = F(\mathbf{x}) + \gamma D_v(\mathbf{x}, \mathbf{z}), \tag{5.5a}$$

$$= \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_1 + \gamma\left(v(\mathbf{x}) - v(\mathbf{z}) - \langle\nabla v(\mathbf{z}), \mathbf{x} - \mathbf{z}\rangle\right), \tag{5.5b}$$

$$\implies F_p(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_1 + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{z}\|_2^2. \tag{5.5c}$$

where (5.5b) follows from the definition of the Bregman distance (5.2), and (5.5c) is the result of the energy function (5.3). The subgradients of the proximal function are

$$\partial_{\mathbf{x}} F_p(\mathbf{x}) = -\mathbf{\Phi}^T\left(\mathbf{y} - \mathbf{\Phi}\mathbf{x}\right) + \mu\partial\|\mathbf{x}\|_1 + \gamma\left(\mathbf{x} - \mathbf{z}\right). \tag{5.6}$$

Following the Hopfield network approach, we can define a network of n computational nodes acting together to find the minimizer of the proximal function F_p in (5.5c). From the network a Bregman integrated dynamical system is constructed as follows: First, define an internal state variable, $\mathbf{u} : \mathbb{R}^n \to \mathbb{R}^n$, which depends on time t. The time derivative of the internal state variable, $\dot{\mathbf{u}}(t)$, is defined to be an element of the negative of subdifferential of F_p in

(5.6):

$$\dot{\mathbf{u}}(t) \in -\partial F_p(\mathbf{x}(t)), \tag{5.7a}$$

$$= \mathbf{\Phi}^T \left(\mathbf{y} - \mathbf{\Phi}\mathbf{x}(t)\right) - \mu\partial\|\mathbf{x}(t)\|_1 - \gamma\left(\mathbf{x}(t) - \mathbf{z}\right). \tag{5.7b}$$

Practically, it is impossible to implement $\partial\|\mathbf{x}\|_1$ since the subdifferential is a set. However, with the introduction of the internal state variable $\mathbf{u}(t)$, as discussed in Section 2.3 of Chapter 3, we can use the relation $\mathbf{u} - \mathbf{x} \in \mu\partial\|\mathbf{x}\|_1$ to replace $\mu\partial\|\mathbf{x}\|_1$ in the differential inclusion (5.7b). Thus, we get

$$\dot{\mathbf{u}}(t) = -\mathbf{u}(t) + \mathbf{\Phi}^T \left(\mathbf{y} - \mathbf{\Phi}\mathbf{x}(t)\right) + \mathbf{x}(t) - \gamma\left(\mathbf{x}(t) - \mathbf{z}\right). \tag{5.8}$$

The optimization variable $\mathbf{x}$ is set to be the output of the network and it is related to the internal state variable $\mathbf{u}$ via the soft-thresholding activation function, $S_\mu(\mathbf{u}(t))$,

$$x_i(t) = S_\mu(u_i(t)) = \begin{cases} 0, & |u_i(t)| \leq \mu, \\ u_i(t) - \mu \cdot \text{sign}(u_i(t)), & |u_i(t)| > \mu, \end{cases} \tag{5.9}$$

where the computations are done component-wise. For the derivation of the soft-thresholding function (5.9) from the relation $\mathbf{u} - \mathbf{x} \in \mu\partial\|\mathbf{x}\|_1$, see the discussion in Section 2.3 of Chapter 3, Figure 3.6. Hence, the complete Bregman integrated dynamical system is defined as

$$\begin{cases} \tau\dot{\mathbf{u}}(t) = -\mathbf{u}(t) - (\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\mathbf{x}(t) + \mathbf{\Phi}^T\mathbf{y} - \gamma\left(\mathbf{x}(t) - \mathbf{z}\right) \\ \mathbf{x}(t) = S_\mu(\mathbf{u}(t)). \end{cases} \tag{5.10}$$

The parameter τ depends on the system implementing the dynamical system (5.10), and it has the influence on the convergence speed (this will be shown later). The Bregman integrated dynamical system (5.10) takes the form of the differential inclusion (4.2), discussed in Chapter 4. The proximal function F_p can be shown to satisfy Assumption 4.1; it is locally Lipschitz continuous, regular and subanalytic on $\mathbb{R}^n$. The activation function S_μ satisfies the Assumption 4.2. Thus, as dissucssed in Chapter 4, the following results regarding the convergence of the Bregman integrated dynamical system (5.10) hold true. The trajectory

of the Bregman integrated dynamical system reaches the steady state in finite length, that is,

$$\lim_{t\to\infty} \mathbf{u}(t) = \mathbf{u}^* \qquad \text{and} \qquad \lim_{t\to\infty} \mathbf{x}(t) = \mathbf{x}^*.$$

Furthermore, with an appropriate choice of the parameter $\mathbf{z}$, the steady state of the Bregman integrated dynamical system corresponds to the optimal solution of the sparse recovery problem (5.1). We summarize the above discussion in the following theorem.

Theorem 5.2. *Suppose the objective function $F_p(\mathbf{x})$ in (5.5c) is Lipschitz continuous, regular and subanalytic on $\mathbb{R}^n$. If the activation function $S_\mu(\mathbf{u})$ in (5.10) satisfies the following relation*

$$\mathbf{u} - \mathbf{x} = \mathbf{u} - S_\mu(\mathbf{u}) \in \mu\partial\|\mathbf{x}\|_1, \tag{5.11}$$

then the equilibrium points of the Bregman integrated system (5.10) are critical points of the objective function (5.5c).

Proof. It follows from (5.11), (5.7b) and (5.8) that the Bregman integrated system satisfy the differential inclusion $\dot{\mathbf{u}}(t) \in -\partial F_p(\mathbf{x}(t))$. Now, for any equilibrium point $\mathbf{u}^*$ of (5.10) satisfies $\dot{\mathbf{u}}(t) = \mathbf{0}$. This implies $\mathbf{0} \in \partial F_p(\mathbf{x}^*)$. Thus, the equilibrium points of (5.10) are critical points of (5.5c). □

The following theorem states that the trajectories of the Bregman integrated dynamical system are global convergent.

Theorem 5.3. *If the activation function $S_\mu(\mathbf{u}(t))$ is locally Lipschitz continuous, odd, non-decreasing, and subanalytic on $\mathbb{R}$, then the state variable $\mathbf{u}(t)$ and the output variable $\mathbf{x}(t)$ of the Bregman integrated dynamical system (5.10) are globally asymptomatically convergent.*

Proof. From Theorem 5.2 it is shown that the Bregman integrated dynamical system (5.10) satisfies the differential inclusion $\dot{\mathbf{u}}(t) \in -\partial F_p(\mathbf{x}(t))$. The theorem statement asserts that the activation function satisfies Assumption 4.2. Then by Theorem 4.9, the trajectories of the output $\mathbf{x}(t)$ of the system (5.10) are globally convergent, and by Theorem 4.10 the trajectories of the state $\mathbf{u}(t)$ are globally convergent. □

Hence, Theorem 5.2 validates the claim that Bregman integrated dynamical system qualify as an appropriate tool to solve the ℓ_1-minimization problem (5.1).

5.3 Properties of the Bregman integrated dynamical system

We discuss the properties of the Bregman integrated dynamical system for solving the ℓ_1-minimization program (5.1). We show that the trajectories of the dynamical system follow an efficient path, which converges to the optimal solution of the ℓ_1-minimization problem. After that, we establish some bounds between the trajectories of the dynamical system and the target signal. The target signal refers to the vector $\mathbf{x}$ that generated the measurements $\mathbf{y} = \mathbf{\Phi x}$. These bounds will be helpful in the proof of the main results in the next section. The discussion follows the works of [54, 56]

5.3.1 The optimality trajectories

The Bregman integrated dynamical system is a linear switched system. The dynamical system is governed by a linear ordinary differential equation that changes every time a component crosses the threshold $|u_i(t)| \leq \mu$ or $|u_i(t)| > \mu$, from the inactive to active set or vice-visa. Switching happens at the time point t_k, and it also happens at the next time point t_{k+1}. The active set $\mathcal{A}$ is unchanged for all the time between the switching times, $t \in [t_k, t_{k+1})$, so is the inactive set $\mathcal{A}_c$. Thus, the Bregman integrated dynamical system can be partially decoupled into the equations corresponding to active and inactive sets.

In the case of active set $\mathcal{A}$, we have the following relation from the activation function:

$$\mathbf{x}_{\mathcal{A}}(t) = \mathbf{u}_{\mathcal{A}}(t) - \mu \cdot \operatorname{sign}\left(\mathbf{u}_{\mathcal{A}}(t)\right). \tag{5.12}$$

Taking the derivative of (5.12) with respect to time, we get relation between the time derivative of the state variable and the output, for example, $\dot{\mathbf{x}}_{\mathcal{A}}(t) = \dot{\mathbf{u}}_{\mathcal{A}}(t)$. Hence, the

dynamical system corresponding to the active set is given by

$$\begin{aligned}
\dot{\mathbf{x}}_{\mathcal{A}}(t) &= \dot{\mathbf{u}}_{\mathcal{A}}(t) \\
&= -\mathbf{u}_{\mathcal{A}} - \left(\mathbf{\Phi}_{\mathcal{A}}^T\mathbf{\Phi}_{\mathcal{A}} - \mathbf{I}\right)\mathbf{x}_{\mathcal{A}}(t) + \mathbf{\Phi}_{\mathcal{A}}^T\mathbf{y} - \gamma(\mathbf{x}_{\mathcal{A}}(t) - \mathbf{z}_{\mathcal{A}}) && \text{(5.13a)} \\
&= -\mathbf{x}_{\mathcal{A}}(t) - \mu\cdot\operatorname{sign}\left(\mathbf{u}_{\mathcal{A}}(t)\right) - \left(\mathbf{\Phi}_{\mathcal{A}}^T\mathbf{\Phi}_{\mathcal{A}} - \mathbf{I}\right)\mathbf{x}_{\mathcal{A}}(t) + \mathbf{\Phi}_{\mathcal{A}}^T\mathbf{y} - \gamma(\mathbf{x}_{\mathcal{A}}(t) - \mathbf{z}_{\mathcal{A}}) && \text{(5.13b)} \\
&= -\mu\cdot\operatorname{sign}\left(\mathbf{u}_{\mathcal{A}}(t)\right) - \mathbf{\Phi}_{\mathcal{A}}^T\mathbf{\Phi}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}(t) + \mathbf{\Phi}_{\mathcal{A}}^T\mathbf{y} - \gamma\left(\mathbf{x}_{\mathcal{A}}(t) - \mathbf{z}_{\mathcal{A}}\right), && \text{(5.13c)} \\
\implies \dot{\mathbf{x}}_{\mathcal{A}}(t) &= -\left(\mathbf{\Phi}_{\mathcal{A}}^T\mathbf{\Phi}_{\mathcal{A}} + \gamma\mathbf{I}\right)\mathbf{x}_{\mathcal{A}}(t) + \mathbf{\Phi}_{\mathcal{A}}^T\mathbf{y} - \mu\mathbf{s}_{\mathcal{A}} + \gamma\mathbf{z}_{\mathcal{A}}(t) && \text{(5.13d)}
\end{aligned}$$

where $\mathbf{s}_{\mathcal{A}} = \operatorname{sign}(\mathbf{u}_{\mathcal{A}}(t))$ is a vector of signs of components of $\mathbf{x}_{\mathcal{A}}$ since $\operatorname{sign}(\mathbf{x}(t)) = \operatorname{sign}(\mathbf{u}(t))$. The equation (5.13a) follows from the definition of the dynamical system (5.10), and (5.13b) follows from active components as defined by the activation function (5.9).

In the case of inactive set $\mathcal{A}_c$, the corresponding dynamical system is

$$\dot{\mathbf{u}}_{\mathcal{A}_c}(t) = -\mathbf{u}_{\mathcal{A}_c} - \mathbf{\Phi}_{\mathcal{A}_c}^T\mathbf{\Phi}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}(t) + \mathbf{\Phi}_{\mathcal{A}_c}^T\mathbf{y}. \tag{5.14}$$

The second term of the differential equation (5.14) is due to the inhibition on inactive components from active components.

Consider the active set $\mathcal{A}$, the solution of (5.13d) between the switching time t_k and t_{k+1} can be explicitly computed using the extension of Gronwall's Lemma 2.18, using the initial condition $\mathbf{x}_{\mathcal{A}}(t_k)$ and integrating from t_k to t,

$$\mathbf{x}_{\mathcal{A}}(t) = e^{-\mathbf{A}(t-t_k)}\mathbf{x}_{\mathcal{A}}(t_k) + \left(\mathbf{I} - e^{-\mathbf{A}(t-t_k)}\right)\mathbf{A}^{-1}(\mathbf{\Phi}_{\mathcal{A}}^T - \mu\mathbf{s}_{\mathcal{A}} + \gamma\mathbf{z}_{\mathcal{A}}), \tag{5.15}$$

where $\mathbf{A} = \mathbf{\Phi}_{\mathcal{A}}^T\mathbf{\Phi}_{\mathcal{A}} + \gamma\mathbf{I}$. The matrix $\mathbf{A}$ is nonsingular and as time evolves the solution settle to a steady state solution

$$\mathbf{x}_{\mathcal{A}}^{st} = \mathbf{A}^{-1}(\mathbf{\Phi}_{\mathcal{A}}^T\mathbf{y} - \mu\mathbf{s}_{\mathcal{A}} + \gamma\mathbf{z}_{\mathcal{A}}).$$

Note that the active set and sign vector $\mathbf{s}_{\mathcal{A}}$ remain unchanged. The point $\mathbf{x}_{\mathcal{A}}^{st}$ will play an important role in the analysis of the Bregman integrated dynamic system.

In the case of inactive set $\mathcal{A}_c$, the solution of (5.14) between the switching times t_k and t_{k+1} is given by

$$\mathbf{u}_{\mathcal{A}_c}(t) = e^{-(t-t_k)}\mathbf{u}_{\mathcal{A}_c}(t_k) + e^{-t}\int_{t_k}^{t} e^{\eta}\rho(\eta)d\eta, \qquad t \in [t_k, t_{k+1}) \tag{5.16}$$

where $\rho(t) = \mathbf{\Phi}_{\mathcal{A}_c}^T\left(\mathbf{y} - \mathbf{\Phi}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}(t)\right)$.

When $t \to \infty$ the output trajectories as described by equation (5.15) converges to the steady-state which is the the equilibrium point $\mathbf{x}^*$, and it is supported on the final active set $\mathcal{A}^*$,

$$\mathbf{x}_{\mathcal{A}^*}^* = \left[\mathbf{\Phi}_{\mathcal{A}^*}\mathbf{\Phi}_{\mathcal{A}^*} + \gamma\mathbf{I}\right]^{-1}\left(\mathbf{\Phi}_{\mathcal{A}^*}^T\mathbf{y} - \mu\mathbf{s}_{\mathcal{A}^*} + \gamma\mathbf{z}_{\mathcal{A}^*}\right). \tag{5.17}$$

Now, using (5.17), we get $\rho(t) \to \mathbf{\Phi}_{\mathcal{A}_c}^T\left(\mathbf{y} - \mathbf{\Phi}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}^*}^*\right)$ as $t \to \infty$. Thus, $\rho(t)$ becomes a constant as the system reaches the equilibrium point. Using this fact, the state trajectories for inactive set, described by equation (5.16), converges to

$$\mathbf{u}_{\mathcal{A}_c^*}^* = \mathbf{\Phi}_{\mathcal{A}_c^*}^T\left(\mathbf{y} - \mathbf{\Phi}_{\mathcal{A}^*}\mathbf{x}_{\mathcal{A}^*}\right). \tag{5.18}$$

Note that if we let $\gamma \to 0$ as $t \to \infty$, the expressions (5.17) and (5.18) respectively reduce to

$$\mathbf{x}_{\mathcal{A}^*}^* = \left[\mathbf{\Phi}_{\mathcal{A}^*}\mathbf{\Phi}_{\mathcal{A}^*}\right]^{-1}\left(\mathbf{\Phi}_{\mathcal{A}^*}^T\mathbf{y} - \mu\mathbf{s}_{\mathcal{A}}\right), \tag{5.19}$$

$$\mathbf{u}_{\mathcal{A}_c^*}^* = \mathbf{\Phi}_{\mathcal{A}_c^*}^T\left(\mathbf{y} - \mathbf{\Phi}_{\mathcal{A}^*}\mathbf{x}_{\mathcal{A}^*}\right). \tag{5.20}$$

Since a component $i \in \mathcal{A}_c^*$ if and only if $|u_i| \leq \mu$, the equations (5.19) and (5.20) respectively translate to

$$\mathbf{x}_{\mathcal{A}^*}^* = \left[\mathbf{\Phi}_{\mathcal{A}^*}\mathbf{\Phi}_{\mathcal{A}^*}\right]^{-1}\left(\mathbf{\Phi}_{\mathcal{A}^*}^T\mathbf{y} - \mu\mathbf{s}_{\mathcal{A}}\right), \tag{5.21}$$

$$\left\|\mathbf{\Phi}_{\mathcal{A}_c^*}^T\left(\mathbf{y} - \mathbf{\Phi}_{\mathcal{A}^*}\mathbf{x}_{\mathcal{A}^*}\right)\right\|_{\infty} \leq \mu. \tag{5.22}$$

Equation (5.21) and inequality (5.22) resemble the optimality conditions for the finite dimensional optimization problem (5.1) obtained using a different approach, see [63]. Hence, the trajectory of the Bregman integrated dynamical system converges to an equilibrium point that corresponds with the optimal solution of the sparse recovery problem (5.1). Also, the above discussion provides a way of updating the parameter γ, that is γ must go to zero as time evolves.

5.3.2 Bounds on trajectories

The trajectory of the Bregman integrated dynamical system converges to the equilibrium point, which is the steady-state solution when $\gamma \to 0$. The results provide bounds on some relevant quantities and will be used to prove that the active set is bounded. Consider the measurement corrupted with noise, $\mathbf{y} = \mathbf{\Phi}\mathbf{x}^* + \mathbf{e}$, where $\mathbf{x}^*$ represents the target signal. The first lemma states that the distance between the steady-state solution $\mathbf{x}^{st}$ and the target signal $\mathbf{x}^*$ is bounded.

Lemma 5.4. *Assume that $\mathbf{x}^{st}$ is a steady-state solution supported on the set $\mathcal{A}$ which satisfies*

$$\mathbf{\Phi}_{\mathcal{A}}^T \mathbf{\Phi}_{\mathcal{A}} \mathbf{x}^{st} = \mathbf{\Phi}_{\mathcal{A}}^T \mathbf{y} - \mu \mathbf{s}_{\mathcal{A}}. \tag{5.23}$$

Assume also that the set $\mathcal{A}$ contains less than p indices. Let $r = |\mathcal{A} \cup \mathcal{A}^*|$ be the number of elements in the support of $(\mathbf{x}^{st} - \mathbf{x}^*)$. If $\mathbf{\Phi}$ satisfies the restricted isometry property with parameters (r, δ), then the following holds:

$$\|\mathbf{x}^{st} - \mathbf{x}^*\|_2 \leq (1-\delta)^{-1} B_\delta(p),$$

where

$$B_\delta(p) = \left(\|\mathbf{x}^*\|_2 + \sqrt{1-\delta}\|\mathbf{e}\|_2 + \mu\sqrt{p} \right). \tag{5.24}$$

Proof. We first note that

$$\begin{aligned} \|\mathbf{x}^{st} - \mathbf{x}^*\|_2 &= \|\mathbf{x}_{\mathcal{A}}^{st} - \mathbf{x}_{\mathcal{A}}^* + \mathbf{x}_{\mathcal{A}_c}^{st} - \mathbf{x}_{\mathcal{A}_c}^*\|_2 \\ &\leq \|\mathbf{x}_{\mathcal{A}}^{st} - \mathbf{x}_{\mathcal{A}}^*\|_2 + \|\mathbf{x}_{\mathcal{A}_c}^*\|_2 \end{aligned} \tag{5.25}$$

where $\mathbf{x}^{st}_{\mathcal{A}_c} = \mathbf{0}$ by definition. We take a close look on the first term on the right hand side of inequality (5.25):

$$\left\|\mathbf{x}^{st}_{\mathcal{A}} - \mathbf{x}^{*}_{\mathcal{A}}\right\|_2 \leq \left\|\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{y} - \mu\mathbf{s}_{\mathcal{A}}\right) - \mathbf{x}^{*}_{\mathcal{A}}\right\|_2 \tag{5.26a}$$

$$= \left\|\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\left(\mathbf{\Phi}^T_{\mathcal{A}}\left(\mathbf{\Phi}\mathbf{x}^{*} + \mathbf{e}\right) - \mu\mathbf{s}_{\mathcal{A}}\right) - \mathbf{x}^{*}_{\mathcal{A}}\right\|_2 \tag{5.26b}$$

$$= \left\|\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\left[\mathbf{\Phi}^T_{\mathcal{A}}\left(\mathbf{\Phi}_{\mathcal{A}_c}\mathbf{x}^{*}_{\mathcal{A}_c} + \mathbf{\Phi}_{\mathcal{A}}\mathbf{x}^{*}_{\mathcal{A}} + \mathbf{e}\right) - \mu\mathbf{s}_{\mathcal{A}}\right] - \mathbf{x}^{*}_{\mathcal{A}}\right\|_2 \tag{5.26c}$$

$$= \Big\|\mathbf{x}^{*}_{\mathcal{A}} + \left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}_c}\mathbf{x}^{*}_{\mathcal{A}_c} + \left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{e}$$
$$-\mu\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\mathbf{s}_{\mathcal{A}} - \mathbf{x}^{*}_{\mathcal{A}}\Big\|_2 \tag{5.26d}$$

$$= \left\|\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}_c}\mathbf{x}^{*}_{\mathcal{A}_c} + \left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{e} - \mu\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\mathbf{s}_{\mathcal{A}}\right\|_2 \tag{5.26e}$$

$$\leq \left\|\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\right\| \left\|\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}^*\cup\mathcal{A}_c}\right\| \|\mathbf{x}^{*}_{\mathcal{A}_c}\|_2 + \left\|\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\mathbf{\Phi}^T_{\mathcal{A}}\right\| \|\mathbf{e}\|_2$$
$$+ \mu\left\|\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\right\| \|\mathbf{s}_{\mathcal{A}}\|_2 \tag{5.26f}$$

The first inequality (5.26a) results from (5.23), (5.26b) is due to the fact that $\mathbf{y} = \mathbf{\Phi}\mathbf{x}^{*} + \mathbf{e}$, and (5.26c) is due to the fact that we can split the target signal as follow, $\mathbf{\Phi}\mathbf{x}^{*} = \mathbf{\Phi}_{\mathbf{A}_c}\mathbf{x}^{*}_{\mathcal{A}_c} + \mathbf{\Phi}_{\mathbf{A}}\mathbf{x}^{*}_{\mathcal{A}}$. The remaining equations (5.26d) and (5.26e), and inequality (5.26f) are due to algebraic manipulations.

Now, since the measurement matrix $\mathbf{\Phi}$ satisfy the restricted isometry property with parameters (r, δ), we recall the following properties from Lemma 3.8 and Lemma 3.9, with $|\mathcal{A}| = p < r$,

$$\left\|\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\right\| \leq \frac{1}{1-\delta} \tag{5.27}$$

$$\left\|\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}^*\cup\mathcal{A}_c}\right\| \leq \delta \tag{5.28}$$

$$\left\|\left(\mathbf{\Phi}^T_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}}\right)^{-1}\mathbf{\Phi}^T_{\mathcal{A}}\right\| \leq \frac{1}{\sqrt{1-\delta}} \tag{5.29}$$

Using the restricted isometry matrix properties bounds (5.27), (5.28) and (5.29) the inequality (5.26f) simplifies to

$$\left\|\mathbf{x}_{\mathcal{A}}^{st} - \mathbf{x}_{\mathcal{A}}^{*}\right\|_2 \leq \frac{\delta}{1-\delta}\left\|\mathbf{x}_{\mathcal{A}_c}^{*}\right\|_2 + \frac{1}{\sqrt{1-\delta}}\|\mathbf{e}\|_2 + \frac{\mu}{1-\delta}\sqrt{p}. \tag{5.30}$$

Finally, substituting the result of inequality (5.30) into (5.25) we obtain

$$\|\mathbf{x}^{st} - \mathbf{x}^{*}\|_2 \leq \frac{\delta}{1-\delta}\left\|\mathbf{x}_{\mathcal{A}_c}^{*}\right\|_2 + \frac{1}{\sqrt{1-\delta}}\|\mathbf{e}\|_2 + \frac{\mu}{1-\delta}\sqrt{p} + \left\|\mathbf{x}_{\mathcal{A}_c}^{*}\right\|_2 \tag{5.31a}$$

$$\leq \frac{1}{1-\delta}\left(\|\mathbf{x}_{\mathcal{A}_c}^{*}\|_2 + \sqrt{1-\delta}\|\mathbf{e}\|_2 + \mu\sqrt{p}\right) \tag{5.31b}$$

$$\leq \frac{1}{1-\delta}\left(\|\mathbf{x}^{*}\|_2 + \sqrt{1-\delta}\|\mathbf{e}\|_2 + \mu\sqrt{p}\right) \tag{5.31c}$$

$$= (1-\delta)^{-1}B(p). \tag{5.31d}$$

□

The above lemma states the bound only for the steady-state solution. It is important that we extend the results to the entire trajectory. The following lemma states that ℓ_2-distance of the output $\mathbf{x}(t)$ to the target signal $\mathbf{x}^*$ remains bounded for all $t \geq 0$.

Lemma 5.5. *Assume that, at the switching time t_k, the current active set $\mathcal{A}(t_k) = \mathcal{A}^k$ contains less than p indices. Furthermore, assume the measurement matrix $\mathbf{\Phi}$ satisfies the restricted isometry property with parameters (r, δ), where $r = |\mathcal{A}^* \cup \mathcal{A}^k|$, and that*

$$\|\mathbf{x}_{\mathcal{A}}(t_k) - \mathbf{x}_{\mathcal{A}}^{*}\|_2 \leq \frac{1+\delta}{(1-\delta)^2}B_\delta(p). \tag{5.32}$$

Then, for all $t \in [t_k, t_{k+1}]$,

$$\|\mathbf{x}(t) - \mathbf{x}^{*}\|_2 \leq \frac{1+\delta}{(1-\delta)^2}B_\delta(p).$$

Proof. Using the solution (5.15) of the dynamical system for the active set, the following argument holds for all $t \in [t_k, t_{k+1})$. Note that we set $\gamma = 0$ to simplify the calculations

$$
\begin{aligned}
\|\mathbf{x}(t) - \mathbf{x}^*_{\mathcal{A}}\|_2 &= \|\mathbf{x}_{\mathcal{A}^k}(t) - \mathbf{x}^*\|_2 \\
&= \left\| e^{-\mathbf{A}(t-t_k)} \mathbf{x}_{\mathcal{A}^k}(t_k) + \left(\mathbf{I} - e^{-\mathbf{A}(t-t_k)}\right) \mathbf{A}^{-1} \left(\mathbf{\Phi}^T_{\mathcal{A}^k} \mathbf{y} - \mu \mathbf{s}_{\mathcal{A}^k}\right) - \mathbf{x}^* \right\|_2 && \text{(5.33a)} \\
&= \left\| e^{-\mathbf{A}(t-t_k)} \mathbf{x}_{\mathcal{A}^k}(t_k) + \left(\mathbf{I} - e^{-\mathbf{A}(t-t_k)}\right) \mathbf{x}^{st}_{\mathcal{A}} - \mathbf{x}^* \right\|_2 && \text{(5.33b)} \\
&= \left\| e^{-\mathbf{A}(t-t_k)} \left(\mathbf{x}_{\mathcal{A}^k}(t_k) - \mathbf{x}^*\right) + \left(\mathbf{I} - e^{-\mathbf{A}(t-t_k)}\right) \left(\mathbf{x}^{st}_{\mathcal{A}^k} - \mathbf{x}^*\right) \right\|_2 && \text{(5.33c)} \\
&\leq \left\| e^{-\mathbf{A}(t-t_k)} \left(\mathbf{x}_{\mathcal{A}^k}(t_k) - \mathbf{x}^*\right) \right\|_2 + \left\| \left(\mathbf{I} - e^{-\mathbf{A}(t-t_k)}\right) \left(\mathbf{x}^{st}_{\mathcal{A}^k} - \mathbf{x}^*\right) \right\|_2 && \text{(5.33d)} \\
&\leq \left\| e^{-\mathbf{A}(t-t_k)} \right\|_2 \|\mathbf{x}_{\mathcal{A}^k}(t_k) - \mathbf{x}^*\|_2 + \left\| \mathbf{I} - e^{-\mathbf{A}(t-t_k)} \right\|_2 \left\| \mathbf{x}^{st}_{\mathcal{A}^k} - \mathbf{x}^* \right\|_2 && \text{(5.33e)} \\
&\leq e^{-(1-\delta)(t-t_k)} \|\mathbf{x}_{\mathcal{A}^k}(t_k) - \mathbf{x}^*\|_2 + \left(1 - e^{-(1+\delta)(t-t_k)}\right) \left\| \mathbf{x}^{st}_{\mathcal{A}^k} - \mathbf{x}^* \right\|_2 && \text{(5.33f)} \\
&\leq e^{-(1-\delta)(t-t_k)} \frac{1+\delta}{(1-\delta)^2} B_\delta(p) + \left(1 - e^{-(1+\delta)(t-t_k)}\right) \frac{1}{1-\delta} B_\delta(p) && \text{(5.33g)} \\
&= \frac{1+\delta}{(1-\delta)^2} B_\delta(p) \left[e^{-(1-\delta)(t-t_k)} + \left(1 - e^{-(1+\delta)(t-t_k)}\right) \frac{1-\delta}{1+\delta} \right] && \text{(5.33h)} \\
&\leq \frac{1+\delta}{(1-\delta)^2} B_\delta(p). && \text{(5.33i)}
\end{aligned}
$$

where $\mathbf{A} = \mathbf{\Phi}^T_{\mathcal{A}} \mathbf{\Phi}_{\mathcal{A}}$. The equation (5.33a) follows from the dynamical system (5.15) with $\gamma = 0$. The equation (5.33b) is due to the fact that the steady-state solution satisfy $\left(\mathbf{\Phi}^T_{\mathcal{A}} \mathbf{\Phi}_{\mathcal{A}}\right) \mathbf{x}^{st}_{\mathcal{A}} = \mathbf{\Phi}^T_{\mathcal{A}} \mathbf{y} - \mu \mathbf{s}_{\mathcal{A}}$. The inequality (5.33f) is due to the fact that $\mathbf{A} = \mathbf{\Phi}^T_{\mathcal{A}} \mathbf{\Phi}_{\mathcal{A}}$ and all the eigenvalues of $\mathbf{\Phi}^T_{\mathcal{A}} \mathbf{\Phi}_{\mathcal{A}}$ lie in the range $[1-\delta, 1+\delta]$. We get inequality (5.33f) using (5.32) and the results of Lemma 5.4. It is easy to show that the expression in the square bracket in equation (5.33h) is in the range $(0, 1)$. Thus, the results of inequality (5.33i) follow from this fact.

Finally, we show that the result does not blow up at t_{k+1}. Since $\mathbf{x}(t) - \mathbf{x}^*$ is continuous in time we have

$$\|\mathbf{x}_{\mathcal{A}^{k+1}}(t_{k+1}) - \mathbf{x}^*\|_2 = \|\mathbf{x}_{\mathcal{A}^k}(t_{k+1}) - \mathbf{x}^*\|_2 \leq \frac{1+\delta}{(1-\delta)^2} B_\delta(p).$$

□

The following lemma will also be useful for the proof of main results. The lemma states that if the energy in the q components with largest magnitudes in $\mathbf{u}(t)$ satisfies a certain inequality then there is more than q active components at the time.

Lemma 5.6. *If $\mathcal{B}$ contains the indices of q components of $\mathbf{u}(t)$ with largest magnitude and*

$$\|\mathbf{u}_\mathcal{B}(t)\|_2 \leq \mu\sqrt{q}, \tag{5.34}$$

then the active set $\mathcal{A}$ corresponding to nonzero components in $\mathbf{x}(t) = S_\mu(\mathbf{u}(t))$ is the subset of $\mathcal{B}$ and contains less than q indices, that is, $\mathcal{A} \subset \mathcal{B}$ and $|\mathcal{A}| \leq q$.

Proof. Since, $\mathcal{B}$ contains the q components with the largest magnitude of $\mathbf{u}(t)$, the smallest component $u_j(t)$ for $j \in \mathcal{B}$ satisfies

$$q|u_j(t)|^2 \leq \sum_{i\in\mathcal{B}} |u_i(t)|^2, \tag{5.35a}$$

$$\implies |u_j| \leq \frac{\|\mathbf{u}\|_\mathcal{B}(t)}{\sqrt{q}}. \tag{5.35b}$$

Now for all the components in $i \in \mathcal{B}_c$ we have

$$|u_i(t)| \leq |u_j(t)|, \tag{5.36a}$$

$$\leq \frac{\|\mathbf{u}_\mathcal{B}\|(t)\|_2}{\sqrt{q}}, \tag{5.36b}$$

$$\leq \mu, \tag{5.36c}$$

where (5.36b) results from (5.35b), and (5.36c) results from (5.34). Thus, the components in $\mathcal{B}_c$ are all below the threshold. Therefore, $\mathcal{A} \subset \mathcal{B}$, and $|\mathcal{A}| \leq |\mathcal{B}| = q$. □

5.4 Bounds of the active set

In this section, we present some results that guarantee that the size of the active set remains bounded throughout convergence. The results are significant when looking for a sparse solution. We first show that under certain conditions, the active set never contains more than s components for s-sparse target signal. In the second part, we show that the size of the active set is bounded by a constant, which makes the dynamical system more practical when solving a sparse recovery problem.

5.4.1 The optimal support bound the active set

Consider the measurements corrupted with noise, $\mathbf{y} = \mathbf{\Phi}\mathbf{x}^* + \mathbf{e}$. Then the following theorem states that under certain conditions based on the restricted isometry property and signal sparsity level, the size of the active set is bounded by the size of the optimal support.

Theorem 5.7. *Assume that the matrix* $\mathbf{\Phi}$ *satisfies the restricted isometry property with parameters* (s, δ) *and the initial active set* $\mathcal{A}(0)$ *is a subset of the optimal support* $\mathcal{A}^*$*. Suppose the threshold* μ*, the sparsity level* s*, the restricted isometry constant* δ*, the target signal* $\mathbf{x}^*$ *and the noise vector* $\mathbf{e}$ *have the follow relation:*

$$\|\mathbf{x}^* - \mathbf{x}(0)\|_2 \leq \frac{1+\delta}{(1-\delta)^2} B_\delta(s), \tag{5.37}$$

$$\frac{1+\delta}{(1-\delta)^2}\delta\left(\|\mathbf{x}^*\|_2 + \sqrt{1-\delta}\|\mathbf{e}\|_2\right) + \left\|\mathbf{\Phi}_{\mathcal{A}_c}^T \mathbf{e}\right\|_\infty \leq \left(1 - \frac{1+\delta}{(1-\delta)^2}\delta\sqrt{s}\right)\mu. \tag{5.38}$$

If the conditions (5.37) and (5.38) are satisfied, then $\mathcal{A}(t) \subseteq \mathcal{A}^*$*, for all* $t \geq 0$*.*

Proof. We proof by induction over switching time t_k that the active set $\mathcal{A}$ is subset of optimal support $\mathcal{A}^*$ for all $t \geq 0$.

For $t = 0$, the theory hypothesis asserts that the support set $\mathcal{A}(0)$ of the initial output $\mathbf{x}(0)$ is a subset of the optimal support set $\mathcal{A}^*$. That is $\mathcal{A}(t) \subset \mathcal{A}^*$ for $t = 0$.

To proceed we make the following inductive hypothesis:

(a) At all switching times t_k, the following holds true

$$|u_i(t)| \leq \mu, \quad \forall i \in \mathcal{A}_c^* \text{ and } \forall t \in [t_k, t_{k+1}). \tag{5.39}$$

The condition (5.39) is conceptualised from the definition of the inactive set. If the condition is satisfied for all $t \geq 0$, the component $i \in \mathcal{A}_c^*$ remain below the threshold μ.

(b) At the switching time t_k, the distance between the output restricted to the active set $\mathbf{x}_{\mathcal{A}^k}$ and the target signal $\mathbf{x}^*$ is bounded, that is,

$$\|\mathbf{x}_{\mathcal{A}^k}(t_k) - \mathbf{x}^*\|_2 \leq \frac{1+\delta}{(1-\delta)^2} B_\delta(s). \tag{5.40}$$

The condition (5.40) is the consequence of Lemma 5.5.

Note that the initial condition $\mathbf{x}(0)$ is set such that the initial support set $\mathcal{A}^0$ is a subset of the optimal support set $\mathcal{A}^*$. Using this fact and condition (5.37), it is clear that the conditions (5.39) and (5.40) of the inductive hypothesis hold true at $t = 0$.

Now, assume that the two inductive hypotheses (a) and (b) hold at the switching time t_k. If there is no more switching after t_k, then the theorem is proven. Otherwise, we need the show that the two inductive hypotheses hold true at the next switching time t_{k+1}. Using the solution of the dynamical system for inactive set, equation (5.16), it follows that for all $i \in \mathcal{A}_c^* \subset \mathcal{A}_c^k$ and for all $t \in [t_k, t_{k+1}]$ we have

$$u_i(t) = e^{-(t-t_k)} u_i(t_k) + e^{-t} \int_{t_k}^{t} e^{\eta} \rho_i(\eta) d\eta,$$

were $\rho_i(t) = \mathbf{\Phi}_i^T (\mathbf{y} - \mathbf{\Phi}_{\mathcal{A}^k} \mathbf{x}_{\mathcal{A}^k}(t))$. The absolute value of the above state variable can be bounded by:

$$|u_i(t)| = \left| e^{-(t-t_k)} u_i(t_k) + e^{-t} \int_{t_k}^{t} e^{\eta} \rho_j(\eta) d\eta \right|, \tag{5.41a}$$

$$\leq e^{-(t-t_k)} |u_i(t_k)| + e^{-t} \int_{t_k}^{t} e^{\eta} |\rho_j(\eta)| \, d\eta, \tag{5.41b}$$

$$\leq e^{-(t-t_k)} |u_i(t_k)| + \left(1 - e^{-(t-t_k)}\right) \sup_{t \in [t_k, t_{k+1}]} |\rho_i(t)|. \tag{5.41c}$$

Since, at time t_k, component $i \in \mathcal{A}_c^*$ is inactive, then $|u_i(t_k)| \leq \mu$. Thus, the condition (5.39) of the first inductive hypothesis is satisfied if

$$\sup_{t\in[t_k,t_{k+1}]} |\rho_i(t)| \leq \mu. \tag{5.42}$$

Hence, we need to show that above expression (5.42) is satisfied.

Since $\mathbf{\Phi}$ satisfies the restricted isometry property with parameters (s, δ), Lemma 3.9 yields that $\|\mathbf{\Phi}_i^T \mathbf{\Phi}_{\mathcal{A}^*}\| \leq \delta$. Then, for all time $t \in [t_k, t_{k+1}]$ and for all components $i \in \mathcal{A}_c^*$,

$$|\rho_i(t)| = \left|\mathbf{\Phi}_i^T(\mathbf{y} - \mathbf{\Phi}_{\mathcal{A}^k}\mathbf{x}_{\mathcal{A}^k})\right|, \tag{5.43a}$$

$$= \left|\mathbf{\Phi}_i^T\left(\mathbf{\Phi}_{\mathcal{A}^*}\mathbf{x}^* + \mathbf{e} - \mathbf{\Phi}_{\mathcal{A}^k}\mathbf{x}_{\mathcal{A}^k}\right)\right|, \tag{5.43b}$$

$$\leq \left|\mathbf{\Phi}_i^T\mathbf{\Phi}_{\mathcal{A}^*}\left(\mathbf{x}_{\mathcal{A}}^* - \mathbf{x}_{\mathcal{A}^k}(t)\right) + \mathbf{\Phi}_i^T\mathbf{e}\right|, \tag{5.43c}$$

$$\leq \left|\mathbf{\Phi}_i^T\mathbf{\Phi}_{\mathcal{A}^*}\left(\mathbf{x}_{\mathcal{A}}^* - \mathbf{x}_{\mathcal{A}^k}(t)\right)\right| + \left|\mathbf{\Phi}_i^T\mathbf{e}\right|, \tag{5.43d}$$

$$\leq \left\|\mathbf{\Phi}_i^T\mathbf{\Phi}_{\mathcal{A}^*}\right\| \left\|\mathbf{x}_{\mathcal{A}^k}(t) - \mathbf{x}_{\mathcal{A}}^*\right\|_2 + \left\|\mathbf{\Phi}_{\mathcal{A}_c^*}^T\mathbf{e}\right\|_\infty, \tag{5.43e}$$

$$\leq \delta\left\|\mathbf{x}_{\mathcal{A}^k}(t) - \mathbf{x}_{\mathcal{A}}^*\right\|_2 + \left\|\mathbf{\Phi}_{\mathcal{A}_c^*}^T\mathbf{e}\right\|_\infty, \tag{5.43f}$$

$$\leq \frac{(1+\delta)\delta}{(1-\delta)^2}B_\delta(s) + \left\|\mathbf{\Phi}_{\mathcal{A}_c^*}^T\mathbf{e}\right\|_\infty, \tag{5.43g}$$

$$= \frac{(1+\delta)\delta}{(1-\delta)^2}\left(\|\mathbf{x}^*\|_2 + \sqrt{1-\delta}\|\mathbf{e}\|_2\right) + \left\|\mathbf{\Phi}_{\mathcal{A}_c^*}^T\mathbf{e}\right\|_\infty + \frac{(1+\delta)\delta}{(1-\delta)^2}\sqrt{s}, \tag{5.43h}$$

$$\leq \mu\left(1 - \frac{(1+\delta)\delta}{(1-\delta)^2}\sqrt{s} + \frac{(1+\delta)}{(1-\delta)^2}\delta\sqrt{s}\right), \tag{5.43i}$$

$$= \mu, \tag{5.43j}$$

where the equation (5.43b) results from $\mathbf{y} = \mathbf{\Phi}\mathbf{x}^* + \mathbf{e}$, and (5.43c) is due to the fact that $\mathcal{A}^k \subset \mathcal{A}^*$, the inequality (5.43g) is due to the application of Lemma 5.5, and the equality (5.43h) results from (5.38). The inequality (5.43i) and equality (5.43j) are due algebraic simplification.

This shows that (5.42) holds true. Thus, the first inductive hypothesis hold true for all $t \in [t_k, t_{k+1}]$, in particular at t_{k+1}.

Now, applying Lemma 5.5 again, we obtain a bound that holds uniformly across time:

$$\|\mathbf{x}(t) - \mathbf{x}^*_{\mathcal{A}}\|_2 \leq \frac{1+\delta}{(1-\delta)^2} B_\delta(s), \qquad \forall t \in [t_k, t_{k+1}].$$

In particular, $\|\mathbf{x}_{\mathcal{A}^{k+1}}(t_{k+1}) - \mathbf{x}^*\|_2 \leq \frac{1+\delta}{(1-\delta)^2} B_\delta(s)$ and the second induction hypothesis (5.40) remains true at time t_{k+1}. □

5.4.2 The active set bounded by the constant

Under certain conditions on the initial condition $\mathbf{u}(0)$ and the threshold μ, more than s components may be active at the optimal solution. This is more useful for practical purpose. We denote the maximum number of active components by q, where q may be larger than s but remains within small deviation from s. Bounding the size of the active set is important for the efficient recovery of a sparse signal by the dynamical system.

Theorem 5.8. *Assume that the measurement matrix satisfies the restricted isometry property with parameters (q, δ) for some $q \geq 0$. If the target signal $\mathbf{x}^*(t)$, the initial state $\mathbf{u}(0)$, the threshold μ, the noise $\mathbf{e}$, the parameter q and the restricted isometry constant δ satisfies the follow conditions*

$$\|\mathbf{u}(0)\|_2 \leq \mu\sqrt{q}, \tag{5.44}$$

$$\frac{1+\delta}{1-3\delta}\frac{1}{\sqrt{q}}\left(\|\mathbf{x}^*\|_2 + \sqrt{1-\delta}\|\mathbf{e}\|_2\right) \leq \mu, \tag{5.45}$$

then the active set $\mathcal{A}$ never contains more than q components, that is, $|\mathcal{A}(t)| \leq q$ for all $t \geq 0$.

Proof. We prove by induction over the switching times t_k. We define the set $\mathcal{B}$ to be the support set of q components of $\mathbf{u}(t)$ with the largest magnitude. We make use of the following inductive hypotheses:

(a) The norm of state variable restricted to $\mathcal{B}$ is bounded as

$$\|\mathbf{u}_{\mathcal{B}}(t)\|_2 \leq \mu\sqrt{q}, \tag{5.46}$$

for all $t \geq 0$. This condition follows from Lemma 5.6.

(b) The distance between the output $\mathbf{x}(t)$ and the target signal $\mathbf{x}^*$ is bounded as

$$\|\mathbf{x}(t) - \mathbf{x}^*\|_2 \leq \frac{1+\delta}{(1-\delta)^2} B_\delta(q), \tag{5.47}$$

for all $t \leq t_k$.

Now, for $t = 0$, condition (5.46) of the first inductive hypothesis holds true. Moreover,

$$\|\mathbf{x}(0) - \mathbf{x}^*\|_2 \leq \|\mathbf{x}(0)\|_2 + \|\mathbf{x}^*\|_2, \tag{5.48a}$$

$$\leq \|\mathbf{u}(0)\|_2 + \|\mathbf{x}^*\|_2, \tag{5.48b}$$

$$\leq \mu\sqrt{q} + \|\mathbf{x}^*\|_2, \tag{5.48c}$$

$$\leq \mu\sqrt{q} + \|\mathbf{x}^*\|_2 + \sqrt{1-\delta}\|\mathbf{e}\|_2, \tag{5.48d}$$

$$\leq \frac{1+\delta}{(1-\delta)^2} B_\delta(q), \tag{5.48e}$$

where (5.48c) results from (5.44), and (5.48e) we used the definition of B_δ in (5.24). Thus, condition (5.47) of the second inductive hypothesis holds true at $t = 0$.

To proceed, we assume that for some switching time t_k, the conditions (5.46) and (5.47) of the inductive hypotheses hold true. If there is no more switching time, then the theorem is proven. Otherwise, we need to show that the inductive hypotheses hold true at next switching time t_{k+1}.

To prove that (5.46) holds true at t_{k+1}, consider the dynamical system (5.10) with $\tau = 1$ on the set $\mathcal{B}$ of q largest components of $\mathbf{u}(t)$:

$$\dot{\mathbf{u}}_{\mathcal{B}}(t) = -\mathbf{u}_{\mathcal{B}}(t) - \left(\mathbf{\Phi}_{\mathcal{B}}^T\mathbf{\Phi}_{\mathcal{B}} - I\right)\mathbf{x}_{\mathcal{B}}(t) + \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{y}. \tag{5.49}$$

Using the extension Gronwall's Lemma (Lemma 2.18 from Section 2.4), the solution of (5.49) for $t \in [t_k, t_{k+1}]$ can be written as:

$$\mathbf{u}_{\mathcal{B}}(t) = e^{-(t-t_k)}\mathbf{u}_{\mathcal{B}}(t_k) + e^{-t}\int_{t_k}^{t} e^{\eta}\beta_{\mathcal{B}}(\eta)d\eta, \tag{5.50}$$

where $\beta_{\mathcal{B}}(t) = \mathbf{x}_{\mathcal{B}}(t) - \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{\Phi}_{\mathcal{B}}\mathbf{x}_{\mathcal{B}}(t) + \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{y}$. The state variable $\mathbf{u}_{\mathcal{B}}(t)$ from (5.50) is bounded as follows:

$$\|\mathbf{u}_{\mathcal{B}}(t)\|_2 \leq e^{-(t-t_k)}\|\mathbf{u}_{\mathcal{B}}(t_k)\|_2 + e^{-t}\int_{t_k}^{t} e^{\eta}\|\beta_{\mathcal{B}}(\eta)\|_2 d\eta, \tag{5.51a}$$

$$\leq e^{-(t-t_k)}\|\mathbf{u}_{\mathcal{B}}(t_k)\|_2 + (1 - e^{-(t-t_k)}) \sup_{t\in[t_k,t_{k+1}]} \|\beta_{\mathcal{B}}(t)\|_2. \tag{5.51b}$$

By the induction hypothesis (a), the results are true at t_k, thus, the first term of the inequality (5.51b) is bounded as:

$$\|\mathbf{u}_{\mathcal{B}}(t_k)\|_2 \leq \mu\sqrt{q}, \tag{5.52}$$

and Lemma 5.6 implies that $\mathcal{A}(t_k) = \mathcal{A}^k \subset \mathcal{B}$ and $\mathcal{A}^k$ contains less than q components. Now we have to obtain a bound of $\beta_{\mathcal{B}}$ for all $t \in [t_k, t_{k+1}]$,

$$\|\beta_{\mathcal{B}}(t)\|_2 = \|\mathbf{x}_{\mathcal{B}}(t) - \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{\Phi}\mathbf{x}(t) + \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{y}\|_2, \tag{5.53a}$$

$$= \|\mathbf{x}^*_{\mathcal{B}} + \left(\mathbf{I} - \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{\Phi}_{\mathcal{B}}\right)(\mathbf{x}_{\mathcal{B}}(t) - \mathbf{x}^*_{\mathcal{B}}) + \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{e}\|_2, \tag{5.53b}$$

$$\leq \|\mathbf{x}^*_{\mathcal{B}}\|_2 + \|\mathbf{I} - \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{\Phi}_{\mathcal{B}}\|_2\|\mathbf{x}_{\mathcal{B}}(t) - \mathbf{x}^*_{\mathcal{B}}\|_2 + \|\mathbf{\Phi}_{\mathcal{B}}^T\mathbf{e}\|_2, \tag{5.53c}$$

$$\leq \|\mathbf{x}^*\|_2 + \|\mathbf{I} - \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{\Phi}_{\mathcal{B}\cup\mathcal{A}^*}\|_2\|\mathbf{x}(t) - \mathbf{x}^*\|_2 + \|\mathbf{\Phi}_{\mathcal{B}}^T\mathbf{e}\|_2, \tag{5.53d}$$

$$\leq \|\mathbf{x}^*\|_2 + \delta\|\mathbf{x}(t) - \mathbf{x}^*\|_2 + (1+\delta)\|\mathbf{e}\|_2, \tag{5.53e}$$

$$\leq \|\mathbf{x}^*\|_2 + \frac{(1+\delta)\delta}{(1-\delta)^2}B_\delta(q) + (1+\delta)\|\mathbf{e}\|_2, \tag{5.53f}$$

$$= \|\mathbf{x}^*\|_2 + \frac{(1+\delta)\delta}{(1-\delta)^2}\left(\|\mathbf{x}^*\|_2 + \sqrt{1-\delta}\|\mathbf{e}\|_2 + \mu\sqrt{p}\right) + (1+\delta)\|\mathbf{e}\|_2, \tag{5.53g}$$

$$= \left(1 + \frac{(1+\delta)\delta}{(1-\delta)^2}\right)\|\mathbf{x}^*\|_2 + (1+\delta)\left(\frac{\delta\sqrt{1-\delta}}{(1-\delta)^2} + 1\right)\|\mathbf{e}\|_2 + \frac{(1+\delta)\delta}{(1-\delta)^2}\mu\sqrt{q}, \tag{5.53h}$$

$$< \frac{1 - 3\delta + (1+\delta)\delta}{(1-\delta)^2}\mu\sqrt{q}, \tag{5.53i}$$

$$\implies \|\beta_{\mathcal{B}}(t)\|_2 = \mu\sqrt{q}. \tag{5.53j}$$

Here (5.53a) results from the definition of $\beta_{\mathcal{B}}$ in equation (5.50). The inequality (5.53e) is due to the fact that the matrix $\mathbf{\Phi}$ satisfies the restricted isometry property with parameters (q, δ) and $\mathcal{A} \subset \mathcal{B}$, property 1 of Lemma 3.9 can be applied to the matrix $\mathbf{I} - \mathbf{\Phi}_{\mathcal{B}}^T\mathbf{\Phi}_{\mathcal{B}\cup\mathcal{A}^*}$, and property 1 of Lemma 3.8 can be applied to last term of (5.53d). The inequality (5.53f) results from Lemma 5.5. The equation (5.53g) results from the definition of $B_\delta(q)$ in (5.24). The inequality (5.53i) results from the hypothesis given in (5.38) of the theorem.

Hence, substituting the inequalities (5.52) and (5.53i) into (5.51b) the state variable on $\mathcal{B}$ is bounded as $\|\mathbf{u}_{\mathcal{B}}\|_2 \leq \mu\sqrt{q}$ for all $t \in [t_k, t_{k+1}]$. In particular the induction condition (5.46) hold at t_{k+1}.

Since condition (5.47) of the second hypothesis hold true at t_k, and the vector $\mathbf{x}(t) - \mathbf{x}^*$ is continuous in time we have

$$\|\mathbf{x}_{\mathcal{A}^{k+1}}(t_{k+1}) - \mathbf{x}^*\|_2 = \|\mathbf{x}_{\mathcal{A}^k}(t_{k+1}) - \mathbf{x}^*\|_2 \leq \frac{(1+\delta)}{(1-\delta)^2} B_\delta(q)$$

Thus the second inductive hypothesis hold true at t_{k+1}. □

These results are important in a sense that the Bregman dynamical system is guaranteed to recover the appropriate support set of the sparse solution.

5.5 Convergence rate

Convergence to the correct solution for any starting state is fundamental for any system that is intended to solve an optimization program. Knowing how fast the trajectories converge to the solution is even more interesting for practical applications. We discuss the convergence rate of the Bregman integrated dynamical system

$$\begin{cases} \tau \dot{\mathbf{u}}(t) = -\mathbf{u}(t) - (\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\mathbf{x}(t) + \mathbf{\Phi}^T\mathbf{y} - \gamma\left(\mathbf{x}(t) - \mathbf{z}\right) \\ \mathbf{x}(t) = S_\mu(\mathbf{u}(t)). \end{cases} \tag{5.54}$$

Here, we introduce the parameter τ similar to the dynamical system of the Locally Competitive Algorithm. The parameter τ plays a role in the convergence rate and depends on the hardware implementing the dynamical system. Furthermore, the expression for convergence speed depends on the Bregman parameter γ, the restricted isometry constant δ of the measurement matrix $\mathbf{\Phi}$, and on the bound U of the subgradients of the activation function.

In order to establish the convergence speed of (5.54), we make use of the following error variables

$$\bar{\mathbf{u}}(t) = \mathbf{u}(t) - \mathbf{u}^* \tag{5.55a}$$

$$\bar{\mathbf{x}}(t) = \mathbf{x}(t) - \mathbf{x}^*. \tag{5.55b}$$

The dynamical system (5.54) can be written in terms of the new variables (5.55a) and (5.55b). Using the fact that $\mathbf{u}^*$ is an equilibrium point, it follows that $\dot{\mathbf{u}}^*(t) = \mathbf{0}$, and the dynamical system (5.54) reduces to

$$\tau\dot{\bar{\mathbf{u}}}(t) = \tau\dot{\mathbf{u}}(t) \tag{5.56a}$$

$$= -\mathbf{u}(t) - \left(\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I}\right)\mathbf{x}(t) + \mathbf{\Phi}^T\mathbf{y} - \gamma(\mathbf{x}(t) - \mathbf{z}) \tag{5.56b}$$

$$= -\bar{\mathbf{u}}(t) - \mathbf{u}^* - \left[\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I}\right](\bar{\mathbf{x}}(t) + \mathbf{x}^*) + \mathbf{\Phi}^T\mathbf{y} - \gamma(\bar{\mathbf{x}}(t) + \mathbf{x}^* - \mathbf{z}) \tag{5.56c}$$

$$= -\bar{\mathbf{u}}(t) - \left[\mathbf{\Phi}^T\mathbf{\Phi} - (1-\gamma)\mathbf{I}\right]\bar{\mathbf{x}}(t) - \mathbf{u}^* - \left(\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I}\right)\mathbf{x}^* + \mathbf{\Phi}^T\mathbf{y} - \gamma(\mathbf{x}^* - \mathbf{z}) \tag{5.56d}$$

$$= -\bar{\mathbf{u}}(t) - \left(\mathbf{\Phi}^T\mathbf{\Phi} - (1-\gamma)\mathbf{I}\right)\bar{\mathbf{x}}(t) + \dot{\mathbf{u}}^*(t) \tag{5.56e}$$

$$\implies \tau\dot{\bar{\mathbf{u}}}(t) = -\bar{\mathbf{u}}(t) - \left(\mathbf{\Phi}^T\mathbf{\Phi} - (1-\gamma)\mathbf{I}\right)\bar{\mathbf{x}}(t). \tag{5.56f}$$

Here, equation (5.56b) follows from the definition of the dynamical system (5.54), equation (5.56c) follows from (5.55a) and (5.55b), and for (5.56f) we use the fact that $\mathbf{u}^*$ is the equilibrium point, therefore $\dot{\mathbf{u}}^* = \mathbf{0}$.

The following lemma will be used in the proof of convergence rate.

Lemma 5.9. *Suppose that the activation function satisfies Assumption 4.2. Then, the variables* $\bar{\mathbf{u}}$ *and* $\bar{\mathbf{x}}$ *satisfies the following property*

$$|\bar{\mathbf{x}}|_2 \leq U|\bar{\mathbf{u}}|_2. \tag{5.57}$$

Now, we are ready to state and prove the convergence rate of the Bregman integrated dynamical system (5.54). The following theorem states that the Bregman integrated system has an exponential convergence rate.

Theorem 5.10. *Suppose the measurement matrix* $\mathbf{\Phi}$ *satisfies the restricted isometry property with parameters* (s, δ) *and the subgradients of the activation function are bounded by* U*, that is, for* $\xi_i \in \partial S_\mu(u_i)$ *we have* $|\xi_i| \leq U$*. Then the convergence rate of the Bregman integrated*

dynamic system (5.54) satisfies:

$$\|\bar{\mathbf{u}}(t)\|_2 \leq C\|\bar{\mathbf{u}}(0)\|_2 \exp\left(-\frac{1-U(\delta+\gamma)}{\tau}t\right) \tag{5.58}$$

for some constant C.

Proof. We establish the convergence rate using the Lyapunov's function method, see the discussion in Section 2.3.2. We make used of the following energy function:

$$E(t) = \frac{1}{2}\|\bar{\mathbf{u}}\|_2^2. \tag{5.59}$$

For all the time $t \geq 0$, the time derivative of the energy function (5.59) along the trajectory of the Bregman integrated dynamical system (5.54) is:

$$\tau\dot{E}(t) = \tau\bar{\mathbf{u}}^T(t)\dot{\bar{\mathbf{u}}}(t), \tag{5.60a}$$

$$= -\bar{\mathbf{u}}^T(t)\left[\bar{\mathbf{u}}(t) + \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} - (1-\gamma)\mathbf{I}\right)\bar{\mathbf{x}}\right], \tag{5.60b}$$

$$= -\|\bar{\mathbf{u}}(t)\|_2^2 - \bar{\mathbf{u}}^T(t)\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} - (1-\gamma)\mathbf{I}\right)\bar{\mathbf{x}}(t), \tag{5.60c}$$

$$\leq -\|\bar{\mathbf{u}}(t)\|_2^2 + \left\|\bar{\mathbf{u}}^T(t)\right\|_2^2 \left\|\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} - (1-\gamma)\mathbf{I}\right)\bar{\mathbf{x}}(t)\right\|_2^2, \tag{5.60d}$$

where (5.60a) is the derivative of (5.59) with respect to time t, (5.60b) results from (5.56f).

Consider the simplification of the following expression from the second term of inequality (5.13d). Since, the matrix $\boldsymbol{\Phi}$ satisfies the restricted isometry property with parameters (s, δ), then the eigenvalues of the matrix $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ lies between $(1-\delta)$ and $(1+\delta)$ and we get

$$\left\|\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} - (1-\gamma)\mathbf{I}\right)\bar{\mathbf{x}}(t)\right\|_2 \leq \left\|\boldsymbol{\Phi}^T\boldsymbol{\Phi} - (1-\gamma)\mathbf{I}\right\|_2 \|\bar{\mathbf{x}}(t)\|_2, \tag{5.61a}$$

$$\leq \max\{(1+\delta)-(1-\gamma), (1-\gamma)-(1-\delta)\}\|\bar{\mathbf{x}}(t)\|_2, \tag{5.61b}$$

$$= (\delta+\gamma)\|\bar{\mathbf{x}}(t)\|_2, \tag{5.61c}$$

Hence substituting the inequality (5.61c) into (5.60d), the time derivative of energy simplifies to the following expression:

$$\tau \dot{E}(t) \leq - \|\bar{\mathbf{u}}(t)\|_2^2 + (\delta + \gamma) \|\bar{\mathbf{u}}(t)\|_2^2 \|\bar{\mathbf{x}}(t)\|_2^2 , \tag{5.62a}$$

$$\leq -\|\bar{\mathbf{u}}(t)\|_2 + U(\delta + \gamma)\|\bar{\mathbf{u}}(t)\|_2^2, \tag{5.62b}$$

$$= - \left[1 - U(\delta + \gamma)\right] \|\bar{\mathbf{u}}(t)\|_2^2 , \tag{5.62c}$$

$$\implies \tau \dot{E}(t) \leq -2 \left[1 - U(\delta + \gamma)\right] E(t), \tag{5.62d}$$

where (5.62c) results from (5.61c) and (5.60d), (5.62b) result from property (5.57) of Lemma 5.9, and (5.62d) results from the definition of the energy function in (5.59).

To proceed, we use Gronwall's Lemma (see Lemma 2.17 in Section 2.4) to solve the deferential inequality (5.62d). Thus for all the time t, in the interval $[t_k, t_{k+1}]$ the energy is bounded as:

$$E(t) = \frac{1}{2}\|\bar{\mathbf{u}}(t)\|_2^2, \tag{5.63a}$$

$$\leq \|\bar{\mathbf{u}}(t_k)\|_2^2 \exp\left(-2\frac{1 - U(\delta + \gamma)}{\tau}(t - t_k)\right). \tag{5.63b}$$

Hence, for all $t \in [t_k, t_{k+1}]$, we get the following bound on $\bar{\mathbf{x}}$:

$$\|\bar{\mathbf{x}}(t)\|_2 \leq U \|\bar{\mathbf{u}}(t)\|_2 , \tag{5.64a}$$

$$\leq U \|\bar{\mathbf{u}}(t_k)\|_2 \exp\left(-\frac{1 - U(\delta + \gamma)}{\tau}(t - t_k)\right), \tag{5.64b}$$

where (5.64a) is property (5.57) of Lemma 5.9, and (5.64b) result from (5.63b).

Now, using the above results on the output $\bar{\mathbf{x}}$, the state variable $\mathbf{u}(t)$ can be shown to converge exponentially fast. Using the extension of Gronwall's Lemma (see Lemma 2.18 of Section 2.4), the solution of the dynamical system (5.56f) can be expressed as follows, for all

$t \in [t_k, t_{k+1}]$:

$$\bar{\mathbf{u}}(t) = e^{-\frac{(t-t_k)}{\tau}} \bar{\mathbf{u}}(t_k) + e^{-\frac{(t-t_k)}{\tau}} \int_{t_k}^{t} e^{-\frac{(\eta-t_k)}{\tau}} \left[(1-\gamma)\mathbf{I} - \mathbf{\Phi}^T\mathbf{\Phi}\right] \bar{\mathbf{x}}(\eta) d\eta. \tag{5.65}$$

Let $z(t)$ denote the second term in the right-hand side of equation (5.65). The norm of the $z(t)$ can be bounded as

$$\|z(t)\|_2 \leq e^{-\frac{(t-t_k)}{\tau}} \int_{t_k}^{t} e^{\frac{(\eta-t_k)}{\tau}} \left\| \left(\mathbf{\Phi}^T\mathbf{\Phi} - (1-\gamma)\mathbf{I}\right) \bar{\mathbf{x}}(\eta) \right\|_2 d\eta, \tag{5.66a}$$

$$\leq e^{-(t-t_k)/\tau} \int_{t_k}^{t} e^{\frac{(\eta-t_k)}{\tau}} \left\| \mathbf{\Phi}^T\mathbf{\Phi} - (1-\gamma)\mathbf{I} \right\|_2 \|\bar{\mathbf{x}}(\eta)\|_2 d\eta, \tag{5.66b}$$

$$\leq e^{-(t-t_k)/\tau} \int_{t_k}^{t} e^{\frac{(\eta-t_k)}{\tau}} C_1 U \|\bar{\mathbf{u}}(t_k)\|_2 e^{-\frac{1-U(\delta+\gamma)}{\tau}(\eta-t_k)} d\eta, \tag{5.66c}$$

$$\leq e^{-(t-t_k)/\tau} \int_{t_k}^{t} C_1 U \|\bar{\mathbf{u}}(t_k)\|_2 e^{U(\delta+\gamma)(\eta-t_k)/\tau} d\eta, \tag{5.66d}$$

$$= \frac{C_1\tau}{\delta+\gamma} \|\bar{\mathbf{u}}(t_k)\|_2 e^{-(t-t_k)/\tau} \left[e^{U(\delta+\gamma)(t-t_k)/\tau} - 1\right], \tag{5.66e}$$

$$\leq C_2 \|\bar{\mathbf{u}}(t_k)\|_2 e^{-\frac{1-U(\delta+\gamma)}{\tau}(t-t_k)}, \tag{5.66f}$$

where (5.66c) results from (5.64b) and the constant $C_1 = \left\| \mathbf{\Phi}^T\mathbf{\Phi} - (1-\gamma)\mathbf{I} \right\|_2$, and in (5.66f) the constant $C_2 = C_1\tau/(\delta+\gamma)$.

To complete the proof, we use the inequality (5.66f) to find the bound on the norm of $\bar{\mathbf{u}}(t)$. Starting with the norm of equation (5.65), the bound on $\bar{\mathbf{u}}(t)$ is obtained as follows:

$$\|\bar{\mathbf{u}}(t)\|_2 = \left\| e^{-\frac{(t-t_k)}{\tau}} \bar{\mathbf{u}}(t_k) + z(t) \right\|_2, \tag{5.67a}$$

$$\leq \|\bar{\mathbf{u}}(t_k)\|_2 \, e^{-\frac{(t-t_k)}{\tau}} + \|z(t)\|_2, \tag{5.67b}$$

$$\leq \|\bar{\mathbf{u}}(t_k)\|_2 \, e^{-\frac{(t-t_k)}{\tau}} + C_2 \, \|\bar{\mathbf{u}}(t_k)\|_2 \, e^{-\frac{1-U(\delta+\gamma)}{\tau}(t-t_k)}, \tag{5.67c}$$

$$\leq (1+C_2) \|\bar{\mathbf{u}}(t_k)\|_2 e^{-\frac{1-U(\delta+\gamma)}{\tau}(t-t_k)}, \tag{5.67d}$$

$$\leq C_3 \|\bar{\mathbf{u}}(t_k)\|_2 e^{-\frac{1-U(\delta+\gamma)}{\tau}(t-t_k)}. \tag{5.67e}$$

Since $\|\bar{\mathbf{u}}(t)\|_2$ is continuous for all time t, it can be showed using induction on t_k that for all $t \geq 0$

$$\|\bar{\mathbf{u}}(t)\|_2 \leq C_3 \|\bar{\mathbf{u}}(0)\|_2 \exp\left(-\frac{1 - U(\delta + \gamma)}{\tau} t\right) \tag{5.68}$$

Thus, the state variable converges exponentially fast to a unique fixed point $\mathbf{u}^*$ with convergence speed $[1 - U(\delta + \gamma)]/\tau$. □

5.6 Chapter Summary

In this chapter, we proposed a new dynamical system to solve the ℓ_1-minimization problems. The Bregman distance is integrated into the proposed dynamical system, and it can accelerate the convergence rate. The mathematical analysis shows that the proposed dynamical system is well suited to solve the ℓ_1-minimization problems. The equilibrium points of the dynamical system correspond with the critical points of the ℓ_1-minimization problem. Most importantly, the recovered solution's active set coincides with the expected support set. The dynamical system is shown to have an exponential convergence rate. Overall, the results suggest that the dynamical system can be used to solve the optimization problem in real time.

Chapter 6

Computational results

In this chapter, we present computational results to support the theoretical work presented in Chapter 4 and most importantly in Chapter 5. In Section 6.1, we set the scene for the computational work done in this chapter. The computational experiments in Section 6.2 confirm the claim that the equilibrium points of the dynamical system are the optimal points of the sparse recovery problem. In Section 6.3, we numerically demonstrate global convergence of the proposed dynamical system. In Section 6.4, we demonstrate the convergence rate of the proposed Bregman integrated dynamical system.

6.1 Introduction

All the experiments are obtained from simulations of the Bregman integrated dynamical system (5.54). Simulations are performed by discretizing the dynamical system using the forward Euler method. The time derivative is replaced by the approximation

$$\dot{\mathbf{u}}(t) \approx \frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t},$$

where $\mathbf{u}^k \approx \mathbf{u}(t_k)$ and $\mathbf{u}^{k+1} \approx \mathbf{u}(t_k + \Delta t)$. The discrete Bregman integrated dynamical system becomes

$$\begin{cases} \mathbf{u}^{k+1} = \mathbf{u}^k - \frac{\Delta t}{\tau}\left[\mathbf{u}^k + (\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\mathbf{x}^k - \mathbf{\Phi}^T\mathbf{y} + \gamma\left(\mathbf{x}^k - \mathbf{z}\right)\right] \\ \mathbf{x}^{k+1} = S_\mu\left(\mathbf{u}^{k+1}\right). \end{cases} \tag{6.1}$$

From the discrete system we have chosen the Bregman point $\mathbf{z}$ to be the previous time point of the output, that is $\mathbf{z} = \mathbf{x}^{k-1}$. With this choice the Bregman distance $D_h(\mathbf{x}^k, \mathbf{x}^{k-1})$ goes to zero as the dynamical system get close to the equilibrium point. Note that dynamical system requires the Bregman parameter $\gamma \to 0$ and $t \to \infty$. Hence, we update the parameter γ as

$$\gamma^{k+1} = (1 - \beta\Delta t)\gamma^k$$

from some initial starting point γ^0 and a constant parameter $\beta > 0$. This update warrants gradual decrease of γ.

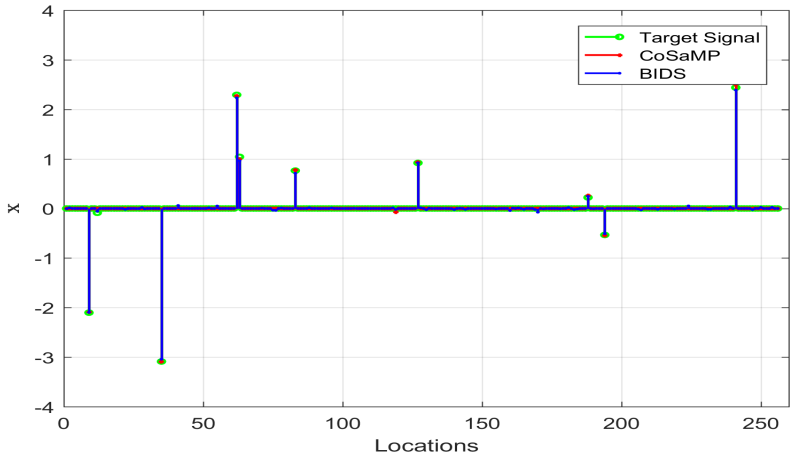
We present several simulations to illustrate the theoretical results presented in the previous chapters. The experimental results are obtained from implementing the discrete dynamical system (6.1) in Matlab using step-size $\Delta t = 0.001$, and a time constant chosen to be equal to $\tau = 0.01$. The dynamical system is given enough time to reach clear convergence. We consider a simple setting to demonstrate the proof of concepts of the proposed Bregman dynamical system. The results are compared with the dynamical system for Locally Competitive Algorithm proposed by Rozell *et.al* [27].

6.2 Equivalence of solution and the equilibrium point

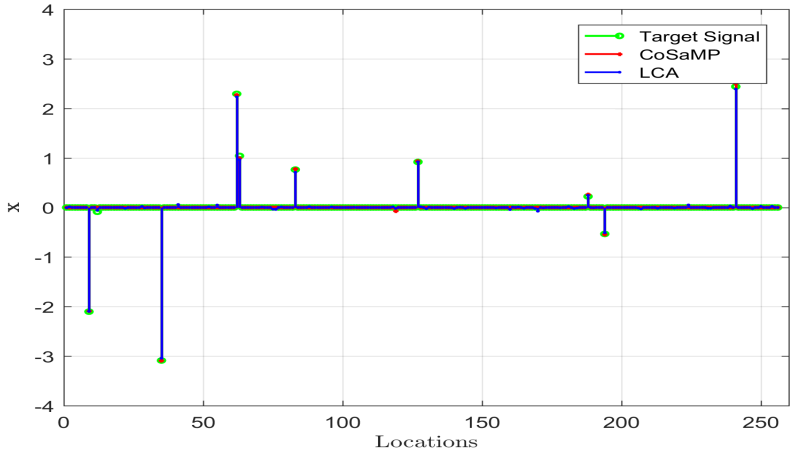
In this section, we illustrate the result of Theorem 5.2, which asserts that the equilibrium point of the Bregman integrated system correspond to the optimal points of ℓ_1-minimization problem. The results are compared with state of the art discrete optimization algorithm for sparse approximation, the Compressive Sampling Matching Pursuit (CoSaMP) algorithm [12]. The numerical experiment has been setup as follows:

Experiment 6.1. *The original sparse signals* $\mathbf{x} \in \mathbb{R}^n$ *with* $n = 256$ *and sparsity* $s = 10$ *are randomly generated. The nonzero components are drawn from a normal distribution,* $\mathcal{N}\left(0, \sigma^2\right)$. *Afterwards, the measurements* $\mathbf{y} \in \mathbb{R}^m$ *with* $m = 80$ *are collected via the projection,* $\mathbf{y} = \mathbf{\Phi x} + \mathbf{e}$. *The measurement matrix* $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ *is also drawn from* $\mathcal{N}\left(0, \sigma^2\right)$. *The columns of* $\mathbf{\Phi}$ *are normalised to have unit norm as required by the dynamical system. The parameter* $\mathbf{e}$ *is the Gaussian random noise with standard deviation* $\sigma = 0.1\|\mathbf{\Phi x}\|_2/\sqrt{m}$, *which is an example of moderate noise. The threshold parameter is set to* $\mu = 0.01$. *The Bregman parameter was initialised with* $\gamma^0 = 95$. *The initial point of the CoSaMP algorithm is set as* $\mathbf{x}(0) = S_\mu(\mathbf{u}(0))$, *where* $\mathbf{u}(0)$ *is the initial point used for the dynamical system.*

The results of Experiment 6.1 are presented in Figure 6.1. The plot shows the target and recovered signals using the Bregman integrated dynamical system and Locally Competitive Algorithm. The results are compared with the recovered signal using the CoSaMP algorithm. Figure 6.1a shows a comparison of the recovered signal using the Bregman integrated dynamical system and the CoSaMP algorithm. Figure 6.1b shows a comparison of signal recovery of the dynamical system of the Locally Competitive Algorithm and CoSaMP algorithm. Furthermore, Table 6.1 presents the statistical summary of the mean squared error between the recovered signal and the target signal. The experiment was repeated for 100 trials using random initialization of $\mathbf{u}(0)$, to generate the mean squared error data.



(a) Comparison of sparse recovery using the proposed Bregman integrated dynamial system (BIDS) and CoSaMP.



(b) Comparison of sparse recovery using Locally Competive Algorithm (LCA) and CoSaMP.

FIGURE 6.1: The output $\mathbf{x}^*$ of the Bregman integrated dynamical system and Locally Competitive Algorithm after convergence. The recovery of the sparse signal using dynamical systems is compared with the optimal solution CoSaMP algorithm.

Mean squared error ($\times 10^{-6}$)			
	LCA	BIDS	CoSaMP
Min	1.3511	1.3511	0.4649
Mean	6.6465	6.1315	1.8834
Std	3.2226	3.0877	1.2751
Max	9.8001	9.8001	5.1002

TABLE 6.1: The statistical summary of the mean squared error between the recovered and the target signal, using Locally Competitive Algorithm, Bregman integrated dynamical system and the CoSaMP algorithm.

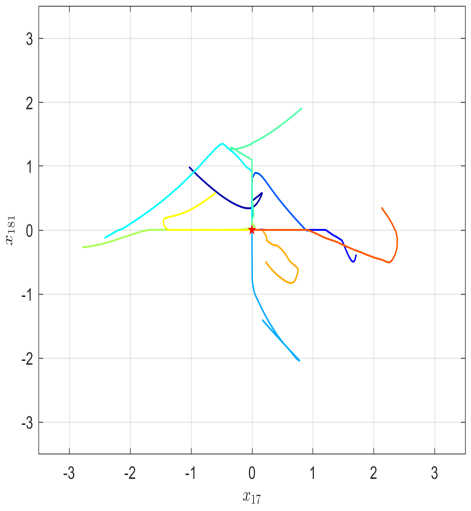
The results in Figure 6.1 and Table 6.1 show that the equilibrium point reached by the Bregman integrated dynamical system is very close to the target signal $\mathbf{x}^*$. The Locally

competitive dynamical system returns similar solutions as the Bregman integrated dynamical system. Notice the slight discrepancy in the recovered signal's amplitude compared to the target signal. We are unlikely to get the exact values due to noise. It is clear from this experiment that the equilibrium point of the Bregman integrated system correspond to the solution of the desired objective function for sparse recovery. Thus, the experimental results confirm the statement of Theorem 5.2.

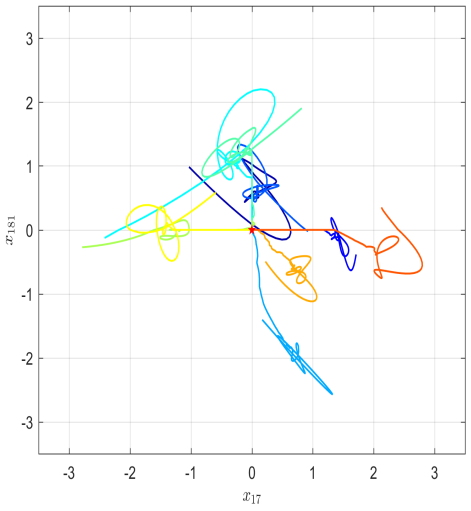
6.3 Global convergence

In this section, we illustrate the consequence of Theorem 5.3, which asserts the trajectories of the Bregman integrated system are globally convergent. The experimental setup is the same as in Experiment 6.1.

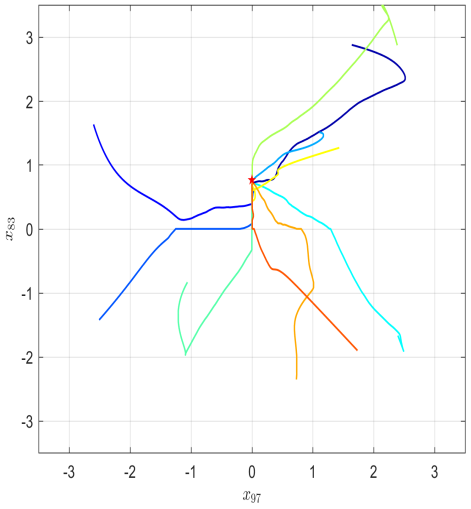
The results of the experiment are presented in Figures 6.2 and 6.3. Figure 6.2 shows a plot of the trajectories of Locally Competitive Algorithm and Bregman integrated dynamical systems for 10 randomly selected different initial points $\mathbf{u}(0)$. Different colours represent different starting points. Convergence is shown for different pairs of active and inactive components of the solution. From any starting point, all the trajectories evolve towards the appropriate components of the optimal solution of the ℓ_1-minimization. A red star in each plot indicates the optimal point. The illustration confirms the global convergence property of the Bregman integrated dynamical system. The results are in agreement with the assertion of Theorem 5.3.



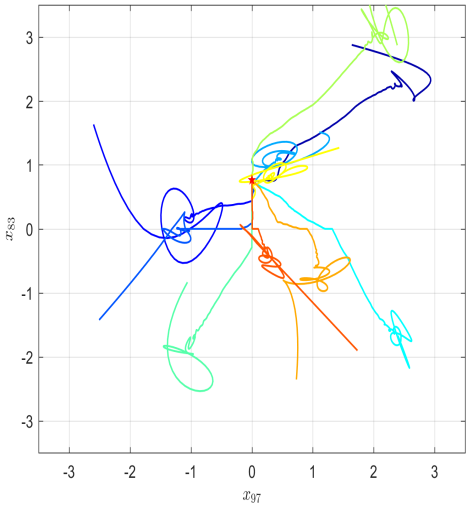
(a) The LCA trajectories for x_{17} vs x_{181}.



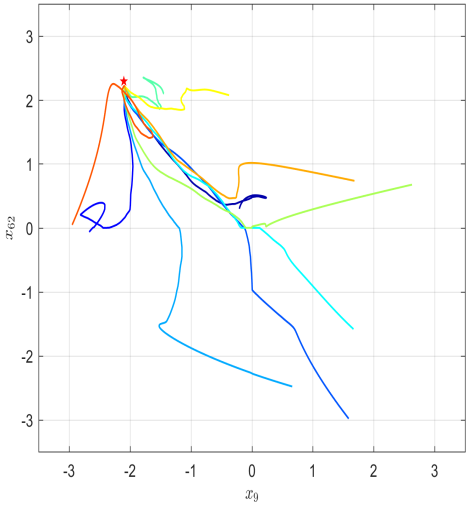
(b) The BIDS trajectories for x_{17} vs x_{181}.



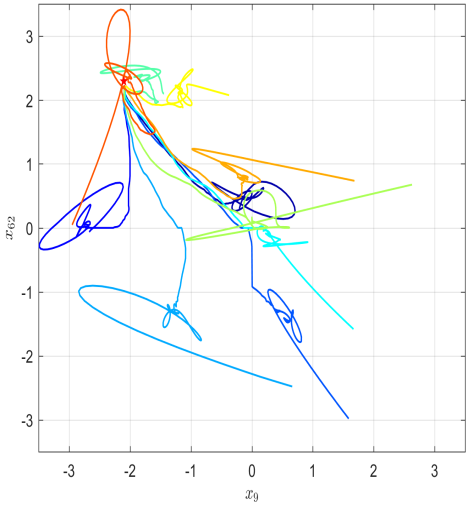
(c) The LCA trajectories for x_{95} vs x_{86}.



(d) The BIDS trajectories for x_{95} vs x_{86}.



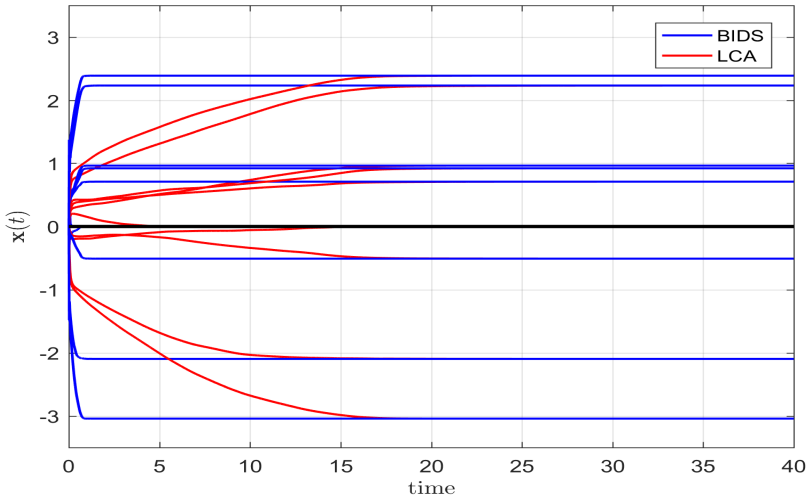
(e) The LCA trajectories for x_9 vs x_{62}.



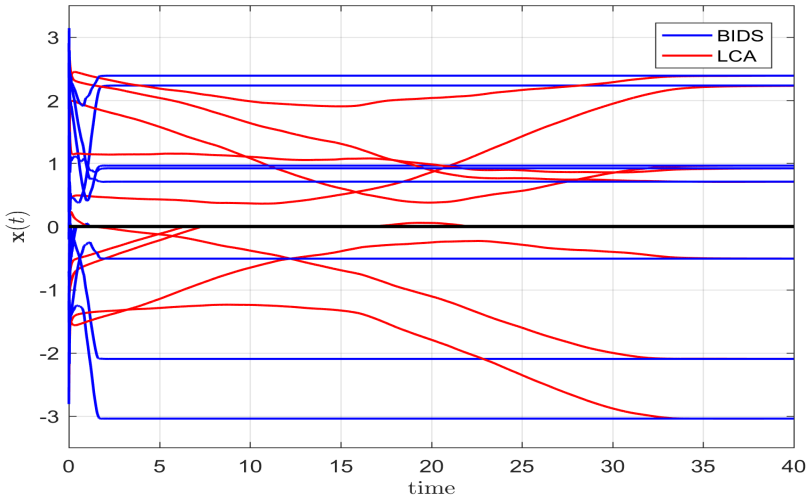
(f) The BIDS trajectories for x_9 vs x_{62}.

FIGURE 6.2: The trajectories of Locally Competitive Algorithm and Bregman integrated dynamical system.

The plots show that the dynamical system of Locally Competitive Algorithm takes a more direct path towards the optimal solution. This is not the case with the proposed Bregman integrated system. Although the trajectories are different for both dynamical systems, they eventually converge to the identical optimal solution as illustrated in Section 6.2. A close look at the behaviour of trajectories shows that the Bregman integrated dynamical system takes an efficient path toward the optimal solution, see Figure 6.3. Figure 6.3 shows a plot of the evolution of several active and inactive components with respect to time for both Locally Competitive Algorithm and Bregman integrated dynamical systems. The values of components of $\mathbf{x}(t)$ are plotted on the y-axis. The initial starting points $\mathbf{x}(0)$ for both systems are identical. The plot clearly shows that the proposed Bregman integrated dynamical system converges faster than the dynamical system of the Locally Competitive Algorithm. More details of the convergence rate are presented in Section 6.4.



(a) Evolution from initial point $\mathbf{u}(0) = \mathbf{0}$



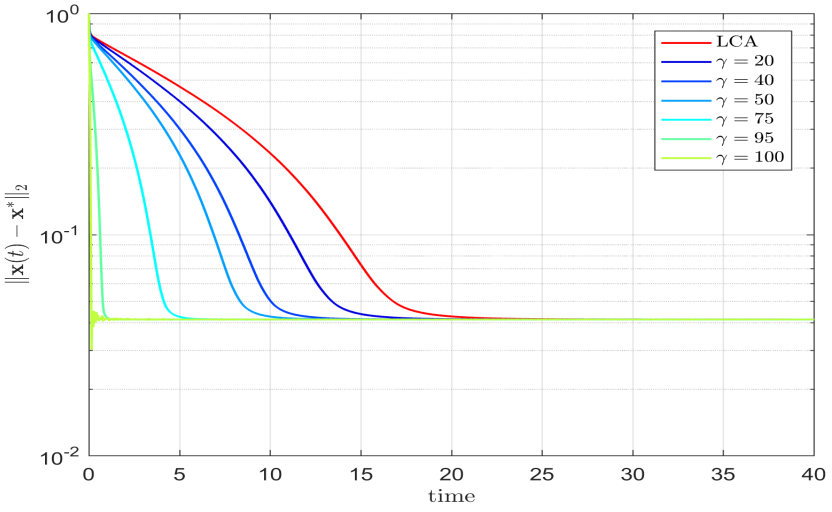
(b) Evolutions from random initial point $\mathbf{u}(0)$.

FIGURE 6.3: The evolution of several randomly selected active and inactive components with respect to time for a Locally Competitive Algorithm and Bregman integrated dynamical system.

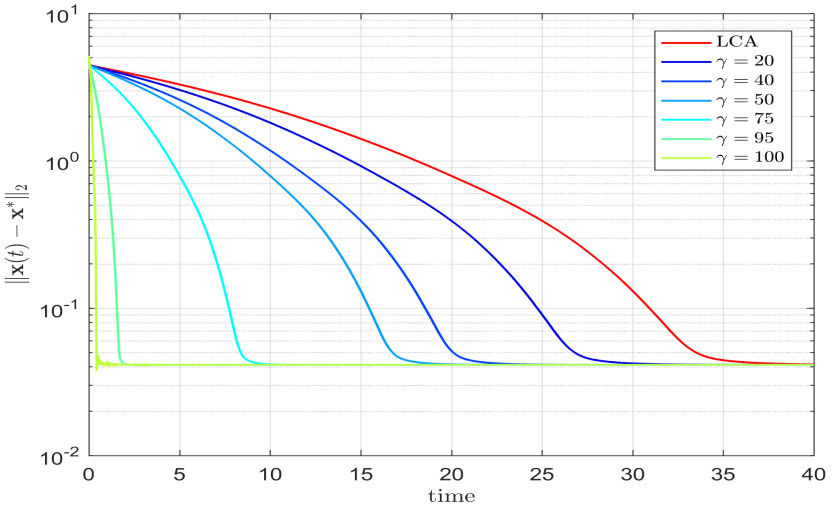
6.4 Convergence rate

In this section, we illustrate the convergence rate of the Bregman integrated system as predicted by Theorem 5.5. We used Experiment 6.1 and focus on solving ℓ_1-minimization problem. In this case, the bound of the subgradient of the soft-thresholding activation is $U = 1$. The evolution of the state error, $\bar{\mathbf{u}}(t) = \mathbf{u}(t) - \mathbf{u}^*$, for both the dynamical system of Locally Competitive Algorithm and Bregman dynamical systems are presented

in Figure 6.4. The results are plotted in a log scale and they exhibit exponential decay. Moreover, it is clear that the proposed Bregman dynamical system converges much faster than the Locally Competitive Algorithm. Moreover, the convergence of the Bregman integrated system depends on the Bregman parameter γ. For $\gamma = 0$, the Bregman integrated dynamical system reduces to the Locally competitive algorithm. The Bregman integrated dynamical system becomes faster as the initial value of γ is increased. The value of γ can only be increased to a certain value, as is evident in Figure 6.4. The dynamical system becomes unstable near the equilibrium for values of $\gamma \gtrsim 100$. The convergence is much faster for systems initialised at rest, that is, $\mathbf{u}(0) = 0$.



(a) Convergence speed with initial solution from rest.



(b) Convergence speed with randomly selected initial point.

FIGURE 6.4: The convergence speed of Bregman integrated dynamical system

Chapter 7

Conclusion

In this chapter, we briefly summarise the work covered in this thesis in Section 7.1 and present ideas to be investigated in future research in Section 7.2.

7.1 Summary

Compressed sensing plays an important role in modern emerging technologies. Specialised discrete algorithms have been proposed to solve recovery in literature. In literature, dynamical systems for solving optimization problems are known to have the potential to alleviate these computational challenges. The focus of this work has been to determine the type of continuous-time dynamical system for solving the sparse recovery problem with applications to compressed sensing. This thesis provides new results that has broadened previous results.

In Chapter 4, we have presented the convergence of trajectories of differential inclusion aimed at solving the sparse recovery problems. The main result is that the trajectories of the differential inclusion are either exponentially convergent or finite-time convergent towards a singleton. These convergence properties are independent of the nature of the set of equilibrium points. Hence, they hold even when the differential inclusion has infinitely many nonisolated equilibrium points. The proof exploits the recently proposed convergence method, based on Łojasiewicz inequality for nonsmooth functions. The results are of interest for efficient real-time computation of solutions to sparse recovery problems.

In Chapter 5, we have presented a new dynamical system to solve the ℓ_1-minimization problems. The Bregman distance is integrated into the proposed dynamical system, accelerating the convergence rate. The mathematical analysis shows that the proposed dynamical system is well suited to solve the ℓ_1-minimization problems. The equilibrium points of the dynamical system correspond with the critical points of the ℓ_1-minimization problem. Most importantly, the active set of the recovered solution coincides with the optimal support set. The dynamical system is shown to have an exponential convergence rate. Overall, the results suggest that the dynamical system can be used to solve sparse recovery problems in real-time.

Computational results have been presented to show the correctness of the developed theory and the good computational performance of the proposed dynamical system. Several comparative experiments on sparse recovery problems demonstrate that the proposed dynamical system approach is efficient and effective.

7.2 Future work

The future work includes the study of the following concepts:

- A complete study of discretization of the Bregman integrated dynamical system is important. It is essential to have a consistent discretization that preserves the properties of the dynamical system. This might result in a new sparse representation algorithm with a faster convergent property.
- Recently, there have been some work on finite-time and fixed time convergence of dynamical system [64]. These are stronger convergence results than the exponential convergence rate. Adopting the finite-time approach to study the proposed dynamical system would be beneficial.
- The value of the regularization parameter μ plays an important role in the speed and accuracy of the dynamical system. The work suggests decreasing the value of μ as the dynamical system process improves the result. This is analogous to continuation in the discrete optimization algorithm. Future work will entail mathematical analysis of the proposed dynamical system coupled with the dynamics describing the behaviour of μ.

Bibliography

[1] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Infornation Theory*, 52(2):489–509, February 2006.

[2] D. L. Donoho. Compressed sensing. *IEEE Transactions on Infornation Theory*, 52(4): 1289–1306, April 2006.

[3] H. Nyquist. Certain topics in telegraph. *Transaction of American Institure of Electrical Engineers*, 2(47):617–644, 1928.

[4] C. E Shannon. Communication in presence of noise. *Proceedings of Institute of Radio Engineers*, 1(37):10–21, 1949.

[5] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[7] E. Candès and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[8] I. Daubechies, O. M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[9] S. Becker and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:182–183, 2009.

[10] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[11] W. Dai and O. Milenkvic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

[12] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26:301–321, 2009.

[13] W. Su, S. Boyd, and E. J. Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *The Journal of Machine Learning Research*, 17(1):5312–5354, 2016.

[14] Y. E. Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Soviet Mathamatics Doklady*, 2(27):372–376, 1983.

[15] Y. E. Nesterov. Introductory lectures on convex optimization: A basic course. *Kluwer Academic Publishers*, 0(1):1–12, 2004.

[16] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79:2554 – 2558, 1982.

[17] J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-stage neurons. *Proceedings of the National Academy of Sciences*, 81: 3088 – 3092, 1984.

[18] J. J. Hopfield and D. W. Tank. Neural computation of decision in optimization problems. *Biological Cybernetics*, 52(3):141 – 152, 1985.

[19] D. Takhar, J. N. Laska, M. B. Wakin, M. F. Duarte, D. Baron, S. Sarvatham, K. F. Kelly, and R. G. Baraniuk. A new compressive imaging camera architecture using optical-domain compression. *Computational Imaging*, 60(65):10–21, 2006.

[20] M. B. Wakin, J. N. Laska, M. F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. F. Kelly, and R. G. Baraniuk. An architecture for compressive imaging. *IEEE International Conference on Image Processing*, 37(1):1273–1276, 2006.

[21] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, April 2008.

[22] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonace in Medicine*, 58(6):1182 – 1195, 2007.

[23] U. Gamper, P. Boesiger, and S. Kozerke. Compressed sensing in dynamic MRI. *Magnetic Resonance in Medicine*, 59(2):365–373, 2008.

[24] M. Lustig, D. Donoho, and M. Santos J. M. Pauly. Compressed sensing MRI. *IEEE Signal Processing Magazine*, 25(2):72 – 82, 2008.

[25] M. Mohtashemi, H. Smith, F. Sutton, D. K. Walburger, and J. Diggans. Sprase sensing DNA microarray-based biosensor: Is it feasible? *IEEE sensors Applications Symposium*, pages 127–130, 2010.

[26] W. Dai, M. A. Sheikh, O. Milenkvic, and R. G. Baraniuk. Compressive sensing dna microarrays. *EURASIP Journal on Bioinformatics and Systems Biology*, 1:2230–2249, 2009.

[27] C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.*, 20(10):2526–2563, 2008.

[28] F. H. Clarke. *Optimization and Nosmooth Analysis*. Society for Industrial Mathematics, New York, 1987.

[29] K. Zajicek. An elementary proof of the one-dimensional radamacher theorem. *Mathematica Bohemica*, 117:133–136, 1994.

[30] A. M. Lyapunov. *The general problem of the Stability of Motion*. Taylar and Francis, New York, 1892.

[31] T. Gronwal. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919.

[32] S. Łojasiewicz. Une propriete topologique des sous-ensembles analytiques reels. *Colloques internationaux du C.N.R.S. Lesequations aux derivees partielles*, 117:87 – 89, 1963.

[33] J. Bolte, N. Daniildis, and N. Liewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal of Optimization*, 17:1205–1223, 2007.

[34] D. L. Donoho and M. Elad. Optimally sparse representation in general dictionaries via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 100(10):2197–2202, 2003.

[35] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the African Mathematical Society*, 22(1):211–231, 2009.

[36] A. M. Tillmann and M. E. Pfetsch. The computational complexity of the restricted isometry property, the null space property and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 2(60):1248–1259, 2014.

[37] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[38] E. Candès. The restricted isometry property and its implications for compressed sensing. *IEEE Transactions on Information Theory*, (12):589–592, 2008.

[39] S. Foucat and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhauser, Basel, 2013.

[40] E. C. Marques, N. Maciel, L. Naviner, H. Cai, and J. Yang. A review of sparse recovery algorithms. *IEEE Access*, 7:2169–3536, 2018.

[41] D. L. Donoho, M. Elad, and V. N. Temlyakow. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52 (1):6–18, April 2006.

[42] Q. Mo. A sharp restricted isometry constant bound orthogonal matching pursuit. *arXiv:1501.01708*, 2015.

[43] S. Becker, J. Bobin, and E. Candès. A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[44] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2008.

[45] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

[46] R. Courant. Variational methods for the solution problems of equilibrium and vibrations. *Bulletin of the American Mathematical Society*, 49(1):1–23, 1943.

[47] J. Schropp. A dynamical systems approach to constrained minimizaton. *Journal of Numerical Functional Analysis and Optimization*, (21):537–551, 2007.

[48] U. Helmke and J. B. Moore. Optimization and dynamical systems. *Proceedings of the IEEE*, 86(6):907, 1996.

[49] D. W. Tank and J. J. Hopfield. Simple neural optimization networks: An A/D converter, signal decision circuit, and a linear programming circuit. *IEEE Transactions on Circuits and Systems*, 33(5):533–541, 1986.

[50] A. N. Michel and D. L. Gray. Analysis and systhesis of neural networks with lower block triangular interconnecting structure. *IEEE Transactions on Circuits and Systems*, 37 (3):1267–1283, 1990.

[51] M. Fort, A. Liberatore, S. Maneti, and M. Marini. Global asymptotic stability for a class of nonsymmetric neural networks. *IEEE International Symposium on Circuits and Systems*, pages 2580–2583, 1995.

[52] H. Yang and T. S. Dillon. Exponential stability and oscillation of hopfield graded response neural network. *IEEE Transactions on Neural Networks*, 5(5):719–729, 1994.

[53] A. F. Filippov. *Differential equations with discontinous righthand sides: Control systems.* Springer, New York, 1988.

[54] A. Balavoine, J. Romberg, and C. J. Rozell. Convergence and rate analysis of neural networks for sparse approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(9):1377–1389, 2012.

[55] A. S. Charles, P. Garrigues, and C. J. Rozell. A common network architecture efficiently implements a variety of sparsity-based inference problems. *Neural Computation*, 24: 3317–3339, 2012.

[56] A. Balavoine, C. J. Rozell, and J. Romberg. Convergence of a neural network for sparse approximation using the nonsmooth łojasiewicz inequality. *IEEE Iternational Joint Conference in Neural Networks*, (9):1–8, 2013.

[57] J. P. Aubin. *Viability Theory*. Birkhauser Boston, MA, New York, 2009.

[58] M. Fort, P. Nistri, and M. Quincampoix. Convergence of neural network for programming problems via a nonsmooth Łojasiewicz inequality. *IEEE Transactions on Neutal Networks*, 17(6):1471–1486, 2006.

[59] W. Lu and J. Wang. Convergence analysis of a class of nonsmooth gradient systems. *IEEE Transactions on Circuits and Systems*, 55(11):3514 – 3527, 2008.

[60] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[61] A. Auslender, M. Teboulle, and S. Ben-Tiba. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal of Optimization*, 16:697–725, 2006.

[62] J. Eckstein. Nonlinear proximal point algorithms using bregam functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

[63] J. J. Fuchs. On sparse representation in arbitrary redundant bases. *IEEE Transactions on Information Therory*, 50:1341–1344, 2004.

[64] A. Polyakov, D. Efimov, and B. Brogliato. Consistent discretization of finite-time and fixed-time stable systems. *SIAM Journal on Control and Optimization*, 57(1):78–103, 2019.