MATHEMATICAL ANALYSIS OF A DYNAMICAL SYSTEM FOR SPARSE RECOVERY

A Dissertation
Presented to
The Academic Faculty

By

Aurèle Balavoine

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
May 2014

MATHEMATICAL ANALYSIS OF A DYNAMICAL SYSTEM FOR SPARSE RECOVERY

Approved by:

Dr. Jennifer 0. Hasler, Committee Chair
Associate Professor, School of ECE
Georgia Institute of Technology

Dr. Justin K. Romberg, Advisor
Associate Professor, School of ECE
Georgia Institute of Technology

Dr. Christopher J. Rozell, Co-Advisor
Assistant Professor, School of ECE
Georgia Institute of Technology

Dr. Anthony J. Yezzi
Chair Professor, School of ECE
Georgia Institute of Technology

Dr. Maria-Florina Balcan
Assistant Professor, School of CS
Georgia Institute of Technology

Dr. Mark A. Davenport
Assistant Professor, School of ECE
Georgia Institute of Technology

Date Approved: May 2014

La Curiosité nous tourmente et nous roule, Comme un Ange cruel qui fouette des soleils. Singulière fortune où le but se déplace, Et, n'étant nulle part, peut être n'importe où ! Où l'homme, dont jamais l'espérance n'est lasse, Pour trouver le repos court toujours comme un fou !

– Charles Baudelaire, Les Fleurs du Mal, 1861

Curiosity tortures and turns us Like a cruel angel whipping the sun.
Whimsical fortune, whose end is out of place, And, being nowhere, can be anywhere!
Where Man, in whom Hope is never weary, Runs ever like a madman searching for repose.

– Geoffrey Wagner, Selected Poems of Charles Baudelaire (NY: Grove Press, 1974)

A mes parents, Sophie et Xavier,

avec tout mon amour toujours et ma reconnaissance.

ACKNOWLEDGMENTS

I am extremely grateful to my two advisors, Dr. Justin Romberg and Dr. Christopher Rozell, and I would like to thank them deeply for their guidance and support throughout my Ph.D. They gave me the freedom to explore new ideas, provided me with valuable feedback and a pleasant work environment. They always knew the right words to inspire me. I learned a lot from them and truly enjoyed working with them. I would also like to thank my thesis committee members, Dr. Hasler and Dr. Yezzi, for their useful feedback that gave me new inspiring ideas for the last chapters of my thesis. I especially thank Dr. Hasler, who was my Master thesis advisor, has expanded my knowledge and has sparked my interest for optimization in analog computing in the first place. Finally, I thank Dr. Maria-Florina Balcan and Dr. Mark Davenport for their valuable input and for serving on my committee.

I wish to thank Adam Charles, Salman Asif, Ali Ahmed and Han Lun Yap who have been reliable sources of help and inspiration when I was struggling, and who have also been amazing friends. I also thank Steve Conover and Abigail Kressner for always having a kind word that carried me a long way. I thank all of my lab-mates, Aditya Joshi, William Mantzel, Ning Tian, Darryl Sale, Mengchen Zhu, Nishant Zachariah, Alireza Aghasi, Sohail Bahmani, Michael Moore, Andrew Massimino, and Nicolas Bertrand for the times spent in lab discussing research and many other topics. I am grateful for the time they spent listening to my work and giving me useful feedback after my presentations. Their help and support has been more important than they can imagine. I am also thankful to some dear friends that made my stay in Atlanta so unforgettable: Nicolas and Laura Dudebout, Taylor and Sam Shapero, Diane Isaacson, Peter and Alex Tuuk, Peter Siy, Sarah Touse, Laura Hansen, Alice Barrett De Sep, Kyle and Stephanie Krueger, Sean Kelly, Jeff Bingham and Jiun-Hong Lai. I also thank Marion Glomot and Marion Garcia for supporting me from overseas.

Je tiens également à remercier de tout mon coeur mes parents, Sophie et Xavier Balavoine, et ma petite soeur Eline. Je dédis cette thèse à mes parents, 27 ans après qu'ils m'ont dédié les leurs. C'est grace à leur soutien sans faille que j'ai eu la force de poursuivre cette aventure, et grace à leurs encouragements que je l'ai continuée à terme. Je sais les sacrifices qu'ils ont acceptés, matériels et immatériels, pour que je puisse faire mes propres choix et je leur suis éternellement reconnaissante. J'espère pouvoir un jour rendre à ma soeur tout l'amour et le soutien qu'elle m'a montré et à mon tour être pilier pour qu'elle avance dans ses propres aventures. I am also grateful for my extended French and American families for being unconditionally loving and supportive.

Last but not least, I thank Christopher Turnes, the love of my life, who has been by my side through all of it. I cannot express how grateful I am for all his support and affection. He has picked me up when my research was not going well, he has cheered with me for every publication, and he has always believed in me. He has made those years invaluable, and I wish to end this chapter of our lives only to open a new one just has thrilling and eventful.

TABLE OF CONTENTS

ACKNOWLEDGMENTS v

LIST OF FIGURES x

LIST OF SYMBOLS OR ABBREVIATIONS xii

SUMMARY xiv

I INTRODUCTION 1
- 1.1 Compressed sensing 1
  - 1.1.1 Sparse representation 1
  - 1.1.2 Restricted isometry property 2
  - 1.1.3 Subgaussian random matrices 3
  - 1.1.4 Sparse signal recovery 3
  - 1.1.5 Recovery guarantees 5
- 1.2 Locally Competitive Algorithm 5
  - 1.2.1 The LCA differential equation 5
  - 1.2.2 A Hopfield-type neural network 6
  - 1.2.3 Limited analysis 7

II BACKGROUND 10
- 2.1 Sparse signal recovery 10
  - 2.1.1 Relaxation methods 11
  - 2.1.2 Greedy algorithms 11
- 2.2 Neural network analysis 12
  - 2.2.1 Stability and convergence 12
  - 2.2.2 Lyapunov's direct method 14
  - 2.2.3 Previous work 15
- 2.3 Nonsmooth neural networks 16
  - 2.3.1 Subgradient 17
  - 2.3.2 Generalized chain rule and calculus 18
  - 2.3.3 Previous work 19
- 2.4 Finite length of trajectories 20
  - 2.4.1 Subanalicity and Łojasiewicz inequality 20
  - 2.4.2 Previous work 21

III PROPERTIES FOR GENERIC SPARSE RECOVERY 23
- 3.1 Nonsmooth objective 24
- 3.2 Fixed points 24
  - 3.2.1 Simulations 26
  - 3.2.2 Proof of Theorem 1 28
- 3.3 Global asymptotic convergence 31

3.3.1 Conditions on the activation function 31
3.3.2 Notation 33
3.3.3 Convergence result 35
3.3.4 Simulations 36
3.3.5 Proof of Theorem 2 37
3.4 Recovery of the support in finite time 39
3.4.1 Simulations 40
3.4.2 Proof of Theorem 3 41
3.5 Finite length of trajectories 42
3.5.1 Simulations 44
3.5.2 Proof of Theorems 4 45
3.5.3 Proof of Theorem 5 50
3.6 Exponential rate of convergence 52
3.6.1 Simulations 54
3.6.2 Proof of Theorem 6 54

IV PROPERTIES FOR CS RECOVERY 58
4.1 Bounding the LCA active set 59
4.1.1 Bounding the active set by the optimal support 59
4.1.2 Bounding the size of the active set by a constant 59
4.1.3 Remarks and consequences on the RIP constant 60
4.2 Application to Compressed Sensing matrices 62
4.2.1 Theorem 7 with CS matrices 62
4.2.2 Theorem 8 with CS matrices 64
4.3 Decreasing threshold 66
4.4 Estimate of the convergence speed 66
4.5 Simulations 67
4.5.1 Effect of the threshold on the size of the active set 68
4.5.2 Decreasing the threshold during convergence 68
4.5.3 Estimate of the convergence speed 72
4.6 Proofs 73
4.6.1 Proof of Theorem 7 73
4.6.2 Proof of Theorem 8 75
4.7 Summary 77

V TRACKING OF TIME-VARYING SIGNALS 80
5.1 Background and related work 80
5.1.1 The ISTA 80
5.1.2 Related work 82
5.2 Tracking a time-varying input 83
5.2.1 Signal model 83
5.2.2 Tracking abilities of ISTA 84
5.2.3 Tracking abilities of the LCA 87
5.3 Simulations 89
5.3.1 Synthetic data 89

5.3.2 Real data 90
5.4 Proofs 92
5.4.1 Lemma 2 92
5.4.2 Proof of Theorem 9 94
5.4.3 Proof of Theorem 10 96
5.5 Summary 99

VI SUMMARY AND FUTURE DIRECTIONS 102
6.1 Summary 102
6.1.1 General performance guarantees 102
6.1.2 Application to CS 103
6.1.3 Tracking of time-varying signals 104
6.2 Comparative overview of results 104
6.3 Future directions 107
6.3.1 Discontinuous activation functions 107
6.3.2 Matrix uncertainty 107

A PROPERTIES OF THE LCA 108
A.1 Properties of the cost and activation functions 108
A.2 Time derivative of the objective function 114
A.3 Boundedness of the objective, the states and outputs 116
A.4 Subanalicity of the objective 119

B DIFFERENTIAL EQUATIONS 122
B.1 Gronwall's Lemma 122
B.2 LCA trajectories for ℓ_1-minimization 124
B.3 LCA inequalities 125

C MATRIX PROPERTIES 129

D DISCONTINUOUS ACTIVATION FUNCTION 131
D.1 Cost function 131
D.2 Solutions to ODEs with a discontinuous right-hand side 136
D.3 Convergence result 138

REFERENCES 139

LIST OF FIGURES

Figure 1 Thesis contributions xvi

Figure 2 Architecture of the LCA 7

Figure 3 Soft-thresholding function 8

Figure 4 Lyapunov stability 13

Figure 5 LCA solution for sparse recovery 28
(a) Using the soft-thresholding function 28
(b) Using the hard-thresholding function 28

Figure 6 LCA solution for Tikhonov regularization 29

Figure 7 Generic activation function 34

Figure 8 Time evolution of trajectories 36

Figure 9 Global convergence of the LCA 37

Figure 10 Average number of switches during convergence 41

Figure 11 Finite length of the LCA trajectories 46

Figure 12 Convergence rate of the LCA 55

Figure 13 Optimal active nodes during convergence 69

Figure 14 Active set size during convergence 69

Figure 15 Decreasing the threshold during convergence 70

Figure 16 Estimation of the convergence rate 71
(a) Effect of the signal length N. 71
(b) Effect of the sparsity level S 71
(c) Effect of the number of measurements M. 71
(d) Effect of the threshold λ. 71

Figure 17 Visualization of the ISTA iteration with a dynamic input 85

Figure 18 Convergence of ISTA with a dynamic input 89
(a) Effect of the parameter P. 89
(b) Effect of the parameter μ. 89

Figure 19 Active set size with a dynamic input 91

Figure 20 Recovering streaming videos with ISTA 93

(a) Average rMSE 93
(b) Average number of products by Φ and Φ^T 93

Figure 21 Construction of the cost function 110
(a) Example of activation function 110
(b) Associated inverse function 110
(c) Associated cost function 110

Figure 22 Construction of the cost function for a discontinuous activation 134
(a) Example of activation function 134
(b) Associated inverse function 134
(c) Associated cost function 134

LIST OF SYMBOLS OR ABBREVIATIONS

a.a.	almost all.
a.e.	almost everywhere.
CoSaMP	Compressive Sampling Matching Pursuit.
CS	Compressed Sensing.
HNN	Hopfield Neural Network.
ISTA	Iterative Soft-Thresholding Algorithm.
LCA	Locally Competitive Algorithm.
ODE	Ordinary Differential Equation.
OMP	Orthogonal Matching Pursuit.
RIP	Restricted Isometry Property.
ROMP	Regularized Orthogonal Matching Pursuit.
δ	RIP constant of the matrix Φ.
$\mathcal{B}_r(x)$	ball of radius r around the point x, *i.e.*, $\overline{x} \in \mathcal{B}_r(x) \;\Leftrightarrow\; \lVert x - \overline{x} \rVert < r$.
$\Pi_{\mathcal{S}}$	projection operator onto the indices in the set $\mathcal{S}$.
s	sign vector of the output vector a.
$u^*,\ a^*$	fixed points of the LCA, respectively for the state and output.
$a^\dagger,\ a^\dagger(t)$	target vector, generating the measurements.
$\Gamma_\dagger,\ \Gamma_\dagger(t)$	optimal support, *i.e.*, set of indices for which the target vector is non-zero.
$\dot{F}(\cdot)$	derivative of the function $F(\cdot)$ with respect to time: $\dot{F}(\cdot) = \dfrac{F(\cdot)}{dt}$.
$\nabla F(\cdot)$	classic gradient of the function $F(\cdot)$.
$\partial F(\cdot)$	subgradient of the function $F(\cdot)$.
Ω_F	set of points where the function $F(\cdot)$ fails to be differentiable.
$\Gamma,\ \Gamma(t)$	active set, *i.e.*, set of indices for which the output is non-zero.
$\Delta_q,\ \Delta,\ \Delta(t)$	set indices corresponding to the q nodes with largest magnitude in the state vector $u(t)$.

$\mathrm{Z}, \mathrm{Z}(t)$ constant set, *i.e.*, set of indices for which the output is in a flat region of the the activation function.

$|\Gamma|$ number of elements in the set Γ.

$\mathcal{Z}$ collection of the intervals in $\mathbb{R}$ for which the activation function is flat.

$\mathcal{U}$ collection of the intervals in $\mathbb{R}$ for which the activation function is strictly increasing.

$\{t_k\}_{k \in \mathbb{R}}$ sequence of switching times between which the active set is constant, *i.e.*, $\Gamma(t) = \Gamma_k, \ \forall t \in [t_k, t_{k+1})$.

SUMMARY

In a large number of scientific and engineering applications, the cost of deploying high-tech sensors and the time required to acquire high-resolution signals have become impractical. To address this issue, the theory of Compressed Sensing (CS) was developed as a new acquisition scheme that can outperform traditional Nyquist rate systems. This approach yields significant savings during acquisition by pushing the computational burden to the processing stage, as the recovery of a signal from its CS measurements can be a computationally expensive problem. One classic approach to this problem, known as sparse recovery problem, consists in solving a complex optimization program. Despite the numerous digital solvers proposed to perform this task, none are currently efficient enough to achieve real-time recovery of very large signals.

Meanwhile, optimization has become a major tool to recast and solve many problems in addition to sparse recovery and across many scientific domains. Developing efficient discrete-time algorithms to solve general classes of optimization programs has driven many research efforts in the digital signal processing community. Despite the many advances in digital technology, the speed and power efficiency of digital computers reaches a bottleneck when the size of the data becomes extremely large. On the contrary, advances in analog technology, such as very-large-scale integrated circuits, have the potential to outperform digital computing, yielding gains in both speed and power efficiency for certain problems of very large size. For this reason, there is a renewed interest in using dynamical systems to solve complex optimization programs.

To answer the need for a fast solver for the sparse recovery problem, a continuous-time dynamical system, called the Locally Competitive Algorithm (LCA), has been proposed. Its evolution is ruled by a set of ordinary differential equations (ODEs) with a highly parallel structure. Implementing this system on a dedicated analog chip has the potential to yield a faster and more power-efficient solver. However, before investing significant time

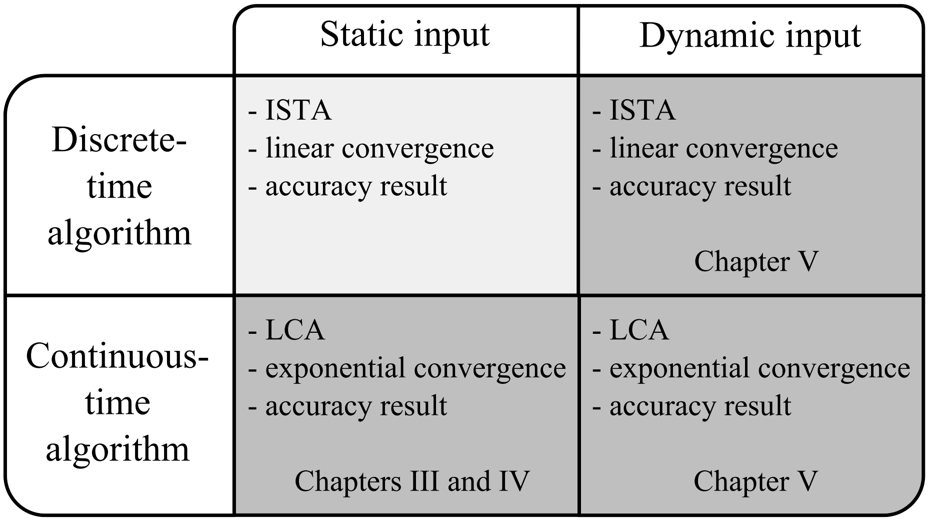
and money to develop and manufacture this circuit, it is important to assess its performance guarantees. The goal of this thesis is to provide a mathematical analysis of the solution provided by the LCA as it is evolving with time. The contributions of this thesis are threefold.

- First, theoretical tools for the analysis of nonlinear neural networks for optimization are developed in a general setting. In particular, new results are presented for the convergence study of a class of networks that extend the current state of research in the field. In Chapter 2, the background material necessary to develop the analytic tools is presented, along with a summary of the previous results in the literature. In Chapter 3, the theoretical findings obtained for an extended class of neural networks are gathered. These findings include a proof of convergence when the fixed points of the system are isolated, a proof of convergence in the case where the fixed points are not isolated, and an analytic expression for the convergence speed.

- Second, in Chapter 4, the previous results are specialized to the case where the network solves the ℓ_1-minimization program to recover a sparse signal. The ℓ_1-minimization program is the most famous optimization program for sparse recovery in CS and comes with strong performance guarantees. The analysis in Chapter 4 shows that the LCA takes an efficient path toward the solution of this program and yields an estimate for the convergence speed that depends only on the problem parameters. Several interesting parallels to properties of digital sparse recovery solvers are brought to light in this study.

- Finally, the convergence properties of the LCA and of the Iterative Soft-Thresholding Algorithm (ISTA) – its discrete-time counterpart – are analyzed in the case where the underlying sparse signal is time-varying and the measurements are streaming. Such a study is of great interest for practical applications that must operate in real-time, such as tracking problems or closed-loop control systems. While convergence guarantees exist for most sparse recovery solvers in the static case, the dynamic case surprisingly lacks theoretical analysis. Of particular interest is the scenario where the number of iterations is constrained by the sampling rate. This situation arises in practical applications, where data are streaming at high rates or the computational resources are limited. The results of this study are presented in Chapter 5, and demonstrate that the LCA and ISTA can efficiently track a time-varying signal from streaming measurements and achieve an error that is essentially optimal.

The contributions of this thesis are organized in Figure 1. Prior to this work, the existing analysis was concentrated on discrete-time algorithms for the recovery of static signals. For instance, the ISTA has been shown to converge with linear rate to the solution of the ℓ_1-minimization recovery problem, which comes with accuracy guarantees. This thesis has provided convergence and accuracy results for the continuous-time LCA for the recovery of static signals, and for the discrete-time ISTA and the continuous-time LCA for the recovery of time-varying (dynamic) signals.



	Static input	Dynamic input
Discrete-time algorithm	- ISTA - linear convergence - accuracy result	- ISTA - linear convergence - accuracy result Chapter V
Continuous-time algorithm	- LCA - exponential convergence - accuracy result Chapters III and IV	- LCA - exponential convergence - accuracy result Chapter V

Figure 1: Visualization of the thesis contributions. The cells in darker grey represent the areas where this thesis has made significant contributions. The cell in lighter grey represents the prior state of knowledge. The cells contain a summary of the results and the chapters where they appear.

CHAPTER I

INTRODUCTION

Optimization plays a key role in many modern signal-processing applications, including image denoising, recovery, and inpainting, data clustering, and more. In the emerging field of CS, a complex optimization program composed of a data fidelity term and a nonlinear sparsity-enforcing term can be used to recover sparse signals from few linear measurements. However, when signals are high-dimensional or streaming at high rates, digital solvers tend to be too slow and computationally intensive to perform real-time recovery. On the contrary, analog networks have a long history as optimization solvers and have been shown to yield significant speed and power improvement over their digital counterparts [1]. The focus of this thesis is to understand what type of continuous-time architectures can be used to solve optimization problems and to analyze their performance mathematically.

1.1 Compressed sensing

Researchers in all fields, ranging from such disparate fields as medical imaging to cosmology, are currently faced with increasing amounts of data. To deal with this problem and reduce this data to a more manageable size, CS theory has proposed a new method for acquiring signals [2, 3]. In place of sampling a high-resolution signal and compressing it as a post-processing step, only a small number of linear measurements are acquired in the CS approach. Thanks to this technique, the number of sensors and acquisition time may be greatly reduced, thus limiting the cost at the front-end of the data stream in a wide range of applications.

1.1.1 Sparse representation

Underlying CS theory is the fact that most signals can be represented by a sparse vector in an appropriate dictionary. A vector $a^\dagger$ in $\mathbb{R}^N$ is called S*-sparse* if it contains only S

non-zero coefficients. Throughout this thesis, the vector $a^\dagger$ is unknown and referred to as *target signal*. The term *optimal support* refers to the support of $a^\dagger$, *i.e.*, the set of indices that correspond to the non-zero entries in $a^\dagger$, and is denoted by $\Gamma_\dagger$. If the location of the non-zero elements are known, the signal $a^\dagger$ can be acquired, represented, transmitted, and stored efficiently. The main advantage of CS is to provide an acquisition scheme that only requires the number of measurements to be on the order of the underlying sparsity S rather than the ambient dimension N, even when the location of the non-zero entries is not known in advance. CS measurements are non-adaptive and take the form

$$y = \Phi a^\dagger \ + \epsilon, \tag{1}$$

where the matrix Φ, called *sensing* or *measurement matrix*, has dimension $M \times N$, where typically $M \ll N$, and ϵ is a noise vector in $\mathbb{R}^N$.

1.1.2 Restricted isometry property

The choice of the measurement matrix is critical to the recovery of the target signal $a^\dagger$ from CS measurements. Intuitively, the vector y obtained via (1) must retain the information contained in $a^\dagger$. For this condition to hold, one possible requirement is for Φ to satisfy the Restricted Isometry Property (RIP) developed in [4].

Definition 1 (Restricted Isometry Property)**.** *The matrix Φ satisfies the RIP of order K if there exists a constant $\delta \in (0, 1)$, such that for any K-sparse vector $x \in \mathbb{R}^N$, the following holds:*

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

If this is the case, the matrix Φ is also said to satisfy the RIP with parameters (K, δ). The RIP-constant δ_K of order K is defined as the smallest positive constant δ satisfying (2).

When δ_K is close to 1, the matrix Φ acts as a near isometry on all K-sparse vectors. When $K = 2S$, the RIP ensures that two distinct S-sparse vectors will remain distinguishable after they have been projected onto the range of Φ. In addition to being used to establish recovery results, the RIP yields several bounds on the eigenvalues of certain submatrices of $\Phi^T\Phi$ that are presented in Appendix C, and are useful to the analysis of both the LCA and its digital counterparts.

1.1.3 Subgaussian random matrices

Some classes of matrices are known to satisfy the RIP with high probability. In particular, Theorem 5.65 in [5] states that if Φ is an $M \times N$ random matrix whose columns Φ_n are independent subgaussian random vectors in $\mathbb{R}^M$ with $\|\Phi_n\|_2 = 1$, then for any sparsity level $1 \leq S \leq N$ and any $\delta \in (0, 1)$, the matrix Φ satisfies the RIP with parameters (S, δ) with high probability, provided

$$M \gtrsim \frac{S}{\delta^2} \log\left(\frac{N}{S}\right),$$

where $\gtrsim$ means that the quantity on the left is greater than the quantity on the right up to a scaling factor. Examples of subgaussian matrices include random matrices with independent and identically distributed Bernoulli columns with unit norm and matrices whose columns are drawn independently and uniformly at random from the unit sphere. In practice, it is unknown how to determine the RIP constant for a given matrix in polynomial time. However, rearranging the terms in the expression above gives a useful estimate:

$$\delta \sim \sqrt{\frac{S \log(N/S)}{M}}, \tag{3}$$

where $\sim$ means "equal up to a constant factor". This estimate is often used to evaluate the number of measurements necessary for a digital solver to recover a sparse signal (see Section 2.1) and will be useful to compare the theoretical guarantees obtained for the LCA to standard digital approaches in Chapter 4.

1.1.4 Sparse signal recovery

Ideally, the following optimization program could be used to recover the target signal $a^\dagger$ from its compressed measurements:

$$\hat{a}^\dagger = \arg\min_{a \in \mathbb{R}^N} \frac{1}{2} \|y - \Phi a\|_2^2 + \lambda \|a\|_0 . \tag{4}$$

The first term is a data fidelity term (the mean-squared error) and the second term is the ℓ_0-pseudo-norm $||\cdot||_0$, which counts the number of non-zero elements. The parameter λ provides a tradeoff between the two objectives. While it could recover the correct solution, this program is NP-hard, meaning that it is unknown if a solution can be attained in polynomial time.

To obtain a solution to the sparse recovery problem in polynomial time, one of the most famous and well-studied approaches is the ℓ_1-minimization problem, which replaces the ℓ_0-regularizer in (4) with its closest convex norm: $||a||_1 = \sum_i |a_i|$. This choice of regularizer yields the following convex program:

$$\hat{a}^\dagger = \arg\min_{a\in\mathbb{R}^N} \frac{1}{2} ||y - \Phi a||_2^2 + \lambda ||a||_1 . \tag{5}$$

This technique is known as convex relaxation. The ℓ_1-norm makes this program easier to solve while still enforcing sparsity on the solution, and yields comparable performances to the ideal sparse recovery problem (4) [6].

Generalizing further, the following objective function can be minimized:

$$V(a) = \frac{1}{2} ||y - \Phi a||_2^2 + \sum_{n=1}^{N} C(a_n), \tag{6}$$

where $C(\cdot) : \mathbb{R} \to \mathbb{R}$ is referred to as *cost function* and is chosen to enforce the sparsity requirement on the solution. For instance, a class of functions called *sparseness measures* was developed in [7] and shown to yield sparse solutions.

Definition 2 (Sparseness Measure)**.** *A function* $f(\cdot) : [0, \infty) \to [0, \infty)$ *is called a* sparseness measure *if it is nondecreasing, not identically zero, with* $f(0) = 0$ *and such that* $x \mapsto f(x)/x$ *is nonincreasing on* $(0, \infty)$*. Then, the associated (sparsity-inducing) cost function is defined for all* $a \in \mathbb{R}^N$ *by*

$$\sum_{n=1}^{N} C(a_n) = \sum_{n=1}^{N} f(|a_n|).$$

In particular, the identity function satisfies the requirements to be a sparness measure, so the ℓ_1-norm is a sparsity-inducing cost functions according to this definition.

1.1.5 Recovery guarantees

The sharpest result obtained for the constrained form of (5) requires the measurement matrix Φ to satisfy the RIP with parameters $(2S, \sqrt{2}-1)$ [8]. In [9], an expression for the error associated with solving (5) under assumptions similar to the RIP is given and can be re-written for the purpose of this thesis as

$$\left\|\hat{a}^\dagger - a^\dagger\right\|_2 \leq C_0 \lambda \sqrt{S} + C_1 \sigma, \tag{7}$$

where C_0 and C_1 are some small constants and σ is a bound on the energy of the noise: $\|\epsilon\|_2 \lesssim \sigma$. These results show that recovery with (5) is *uniform* and *stable*. Uniform recovery means that one choice of a measurement matrix Φ can recover *every* sparse signal, while stable recovery means that the error scales slowly with the noise level.

1.2 Locally Competitive Algorithm

The Locally Competitive Algorithm (LCA) proposed by Rozell and al. in 2008 [10] is a continuous-time dynamical system that is designed to solve sparse recovery problems in the form of (6). This algorithm can be viewed as a network of nodes whose evolution is described by a first-order ordinary differential equation.

1.2.1 The LCA differential equation

To each column Φ_n of the matrix Φ is associated a node or "neuron" in the network, whose internal state is modeled by a continuous-time variable $u_n(t)$, for $n = 1, \ldots, N$. The evolution of the state variables with time is governed by a set of coupled nonlinear ODEs of the form

$$\begin{cases} \tau\dot{u}(t) & = -u(t) - (\Phi^T\Phi - I)a(t) + \Phi^T y \\ a(t) & = T_\lambda(u(t)) \end{cases}. \tag{8}$$

The notation $\dot{F}(\cdot)$ refers to the derivative with respect to time: $\dot{F}(t) = \dfrac{dF(t)}{dt}$. The columns of Φ are assumed to be normalized to one: $\|\Phi_n\|_2 = 1$. The input to the network is a vector of measurements y in $\mathbb{R}^M$ whose projection onto Φ generates the set of *driving* *inputs* $\Phi_n^T y$, for $n = 1, \ldots, N$. These scalar values reflect how well the input y matches each dictionary element. The state variables produce outputs a_n, for $n = 1, \ldots, N$ through the activation function $T_\lambda(\cdot) : \mathbb{R} \mapsto \mathbb{R}$. By a slight abuse of notation, $T_\lambda(\cdot)$ applied to a vector $u \in \mathbb{R}^N$ means that the function is applied entry-wise. Each output a_m feeds back into each node u_n proportionally to the corresponding feedback weight $W_{n,m}$ of the interconnection matrix $W = \Phi^T \Phi - I$ (*i.e.*, a modified Grammian matrix for the dictionary). There is no self-feedback, as the diagonal elements of W are zero. When two nodes overlap (resulting in a large inner product), the corresponding feedback weight is close to one, while it is zero for orthogonal dictionary elements. Though the interconnection matrix is symmetric, the total inhibition is not because it is also modulated by the activity of each individual node. This feedback structure ensures that nodes that carry the same information about the signal inhibit each other. The parameter τ represents the time constant of the analog system implementing the ODE, and is characterized by the physical properties of the system (such as capacitance, resistance and level of bias current). Since it does not affect the mathematical analyses of the LCA, it is assumed in the following that $\tau = 1$ except when its influence on the convergence speed is made explicit. The architecture of the LCA is shown in Figure 2.

1.2.2 A Hopfield-type neural network

Due to its feedback structure combined with an activation function before the output stage, the LCA is a type of Hopfield neural network (HNN), a pioneering system of analog computing. The first HNN introduced by John J. Hopfield in the early 1980s is a network of simple computing units that can take on one of two values [11]. Contrary to earlier neural networks, such as the perceptron [12], the HNN contains feedback from every output to every input variable. This structure is characteristic of what are now known as *recurrent neural networks*. In a later paper, Hopfield proposed the same network structure with neurons that have graded, rather than binary, responses [13]. In both cases, Hopfield shows that a global behavior emerges from this intricate structure. More specifically, Hopfield shows

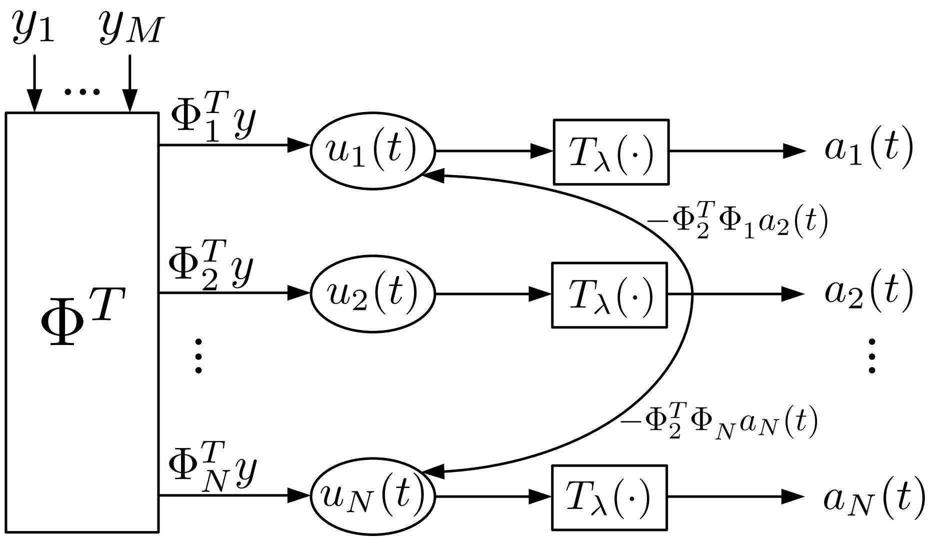


Figure 2: The LCA neural network is designed to solve sparse recovery problems. The activation function may be unbounded and not strictly increasing. The matrix Φ has dimension $M \times N$ with $M \ll N$, so the interconnection matrix W may be singular and have both positive and negative eigenvalues.

that the state variables of the network evolve to approach a global equilibrium point.

1.2.3 Limited analysis

An important contribution of Hopfield's work resides in proving theoretically that the outputs evolve towards an equilibrium. For this proof, Hopfield defines a function that represents a notion of energy for the system known as a *Lyapunov function* that is presented in more detail in Chapter 2. If this energy function is always strictly decreasing as the network evolves, then the output trajectories evolve towards a stable fixed point. Similarly, Rozell and al. showed in [10] that the objective function in (6) is decreasing along the LCA trajectories provided that $T'_\lambda(u) \geq 0$ on $\mathbb{R}$ and that the following relationship between the cost penalty term $C(\cdot)$ and the activation function $T_\lambda(\cdot)$ is satisfied for all $a_n \in \mathbb{R}$ such that $a_n \neq 0$:

$$\lambda \frac{dC(a_n)}{da_n} = u_n - a_n = u_n - T_\lambda(u_n).$$

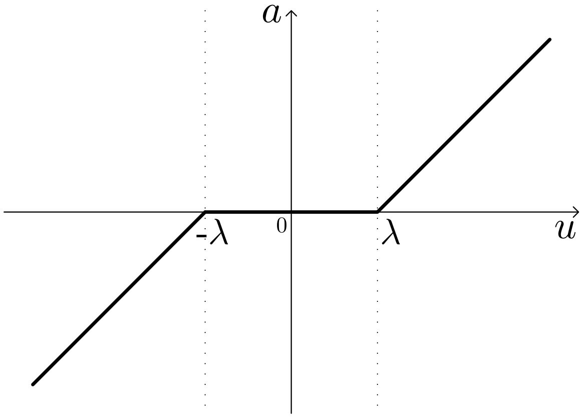


Figure 3: Plot of the soft-thresholding activation function. When this function is used as the activation function $T_\lambda(\cdot)$, the LCA solves the classic ℓ_1-minimization optimization problem used in many sparse approximation applications.

In the case of ℓ_1-minimization, the cost penalty is $C(x) = |x|$ and the associated activation function is the soft-thresholding function shown in Figure 3 and defined by:

$$a_n(t) = T_\lambda(u_n(t)) = \begin{cases} 0, & |u_n(t)| \leq \lambda \\ u_n(t) - \lambda \,\mathrm{sign}(u_n(t)), & |u_n(t)| > \lambda \end{cases}. \tag{9}$$

While a nonincreasing objective function is a necessary property for a network that solves an optimization problem, it is not sufficient to conclude that the state converges to a fixed point (or to a subset of fixed points). To prove convergence, the Lyapunov approach requires that the objective be strictly decreasing on non-stationary trajectories. Moreover, one must show that the fixed points correspond to actual solutions of the optimization program. Both of these results are necessary guarantees to make before relying on a system in engineering applications. Such guarantees have been obtained in the literature for related networks and are presented in Chapter 2. However, several characteristics distinguish the LCA from previous studies and make its analysis particularly challenging. First, the LCA activation function is often nonlinear and unbounded for problems of interest in CS. In fact, the LCA objective and activation functions are usually not differentiable everywhere. Second, the LCA interconnection matrix W has a potentially large nullspace, since $M \ll N$ in CS applications. For these reasons, it has been difficult to provide convergence guarantees for the network. The mathematical tools necessary to develop the analysis in this thesis and previous results in the literature are the focus of the next chapter.

CHAPTER II

BACKGROUND

The impetus of this thesis is the efficient recovery of sparse signals from compressed measurements using a continuous-time solver. This chapter begins with a review of standard methods and solvers for sparse recovery. In addition to classic methods and results for the analysis of neural networks, more recent breakthroughs and their limitations are discussed. Also introduced as needed are the analytic tools necessary to obtain the results presented in this chapter and the findings in later chapters.

2.1 *Sparse signal recovery*

Significant efforts have been put into developing algorithms that can solve the sparse recovery problem (4) efficiently. These algorithms can broadly be divided into two categories: relaxation methods that solve an optimization program and greedy algorithms that recover the locations of the non-zero coefficients iteratively. The LCA belongs to the class of relaxation methods. Convergence and accuracy results have been obtained in the digital community via the RIP in (2) for many algorithms. There are typically two kinds of requirements that emerge from those studies: either δ must scale as $1/\sqrt{S}$ or δ is a small constant independent of the sparsity level S. Using the estimate for δ in (3) for subgaussian random matrices, the corresponding number of measurements are $\mathcal{O}\left(S^2 \log(N/S)\right)$ and $\mathcal{O}(S \log(N/S))$, respectively. In practice, the results obtained for a number of measurements $M = \mathcal{O}\left(S^2 \log(N/S)\right)$ are stronger, but it is more desirable to obtain guarantees that require a smaller number of measurements $\mathcal{O}(S \log(N/S))$. Presented below are several recovery algorithms that, while digital, show interesting parallels to the properties of the LCA that arise from the analysis in this thesis.

2.1.1 Relaxation methods

While the accuracy results associated with the ℓ_1-minimization program in (5) presented in Section 1.1.5 are the most favorable to date, current digital solvers for ℓ_1-minimization suffer from high computational costs and tend to lack convergence-time guarantees. Some state-of-the art solvers (e.g., [14–18]) can handle large-scale problems, but they usually do not have strong guarantees about their running time. On the other hand, iterative thresholding schemes (e.g. [19,20]) are simple and come with guarantees on the number of iterations needed to achieve a certain accuracy. Unfortunately, this number may be large. Homotopy-based schemes solve (5) by tracing a piecewise-linear solution path as the tradeoff parameter λ is varied [21, 22]. If the solution is very sparse and the number of measurements is large enough, these approaches can converge in exactly S iterations, known as the *S-step property*. For instance, for subgaussian random matrices that satisfy (3), the homotopy for (5) converges in S-steps for a number of measurements $M \sim \mathcal{O}\left(S^2 \log(N/S)\right)$ [23]. While the accuracy guarantees for the above algorithms lead to state-of-the-art results, their complexity prevents their use in real-time applications for very large signals or data sampled at very high rates.

2.1.2 Greedy algorithms

Greedy algorithms solve (4) by recovering the support of the original signal iteratively. These solvers are faster than relaxation methods in general, but have less sharp performance guarantees. The most basic greedy algorithm is Orthogonal Matching Pursuit (OMP), which adds to the support the element that has the strongest correlation with the residual at each iteration. Conditions on the RIP for OMP to recover the target signal accurately in $\mathcal{O}(S)$ iterations were obtained in the noiseless [24] and noisy cases [25]. In the case of subgaussian random matrices satisfying (3), the corresponding number of measurements is on the order of $\mathcal{O}(S \log(N/S))$. Recent work has also shown that OMP can recover an S-sparse signal in exactly S iterations (*i.e.*, has the S-step property) [26]. The corresponding number of noiseless measurements is on the order of $\mathcal{O}(S^2 \log(N/S))$ for random matrices satisfying (3). In contrast to OMP, Regularized Orthogonal Matching Pursuit (ROMP) [27] and Compressive Sampling Matching Pursuit (CoSaMP) [28] add a set of nodes at each iteration. Both ROMP and CoSaMP guarantee uniform and stable recovery in $\mathcal{O}(S)$ iterations from only $\mathcal{O}(S \log(N/S))$ measurements for random matrices satisfying (3). However, the requirements on the RIP constant for these algorithms are slightly stricter than necessary for ℓ_1-minimization: $\delta_{8S} \leq \frac{0.01}{\sqrt{\log S}}$ and $\delta_{4S} \leq 0.1$ for ROMP and CoSaMP respectively.

2.2 Neural network analysis

Contrary to the solvers discussed in the previous section, the LCA is a continuous-time algorithm, and belongs to the class of Hopfield-type neural networks. The convergence analysis for a continuous-time system cannot be done in terms of number of iterations as in the digital case. When analyzing a neural network, the main goal is to determine the behavior of the state and output variables with respect to time. In particular, the outputs need to settle to an appropriate equilibrium point for the network to be suited for optimization, and an estimate for the convergence speed needs to be determined. This section presents some definitions and some tools for determining the stability of a network.

2.2.1 Stability and convergence

For any function $F(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$, a *fixed point* of the ODE

$$\dot{x}(t) = F(x(t)), \qquad \forall t \geq 0, \tag{10}$$

is a constant vector $x^* \in \mathbb{R}^N$ such that $F(x^*) = 0$. There exist several notions of stability that describe the evolution of trajectories both locally and globally near a fixed point. First, the notion of Lyapunov stability describes the behavior of the trajectories locally around a fixed point and guarantees that if $x(t)$ starts close to a fixed point, it will remain nearby. Formally, a fixed point x^* of (10) is (Lyapunov) *stable* if for each $\varepsilon > 0$, there exists an $R > 0$ such that, for all starting points x_0 with $\|x_0 - x^*\| < R$ (*i.e.*, $x_0 \in \mathcal{B}_R(x^*)$) and all

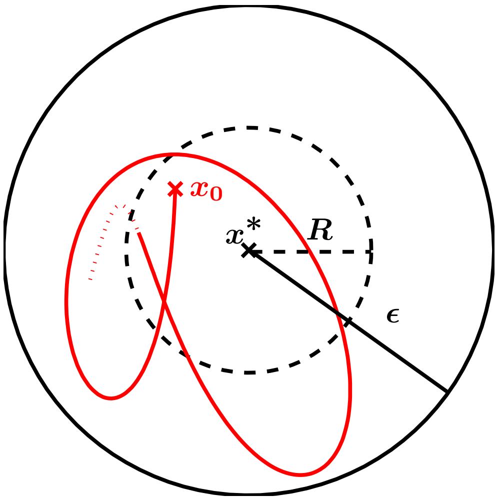


Figure 4: A point is (Lyapunov) stable if a trajectory that starts nearby (within a ball of radius R) remains nearby (within a ball of radius ε).

solutions $x(\cdot) : \mathbb{R} \to \mathbb{R}^N$ with $x(0) = x_0$,

$$\|x(t) - x^*\| < \varepsilon, \qquad \forall t > 0. \tag{11}$$

This property is illustrated in Figure 4. As can be seen in the figure, this type of stability does not guarantee that trajectories approach a fixed point as time goes to infinity. However, a fixed point x^* is called *asymptotically stable* if, for any initial states $x_0 \in \mathbb{R}^N$ such that x_0 is in a neighborhood $\mathcal{B}_R(x^*)$ of the fixed point for some $R > 0$, the solutions $x(\cdot) : \mathbb{R} \to \mathbb{R}^N$ with $x(0) = x_0$ satisfy $\lim\limits_{t\to+\infty} x(t) = x^*$. It is *globally asymptotically stable* if this limit holds for any $x_0 \in \mathbb{R}^N$. In this case, every trajectory is guaranteed to approach a unique fixed point as t goes to infinity.

These notions extend to neural networks. The network (10) is said to be *globally convergent*, or equivalently *globally asymptotically stable*, if there exists a unique fixed point x^* that is globally asymptotically stable. On the other hand, if the trajectories can only be shown to approach a **set** of stable fixed points, then the neural network is called *quasi-convergent*.

In addition to the question of stability, it is essential to know how fast trajectories converge for real-time applications. Of interest in this thesis is the notion of an exponential rate of convergence. The network (10) is called *exponentially convergent* to a fixed point x^* if there exists a constant $c > 0$ such that for any initial point $x(0)$, there exists a constant $\kappa_0 > 0$ (which may depend on $x(0)$) for which the solutions $x(\cdot) : \mathbb{R} \to \mathbb{R}^N$ of (8) with $x(0) = x_0$ satisfy

$$\|x(t) - x^*\| \leq \kappa_0 e^{-ct}, \qquad \forall t \geq 0.$$

The constant c is referred to as *convergence speed* of the system. When a network is exponentially convergent, the distance to the fixed point decays rapidly and can be considered small for $t \sim 1/c$. When c is known, an exact time to achieve a specified error can be computed.

2.2.2 Lyapunov's direct method

Lyapunov's direct method, developed by Russian mathematician Aleksandr M. Lyapunov in the late nineteenth century, makes the mathematical analysis of the stability and convergence of some neural networks easier [29]. The key to this method resides in finding a positive-definite function that represents a notion of energy for the dynamical system (10). If the energy function is nonincreasing along the system's trajectories, then the fixed points are stable. If in addition the energy function is strictly decreasing along all the nonstationary trajectories, the fixed points are also asymptotically stable.

Theorem (Lyapunov's Direct method)**.** *If there exists an open set* Ω *that contains* 0 *and a function* $V(\cdot) : \Omega \longrightarrow \mathbb{R}$ *which is continuous and positive definite on* Ω*, with* $\dot{V}(x) \leq 0$ *for all* $x \in \Omega$*, then* $V(\cdot)$ *is called a weak Lyapunov function for* (10) *and, the solution* $x(t) = 0$ *is a stable fixed point of* (10)*.*

If, in addition, $\dot{V}(x) < 0$ *for all* $x \in \Omega \backslash \{0\}$*, then* $V(\cdot)$ *is called a Lyapunov function or strict Lyapunov function for* (10) *and the solution* $x(t) = 0$ *is asymptotically stable.*

The time derivative $\dot{V}(\cdot)$ can be computed using the classic chain rule:

$$\dot{V}(x(t)) = \nabla_x V(x)^T \dot{x}(t) = \nabla_x V(x)^T F(x(t)),$$

where $\nabla_x V(x)$ denotes the usual gradient of $V(\cdot)$ at x. Since the term $x(t)$ does not appear in the above expression, it is neither necessary to solve the differential equation (10) explicitly nor to apply the definition of stability directly to determine the stability of the fixed point. These simplifications make the Lyapunov method extremely useful and powerful for the convergence analysis of neural networks.

The theorem can easily be adapted to a fixed point x^* different from $x(t) = 0$. For this case, it suffices to apply the theorem to the translated differential equation:

$$\dot{u}(t) = F(u(t) + x^*),$$

where $u(t) = x(t) - x^*$. Using the translated Lyapunov function $W(t) = V(u(t) + x^*)$, $u(t) = 0$ is a solution of the above ODE that has the same stability as $x(t) = x^*$.

2.2.3 Previous work

Using Lyapunov's direct method, Hopfield showed that the HNN converges to a stable fixed point that corresponds to the minimum of the energy function [13]. Later, these ideas naturally led Hopfield to consider the reverse problem. Starting from an objective function to minimize, he showed how to choose the neural network parameters to perform the desired computation [30]. He applied this technique to the traveling-salesman problem in [30] and to linear programming in [31]. These were pioneering steps in the field of analog computing that paved the way for many extensions. In particular, the LCA descends from this lineage of neural networks designed for a specific optimization.

Unfortunately, not all optimization programs have the necessary properties for Lyapunov's method to apply. Specifically, Hopfield's paper on linear programming [31] restricts the matrix W to be symmetric with zeroes on the diagonal, the activation function to be nondecreasing everywhere, and the activation and objective functions to be smooth and accept a derivative everywhere. The need for neural networks that can solve more complex optimization programs has led researchers to analyze neural networks that extend the classic HNN.

To remove the symmetry condition on the interconnection matrix, the authors of [32] prove global asymptotic convergence when the interconnection matrix is lower triangular. In [33], the interconnection can be non-symmetric but must have symmetric and positive semidefinite submatrices. The result in [34] removes the symmetry assumption altogether. However, these results require the activation function to be bounded and strictly increasing. In [35], this condition is also removed by letting the activation function be possibly unbounded and with slope zero. This result is particularly interesting for the LCA, whose activation function contains a thresholding region where the outputs are exactly zero over some interval. Unfortunately, to show global asymptotic convergence of such a system, the authors of [35] develop the notion of a Lyapunov Diagonally-Stable matrix, which requires that the interconnection matrix W be nonsingular. As stated before, for problems in CS recovery, the interconnection matrix W may have a large nullspace.

2.3 Nonsmooth neural networks

While the LCA architecture is a type of HNN, its objective function does not satisfy the smoothness requirement of the traditional Lyapunov approach. In an effort to extend neural networks to more general classes of optimization, several papers have considered nonsmooth objective functions. Their analysis relies on the notion of subgradient developed by Clarke [36], and on the theory of differential inclusions as studied by Filippov [37]. The typical approach considers a network that satisfies a differential inclusion of the form

$$\dot{x}(t) \in -\partial F(x(t)),$$

where $\partial F(x)$ represents the subgradient of $F(\cdot)$ at x. The next sections will introduce the notions of subgradient, regularity and some useful calculus results.

2.3.1 Subgradient

The subgradient extends the traditional notion of gradient to functions that are locally Lipschitz but not necessarily differentiable. The definition of *subgradient*, also called *generalized gradient*, developed in [36] and [38] is used in this thesis.

A function $F(\cdot) : \mathbb{R}^N \to \mathbb{R}$ is called *locally Lipschitz* at $x \in \mathbb{R}^N$ if there exist $\varepsilon > 0$ and $K > 0$ such that for all $x_1,\ x_2 \in \mathcal{B}_\varepsilon(x)$ (*i.e.*, in an ε-neighborhood of x), one has $|F(x_1) - F(x_2)| \leq K \|x_1 - x_2\|$. A locally Lipschitz function is not necessarily differentiable. However, Rademacher's theorem implies that a locally Lipschitz function $F(\cdot)$ is differentiable almost everywhere (a.e.) on any neighborhood of x in which $F(\cdot)$ is Lipschitz.

For a function $F(\cdot) : \mathbb{R}^N \to \mathbb{R}$ locally Lipschitz at $x \in \mathbb{R}^N$, there exist several definitions related to the standard notion of directional derivative. The usual *one-sided directional derivative* of $F(\cdot)$ at $x \in \mathbb{R}^N$ in the direction $v \in \mathbb{R}^N$ is

$$F'(x; v) = \lim_{t \downarrow 0} \frac{F(x + tv) - F(x)}{t}.$$

Since some nonsmooth functions may fail to admit one-sided derivatives, this definition can be relaxed to the following notion of *generalized directional derivative*:

$$F^\circ(x; v) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + tv) - F(y)}{t}.$$

With this definition, the existence of directional derivatives of $F(\cdot)$ at x are not necessary. For instance, the quantity $F^\circ(x; v)$ is well-defined when $F(\cdot)$ is only locally Lipschitz. This notion has been generalized even further to functions that are only directionally Lipschitz [39], but this is beyond the scope of this thesis.

The *subgradient* of $F(\cdot)$ at x is the subset of $\mathbb{R}^N$ defined by

$$\partial F(x) = \left\{ \xi \in \mathbb{R}^N \ s.t.\ F^\circ(x; v) \geq \xi^T v, \ \ \forall v \in \mathbb{R}^N \right\}.$$

For a locally Lipschitz function, this set is well-defined, nonempty and convex. Since Rademacher's theorem implies that $F(\cdot)$ is differentiable a.e., the set Ω_F of points where

$F(\cdot)$ fails to be differentiable has Lebesgue measure zero. Then, the subgradient simplifies to the following definition [36]:

$$\partial F(x) = co \left\{ \lim_{i \to \infty} \nabla F(x_i) : x_i \to x, x_i \notin \mathcal{S}, x_i \notin \Omega_F \right\},$$

where co is the convex hull, and $\mathcal{S}$ is any set of Lebesgue measure 0 in $\mathbb{R}^N$. In other words, $\partial F(x)$ is the smallest convex set containing the limit points of the gradients along any sequence of points $\{x_i\}$ approaching x while avoiding $\Omega_F \cup \mathcal{S}$. When $F(\cdot)$ is smooth at x, $\partial F(x)$ is a singleton that coincides with the classic notion of gradient $\partial F(x) = \{\nabla F(x)\}$. For a convex function, this notion also coincides with the notion of subgradient in convex analysis.

2.3.2 Generalized chain rule and calculus

The notion of regularity is essential to easily compute the subgradient and apply the chain rule to nonsmooth functions with equality[1]. The function $F(\cdot) : \mathbb{R}^N \mapsto \mathbb{R}$ is *regular* at x if $F'(x; v)$ exists and $F'(x; v) = F^\circ(x; v)$ for all $v \in \mathbb{R}^N$ [36, Def. 2.3.4]. The following chain rule concerns functions of the form $G(\cdot) = F \circ H(\cdot)$, where $H(\cdot) : \mathbb{R}^M \to \mathbb{R}^N$ and $F(\cdot) : \mathbb{R}^N \to \mathbb{R}$.

Theorem (Chain Rule I)**.** *Assume that the function $H(\cdot) : \mathbb{R}^M \to \mathbb{R}^N$ has component functions $H_n(\cdot) : \mathbb{R}^M \to \mathbb{R}$ for $n = 1, \ldots, N$, that each $H_n(\cdot)$ is locally Lipschitz and differentiable at $x \in \mathbb{R}^M$, and $F(\cdot) : \mathbb{R}^N \to \mathbb{R}$ is locally Lipschitz and regular at $H(x)$. Then, the function $G(\cdot) = F \circ H(\cdot)$ is locally Lipschitz and regular at x and*

$$\partial G(x) = \partial F \circ H(x) = \left\{ \sum_{n=1}^{N} \zeta_n \nabla H_n(x) \ s.t. \ \zeta \in \partial F(H(x)) \ and \ \zeta = (\zeta_1, \ldots, \zeta_N) \right\}. \tag{12}$$

In addition, Corollary 3 of Propositions 2.3.1 and 2.3.3 of [36] imply that if N functions $F_n(\cdot) : \mathbb{R}^N \mapsto \mathbb{R}$ for $n = 1, \ldots, N$ are locally Lipschitz and regular at x, and $\alpha_n \geq 0$ for $n = 1, \ldots, N$, then

$$\partial \left(\sum_{n=1}^{N} \alpha_n F_n \right)(x) = \sum_{n=1}^{N} \alpha_n \partial F_n(x). \tag{13}$$

[1] Without the notion of regularity, most of the properties in this section hold only with an inclusion in one direction rather than an equality between two sets.

Finally, if $G(\cdot) = F \circ x(\cdot)$ with $x(\cdot) : [0, +\infty) \to \mathbb{R}^N$, then Theorem 2.3.10 in [36] given below yields a special case of the chain rule that is often used in the study of the LCA trajectories.

Theorem (Chain Rule II)**.** *If* $F(\cdot) : \mathbb{R}^N \to \mathbb{R}$ *is locally Lipschitz and regular on* $\mathbb{R}^N$ *and* $x(\cdot) : [0, +\infty) \to \mathbb{R}^N$ *is differentiable on* $[0, +\infty)$*, then* $F(x(\cdot))$ *is also locally Lipschitz and regular on* $[0, +\infty)$*, its time derivative* $\dot{F}(x(t))$ *exists for almost all (a.a.)* $t \geq 0$ *and satisfies*

$$\dot{F}(x(t)) = \zeta^T \dot{x}(t), \qquad \forall \zeta \in \partial F(x(t)). \tag{14}$$

This theorem states that any element ζ in the subgradient can be used to compute the time derivative of $F(x(t))$.

2.3.3 Previous work

Using these new tools, several papers have given convergence results for nonsmooth neural networks. Unfortunately, several characteristics distinguish the LCA from previous studies. In [40], the objective function is linear, while it is piecewise linear in [41], and nonlinear but increasing and bounded in [34, 42, 43]. On the contrary, for cases of interest, the LCA activation function is nonlinear, unbounded, and exactly zero on the interval $[-\lambda, \lambda]$. Furthermore, contrary to previous work on nonsmooth systems, Lemma 1 shows that the LCA dynamics satisfy

$$\dot{u}(t) \in -\partial_a V(a(t)),$$

rather than $\dot{u}(t) \in -\partial_u V(u(t))$ or $\dot{a}(t) \in -\partial_a V(a(t))$ as in [44–46]. This difference is significant since the state variables $u(t)$ could still be evolving while the objective $V(a(t))$ remains constant. Finally, the LCA interconnection matrix W has a potentially large nullspace since $M \ll N$, whereas other analyses assume the interconnection matrix to be positive definite [41] or nonsingular [33, 35]. For these reasons, it has been difficult to provide convergence guarantees for the LCA network.

2.4 Finite length of trajectories

Techniques based on Lyapunov functions only prove convergence to a set of fixed points. If there exists a set of connected fixed points, the trajectories are only guaranteed to evolve towards this set, but there is no certainty that they will converge towards one unique point in the set. In other words, trajectories are not prevented from growing unbounded or oscillating indefinitely as they approach the solution set. Recent papers have developed a new technique based on the Łojasiewicz inequality [47] to overcome this limitation. While the results obtained with this technique are strong, they do not readily apply to the LCA specifics.

2.4.1 Subanalicity and Łojasiewicz inequality

The Łojasiewicz (gradient) inequality relies on geometric properties of a function, and relates differences of a function near a point to the value of its gradient at that point [47]. Formally, it states that for a real-analytic function $F(\cdot) : \mathbb{R}^N \to \mathbb{R}$ and for all $\overline{x} \in \mathbb{R}^N$, there exists $\nu \in [0, 1)$, $C > 0$ and $\Delta > 0$ such that the function $F(\cdot)$ satisfies

$$|F(x) - F(\overline{x})|^{\nu} \leq C \left\|\nabla F(x)\right\|, \qquad \forall x \in B_{\Delta}(\overline{x}).$$

Using this inequality, Łojasiewicz showed that the trajectories of networks of the form $\dot{x}(t) = -\nabla F(x(t))$ have finite length, thus ensuring their convergence to a singleton even when the fixed points are not isolated [47].

Recently, an extension of the Łojasiewicz inequality was developed for nonsmooth functions [44, Th 3.1.]. The gradient in the original formulation is replaced by the nonsmooth slope, which represents the smallest norm of any vector in the set $\partial F(x)$.

Theorem (Nonsmooth Łojasiewicz inequality)**.** *Suppose that a function $F(\cdot) : \mathbb{R}^N \to \mathbb{R}$ is subanalytic and continuous on $\mathbb{R}^N$. Then, for any $\overline{x} \in \mathbb{R}^N$, there exist $\nu \in [0, 1)$, $C > 0$, and $\Delta > 0$ such that*

$$|F(x) - F(\overline{x})|^{\nu} \leq C\, m(\partial F(x)), \qquad \forall x \in B_{\Delta}(\overline{x}),$$

where the nonsmooth slope *of $F(\cdot)$ at $x \in \mathbb{R}^N$ is defined as*

$$m(\partial F(x)) = \inf\Big\{\|\xi\|_2\,,\ \xi \in \partial F(x)\Big\}. \tag{15}$$

The nonsmooth Łojasiewicz inequality requires the function $F(\cdot)$ to be *subanalytic*. This property does not require the function to be differentiable, but it involves geometric properties of the graph, such as algebraic manipulations (unions and intersections) of sets defined by real-analytic equations and inequalities. More precisely, a set $A \subset \mathbb{R}^N$ is said to be *semianalytic* if each point $x \in \mathbb{R}^N$ admits a neighborhood $\mathcal{N}$ for which

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

where $f_{ij}(\cdot), g_{ij}(\cdot) : \mathcal{N} \to \mathbb{R}$ 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 B is said to be *subanalytic* if it is locally the projection of a semianalytic set, *i.e.*, each point $x \in \mathbb{R}^N$ admits a neighborhood $\mathcal{N}$ such that $B \cap \mathcal{N} = \Big\{x \in \mathbb{R}^N,\ \ (x, y) \in A\Big\}$, where A is a bounded semianalytic subset of $\mathbb{R}^N \times \mathbb{R}^M$ for some $M \geq 1$. Finally, a function $F(\cdot) : \mathbb{R}^N \to \mathbb{R}$ is said to be *subanalytic* if its graph, $\mathrm{Graf}F = \{(x, y)\ s.t.\ y = F(x)\}$, is a subanalytic subset of $\mathbb{R}^N \times \mathbb{R}$.

2.4.2 Previous work

Several recent papers have used the Łojasiewicz inequality to show convergence of specific neural networks to a single point even when the fixed points are not isolated. In [42], a general approach is taken where the network's equation has the form

$$\begin{cases} \dot{u}(t) = -Du(t) - \nabla F(a(t)) \\ a(t) = T(u(t)) \end{cases}.$$

In this paper, the functions $F(\cdot)$ and $T(\cdot)$ are assumed to be analytic (which implies the existence of derivatives of any order), and the activation function $T(\cdot)$ is required to be bounded and strictly increasing. Thanks to the extension of the Łojasiewicz inequality to nonsmooth functions, the result of this paper was extended to the nonsmooth case in [45]. However, in this later paper only quadratic programming with linear constraints is considered. In [44], the authors show how a network satisfying the differential inclusion $\dot{u}(t) \in -\partial F(u(t))$ has finite-length trajectories if $F(\cdot)$ is subanalytic and either lower semicontinuous convex or lower-C^2. Unfortunately, as explained in section 2.3.3, the LCA cannot be put in this form as it satisfies a different inclusion, namely $\dot{u}(t) \in -\partial_a V(a(t))$ (see Lemma 1). Finally, the authors of [43] make use of the nonsmooth Łojasiewicz inequality to prove that a network of the form

$$\begin{cases} \dot{u}(t) \in -Du(t) - \partial V(a(t)) + \theta \\ a(t) = T(u(t)) \end{cases}$$

converges to a singleton. However, the activation function must be bounded and the matrix D must be diagonal with strictly positive entries, which does not comply with the specifics of the LCA.

CHAPTER III

PROPERTIES FOR GENERIC SPARSE RECOVERY

For any engineering application, it is essential to obtain theoretical guarantees on the behavior of a system before its deployment. The previous chapter showed that significant advances have been made in the field of neural network analysis. Unfortunately, the specifics of the LCA neural network do not fit the necessary criteria for any of the existing approaches. In this chapter, theoretical results that extend previous requirements for neural network analysis (in particular on the activation function and the feedback matrix) are presented. Using tools from nonsmooth analysis, the results of this chapter prove that the LCA is well-suited for solving a wide class of nonsmooth optimization programs by showing that

- the fixed points of the neural network correspond to critical points of the desired objective function,
- the network trajectories converge to a fixed point from any initial state when the fixed points are isolated,
- the network trajectories converge to a fixed point from any initial state even when the fixed points are not isolated,
- the support of the solution is recovered in finite time, and
- the network trajectories converge exponentially fast from any initial state.

These guarantees are essential for a dynamical system designed to solve an optimization program in real-world applications. The resulting class of neural networks extends the neural networks previously studied in the literature and on the results published in [48–50].

3.1 Nonsmooth objective

In the most generic case, the LCA is designed to solve optimization programs with an objective function of the form

$$V(a) = \frac{1}{2}\|y - \Phi a\|_2^2 + \mathbf{C}(a), \tag{16}$$

where $\mathbf{C}(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$. For many practical applications, the cost function $\mathbf{C}(\cdot)$ is separable into its component in each dimension:

$$\mathbf{C}(a) = \sum_{n=1}^{N} C(a_n), \qquad \forall a = (a_1, \ldots, a_N) \in \mathbb{R}^N. \tag{17}$$

For instance, the cost function in the ℓ_1-minimization objective takes the form of (17) with $C(a_n) = \lambda |a_n|$ for some $\lambda > 0$. Sparseness measures presented in Definition 2 provide another example of such separable cost functions. The main difficulty for the analysis is the fact that the objective function is not necessarily differentiable everywhere. For instance, the ℓ_1-minimization objective is not differentiable at points $a \in \mathbb{R}^N$ that have one or more entries equal to zero. The theoretical results in the following sections show under what conditions on the activation function the LCA can be used to solve optimization programs of the form (16). The conditions imposed on the activation function are general and encompass a wide variety of objective functions used in sparse recovery.

3.2 Fixed points

The first result of this chapter concerns the fixed points of the LCA neural network and presents a condition for them to correspond to solutions of the desired objective. The condition is general and assumes nothing about the form of the activation function.

Theorem 1. *Assume that the cost function* $\mathbf{C}(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$ *in* (16) *is locally Lipschitz and regular on* $\mathbb{R}^N$. *If the activation function* $T_\lambda(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$ *in* (8) *and the cost function satisfy, for all* $a \in \mathbb{R}^N$,

$$u - a = u - T_\lambda(u) \in \partial\mathbf{C}(a), \tag{18}$$

then the fixed points of the LCA are critical points of the objective function.

Critical points of the objective function in (16) are defined as points $a^* \in \mathbb{R}^N$ that satisfy $0 \in \partial V(a^*)$. The set of critical points includes the local minima and maxima of $V(\cdot)$. If the objective is convex, then all of the critical points are local minima. If it is strictly convex, then there is a unique minimum and, as a consequence, the LCA has a unique fixed point. In Theorem 1, the cost function is only required to be locally Lipschitz and regular, which is a weak requirement satisfied by many functions that are used in practice. For instance, the cost function does not need to be differentiable. The ℓ_1-norm satisfies this condition, and it will be shown later that the soft-thresholding satisfies the relationship (18) for this cost function. For this theorem, the activation does not necessarily have to be continuous. In the following, a cost function satisfying (19) is derived for the famous hard-thresholding function, which is not continuous. Other activation functions satisfying (19) for several sparsity-inducing cost functions of interest can be found in [51]. In the special case where the cost function $\mathbf{C}(\cdot)$ is separable and takes the form of (17), the following corollary holds.

Corollary 1. *Assume that the cost function $C(\cdot) : \mathbb{R} \to \mathbb{R}$ in* (6) *is locally Lipschitz and regular on $\mathbb{R}^N$. If the activation function $T_\lambda(\cdot) : \mathbb{R} \to \mathbb{R}$ in* (8) *and the cost function $C(\cdot)$ satisfy, for all $u_n \in \mathbb{R}$,*

$$u_n - a_n = u_n - T_\lambda(u_n) \in \partial C(a_n), \tag{19}$$

then the fixed points of the LCA are critical points of the objective function.

For example, when solving ℓ_1-minimization, since the left and right derivative of the cost function $C(a_n) = \lambda\,|a_n|$ exist for all $a_n \in \mathbb{R}$, the cost function is obviously locally Lipschitz (with Lipschitz constant λ) and regular, and it is easy to check that its subgradient is

$$\partial C(a_n) = \begin{cases} \lambda \operatorname{sign}(a_n), & \text{for } a_n \neq 0 \\ [-\lambda, \lambda]\,, & \text{for } a_n = 0 \end{cases}.$$

Since the soft-thresholding function (9) satisfies

$$u_n - a_n = u_n - T_\lambda(u_n) = \begin{cases} \lambda\,\text{sign}(u_n) = \lambda\,\text{sign}(a_n), & \text{if } |u_n| > \lambda \text{ (for which } |a_n| > 0) \\ u_n, & \text{if } u_n \in [-\lambda, \lambda] \text{ (for which } a_n = 0) \end{cases}$$

by Corollary 1, this expression shows that the soft-thresholding function in (9) satisfies (19) and can be used to solve the ℓ_1-minimization program (5).

3.2.1 Simulations

Unless stated otherwise, all of the experimental results in this thesis are obtained from simulating the LCA dynamical equations (8) in Matlab using a first-order discrete approximation with a step size of 0.001 and a time constant chosen to be equal to $\tau = 0.01$. The internal states are started at rest (*i.e.*, $u(0) = 0$) and the system is given enough time to converge.

To illustrate the fact that the fixed points of the LCA correspond to critical points of the desired objective function, three examples are studied. In the first case, the soft- and hard-thresholding functions are used to recover a sparse signal from CS measurements. In the second case, Tikhonov regularization is used as the objective function. The fixed point reached by the LCA is compared to the solution of a digital solver for sparse approximation, called SpaRSA [52]. In addition to being a state-of-the-art solver, SpaRSA is used for comparison because it can take as an argument the specific cost function $C(\cdot)$ to be used in the optimization, while most other existing solvers only handle ℓ_1-minimization.

3.2.1.1 Sparse recovery

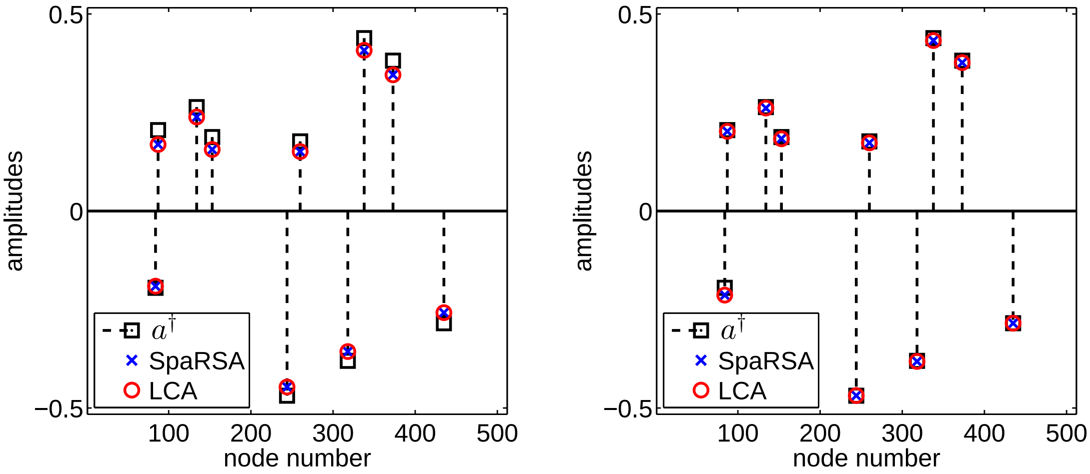
Two optimization programs for sparse recovery are considered. First, the soft-thresholding function in (9) is used. Since the ℓ_1-norm was shown to satisfy (19), Theorem 1 implies that the LCA should solve the ℓ_1-minimization program in (5). Second, the hard-thresholding function defined by $T_\lambda(u) = u$ if $|u| > \lambda$ and $T_\lambda(u) = 0$ otherwise is considered. In Appendix D, it is shown how to construct an associated cost function that satisfies (19) when the activation function has discontinuities. While the corresponding cost function does not exactly correspond to the ℓ_0-pseudo norm (which is not locally Lipschitz at 0), it has been shown in [20] that the hard-thresholding function can be used to approximately recover the solution to the ideal ℓ_0-minimization program (4) (with a tradeoff parameter of $\lambda^2/2$).

To test these statements, a vector $a^\dagger$ of length $N = 512$ is generated by selecting $S = 10$ non-zero entries uniformly at random. Amplitudes for the non-zero entries are drawn from a uniform distribution on $[1, 3]$ and $a^\dagger$ is normalized to have unit norm. The dictionary Φ is a union of the canonical basis and a sinusoidal basis having dimensions $M \times N$ with $M = 256$. The vector of measurements is $y = \Phi a^\dagger + \epsilon$, where ϵ is a Gaussian random noise vector with standard deviation $\sigma = 0.1 \left\|\Phi a^\dagger\right\|_2 / \sqrt{M}$ (which is a moderate level of noise). The threshold for the activation function is $\lambda = 0.025$. Figure 5 shows that the fixed point a^* reached by the LCA is indeed close to the target vector $a^\dagger$ in both cases, though the amplitudes cannot be exactly recovered because of the noise, as predicted by CS theory. The solutions reached by the network are close to those produced by the digital solver SpaRSA used with the ℓ_1-norm and ℓ_0-pseudo norm, respectively. This experiment confirms that the fixed points of the LCA correspond to solutions of the desired objective function as predicted by Theorem 1.

3.2.1.2 Tikhonov regularization

For the second example, the target signal does not need to be sparse. In Tikhonov regularization, the cost function in (6) is $C(a_n) = \lambda |a_n|^2$. An activation function satisfying (19) can be easily checked to be $T_\lambda(u_n) = u_n/(1 + 2\lambda)$. The parameter λ is chosen to regularize the solution when the matrix Φ is ill-conditioned.

To illustrate this program, a Gaussian random matrix Φ of size 256×256 is generated. After taking a singular value decomposition, the last 50 singular values of Φ are set to a small value by multiplying them by 10^{-10}. The columns of Φ are then normalized to have unit norm. A vector $a^\dagger$ of length $N = 256$ is obtained by generating a random linear combination of the 20 first right singular vectors. Coefficients of the linear combination are drawn from a standard Gaussian distribution. The vector of measurements is $y = \Phi a^\dagger + \epsilon$, where



(a) Using the soft-thresholding function

(b) Using the hard-thresholding function

Figure 5: Output a^* of the LCA after convergence. Only non-zero elements are plotted. The fixed point reached by the system is close to the initial sparse vector used to create the measurements (it cannot be exact due to noise). The solution is also close to the solution of the standard digital solver SpaRSA.

ϵ is random Gaussian noise with standard deviation $\sigma = 0.1 \left\|\Phi a^\dagger\right\|_2 / \sqrt{M}$. The regularizing parameter is set to $\lambda = 0.25$. The closed-form solution to the Tikhonov regularization problem can be computed explicitly as

$$a^{\text{Tik}} = \left(\Phi^T\Phi + 2\lambda I\right)^{-1} \Phi^T y.$$

In Figure 6, the absolute error in the Tikhonov solution $\left|a_k^{\text{Tik}} - a_k\right|$ is plotted for the fixed points of LCA and SpaRSA. Both algorithms yield an output that is very close to the true closed-form solution. This experiment again agrees with the conclusion of Theorem 1.

3.2.2 Proof of Theorem 1

Before proving the main theorem, the fundamental lemma below redefines the LCA as a differential inclusion.

Lemma 1. *Assume that the cost function* $\mathbf{C}(\cdot)$ *in* (16) *is locally Lipschitz and regular on* $\mathbb{R}^N$ *and that the LCA activation function in* (8) *satisfies* (19). *Then the LCA trajectories satisfy the following differential inclusion:*

$$-\dot{u}(t) \in \partial V(a(t)). \tag{20}$$

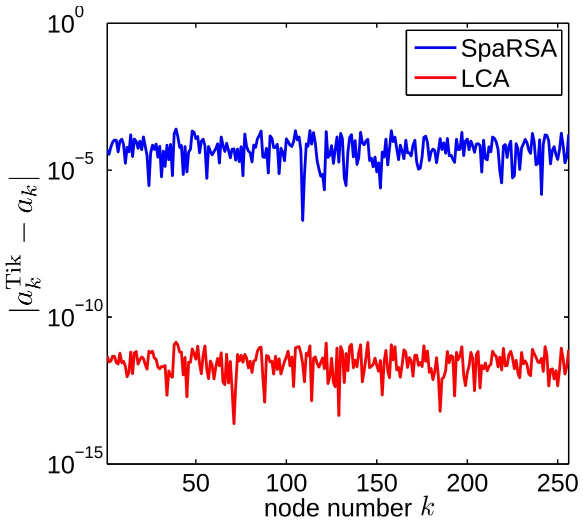


Figure 6: Absolute error of the outputs of the LCA and SpaRSA relative to the Tikhonov solution.

Proof. The objective function $V(\cdot)$ is locally Lipschitz and regular on $\mathbb{R}^N$ as the sum of $\mathbf{C}(\cdot)$, which is locally Lipschitz and regular on $\mathbb{R}^N$ by assumption, and a quadratic form, which is Lipschitz and regular on $\mathbb{R}^N$. Consequently, the rules of calculus for subgradients presented in Section 3.1 imply that

$$\partial V(a(t)) = -\Phi^T y + \Phi^T \Phi a(t) + \partial \mathbf{C}(a(t)). \tag{21}$$

Since $\dot{u}(t)$ satisfies (8), condition (19) yields

$$\begin{aligned} -\dot{u}(t) &= u(t) - a(t) + \Phi^T \Phi a(t) - \Phi^T y \\ &\in \partial \mathbf{C}(a(t)) + \Phi^T \Phi a(t) - \Phi^T y \\ &\in \partial V(a(t)) \end{aligned} \qquad \square$$

The proof of the theorem follows trivially.

Proof of Theorem 1. Any fixed point u^* of (8) satisfies

$$\dot{u}^*(t) = 0.$$

Applying Lemma 1, this equality implies that

$$0 \in \partial V(a^*),$$

where $a^* = T_\lambda(u^*)$. This equation is exactly the condition for a^* to be a critical point of $V(\cdot)$. □

The corollary can be proven easily by explicitly computing the subgradient of a separable cost function in the form of (17).

Proof of Corollary 1. A separable cost function $\mathbf{C}(\cdot)$ in the form of (17) can be rewritten as

$$\mathbf{C}(a) = \sum_{n=1}^{N} \mathbf{C}_n(a),$$

where $\mathbf{C}_n(\cdot) : \mathbb{R}^N \to \mathbb{R}$ is defined by $\mathbf{C}_n(a) = C(a_n)$ for $n = 1, \ldots, N$. Since $\mathbf{C}(\cdot)$ is locally Lipschitz and regular on $\mathbb{R}^N$, equality (13) implies that $\forall a \in \mathbb{R}^N$

$$\partial\mathbf{C}(a) = \sum_{n=1}^{N} \partial\mathbf{C}_n(a).$$

Each $\mathbf{C}_n(\cdot)$ can be viewed as the composition of $C(\cdot)$ and the projection $\Pi_n(\cdot) : \mathbb{R}^N \to \mathbb{R}$ that returns the n^{th} component of a vector. The projection operator $\Pi_n(\cdot)$ is differentiable on $\mathbb{R}^N$ and it is simple to compute its gradient $\nabla\Pi_n(a) = (0, \ldots, 1, \ldots, 0)^T$, where the 1 is at position n. Then the chain rule in (12) yields

$$\begin{aligned}\partial\mathbf{C}_n(a) &= \left\{\xi_n \nabla\Pi_n(a) \;\; s.t. \;\; \xi_n \in \partial C(\Pi_n(a))\right\} \\ &= \left\{(0, \ldots, \xi_n, \ldots, 0)^T \;\; s.t. \;\; \xi_n \in \partial C(a_n)\right\}.\end{aligned}$$

Putting everything together,

$$\partial\mathbf{C}(a) = \left\{\xi \;\; s.t. \;\; \xi = (\xi_1, \ldots, \xi_N)^T \;\; \text{and} \;\; \xi_n \in \partial C(a_n)\right\}. \tag{22}$$

As a consequence, if hypothesis (19) holds, then the subgradient satisfies, for all $a \in \mathbb{R}^N$ and all $n = 1, \ldots, N$,

$$\xi_n = u_n - a_n \in \partial C(a_n).$$

Thus, $u - a = (\xi_1, \ldots, \xi_N) \in \partial\mathbf{C}(a)$, and applying Theorem 1 finishes the proof. □

3.3 Global asymptotic convergence

The result on the fixed points presented in the previous section is general, and the requirements on the activation and cost functions are minimal. The hard-thresholding function, for instance, is not continuous, but there exists a locally Lipschitz cost function that satisfies relationship (19). Despite its large scope, Theorem 1 only guarantees that the fixed points are critical points of the corresponding objective; nothing can yet be said about how the trajectories evolve with time. In this section, a Lyapunov-type approach is taken to prove that the LCA network converges to a set of fixed points. This property is known as quasi-convergence.

3.3.1 Conditions on the activation function

To give the first convergence result for the LCA, the activation function needs to satisfy several requirements. The first natural condition is for the activation function to be nondecreasing everywhere. This property is necessary for the objective function to be nonincreasing almost everywhere as the system evolves and to be a candidate Lyapunov function for the network. A second requirement is for the activation function to be continuous, which ensures that the objective function is also continuous. This requirement prevents scenarios where the objective is decreasing for almost all time but returns to a high value at points of discontinuity and thus never reaches a stable minimum. If, in addition, the activation function is locally Lipschitz, its slope is bounded on bounded intervals and results from nonsmooth analysis apply. The form of the activation function and complete list of necessary conditions are summarized below.

Assumption 1. *The activation function $T_\lambda(\cdot)$ is locally Lipschitz continuous, odd and nondecreasing on $\mathbb{R}$. In addition, there exist $\lambda \geq 0$, and locally finitely many $\{(v_k, w_k, z_k)\}_{k \in \mathcal{K}}$ in*

$\mathbb{R}\times\mathbb{R}\times\mathbb{R}$, with $v_k < w_k$, such that $T_\lambda(\cdot)$ has the form

$$a_n = T_\lambda(u_n) = \begin{cases} 0, & |u_n| \leq \lambda \\ z_k, & u_n \in \bigcup_{k\in\mathcal{K}} [v_k, w_k] := \mathcal{Z} \\ \textit{is strictly increasing otherwise with } \zeta_n > 0,\ \forall \zeta_n \in \partial T_\lambda(u_n) \end{cases} \tag{23}$$

and satisfies

$$|T_\lambda(u_n)| \leq |u_n|\,, \qquad \forall u_n \in \mathbb{R}. \tag{24}$$

Explicitly, the form in (23) *means that $T_\lambda(\cdot)$*

- *is exactly zero on the interval $[-\lambda, \lambda]$,*

- *is constant on a countable and locally finite number of intervals denoted by $\mathcal{Z}$ (which include the interval $[-\lambda, \lambda]$ and potentially the case where w_k is equal to infinity for some k), and*

- *is otherwise strictly increasing on any open interval $\mathcal{U}$ in $\mathbb{R}\backslash\mathcal{Z}$ (where $T_\lambda(\cdot)$ is not constant) with strictly positive subgradients.*

For any $\lambda \geq 0$, it is guaranteed that $T_\lambda(u) > 0$ for all $u > \lambda$. In the case where $\lambda > 0$, the activation function is exactly zero on the nontrivial interval $[-\lambda, \lambda]$. This form is common for activation functions used in sparse recovery problems. Intuitively, many elements with small amplitude are forced to zero, thus promoting a sparse output. The case where $\lambda = 0$ is less interesting for sparse recovery problems, as it does not yield sparse outputs, but it encompasses other types of regularizers such as $C(a_n) = \lambda\,|a_n|^2$ for Tikhonov regularization (whose associated activation function is $T_\lambda(u_n) = u_n/(1 + 2\lambda)$).

If one of the interval limits w_k is equal to $+\infty$, the resulting activation function is constant on an interval of the form $[v_k, +\infty)$ and is obviously bounded. However, the form (23) allows for activation functions that grow unbounded as $u_n \to \infty$.

The last requirement (24) is less intuitive. To understand it, it is necessary to construct a cost function associated with $T_\lambda(\cdot)$ such that the relationship (19) in Theorem 1 is satisfied.

It is possible to build such a cost function $C(\cdot)$ that is continuous, even and nondecreasing on $\mathbb{R}$ (see Lemma 4 in Appendix A). Looking at the form (75) of the cost function constructed in the proof of Lemma 4, it is clear that condition (24) forces $C(\cdot)$ to increase with the absolute value of the coefficients. This property is essential for solving sparse recovery problems as it again encourages zero and small coefficients in the solution. If condition (24) was replaced by the stronger condition $\zeta_n \leq 1$ for all $u_n \in \mathbb{R}$ and $\zeta_n \in \partial T_\lambda(u_n)$ (which simplifies to $T'_\lambda(u_n) \leq 1$ when $T_\lambda(\cdot)$ is differentiable), then the function $C(u_n)/u_n$ would be nonincreasing on $(0, \infty)$. The resulting cost function would satisfy all of the requirements to be a sparseness measure (see Definition 2). However, the weaker condition (24) is similar in nature and sufficient to prove convergence of the LCA network.

The soft-thresholding function in (9) is one of the main focuses in this thesis and satisfies all of the requirements. More generally, activation functions $T_\lambda(\cdot)$ satisfying Assumption 1 correspond to a large class of cost functions that are often used in practice [51]. A generic stylized activation function that satisfies these conditions is shown in Figure 7.

3.3.2 Notation

This section introduces some notations that will be used throughout this thesis. For an activation function of the form (23), the LCA nodes can be split into several sets.

- The *active set* $\Gamma(t)$ contains indices such that

  $$n \in \Gamma(t) \qquad \Leftrightarrow \qquad |u_n(t)| > \lambda \ \text{ and } \ |a_n(t)| > 0.$$

  Indeed, outside of the intervals in $\mathcal{Z}$ where it is constant, the activation function is strictly increasing. As a consequence, state variables that satisfy $|u_n(t)| > \lambda$ generate outputs that satisfy $|a_n(t)| > 0$. These nodes are called *active nodes*. On the contrary, state variables that satisfy $|u_n(t)| \leq \lambda$ generate outputs $a_n(t) = 0$ and are called *inactive nodes*. Their indices belong to the *inactive set* $\Gamma^c(t)$.

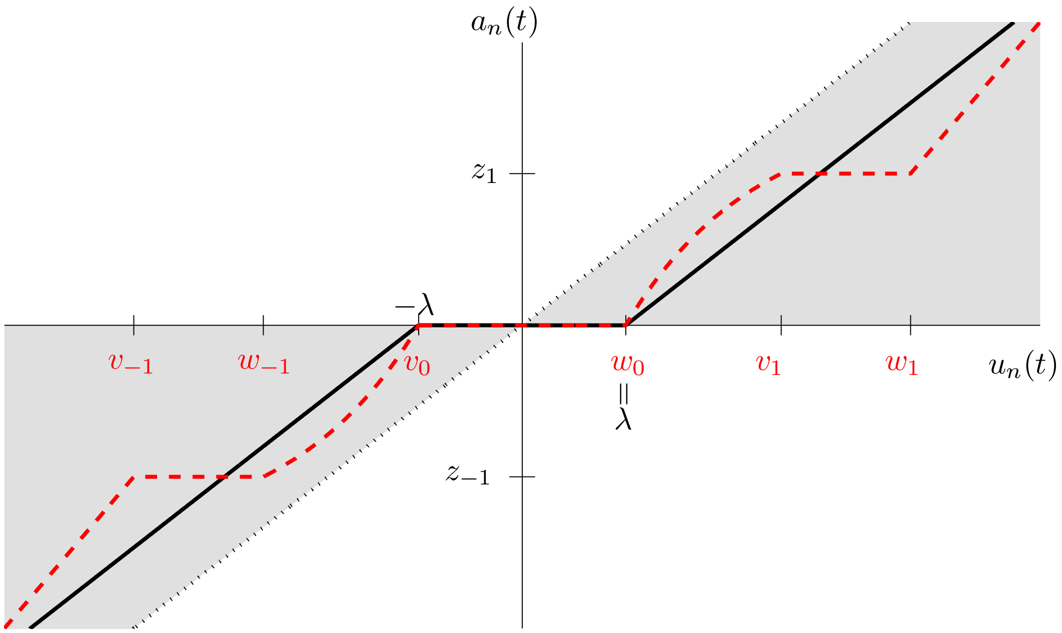


Figure 7: The dashed red curve is a generic activation function satisfying Assumption 1. It has three intervals of the form $[v_k, w_k]$ where it is constant. The area in gray represents where the activation function must lie to satisfy condition (24). The function in black is the soft-thresholding activation function used for ℓ_1-minimization.

- The *constant set* $\mathrm{Z}(t)$ contains indices such that

  $$n \in \mathrm{Z}(t) \qquad \Leftrightarrow \qquad u_n(t) \in \mathcal{Z} = \bigcup_{k \in \mathcal{K}} [v_k, w_k] \ \text{ and } \ a_n(t) = z_k.$$

  For nodes in this set, the output is a constant and $\dot{a}_n(t) = 0$. This set includes the inactive set for which $z_k = 0$, and nodes in this set are referred to as *constant nodes*. On the contrary, for a node n in the complement $\mathrm{Z}^c(t)$, the state variable $u_n(t)$ belongs to an interval $\mathcal{U} \subset \mathbb{R} \backslash \mathcal{Z}$, the output $a_n(t)$ is not constant and every subgradient $\zeta_n \in \partial T_\lambda(u_n(t))$ is strictly positive $\zeta_n > 0$.

- The set $\Delta_q(t)$ contains the q indices with largest magnitude in $u(t)$.

As the system evolves with time according to the ODE in (8), the nodes may switch from one set to its complement and back. As a consequence, the three sets defined above will depend on the specific time t. Nevertheless, their dependence on time is often omitted in the notation for the sake of readability. When it is clear from the context, they will simply be denoted as Γ, Z and Δ.

The notation $\Phi_{\mathcal{S}}$ represents the matrix composed of the columns of Φ indexed by the set $\mathcal{S}$, setting all the other entries to zero. Similarly, $u_{\mathcal{S}}$ and $a_{\mathcal{S}}$ refer to the elements in the vectors u and a indexed by $\mathcal{S}$ setting other entries to zero.

Finally, the sequence $\{t_k\}_{k\in\mathbb{N}}$ of *switching times* is defined such that the set $\Gamma(t) = \Gamma_k$ is constant $\forall t \in [t_k, t_{k+1})$. In other words, a *switch* occurs if a node either leaves or enters the support of the output $a(t)$.

3.3.3 Convergence result

The theorem below summarizes the first convergence result for the LCA network obtained via a Lyapunov approach. It extends the results published in [49], where the activation function may only be constant on the interval $[-\lambda, \lambda]$, while being strictly increasing and differentiable otherwise. Under the more general conditions in Assumption 1, the following convergence result holds.

Theorem 2. *If the LCA system defined by the ODE in* (8) *has an activation function satisfying Assumption 1, then*

1. *the output is globally quasi-convergent in the sense that it converges to the set E of fixed points for any initial state $u(0) \in \mathbb{R}^N$:*

$$a(t) \underset{t\to\infty}{\longrightarrow} E := \left\{a^* \in \mathbb{R}^N \ \ s.t. \ \ \dot{a}^*(t) = 0\right\};$$

2. *if in addition the fixed points are isolated, the output and state variables are globally convergent, i.e., $\forall u(0) \in \mathbb{R}^N$, $\exists!\ a^*, u^* \in \mathbb{R}^N$ such that*

$$a(t) \underset{t\to\infty}{\longrightarrow} a^* \qquad \text{and} \qquad u(t) \underset{t\to\infty}{\longrightarrow} u^*.$$

When the fixed points are isolated, this theorem says that the system converges to a unique fixed point. This is the case for a strictly convex objective, for instance. The proof

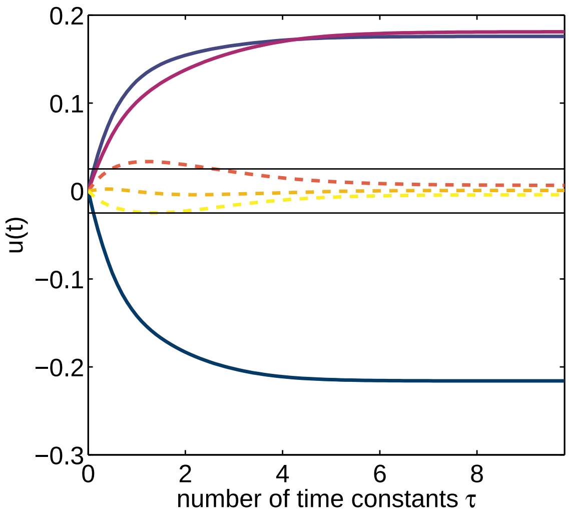


Figure 8: Plot of the evolution over time of several LCA nodes $u_k(t)$. The plain lines correspond to nodes that are active in the final solution and the dashed lines correspond to nodes that are inactive in the final solution.

of the theorem in Section 3.3.5 relies on splitting the ODEs into two sets of differential equations that are partially decoupled.

3.3.4 Simulations

The example of Section 3.2.1 is reused, focusing on the case where the LCA solves the ℓ_1-minimization program (5) with a unique minimum. In this scenario, Theorem 1 guarantees that the LCA has a unique fixed point, and Theorem 2 that it is globally convergent. Figure 8 shows the evolution of a few nodes $u_n(t)$ selected at random from the active and inactive sets of the solution. Both active and inactive nodes converge to their final value in only a few time constants. Figure 9 illustrates the global convergence behavior by showing the evolution over time of several trajectories for different initial points. Two nodes in the support of a^* are chosen at random and their evolution over time is plotted in the state-space defined by those two nodes for 30 random initial points. The color gradient in each curve represents the evolution of time, a lighter gray corresponding to times closer to zero. All resulting trajectories evolve towards a single point in concurrence with Theorem 2.

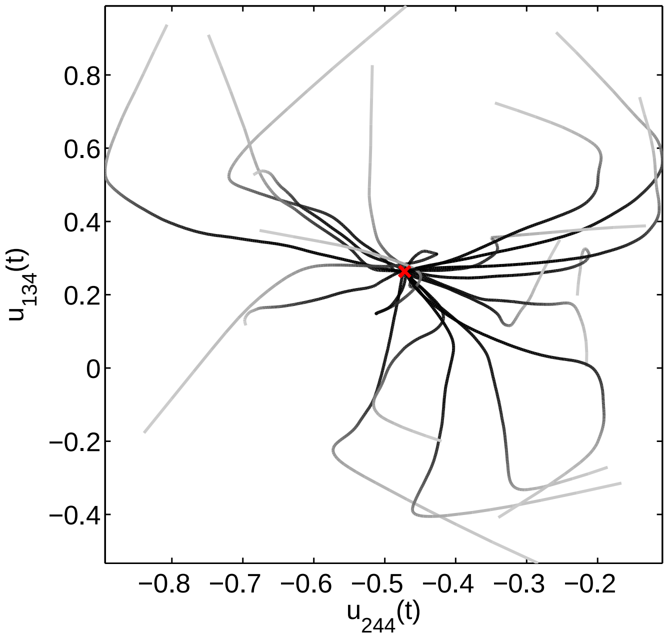


Figure 9: Trajectories $u(t)$ of the LCA for 30 random initial points, projected on the plane defined by two active nodes chosen at random. The color gradient indicates the time evolution. The red cross indicates the fixed point of the network.

3.3.5 Proof of Theorem 2

Proof. First, a cost function $C(\cdot)$ associated to $T_\lambda(\cdot)$ is constructed as in Lemma 4 in Appendix A. By Lemma 6 and Corollary 2, the resulting objective $V(a(\cdot))$ is continuous and regular on $\mathbb{R}^+$ and converges to a constant value V^* as $t \to \infty$. Thus, its time derivative $\dot{V}(a(t))$ tends to zero as $t \to \infty$. Using equation (77) that was derived using the chain rule (14), the following holds for a.a. $t \geq 0$:

$$\dot{V}(a(t)) = -\sum_{n \notin Z} \frac{1}{\zeta_n} |\dot{a}_n(t)|^2$$

for any $\zeta_n \in \partial T_\lambda(u_n)$. Since $\zeta_n > 0$ for $n \notin Z$ and $\dot{a}_n(t) = 0$ for $n \in Z$, the previous observations imply that $\lim\limits_{t \to +\infty} \|\dot{a}(t)\|_2 = 0$. This limit shows that the outputs converge to the set $E = \{a : \text{ s.t. } \dot{a}(t) = 0\}$, which proves that the LCA outputs are quasi-convergent.

Moving on to the second part of the theorem, the fixed points are assumed to be isolated. In this case, the theorem states that both active and inactive nodes converge to a single fixed point. The first part of the proof showed that the outputs converge to the set of fixed points

E. Thus, for any small value $R > 0$, there exists a time t_p and a fixed point $a^* \in E$ such that $a(t_p) \in \mathcal{B}_R(a^*)$, *i.e.*, the output is within a ball of radius R around a^*. Since the fixed points are isolated, there must exist a ball $\mathcal{B}_\varepsilon(a^*)$ of radius $\varepsilon > 0$ around a^* that does not contain any other fixed point:

$$a^* \in E, \qquad \text{and} \qquad \forall a \in \mathcal{B}_\varepsilon(a^*),\ a \neq a^* \Rightarrow a \notin E.$$

Since $\dot{V}(a(t)) \leq 0$ for a.a. $t \geq 0$, Lyapunov's direct method states that the network is stable. As a consequence, by the definition of (Lyapunov) stability in (11), for $\varepsilon/2$ there must exist an R such that, if $a(t_0) \in \mathcal{B}_R(a^*)$, then $a(t) \in \mathcal{B}_{\varepsilon/2}(a^*)$ for all $t \geq t_0$. It was shown earlier that such a time t_0 exists for any $R > 0$. As a consequence, once the trajectory is close enough to one element a^* in E, it must converge to the point a^*, *i.e.*,

$$\lim_{t \to +\infty} a(t) = a^*. \tag{25}$$

Letting

$$u^* = -\Phi^T \Phi a^* + \Phi^T y + a^*,$$

the LCA ODE (8) can be rewritten in terms of the distance $\widetilde{a}(t) = a(t) - a^*$ as

$$\begin{aligned}\dot{u}(t) &= -u(t) - \Phi^T \Phi a^* + \Phi^T y + a^* - \Phi^T \Phi \widetilde{a}(t) + \widetilde{a}(t) \\ &= -u(t) + u^* - \left(\Phi^T \Phi - I\right) \widetilde{a}(t).\end{aligned}$$

Solving this ODE (see Appendix B) yields, for all $t \geq 0$,

$$u(t) = u^* + e^{-t} \left(u(0) - u^*\right) + e^{-t} \int_0^t e^s \left(\Phi^T \Phi - I\right) \widetilde{a}(s) ds.$$

While it is difficult to say anything directly about the trajectory of the system, it is helpful to consider a surrogate trajectory that is a straight line in the state-space: $u^* + e^{-t}\left(u(0) - u^*\right)$. This linear path obviously converges to the fixed point u^*. If the actual trajectory $u(t)$ asymptotically approaches this idealized linear path, then the system is guaranteed to converge to u^*. Taking this approach, the norm of the quantity

$$h(t) = u(t) - u^* - e^{-t} \left(u(0) - u^*\right),$$

which is the deviation from the linear path, can be bounded as follows:

$$\begin{aligned}\|h(t)\|_2 &= \left\|u(t) - u^* - e^{-t}\,(u(0) - u^*)\right\|_2 \\ &= \left\|e^{-t}\int_0^t e^s\left(\Phi^T\Phi - I\right)\widetilde{a}(s)ds\right\|_2 \\ &\leq e^{-t}\left\|\Phi^T\Phi - I\right\|_2 \int_0^t e^s\,\|\widetilde{a}(s)\|_2\,ds\end{aligned}$$

and converges to zero. Since $\widetilde{a}(t) \xrightarrow[t\to+\infty]{} 0$, for any $\widetilde{\varepsilon} > 0$ there exists a time $t_c \geq 0$ such that $\forall t \geq t_c$, $\|\widetilde{a}(t)\|_2 \leq \widetilde{\varepsilon}$. Moreover, since $\|\widetilde{a}(t)\|_2$ is continuous and goes to zero as t goes to infinity, it admits an upper bound $\forall t \geq 0$. Thus, there exists $\mu > 0$ such that $\|\widetilde{a}(t)\|_2 \leq \mu$ for all $t \geq 0$. Thus, for all $t \geq 2t_c$, the integral can be split into two parts to obtain

$$\begin{aligned}\|h(t)\|_2 &\leq e^{-t}\left\|\Phi^T\Phi - I\right\|_2 \mu \int_0^{t_c} e^s ds \;+\; e^{-t}\left\|\Phi^T\Phi - I\right\|_2 \widetilde{\varepsilon}\int_{t_c}^t e^s ds \\ &\leq \left\|\Phi^T\Phi - I\right\|_2 \mu\,[e^{-t+t_c} - e^{-t}] \;+\; \left\|\Phi^T\Phi - I\right\|_2 \widetilde{\varepsilon}\,[1 - e^{-t+t_c}] \\ &\leq \left\|\Phi^T\Phi - I\right\|_2 \mu\left[e^{-t/2} - e^{-t}\right] \;+\; \left\|\Phi^T\Phi - I\right\|_2 \widetilde{\varepsilon}.\end{aligned}$$

The first term in the right-hand side converges to zero as $t \to \infty$, while $\widetilde{\varepsilon}$ can be chosen to be arbitrarily small. Thus the deviation $\|h(t)\|_2$ converges to zero and the trajectory $u(t)$ converges to the trajectory $u^* + e^{-t}\,(u(0) - u^*)$ as t goes to infinity. This result shows that $u(t) \xrightarrow[t\to+\infty]{} u^*$. Therefore, both the output and state variables converge for any initial state, which concludes the proof that the system is globally convergent. □

3.4 Recovery of the support in finite time

As the system evolves, nodes can cross from the active set to the inactive set and *vice versa*. Even if the LCA converges to a unique fixed point u^*, there could be infinitely many such switches. Some mild assumptions on the fixed point guarantee that nodes switch only a finite number of times between the active and inactive sets. Equivalently, this result states that the support Γ_* of the fixed point a^* is recovered in finite time. For this result to hold, the entries of u^* must lie outside of a margin of width $2r$ around the threshold λ. The margin

r must be strictly positive, but can be arbitrarily small. One expects this condition to hold with near certainty for any signal that was not pathologically constructed.

Theorem 3. *If the system* (8) *converges to a fixed point u^* such that there exists $r > 0$ that satisfies*

$$\begin{aligned} \left|u_n^*\right| &\geq \lambda + r, \qquad \forall n \in \Gamma_*, \\ \left|u_n^*\right| &\leq \lambda - r, \qquad \forall n \in \Gamma_*^c, \end{aligned}$$

then the support of a^* is recovered in finite time.

3.4.1 Simulations

To illustrate the number of switches that occur during convergence, the same matrix as in example in Section 3.2.1 is reused. However, this time the sparsity level S of the target is varied from 2 to 50 and the threshold λ from 0.02 to 0.2. For each pair (S, λ), 10 sparse vectors $a^\dagger$ and measurements y are generated as in Section 3.2.1, and the number of switches that occur during convergence is recorded. Figure 10 is a plot of the average number of switches for each pair. Also plotted is the best linear approximation of the minimum threshold value for the number of switches to be less than the sparsity S. For a threshold λ below this line, the system makes more than S switches during convergence. Above the line, the system makes fewer switches. This experiment illustrates that the number of switches is finite and on the order of the sparsity. For small values of the threshold, more nodes are expected to become active, which corresponds to the bottom half of the figure. In addition, this experiment reveals that for a reasonable choice of the threshold (on the order of the noise variance), the number of switches is smaller than the number of active nodes S in the optimal support. This situation can only happen if the nodes in the final solution enter the active set one at a time and never leave the active set. In the optimization literature, this property is referred to as the *S-step property* [23,53], and is characteristic of a solver taking an efficient path toward the solution. Conditions that provide guarantees on the size of the

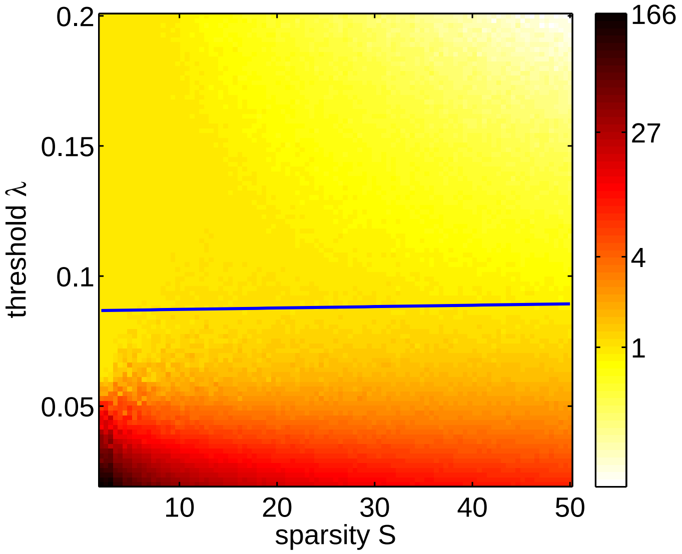


Figure 10: Number of switches during convergence of the LCA network for various values of the sparsity S and threshold λ, averaged over 10 trials for each pair. The blue line represents the best linear approximation to the minimum value of the threshold above which the system makes less than S switches during convergence.

active set during convergence are studied theoretically for the ℓ_1-minimization program in Chapter 4.

3.4.2 Proof of Theorem 3

The proof uses the fact that, if the fixed point does not lie exactly on the transition surface between an active and inactive set for any node, there cannot be more switches after some long enough period of time.

Proof. Let Γ_* be the set of active nodes in u^*. By contradiction, assume that the sequence of switching times $\{t_k\}_{k\in\mathbb{N}}$ is infinite. Since the LCA converges to u^*, then

$$u(t_k) \xrightarrow[k\to+\infty]{} u^*.$$

As a consequence, for $r > 0$, there exists $K \in \mathbb{N}$ such that $\forall k \geq K$, $\|u(t_k) - u^*\|_2 < r$. The following shows that for all $k \geq K$, the state variables $u(t_k)$ are in the subsystem Γ_*. There are two possible cases:

- A node n is active in u^*. In this case, n is also active in $u(t_k)$. Indeed, $\forall n \in \Gamma_*$,

  $$r > \left|u_n(t_k) - u_n^*\right| \geq \left|u_n^*\right| - |u_n(t_k)| \geq \lambda + r - |u_n(t_k)|$$

  $$\Rightarrow \; |u_n(t_k)| > \lambda.$$

  Moreover, nodes n in Γ_* are active with the correct sign in $u(t_k)$, otherwise,

  $$r > \left|u_n(t_k) - u_n^*\right| = |u_n(t_k)| + \left|u_n^*\right| > \lambda + \lambda + r$$

  $$\Rightarrow \; 0 > \lambda,$$

  which is a contradiction.

- A node n is inactive in u^*, in which case it is also inactive in $u(t_k)$. Indeed, $\forall n \in \Gamma_*^c$,

  $$|u_n(t_k)| - \lambda \leq |u_n(t_k)| - \left|u_n^*\right| - r \leq \left|u_n(t_k) - u_n^*\right| - r < 0$$

  $$\Rightarrow \; |u_n(t_k)| < \lambda.$$

As a consequence, $\Gamma_k = \Gamma_*$ for all $k \geq K$. However, Γ_k and Γ_{k+1} must be different to define the switching time t_{k+1}, which yields a contradiction. This contradiction proves that after a finite number of switches K, there cannot be any switching out of the subsystem Γ_*. □

3.5 Finite length of trajectories

While the previous three sections show the potential of the LCA as an efficient solver for optimization problems of the form (6), the results obtained so far are insufficient to prove that the outputs converge to a single point when solutions of (6) are not isolated. This limitation is characteristic of convergence results obtained via a Lyapunov approach. Intuitively, if the solution set is continuously connected, the trajectories could oscillate indefinitely or grow unbounded (if the set of fixed points is unbounded for instance) even though they are getting closer to the set of solutions. Recently, several papers presented in Section 2.4.2 have used a new technique to overcome this problem. Using the Łojasiewicz inequality [47], the authors show that the outputs of certain networks converge to a singleton even when the fixed points are not isolated. Unfortunately, the specifics of the LCA network do not fit any of the previous studies directly. In particular, the LCA activation function is zero on some interval and may be unbounded. In addition, the interconnection matrix W may be singular (see the discussion in Section 2.4.1). To utilize the Łojasiewicz inequality, the following additional requirements on the activation function are necessary.

Assumption 2. *The activation function $T_\lambda(\cdot)$ is subanalytic on $\mathbb{R}$ and for all open and bounded intervals $\mathcal{U} \subset \mathbb{R}\backslash\mathcal{Z}$ where $T_\lambda(\cdot)$ is not constant, there exists a constant $\beta_U > 0$ (that may depend on $\mathcal{U}$) such that*

$$0 < \beta_U \leq \zeta_n, \qquad \forall u_n \in \mathcal{U} \text{ and } \forall \zeta_n \in \partial T_\lambda(u_n). \tag{26}$$

The first condition ensures that the cost $C(\cdot)$ and thus the objective $V(\cdot)$ are subanalytic, which is necessary to apply the Łojasiewicz inequality. This notion was presented in Section 2.4.1 and does not require the function to be continuous. Using results on piecewise analytic functions [47], it is possible to check that if the activation function has the form of (23) and is analytic on the intervals $\mathcal{U} \subset \mathbb{R}\backslash\mathcal{Z}$ where it is not constant, then it is subanalytic. For instance, the soft-thresholding function in (9) is subanalytic. Condition (26) is slightly stronger than the previous condition on the subgradients ($\zeta_n > 0$) in Assumption 1, and requires the existence of a strictly positive lower bound on the subgradients ($\zeta_n > \beta_U$) on bounded intervals where the activation function is not constant. For the soft-thresholding function, this condition holds with $\beta_U = 1$ for all open (even unbounded) intervals $\mathcal{U} \subset \mathbb{R}\backslash\mathcal{Z}$. Because this requirement is only for bounded intervals, it does not prevent the subgradients from tending to zero as $u_n \to \infty$. For instance, an activation function equal to $\sqrt{u_n - \lambda}$ for all $u_n \geq \lambda$ does not satisfy condition (26) on open intervals of the form $(u_0, +\infty)$ since its derivative tends to 0 as $t \to \infty$. However, on any bounded interval of the form $(u_0, u_1) \subset (\lambda, +\infty)$, the derivative admits a strictly positive lower bound (namely $1/2\sqrt{u_1}$), and thus this function satisfies Assumption 2.

The main contribution of this section is to apply a variation of the Łojasiewicz inequality for nonsmooth functions [44] to show two results. First, the output $a(t)$ of the network converges to a single fixed point when starting from any initial point, even if the fixed points are not isolated; *i.e.*, $a(t)$ is *globally asymptotically convergent*.

Theorem 4. *If the activation function $T_\lambda(\cdot)$ satisfies Assumptions 1 and 2, the output $a(t)$ of* (8) *is globally asymptotically convergent; i.e., for all $u(0) \in \mathbb{R}^N$, there exists a unique $a^* \in \mathbb{R}^N$ such that*

$$a(t) \underset{t\to\infty}{\longrightarrow} a^*.$$

Second, the state $u(t)$ also converges to a single fixed point even if the fixed points are not isolated. As a consequence, the LCA network is *globally asymptotically convergent*.

Theorem 5. *If the activation function $T_\lambda(\cdot)$ satisfies Assumptions 1 and 2, the state $u(t)$ of* (8) *is globally asymptotically convergent, i.e., for all $u(0) \in \mathbb{R}^N$, there exists a unique $u^* \in \mathbb{R}^N$ such that*

$$u(t) \underset{t\to\infty}{\longrightarrow} u^*.$$

These two theorems extend the analysis published in [50], where the activation function was not allowed to be constant outside of the interval $[-\lambda, \lambda]$.

With a little more work, it seems possible to extend the results of this section to the case where the activation function is discontinuous. For this case, it is necessary to carefully redefine the cost function (75) associated with $T_\lambda(\cdot)$, which is done in Appendix D. The associated cost function $C(\cdot)$ is still Lipschitz continuous, and so Theorems 4 and 5 still hold in that case.

3.5.1 Simulations

To illustrate the two theoretical results above, an example of an ℓ_1-minimization problem for which there exists a subspace of non-isolated solutions is created. The matrix Φ has dimension $M = 256$ by $N = 512$ and is generated uniformly at random from a Gaussian distribution (then normalized to have columns with unit norm). A sparse vector $a^\dagger$ is generated by selecting uniformly at random the location of 5 non-zero entries. Their amplitudes are generated from a normal distribution and normalized to have norm one. One column of Φ corresponding to one of the 5 non-zero entries in $a^\dagger$ is replaced by a random linear combination of the other 4 columns and re-normalized to 1. The measurements are $y = \Phi a^\dagger + \epsilon$, where ϵ is a Gaussian noise vector with standard deviation $\sigma = 0.01$. The threshold is set to $\lambda = 0.03$. Since the target vector $a^\dagger$ belongs to a set of 5 linearly dependent columns of Φ, there exists an infinite subspace of solutions to the corresponding ℓ_1-minimization problem. The trajectories for 20 random starting points projected onto the space spanned by two randomly selected nodes in the support of $a^\dagger$ are plotted in Figure 11. Despite the solutions being non-isolated and lying on an (unbounded and connected) linear subspace, the system converges and reaches a single fixed point for every starting point in concurrence with the theorem's claims.

3.5.2 Proof of Theorems 4

First, the Łosajiewicz inequality is used on $V(\cdot)$ to show that the output trajectories necessarily converge to a single fixed point a^*. The following proof extends the proof in [50] where $\mathcal{Z}$ was assumed to reduce to $[-\lambda, \lambda]$ and $\mathrm{Z} = \Gamma$. On the contrary, in the following, the activation function has the form of (23) and is allowed to be constant on a locally finite number of intervals in $\mathbb{R}$.

Proof. The cost function $C(\cdot)$ associated with $T_\lambda(\cdot)$ is again constructed as in Lemma 4. Applying Corollary 2, the corresponding objective function $V(a(\cdot))$ converges to a constant $V^* \geq 0$ as $t \to \infty$. In addition, by Lemma 7, $a(t)$ is bounded for all $t \geq 0$. Applying the Bolzano-Weierstrass theorem, there exists a sequence of increasing times $\{t_k\}_{k\in\mathbb{N}}$ such that $\{a(t_k)\}_{k\in\mathbb{N}}$ converges as $k \to \infty$. Let a^* be the limit point of this converging sequence. The following shows that the output $a(t)$ converges to a^* by contradiction. By the continuity of $V(a(\cdot))$ with respect to time, the limit of the sequence $\{V(a(t_k))\}_{k\in\mathbb{N}}$ satisfies $V(a^*) = V^*$.

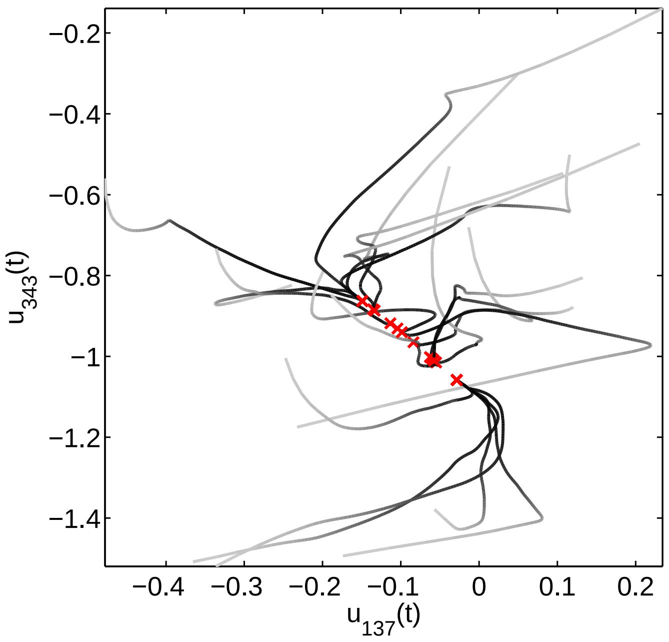


Figure 11: Convergence of LCA trajectories obtained for 20 different initial points and projected onto the space spanned by two randomly chosen non-zero entries in the support of $a^{\dagger}$. The color gradient indicates the time evolution. A red cross indicates the fixed point reached by the system.

Since $u(t)$ is bounded for all $t \geq 0$ by Lemma 7, and since there are only finitely many intervals of the form $[v_k, w_k]$ on any bounded set of $\mathcal{Z}$ by Assumption 1, $u(t)$ visits a finite number L of constant sets Z_l for all $t \geq 0$. For all $l = 1, \ldots, L$, the function $W_l(\cdot)$ is defined by

$$W_l(a_{\mathrm{Z}_l^c}) = V(a), \qquad \forall a \in \mathbb{R}^N.$$

Since $V(\cdot)$ is subanalytic on $\mathbb{R}^N$, $W_l(\cdot)$ is also subanalytic. For all $l = 1, \ldots, L$, applying the nonsmooth Łojasiewicz inequality in Theorem 2.4.1 to $W_l(\cdot)$ at a^*, there exist $\nu_l \in [0, 1)$, $C_l > 0$, and $\Delta_l > 0$ such that

$$|V(a) - V^*|^{\nu_l} = \left| W_l(a_{\mathrm{Z}_l^c}) - W_l(a^*_{\mathrm{Z}_l^c}) \right|^{\nu_l} \leq C_l \; m(\partial W_l(a_{\mathrm{Z}_l^c})), \qquad \forall a \in \mathcal{B}_{\Delta_l}(a^*). \tag{27}$$

Define

$$\begin{aligned} \nu &= \min_{l=1,\ldots,L} \nu_l \quad \in [0, 1)\,, \\ C &= \max_{l=1,\ldots,L} C_l \quad > 0, \\ \Delta &= min_{l=1,\ldots,L} \Delta_l \quad > 0. \end{aligned} \tag{28}$$

Fix a $\delta \in (0, \Delta]$. Since $\{a(t_k)\}_{k \in \mathbb{N}}$ converges to a^*, there exists $K \in \mathbb{N}$ such that

$$\|a(t_k) - a^*\|_2 < \frac{\delta}{4}, \qquad \forall k \geq K. \tag{29}$$

Since $V(a(\cdot))$ is decreasing and converges to V^*, there exist $T_1, T_2 \geq 0$ such that

$$0 \leq V(a(t)) - V^* < 1, \qquad \forall t \geq T_1, \tag{30}$$

and

$$0 \leq V(a(t)) - V^* \leq \left[\frac{\beta\delta\,(1 - \nu)}{4C\alpha} \right]^{\frac{1}{1-\nu}}, \qquad \forall t \geq T_2, \tag{31}$$

where C and ν are defined in (28) and α, β are defined in Corollary 3.

Letting $T = \max(T_1, T_2)$, there exists a time index

$$p = \min\{k \in \mathbb{N} \; s.t. \; k \geq K \text{ and } t_k \geq T\}\,.$$

Time t_p exists, since the sequence of times $\{t_k\}_{k \in \mathbb{N}}$ is increasing and goes to infinity. In addition, t_p is such that it satisfies (29), (30) and (31). Finally, define the following time:

$$t_q = \sup \left\{ t \geq t_p \ s.t. \ \forall s \in \left[t_p, t \right) \ \ \|a(s) - a^*\|_2 < \delta \right\}.$$

If $t_q = +\infty$, then for all time $t \geq t_p$, $\|a(t) - a^*\|_2 \leq \delta$. Since δ can be chosen arbitrarily small, this inequality proves that the output $a(t)$ converges to the single fixed point a^*.

By contradiction, assume that $t_q < +\infty$. This condition implies that for all time $s \in \left[t_p, t_q \right)$, the output trajectory remains within a ball of radius δ around the fixed point, *i.e.*, $\|a(s) - a^*\|_2 < \delta$, but leaves this ball at time t_q, *i.e.*, $\left\| a(t_q) - a^* \right\|_2 = \delta$. According to the inequality involving β in Corollary 3 and using the chain rule (14), the following holds for a.a. $t \geq 0$, for all $n \notin \mathrm{Z}$ and any $\zeta_n \in \partial T_\lambda(u_n(t))$:

$$\|\dot{a}_n(t)\|_2 = \|\zeta_n \dot{u}_n(t)\|_2 \geq \beta \, \|\dot{u}_n(t)\|_2 \, .$$

In addition, to compute $\partial W_l(a_{\mathrm{Z}_l^c})$ more easily, observe that

$$W_l(a_{\mathrm{Z}_l^c}) = V(a) = V(a_{\mathrm{Z}_l^c} + z_{\mathrm{Z}_l}), \qquad \forall a \in \mathbb{R}^N,$$

where z_{Z_l} is the value taken on by the constant outputs a_{Z_l} (see the form of the activation function in (23)). As a consequence, defining the function $H_l(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$ by

$$H_l(a) = a_{\mathrm{Z}_l^c} + z_{\mathrm{Z}_l}, \qquad \forall a \in \mathbb{R}^N,$$

implies that $W_l(a_{\mathrm{Z}_l^c}) = (V \circ H_l)(a)$. Applying the chain rule (12), it is easy to check that

$$\partial W_l(a_{\mathrm{Z}_l^c}) := \partial_{a_{\mathrm{Z}_l^c}} W_l(a_{\mathrm{Z}_l^c}) = \Pi_{\mathrm{Z}_l^c} \partial W_l(a_{\mathrm{Z}_l^c}) = \Pi_{\mathrm{Z}_l^c} \partial V(a),$$

where $\Pi_{\mathrm{Z}^c}(\cdot)$ is the projection onto the set of indices Z^c. Finally, since the LCA satisfies the differential inclusion $-\dot{u}(t) \in \partial V(a(t))$ by Lemma 1, then

$$-\dot{u}_{\mathrm{Z}_l^c}(t) \in \Pi_{\mathrm{Z}_l^c} \partial V(a(t)) = \partial W_l(a_{\mathrm{Z}_l^c}(t)).$$

Putting everything together,

$$\begin{aligned}\|\dot{a}(t)\|_2 &= \|\dot{a}_{Z^c}(t)\|_2 \\ &\geq \beta \, \|\dot{u}_{Z^c}(t)\|_2 \\ &\geq \beta \left\|\dot{u}_{Z_l^c}(t)\right\|_2 \\ &\geq \beta \, m(\partial W_l(a_{Z_l^c})).\end{aligned}$$

for some l between 0 and L.

Furthermore, combining (74) and (77) implies that for a.a. $t \geq 0$ and any $\zeta_n \in \partial T_\lambda(u_n(t))$

$$\dot{V}(a(t)) = -\sum_{n \in Z^c} \frac{1}{\zeta_n} |\dot{a}_n(t)|^2 \leq -\frac{1}{\alpha} \|\dot{a}(t)\|_2^2 \,.$$

This inequality shows that

$$\dot{V}(a(t)) \leq -\frac{1}{\alpha} \|\dot{a}(t)\|_2^2 \leq -\frac{\beta}{\alpha} \|\dot{a}(t)\|_2 \, m\left(\partial W_l(a_{Z_l^c})\right).$$

By definition of t_p and t_q, and since $\delta < \Delta < \Delta_l$, the output satisfies $a(t) \in \mathcal{B}_\delta(a^*) \subset \mathcal{B}_{\Delta_l}(a^*)$ for all $t \in \left(t_p, t_q\right)$, and so (27) yields

$$\begin{aligned}\dot{V}(a(t)) &\leq -\frac{\beta}{\alpha} \|\dot{a}(t)\|_2 \, m\left(\partial W_l(a_{Z_l^c})\right) \\ &\leq -\frac{\beta}{\alpha C_l} \|\dot{a}(t)\|_2 \, (V(a(t)) - V^*)^{\nu_l} \\ &\leq -\frac{\beta}{\alpha C} \|\dot{a}(t)\|_2 \, (V(a(t)) - V^*)^{\nu} \,,\end{aligned}$$

where the last inequality comes from the definition of C and ν in (28) and the fact that $0 < V(a(t)) - V^* < 1$, for all $t \geq t_p$, by (30). Rearranging the terms yields

$$\|\dot{a}(t)\|_2 \leq \frac{\alpha C}{\beta} \frac{-\dot{V}(a(t))}{(V(a(t)) - V^*)^{\nu}}.$$

This result yields a bound on the following integral:

$$\begin{aligned}
\left\|a(t_q) - a(t_p)\right\|_2 &= \left\|\int_{t_p}^{t_q} \dot{a}(s)ds\right\|_2 \\
&\leq \int_{t_p}^{t_q} \|\dot{a}(s)\|_2 \, ds \\
&\leq -\frac{\alpha C}{\beta} \int_{t_p}^{t_q} \frac{\dot{V}(a(s))}{(V(a(s)) - V^*)^\nu} ds \\
&= -\frac{\alpha C}{\beta} \int_{V(a(t_p))}^{V(a(t_q))} \frac{dV}{(V - V^*)^\nu} \\
&= \frac{\alpha C}{\beta(1-\nu)} \left[\left(V(a(t_p)) - V^*\right)^{1-\nu} - \left(V(a(t_q)) - V^*\right)^{1-\nu}\right] \\
&\leq \frac{\alpha C}{\beta(1-\nu)} \left(V(a(t_p)) - V^*\right)^{1-\nu} \\
&\leq \frac{\delta}{4}. && \text{(from (31))}
\end{aligned}$$

Finally, the derivation above shows that

$$\begin{aligned}
\delta = \left\|a(t_q) - a^*\right\|_2 &\leq \left\|a(t_q) - a(t_p)\right\|_2 + \left\|a(t_p) - a^*\right\|_2 \\
&\leq \frac{\delta}{4} + \frac{\delta}{4} \\
&= \frac{\delta}{2} < \delta.
\end{aligned}$$

This is a contradiction, which proves that $t_q = +\infty$. Consequently, it must hold that $\|a(t) - a^*\|_2 \leq \delta$ for all $t \geq t_p$. Since δ can be chosen arbitrarily small, this derivation shows that $\lim_{t \to +\infty} a(t) = a^*$, and thus the output converges. □

3.5.3 Proof of Theorem 5

The following proof shows that the state variables also converge to a single fixed point u^*.

Proof. By Theorem 4, the output converges to some fixed point $a^* \in \mathbb{R}^N$. The dynamical equation (8) can be written in terms of the distance $\widetilde{a}(t) = a(t) - a^*$ of the output to the fixed point:

$$\dot{u}(t) = -u(t) - \Phi^T \Phi a^* + \Phi^T y + a^* - \Phi^T \Phi \widetilde{a}(t) + \widetilde{a}(t).$$

Defining $u^* = -\Phi^T\Phi a^* + \Phi^T y + a^*$ yields the following equation:

$$\dot{u}(t) = -u(t) + u^* - \left(\Phi^T\Phi - I\right)\widetilde{a}(t).$$

The solutions of this differential equation have a known form (see Appendix B) $\forall t \geq 0$:

$$u(t) = u^* + e^{-t}\left(u(0) - u^*\right) + e^{-t}\int_0^t e^s\left(\Phi^T\Phi - I\right)\widetilde{a}(s)ds.$$

The term $e^{-t}\left(u(0) - u^*\right)$ obviously converges to zero as t goes to infinity. To prove that $u(t)$ converges to u^*, it remains to show that the last term in the right-hand side also converges to zero. Denoting this integral term by $h(t)$ and analyzing its norm yields

$$\begin{aligned}
\|h(t)\|_2 &= \left\|e^{-t}\int_0^t e^s\left(\Phi^T\Phi - I\right)\widetilde{a}(s)ds\right\|_2 \\
&= e^{-t}\int_0^t e^s\left\|\left(\Phi^T\Phi - I\right)\widetilde{a}(s)\right\|_2 ds \\
&\leq e^{-t}\sigma_1\int_0^t e^s\|\widetilde{a}(s)\|_2\, ds,
\end{aligned}$$

where $\sigma_1 \geq 0$ is the largest eigenvalue of the interconnection matrix $W = \Phi\Phi^T - I$. To show convergence to zero, the integral is split into two parts. Since $a(t)$ converges to a^*, $\widetilde{a}(t)$ converges to 0 as $t \to +\infty$. Thus, for any $\widetilde{\epsilon} > 0$ there exists a time $t_c \geq 0$ such that $\|\widetilde{a}(t)\|_2 \leq \widetilde{\epsilon}$, $\forall t \geq t_c$. Moreover, since $\|\widetilde{a}(t)\|_2$ is continuous and goes to zero as t goes to infinity, it admits a maximum μ, $\forall t \in \mathbb{R}$. These two bounds yield, for all $t \geq 2t_c$,

$$\begin{aligned}
\|h(t)\|_2 &\leq e^{-t}\left\|\Phi^T\Phi - I\right\|_2\mu\int_0^{t_c} e^s ds \;+\; e^{-t}\left\|\Phi^T\Phi - I\right\|_2\widetilde{\epsilon}\int_{t_c}^t e^s ds \\
&\leq \left\|\Phi^T\Phi - I\right\|_2\mu\left[e^{-t+t_c} - e^{-t}\right] \;+\; \left\|\Phi^T\Phi - I\right\|_2\widetilde{\epsilon}\left[1 - e^{-t+t_c}\right] \\
&\leq \left\|\Phi^T\Phi - I\right\|_2\mu\left[e^{-t/2} - e^{-t}\right] \;+\; \left\|\Phi^T\Phi - I\right\|_2\widetilde{\epsilon}.
\end{aligned}$$

Since the left term converges to 0 and $\widetilde{\epsilon}$ can be chosen to be arbitrarily small, this computation shows that the trajectory $u(t)$ converges to the trajectory $u^* + e^{-t}\left(u(0) - u^*\right)$ as t goes to infinity, and thus $u(t) \xrightarrow[t\to+\infty]{} u^*$. □

3.6 Exponential rate of convergence

Convergence to the correct solution for any starting state is a fundamental property for any system intended to solve an optimization program. Even more interesting for practical applications is knowing how fast the trajectories converge to the solution. In this section, the LCA network is shown to converge *exponentially fast* to a unique fixed point u^* under some condition on the matrix Φ. Furthermore, an analytic bound for the convergence speed is derived.[1] Such a bound is especially important for implementations in real-world applications, which must guarantee solution times. The results obtained in this section extend further those published in [49], where the activation function was assumed to be strictly increasing and differentiable outside of the interval $[-\lambda, \lambda]$.

To state the theorem regarding the convergence speed of the LCA, there must exist a bound on the eigenvalues of the matrix $\Phi^T\Phi$ when it is applied to certain vectors.

Assumption 3. *There exists a constant* $0 < d < 1$ *such that*

$$(1-d)\,\|\widetilde{a}(t)\|_2^2 \leq \|\Phi\widetilde{a}(t)\|_2^2 \leq (1+d)\,\|\widetilde{a}(t)\|_2^2\,, \tag{32}$$

for all $t \geq 0$ *and all output trajectories* $\widetilde{a}(t) = a(t) - a^*$.

The constant d depends on the singular values of the matrix $\Phi_{\widetilde{\Gamma}(t)}$ and, as a consequence, on the sequence of active sets $\{\Gamma_k\}_{k\in\mathbb{N}}$ visited by the system. This constant may not be well defined for every matrix Φ or input y. However, in many interesting cases in CS, the constant d is close to 0 and the dictionary elements are almost orthogonal for any small enough active set [54]. If condition (32) is satisfied, the LCA can be shown to converge exponentially fast to a unique fixed point. The expression for the convergence speed depends on d and on the bound α on the subgradients of the activation function; *i.e.*, α is such that for all $u_n \in \mathbb{R}$ and for all $\zeta_n \in \partial T_\lambda(u_n)$

$$|\zeta_n| \leq \alpha.$$

[1] The time constant τ is reintroduced in this discussion to make its effect on the convergence speed explicit.

The existence of the constant α is guaranteed by Lemma 3. In the following theorem, the two constants d and α are directly related to the convergence speed of the network.

Theorem 6. *Assume that the activation function $T_\lambda(\cdot)$ satisfies Assumption 1 and that the constant d in Assumption 3 is well-defined. If the constants α and d, defined in* (74) *and* (32)*, respectively, satisfy*

$$\alpha d < 1, \tag{33}$$

then the LCA network in (8) *is globally exponentially convergent to a unique fixed point with convergence speed*

$$c = \frac{1 - \alpha d}{\tau}.$$

Explicitly, for all $u(0) \in \mathbb{R}^N$, there exist a unique $u^* \in \mathbb{R}^N$ and a constant $\kappa \geq 0$ (that may depend on $u(0)$) such that

$$\|u(t) - u^*\|_2 \leq \kappa e^{-(1 - \alpha d)t/\tau}, \qquad \forall t \geq 0. \tag{34}$$

Condition (33) is necessary to ensure that the convergence speed is positive and meaningful. The time constant τ of the physical solver implementing the LCA neural network appears in the expression for the speed of convergence. The smaller the time constant τ, the faster the system converges. In general, analog systems have smaller time constants than their digital counterparts and scale better with the problem size [1]. In the case of ℓ_1-minimization, the bound on the subgradients of the soft-thresholding function is $\alpha = 1$, and condition (33) reduces to $d < 1$. Assuming that the active set $\Gamma(t)$ has only a small number of active components for all time $t \geq 0$, Assumption 3 with $d < 1$ corresponds exactly to the RIP condition for Φ. Unfortunately, the sequence of active sets visited by the network is signal dependent and cannot be predicted in advance. Nevertheless, a set of conditions for the active set to remain bounded throughout convergence is the object of the next chapter.

3.6.1 Simulations

To illustrate the result on the convergence rate, the example using ℓ_1-minimization (for which $\alpha = 1$) for sparse recovery in Section 3.2.1 is again used. The evolution of the normalized ℓ_2-distance from the LCA trajectories to the fixed point

$$\frac{\|u(t) - u^*\|_2}{\|u^*\|_2}$$

is recorded. This quantity is equal to 1 for $t = 0$ since the system starts at rest. To demonstrate the validity of the theoretical expression (34) for the decay, it is necessary to estimate what the constant d in (32) is. The constant d must be an upper bound for the eigenvalues of the matrix $\Phi_{\widetilde{\Gamma}}^T \Phi_{\widetilde{\Gamma}}$, where $\widetilde{\Gamma} = \Gamma \cup \Gamma_*$. As a consequence, the largest support $\Gamma_{\max}$ reached by the network during convergence is also recorded. Two estimates for d are tested: d_* corresponds the largest eigenvalue of the matrix $\Phi_{\Gamma_*}^T \Phi_{\Gamma_*} - I$, and $d_{\max}$ corresponds to the largest eigenvalue of the matrix $\Phi_{\Gamma_{\max}}^T \Phi_{\Gamma_{\max}} - I$. Since there may be many more nodes entering the active set during convergence than in the final support, $d_{\max}$ is expected to be larger than d_*, and the corresponding bound on the decay to be less tight. This hypothesis matches what is observed in Figure 12. The two colored dashed lines correspond to the theoretical decay $e^{-(1-d)t/\tau}$ for the two estimated values of d. As expected, the theoretical decay computed with $d_{\max}$ is an upper bound for the convergence speed. However, this estimate seems too conservative, and the bound computed with d_* is a better estimate for the experimental decay. This simulation illustrates that the theoretical exponential convergence appears to capture the essential system behavior.

3.6.2 Proof of Theorem 6

Below is a proof of the exponential convergence of the LCA trajectories.

Proof. The expression of the convergence speed is established via the study of the following energy function:

$$E(t) = \frac{1}{2} \|\widetilde{u}(t)\|_2^2 , \tag{35}$$

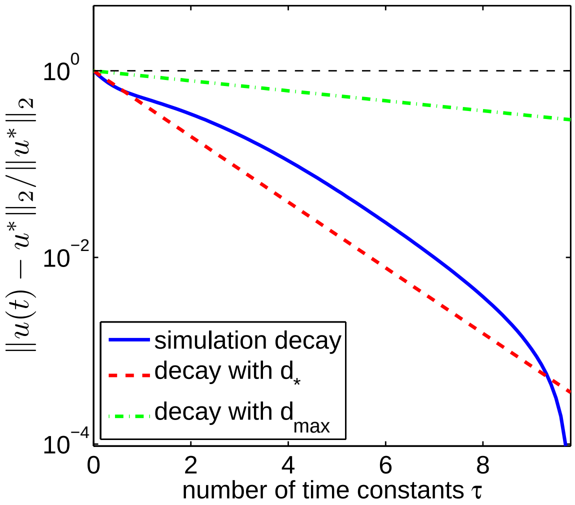


Figure 12: Convergence of the experimental normalized ℓ_2-distance from the state variables to the fixed point. Also plotted is the theoretical decay in (34) for two estimated values of the constant d: d_* is computed by using the final solution support, and $d_{\max}$ is computed on the largest active set visited.

where the variables $\widetilde{u}$ and $\widetilde{a}$ measure the distance of the states and outputs from any arbitrary fixed point u^* and $a^* = T_\lambda(u^*)$ of (8):

$$\begin{aligned} \widetilde{u}_n(t) &= u_n(t) - u_n^*, \\ \widetilde{a}_n(t) &= a_n(t) - a_n^* = T_\lambda(\widetilde{u}_n(t) + u_n^*) - T_\lambda(u_n^*). \end{aligned} \tag{36}$$

The set $\widetilde{\Gamma}$ denotes the support of $\widetilde{a}$ and is equal to $\widetilde{\Gamma} = \Gamma \cup \Gamma_*$, where Γ_* is the support of the fixed point a^*. Like Γ, the set $\widetilde{\Gamma}$ depends on time, but the time index is omitted from the notation to increase readability. To show that the function (35) converges exponentially fast to zero, it is first analyzed on the set of indices $\widetilde{\Gamma}$.

Using the fact that u^* is a fixed point of (8) (*i.e.*, $\dot{u}^*(t) = 0$), rewriting the dynamics in terms of the new variables in (36) yields

$$\tau\dot{\widetilde{u}}(t) = -\widetilde{u}(t) - \left(\Phi^T\Phi - I\right)\widetilde{a}(t). \tag{37}$$

First, the partial energy function $E_{\widetilde{\Gamma}}(t) = \frac{1}{2}\left\|\widetilde{u}_{\widetilde{\Gamma}}(t)\right\|_2^2$ is shown to converge exponentially fast. From this result, the behavior of the outputs can be deduced. Then, the result on the output is used to prove the convergence of the entire state vector to the fixed point u^*.

Using the chain rule, the time derivative of $E_{\widetilde{\Gamma}}(t)$ along the network trajectory is for a.a. $t \geq 0$

$$\begin{aligned}
\tau \dot{E}_{\widetilde{\Gamma}}(t) &= \tau \dot{\widetilde{u}}_{\widetilde{\Gamma}}^{T}(t) \widetilde{u}_{\widetilde{\Gamma}}(t) \\
&= -\widetilde{u}_{\widetilde{\Gamma}}^{T}(t) \left(\widetilde{u}_{\widetilde{\Gamma}}(t) + \left(\Phi_{\widetilde{\Gamma}}^{T} \Phi_{\widetilde{\Gamma}} - I_{\widetilde{\Gamma}} \right) \widetilde{a}_{\widetilde{\Gamma}}(t) \right) \\
&= - \left\| \widetilde{u}_{\widetilde{\Gamma}}(t) \right\|_2^2 - \widetilde{u}_{\widetilde{\Gamma}}^{T}(t) \left(\Phi_{\widetilde{\Gamma}}^{T} \Phi_{\widetilde{\Gamma}} - I_{\widetilde{\Gamma}} \right) \widetilde{a}_{\widetilde{\Gamma}}(t).
\end{aligned}$$

Assumption 3 implies that the eigenvalues of $\Phi_{\widetilde{\Gamma}}^{T} \Phi_{\widetilde{\Gamma}}$ lie between $(1-d)$ and $(1+d)$ and so:

$$\begin{aligned}
\left\| \left(\Phi_{\widetilde{\Gamma}}^{T} \Phi_{\widetilde{\Gamma}} - I_{\widetilde{\Gamma}} \right) \widetilde{a}_{\widetilde{\Gamma}} \right\|_2 &\leq \left\| \Phi_{\widetilde{\Gamma}}^{T} \Phi_{\widetilde{\Gamma}} - I_{\widetilde{\Gamma}} \right\| \left\| \widetilde{a}_{\widetilde{\Gamma}} \right\|_2 \\
&\leq \max \left\{ (1+d) - 1,\ 1 - (1-d) \right\} \left\| \widetilde{a}_{\widetilde{\Gamma}} \right\|_2 \\
&= d \left\| \widetilde{a}_{\widetilde{\Gamma}} \right\|_2 .
\end{aligned}$$

Finally, property (*iii*) of Lemma 5 states that for any set $\mathcal{T}$, $\|\widetilde{a}_{\mathcal{T}}\|_2^2 \leq \alpha^2 \|\widetilde{u}_{\mathcal{T}}\|_2^2$. Using the Cauchy-Schwartz inequality and putting everything together,

$$\begin{aligned}
\left| \widetilde{u}_{\widetilde{\Gamma}}^{T} \left(\Phi_{\widetilde{\Gamma}}^{T} \Phi_{\widetilde{\Gamma}} - I_{\widetilde{\Gamma}} \right) \widetilde{a}_{\widetilde{\Gamma}} \right| &\leq \left\| \widetilde{u}_{\widetilde{\Gamma}} \right\|_2 \left\| \left(\Phi_{\widetilde{\Gamma}}^{T} \Phi_{\widetilde{\Gamma}} - I_{\widetilde{\Gamma}} \right) \widetilde{a}_{\widetilde{\Gamma}} \right\|_2 \\
&\leq \left\| \widetilde{u}_{\widetilde{\Gamma}} \right\|_2 d \left\| \widetilde{a}_{\widetilde{\Gamma}} \right\|_2 \\
&\leq \alpha d \left\| \widetilde{u}_{\widetilde{\Gamma}} \right\|_2^2 .
\end{aligned}$$

As a consequence, the time derivative of the partial energy function satisfies

$$\begin{aligned}
\tau \dot{E}_{\widetilde{\Gamma}}(t) &\leq - \left\| \widetilde{u}_{\widetilde{\Gamma}}(t) \right\|_2^2 + \alpha d \left\| \widetilde{u}_{\widetilde{\Gamma}}(t) \right\|_2^2 \\
&\leq -2 \left(1 - \alpha d \right) E_{\widetilde{\Gamma}}(t).
\end{aligned}$$

Using Gronwall's inequality in Appendix B on the interval $[t_k, t_{k+1}]$ where $\widetilde{\Gamma}$ is constant yields

$$E_{\widetilde{\Gamma}}(t) = \frac{1}{2} \left\| \widetilde{u}_{\widetilde{\Gamma}}(t) \right\|_2^2 \leq \frac{1}{2} \left\| \widetilde{u}_{\widetilde{\Gamma}}(t_k) \right\|_2^2 e^{-2(1-\alpha d)(t-t_k)/\tau}.$$

Since $\|\widetilde{a}(t)\|_2 \leq \alpha \left\| \widetilde{u}_{\widetilde{\Gamma}}(t) \right\|_2$, $\forall t \in [t_k, t_{k+1}]$:

$$\|\widetilde{a}(t)\|_2 \leq \alpha \left\| \widetilde{u}_{\widetilde{\Gamma}}(t_k) \right\|_2 e^{-(1-\alpha d)(t-t_k)/\tau}. \tag{38}$$

Using this result on the output, the state $u(t)$ can now be shown to converge exponentially fast. Using the equality form of Gronwall's Lemma in Appendix B, the solution to (37) can be expressed as follows $\forall t \in [t_k, t_{k+1}]$:

$$\widetilde{u}(t) = e^{-(t-t_k)/\tau}\, \widetilde{u}(t_k) \ + \ e^{-(t-t_k)/\tau} \int_{t_k}^{t} e^{(\nu-t_k)/\tau} \left(I - \Phi^T \Phi\right) \widetilde{a}(\nu) d\nu.$$

Denoting by $h(t)$ the second term in the right-hand side, and plugging in (38), the norm of $h(t)$ can be bounded by

$$\begin{aligned}
\|h(t)\|_2 &\le e^{-(t-t_k)/\tau} \int_{t_k}^{t} e^{(\nu-t_k)/\tau} \left\| \left(\Phi^T \Phi - I\right) \widetilde{a}(\nu) \right\|_2 d\nu \\
&\le e^{-(t-t_k)/\tau} \int_{t_k}^{t} \underbrace{\left\|\Phi^T \Phi - I\right\|_2}_{=C_1} e^{(\nu-t_k)/\tau} \left\|\widetilde{a_{\Gamma}}(\nu)\right\|_2 d\nu \\
&\le e^{-(t-t_k)/\tau} \int_{t_k}^{t} C_1 \alpha \left\|\widetilde{u_{\Gamma}}(t_k)\right\|_2 e^{\alpha d(\nu-t_k)/\tau} d\nu \\
&= \frac{C_1 \tau}{d} \left\|\widetilde{u_{\Gamma}}(t_k)\right\|_2 e^{-(t-t_k)/\tau} \left[e^{\alpha d(t-t_k)/\tau} - 1\right] \\
&\le C_2 \left\|\widetilde{u_{\Gamma}}(t_k)\right\|_2 \ e^{-(1-\alpha d)(t-t_k)/\tau} \\
&\le C_2 \left\|\widetilde{u}(t_k)\right\|_2 \ e^{-(1-\alpha d)(t-t_k)/\tau},
\end{aligned}$$

where $C_2 = \left(\left\|\Phi^T \Phi - I\right\|_2 \tau/d\right)$. Plugging this bound back in the expression for $\widetilde{u}(t)$ yields

$$\begin{aligned}
\|\widetilde{u}(t)\|_2 &= \left\| e^{-(t-t_k)/\tau} \widetilde{u}(t_k) + h(t) \right\|_2 \\
&\le \|\widetilde{u}(t_k)\|_2 \, e^{-(t-t_k)/\tau} + \|h(t)\|_2 \\
&\le \|\widetilde{u}(t_k)\|_2 \, e^{-(t-t_k)/\tau} + C_2 \, \|\widetilde{u}(t_k)\|_2 \, e^{-(1-\alpha d)(t-t_k)/\tau} \\
&\le (1 + C_2) \, \|\widetilde{u}(t_k)\|_2 \, e^{-(1-\alpha d)(t-t_k)/\tau} \\
&= C_3 \, \|\widetilde{u}(t_k)\|_2 \, e^{-(1-\alpha d)(t-t_k)/\tau},
\end{aligned}$$

where $C_3 = 1 + C_2$. Since $\|\widetilde{u}(t)\|_2$ is continuous for all time t, it is easy to show (by induction on t_k) that $\forall t \geq 0$

$$\|\widetilde{u}(t)\|_2 \le e^{-(1-\alpha d)t/\tau} C_3 \, \|\widetilde{u}(0)\|_2 \, . \tag{39}$$

This last inequality shows that the state variable converges exponentially fast to a unique fixed point u^* with convergence speed $(1 - \alpha d)/\tau$. □

CHAPTER IV

PROPERTIES FOR CS RECOVERY

The exponential rate of convergence of the LCA in Theorem 6 is difficult to interpret in general because it requires the existence and an estimate for the constant d in (32). This constant depends on the singular values of submatrices of Φ and on the specific path taken by the network trajectories, which is signal-dependent. In the context of CS recovery, however, the well-known RIP in (2) guarantees the existence of such a constant for any vector that is sufficiently sparse. The objective of this chapter is to make use of the RIP to provide convergence guarantees in the special case where the LCA solves the ℓ_1-minimization problem (5) to recover a sparse signal. In this case, the constant α in (74) is equal to 1, and the constant d can be estimated. The contributions of this chapter are

- two theorems that guarantee that the size of the active set remains bounded throughout convergence using the RIP,
- application of the results to the special case of CS random matrices and comparisons to well-known digital solvers in terms of number of measurements,
- to obtain an estimate for the convergence speed that depends only on the problem parameters and is signal-independent, using known estimates for the RIP constant,
- some intuition on the advantage of using a decreasing threshold when solving ℓ_1-minimization.

Moreover, the effect of the noise vector ϵ appears clearly in the results, making the noiseless setting a special case of this study. The results in this chapter were published in [55].

4.1 Bounding the LCA active set

In this section, two theorems are presented that give guarantees on the size of the active set throughout convergence under different conditions on the problem parameters. The following quantities appear in both theorems and their proofs:

$$b = b_\delta = (1+\delta)(1-\delta)^{-2},$$
$$B_\delta(p) = b\left(\left\|a^\dagger\right\|_2 + \sqrt{1-\delta}\,\|\epsilon\|_2 + \lambda\sqrt{p}\right).$$

When δ is an RIP constant, then $0 < \delta < 1$, and it follows that $b > 1$.

4.1.1 Bounding the active set by the optimal support

The first result contains a set of conditions for the active set Γ to be a subset of the optimal support $\Gamma_\dagger$ throughout convergence. This result ensures that the active set never contains more than the S optimal nodes, and thus is always bounded by S in size.

Theorem 7. *Assume that the dictionary Φ satisfies the RIP with parameters $(S+1,\delta)$ and that the support $\Gamma(0)$ of the initial output $a(0)$ is a subset of the optimal support (i.e., $\Gamma(0) \subset \Gamma_\dagger$). If the following two conditions between the original signal $a^\dagger$, the threshold λ, the noise ϵ, the sparsity S and the RIP constant δ are satisfied:*

$$\left\|a^\dagger - a(0)\right\|_2 \leq B_\delta(S), \tag{40}$$

$$\left(1 - b\delta\sqrt{S}\right)\lambda \geq b\delta\left(\left\|a^\dagger\right\|_2 + \sqrt{1-\delta}\,\|\epsilon\|_2\right) + \left\|\Phi^T_{\Gamma^c_\dagger}\epsilon\right\|_\infty, \tag{41}$$

then nodes in $\Gamma^c_\dagger$ never cross threshold (i.e., $\Gamma(t) \subset \Gamma_\dagger,\ \forall t \geq 0$).

This first result provides guarantees similar to the S-step property in that only the S nodes that belong to the optimal support $\Gamma_\dagger$ become active. In addition, it is shown in Section 4.2 that the requirements on the RIP constant are similar to those for the S-step property to hold for several digital solvers.

4.1.2 Bounding the size of the active set by a constant

Similar to the analysis of some digital solvers discussed in Section 2.1, weaker requirements on the RIP constant still yield interesting convergence results. In this section, the outcome of Theorem 7 is relaxed in that more than the S optimal nodes may become active. The maximum number of active nodes is denoted by q, where q may be larger than S but remains small. In contrast to the analysis for digital solvers, these conditions do not result in a bound on the number of "steps" or iterations to achieve a certain error, but bounding the size of the active set yields an explicit estimate for the exponential convergence speed of the network.

Theorem 8. *Assume that the dictionary* Φ *satisfies the RIP with parameters* $\left(S+q, \bar{\delta}\right)$ *for some* $q \geq 0$. *If the original signal* $a^{\dagger}$*, the initial state* $u(0)$*, the threshold* λ*, the noise* ϵ*, the parameter* q *and the RIP constant* $\bar{\delta}$ *satisfy*

$$\|u(0)\|_2 \leq \lambda\sqrt{q}, \tag{42}$$

$$\lambda \geq \frac{1+\bar{\delta}}{1-3\bar{\delta}}\frac{1}{\sqrt{q}}\left(\left\|a^{\dagger}\right\|_2 + \sqrt{1-\bar{\delta}}\,\|\epsilon\|_2\right), \tag{43}$$

then the active set Γ *never contains more than* q *nodes (i.e.,* $|\Gamma(t)| \leq q, \ \forall t \geq 0$*).*

The simulations in Section 4.5 show that useful values for q are typically small multiples of S. In the next section, the implications of the two theorems on the RIP constant are studied. In concurrence with results for digital solvers presented in Section 2.1, the requirements of Theorem 8 on the RIP constant are weaker than for Theorem 7.

4.1.3 Remarks and consequences on the RIP constant

Conditions (40), (41), (42) and (43) in Theorems 7 and 8 involve complex relationships between the various problem parameters. Below are a few observations and an analysis of their implication on the RIP constant.

First, condition (40) of Theorem 7 constrains the starting point to be reasonably close to the optimum $a^{\dagger}$. When the system starts at rest, $u(0) = 0$ and condition (40) becomes

$$\left\|a^{\dagger}\right\|_2 \leq b\left(\left\|a^{\dagger}\right\|_2 + \sqrt{1-\delta}\,\|\epsilon\|_2 + \lambda\sqrt{S}\right),$$

which always holds since $b \geq 1$. Similarly, if the system starts at rest, condition (42) obviously holds. Thus, no *a priori* information on the signal is necessary for the two theorems to apply.

To analyze the requirements of the theorems on the RIP more easily, the target signal is assumed to have unit norm (*i.e.*, $\left\|a^\dagger\right\|_2 = 1$) without loss of generality. For now, it is also assumed that there is no noise (*i.e.*, $\epsilon = 0$); the noise level is addressed in Section 4.2. It is also instructive to look at the scenario where all of the non-zero entries in $a^\dagger$ have the same magnitude. Indeed, from (85), the solution a^* is a thresholded version of $a^\dagger$:

$$a^*_{\Gamma_*} = a^\dagger_{\Gamma_*} - \lambda \left(\Phi^T_{\Gamma_*}\Phi_{\Gamma_*}\right)^{-1} z_{\Gamma_*}.$$

If some nodes in $\Gamma_\dagger$ have small amplitudes, they do not contribute much to the signal energy and setting them to zero in a^* may be acceptable. When the nodes in $a^\dagger$ have the same magnitude, however, they contribute equally to the target signal's energy and it is important to recover them all. If the threshold is too large, the outputs of the LCA simply remain zero. When $\left\|a^\dagger\right\|_2 = 1$, each non-zero element of $a^\dagger$ is equal to $\pm 1/\sqrt{S}$, thus the threshold must be smaller than $1/\sqrt{S}$. On the other hand, conditions (41) and (43) of the two theorems require the threshold to be sufficiently large. Taking $\lambda = r/\sqrt{S}$, for some $0 < r < 1$, and rearranging the terms in (41) in Theorem 7 yields the following condition on the RIP constant:

$$\delta \leq \frac{r}{(1+r)\, b\sqrt{S}}. \tag{44}$$

Consequently, for the active set to remain a subset of the optimal support, the RIP constant needs to scale with $1/\sqrt{S}$.

Since q is typically a small multiple of S, taking $q = \beta S$ in Theorem 8 for a small constant β, inequality (43) becomes

$$\bar{\delta} \leq \frac{r\sqrt{\beta} - 1}{3r\sqrt{\beta} + 1}. \tag{45}$$

Thus, for the active set to contain less than q nodes, the RIP constant needs only to be bounded by a small constant that does not depend on S anymore, which is more favorable than condition (44) as will be discussed in the next section.

4.2 Application to Compressed Sensing matrices

Theorems 7 and 8 are deterministic. However, when the matrix Φ is a classic CS random matrix, the results can be interpreted using a known estimate for the RIP constant. For instance, assuming Φ is a subgaussian random matrix as in Section 1.1.3, a good estimate for the RIP constant is given by (3):

$$\delta \sim \sqrt{\frac{S\log(N/S)}{M}}.$$

4.2.1 Theorem 7 with CS matrices

The implications of Theorem 7 on the problem parameters, specifically the number of measurements and noise level, are examined first.

4.2.1.1 Measurements

Plugging the estimate (3) for δ in (44) yields

$$\sqrt{M} \gtrsim S\sqrt{\log(N/S)}\,\frac{(1+r)b}{r},$$

where the notation $\gtrsim$ means greater up to a constant factor. When $S \ll M$, δ is small and $b \sim 1$, so the term $(1+r)b/r$ is a small constant. As a reference:

- if $r = 0.95$ and $\delta \leq 0.5$, then $b \leq 6$ and $\dfrac{(1+r)b}{r} \leq 13$,
- if $r = 0.95$ and $\delta \leq 0.1$, then $b \leq 1.358$ and $\dfrac{(1+r)b}{r} \leq 3$.

This estimate shows that the number M of measurements for a subgaussian random matrix Φ must be on the order of $S^2 \log(N/S)$.

This result strongly resembles the condition for the Homotopy algorithm to satisfy the S-step property [23], which requires that $S \leq \left(1+\mu^{-1}\right)/2$, where μ is the mutual coherence [56] and leads to the same number of measurements. For $M \sim O\left(S^2\log(N/S)\right)$, the Homotopy algorithm on the parameter λ behaves like a pursuit algorithm, where nodes are added to or removed from the active set and the solution evolves in a piecewise-linear manner. Likewise, the LCA solution evolves according to a continuous switched linear system of ODE and nodes enter or leave the active set until the solution is reached. Both results ensure that only nodes present in the optimal support enter the active set. The OMP solver, which is a greedy algorithm, was also shown to recover an S-sparse signal in exactly S steps provided that Φ satisfies the RIP with $\delta_{S+1} \leq 1/\left(3\sqrt{S}\right)$. This bound is similar to that of (44) obtained for the LCA. Likewise, this result for OMP leads to $O\left(S^2 \log(N/S)\right)$ measurements [26]. Consequently, despite the continuous-time nature of the LCA trajectories, bounds on the RIP constant comparable to those obtained for the analysis of digital solvers emerge from this study.

4.2.1.2 Noise level

When ϵ is a Gaussian white noise vector whose entries have variance σ^2, the terms due to the noise in (41) become $\|\epsilon\|_2 \sim \sqrt{M}\sigma$ and $\left\|\Phi^T\epsilon\right\|_\infty \sim \sqrt{\log N}\sigma$ with high probability. Taking these terms into account in (41) does not change the bound on δ in (44) by more than a constant if the following is true:

$$b\delta\sqrt{1-\delta}\,\|\epsilon\|_2 + \left\|\Phi^T_{\Gamma^c_\dagger}\epsilon\right\|_2 = \kappa b\delta$$

for some constant $\kappa > 0$. Using the estimate (3), along with $S \ll N$, $M \sim S^2\log(N/S)$, $b \sim 1$, $b\sqrt{1-\delta} \sim 1$, and reorganizing the terms yield a noise variance of

$$\begin{aligned}
\sigma &\sim \frac{b\delta\kappa}{b\delta\sqrt{1-\delta}\sqrt{M} + \sqrt{\log N}} \\
&\sim \frac{\kappa\sqrt{\frac{S\log(N/S)}{M}}}{\sqrt{S\log(N/S)} + \sqrt{\log N}} \\
&\sim \frac{\kappa}{1 + \sqrt{\frac{\log N}{S\log(N/S)}}}\frac{1}{\sqrt{M}} \\
&\sim \frac{\kappa}{1 + \frac{1}{\sqrt{S}}}\frac{1}{\sqrt{M}}.
\end{aligned}$$

Thus, the total energy allowed in the noise vector is on the order of $\|\epsilon\|_2 \sim \left(1 + 1/\sqrt{S}\right)^{-1}$, which is approximately on the same order as the energy of the signal.

This result can be improved upon. Theorem 7 is stated for any fixed noise vector ϵ. In the case where the noise ϵ is assumed to be a Gaussian random vector, the proof of

Lemma 11 in Appendix B hints that the bound used for $\left\|a^\infty - a^\dagger\right\|_2$ can be improved. An essential step in the proof is to bound $\left\|(\Phi_\Gamma^T\Phi_\Gamma)^{-1}\Phi_\Gamma^T\epsilon\right\|_2$. It is a simple calculation to show that

$$\mathbb{E}\left\{\left\|(\Phi_\Gamma^T\Phi_\Gamma)^{-1}\Phi_\Gamma^T\epsilon\right\|_2^2\right\} = \sigma^2 \operatorname{Trace}\left((\Phi_\Gamma^T\Phi_\Gamma)^{-1}\right) \leq \frac{S\sigma^2}{1-\delta}.$$

Moreover, standard tail inequalities [5] show that this random variable concentrates around its mean. Thus, when the noise is Gaussian, $\sqrt{1-\delta}\,\|\epsilon\|_2$ can be replaced by $\sqrt{S}\sigma$ with high probability in (41). From the equations in the previous paragraph, the noise variance has the form

$$\sigma \sim \frac{b\delta\kappa}{b\delta\sqrt{S} + \sqrt{\log N}} \sim \frac{\kappa}{1+\sqrt{\log N}}\frac{1}{\sqrt{S}}.$$

The total energy allowed in the noise vector becomes $\|\epsilon\|_2 \sim \sqrt{M/S}\Big/\left(1+\sqrt{\log N}\right)$, which increases with the number of measurements M.

4.2.2 Theorem 8 with CS matrices

The implications of the second theorem on the problem parameters are now studied. The following shows that the number of necessary measurements is smaller than for Theorem 7 and again matches conclusions drawn for digital solvers.

4.2.2.1 Measurements

For subgaussian random matrices, using the estimate (3) for the RIP constant $\bar{\delta}$ of order $S+q = (1+\beta)S$ in (45) yields

$$\sqrt{M} \gtrsim \sqrt{(1+\beta)S\log\left(\frac{N}{(1+\beta)S}\right)}\,\frac{3r\sqrt{\beta}+1}{r\sqrt{\beta}-1}.$$

If β is a small constant, the number of measurements is on the order of $O(S\log(N/S))$. For reference, if $\beta = 30$ and $r = 0.95$, then (45) yields $\delta_{31S} \leq 0.25$. In comparison, OMP has been shown to converge for $\delta_{31S} \leq 1/3$ [25]. The result obtained for ROMP in [57] has a slightly worse form since it depends on the sparsity S with $\delta_{8S} \leq 0.01/\sqrt{\log S}$. Finally, CoSaMP was shown to converge for $\delta_{4S} \leq 0.1$ in [28]. For all of these algorithms, the reported RIP constants lead to the same order of measurements $O(S\log(N/S))$. This observation brings to light another interesting parallel between the LCA and its digital equivalents. Letting more than the S optimal nodes enter the active set still yields good convergence results, while giving better scaling on the RIP constant and number of measurements. Contrary to the digital solvers, however, the conditions are only necessary to guarantee a bound on the exponential speed of convergence of the LCA, and not to prove convergence. Theorems 2 and 3 guarantee that the LCA converges to the solution of (5) without any requirement on the RIP constant. In addition, the error achieved by the LCA is linked to the performance guarantees associated with ℓ_1-minimization, as discussed in Section 1.1.5.

4.2.2.2 Noise level

The influence of the noise appears clearly in the results. The noise vector in (43) does not affect the bound on $\bar{\delta}$ in (45) by more than a constant if

$$\sqrt{1-\bar{\delta}}\,\|\epsilon\|_2 = \bar{\kappa}$$

for some $\bar{\kappa} > 0$. Assuming again that ϵ is a Gaussian white noise vector, whose entries have variance σ^2, and that $\left\|a^\dagger\right\|_2 = 1$ yields a noise variance of

$$\sigma \sim \frac{\bar{\kappa}}{\sqrt{1-\delta}}\frac{1}{\sqrt{M}} \sim \frac{1}{\sqrt{M}}.$$

As a consequence, the total energy $\|\epsilon\|_2$ allowed in the noise vector is $O(1)$, which is the same order as the energy of the signal. Here again, assuming that the noise is Gaussian in the proof of the theorem itself leads to a sharper bound. Using the same concentration argument as before, the term $\sqrt{1-\delta}\,\|\epsilon\|_2$ can be replaced by $\sqrt{q}\sigma$ with high probability in (43). This analysis yields a new noise variance of the form $\sigma \sim \kappa/\sqrt{q}$ and the energy in the noise vector becomes $\|\epsilon\|_2 \sim \sqrt{M/q}$. This result again shows that the noise variance can increase as the number of measurements increases without changing the condition on the RIP constant too much.

4.3 Decreasing threshold

An interesting observation emerges from the analysis in this chapter. The proofs of Theorems 7 and 8 in Section 4.6 hint that the results possibly still hold when the threshold is not constant but instead decreases at an exponential rate. In the proof of Theorem 7, the lower bound on the threshold λ depends on the quantity $\left\|a(t) - a^{\dagger}\right\|_2$, while it depends on $\|u(t) - u^{*}\|_2$ in the proof of Theorem 8. If the network is exponentially convergent, both quantities should decrease exponentially fast over time. Thus, the threshold λ could be decreased according to an exponential decay while still satisfying the inequalities in the two theorems. Decreasing the threshold would allow the system to potentially recover more nodes from $a^{\dagger}$ while keeping the size of the active set bounded and yielding faster convergence. This hypothesis is confirmed in simulation (see Section 4.5). Interestingly, similar observations have been made for digital solvers (e.g. in [58], the threshold is decreased according to a geometric progression to speed up recovery). However, there has been no analytic justification for the observed increase in speed or for how to choose the decay rate. While our analysis suggests the potential advantage of decreasing the threshold at an exponential rate, the additional dynamics on the threshold would drastically change the nature of the analysis, starting with the proof of convergence in Chapter 3, where the threshold is considered to be constant throughout. This open problem is potentially an object of future research.

4.4 Estimate of the convergence speed

In Theorem 7, the active set visited during convergence was shown to never contain more than the S optimal nodes under some strong condition on the RIP. This result was generalized in Theorem 8 to allowing no more than q nodes to become active, where q is typically a small multiple of S. With such guarantees, Assumption 3 closely resembles the RIP. Indeed, if the theorems hold, the active set never contains more that q nodes and the matrix Φ

satisfies the RIP with parameters $(\delta, S+q)$, then the constant d exists and can be approximated by the RIP constant δ. In practice, it is reasonable to expect that S of the q nodes that do become active belong to the optimal support $\Gamma_{\dagger}$. As a consequence, the RIP of Φ can be relaxed to only hold with parameters (δ, q) in practice. For instance, when Φ is a subgaussian random matrix, d may be approximated by the classic estimate $\sqrt{S\log(N/S)/M}$ as in (3). The convergence in (34) being exponential (specifically the ℓ_2-distance to the fixed point u^* is bounded by $e^{-(1-d)t/\tau}$), it is clear that a small multiple of $\tau/(1-d)$ will make the ℓ_2-distance infinitesimally close to 0. As a consequence, the convergence time of the LCA is on the order of

$$O\left(\frac{\tau}{1-\sqrt{S\log(N/S)/M}}\right),$$

where τ is the time constant of the physical solver implementing the ODE.

For comparison, the digital solvers Homotopy, OMP, ROMP and CoSaMP have been proven to have running times on the order of $O(SMN)$ floating point operations (flops) when the number of iterations is finite [23, 25, 28, 57]. This estimate can typically be reduced if a fast multiply for Φ and Φ^T is available. It is important to keep in mind that the time constant τ for the LCA has the potential to be much smaller than the time for a digital solver to perform a single matrix multiply [59]. As a consequence, the scaling properties of the LCA seem more favorable for large problems than those of digital algorithms.

4.5 Simulations

The simulations in this section illustrate the previous theoretical findings[1]. As an example, a sparse vector $a^{\dagger}$ of length $N = 400$ with S non-zero entries is generated by selecting S indices uniformly at random, drawing S amplitudes from a normal distribution and normalizing them so that $\left\|a^{\dagger}\right\|_2 = 1$. A number $M = 200$ of measurements are generated via a Gaussian random matrix Φ of size 200 × 400, with entries drawn independently from a normal distribution and columns normalized to have unit norm. A Gaussian white noise

[1] Matlab code for running the experiments in this section can be downloaded from `http://users.ece.gatech.edu/~abalavoine3/code/LCA_CS_exp.zip`

vector with standard deviation $\sigma = 0.025$ is added to the measurements so that $y = \Phi a^{\dagger} + \epsilon$. The LCA is always started at rest with $u(0) = 0$.

4.5.1 Effect of the threshold on the size of the active set

First, the effect of the threshold λ on the size of the active set during convergence is explored. Figures 13 and 14 illustrate the theoretical findings of Theorems 7 and 8 respectively. For each pixel on the figures, 100 random draws of a sparse vector $a^{\dagger}$ and a measurement matrix Φ are simulated, assuming that no noise is present.

In Figure 13, the percentage of the 100 trials where only nodes that are part of the optimal support $\Gamma_{\dagger}$ become active is shown. For large S (approximately $S > 28$), the transition phase for λ follows a curve that resembles $1/\sqrt{S}$. For small S, the behavior appears qualitatively different. Both follow the general prediction from (41):

$$\lambda \gtrsim \frac{b\delta}{1 - b\delta\sqrt{S}}.$$

Above this value, in the white region, only nodes in the optimal support $\Gamma_{\dagger}$ become active for all 100 trials. In the black region, one or more nodes outside the optimal support became active for all 100 trials. The transition between the two regions seems to sharpen as S increases.

In Figure 14, the color coding represents the ratio of the maximum number of active elements q during convergence over the sparsity level S. The phase transition on this plot follows a $1/\sqrt{S}$ behavior, as expected from (43). For most of the pairs (λ, S), the maximum number of active nodes q is contained between $1S$ and $10S$, illustrating that q is typically a small multiple of S and that the active set remains bounded. When the threshold is too high, no nodes become active. The results shown in Figure 13 and Figure 14 thus confirm the qualitative behavior of the bounds derived in Theorems 7 and 8.

4.5.2 Decreasing the threshold during convergence

As mentioned in Section 4.3, the proofs of Theorems 7 and 8 suggest that the active set remains bounded even when the threshold is decreased at an exponential rate, while yielding

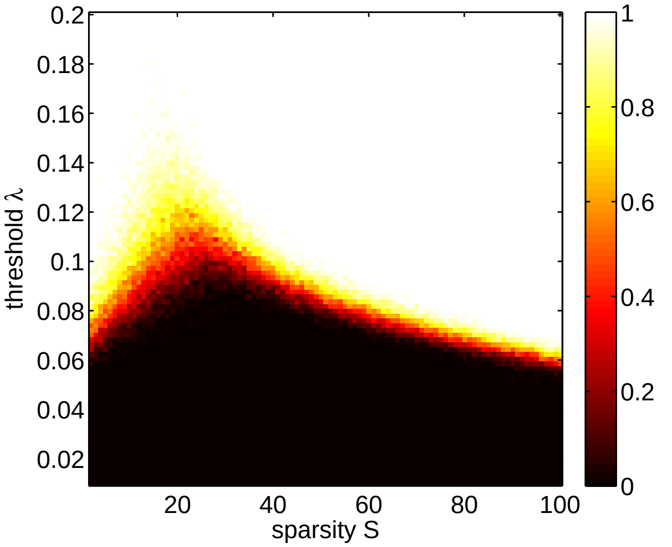


Figure 13: Percentage of the trials where no more than the S nodes from the optimal support $\Gamma_\dagger$ become active during convergence, *i.e.*, $\Gamma(t) \subset \Gamma_\dagger$, $\forall t \geq 0$. The value 1 means that 100% of the trials satisfied this condition.

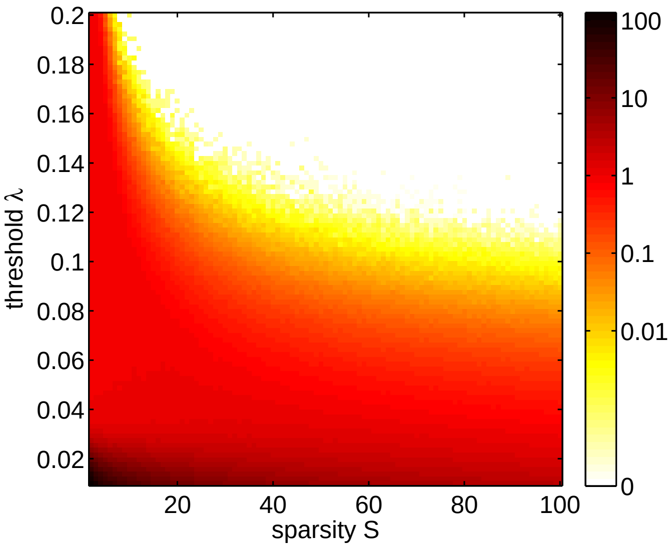


Figure 14: Ratio of the maximum number of active elements q during convergence over the sparsity level S, *i.e.*, $\max_{t\geq 0} |\Gamma(t)|/S$. For instance, a value of 10 in the color bar means that the largest active set during convergence contains $10S$ active elements.

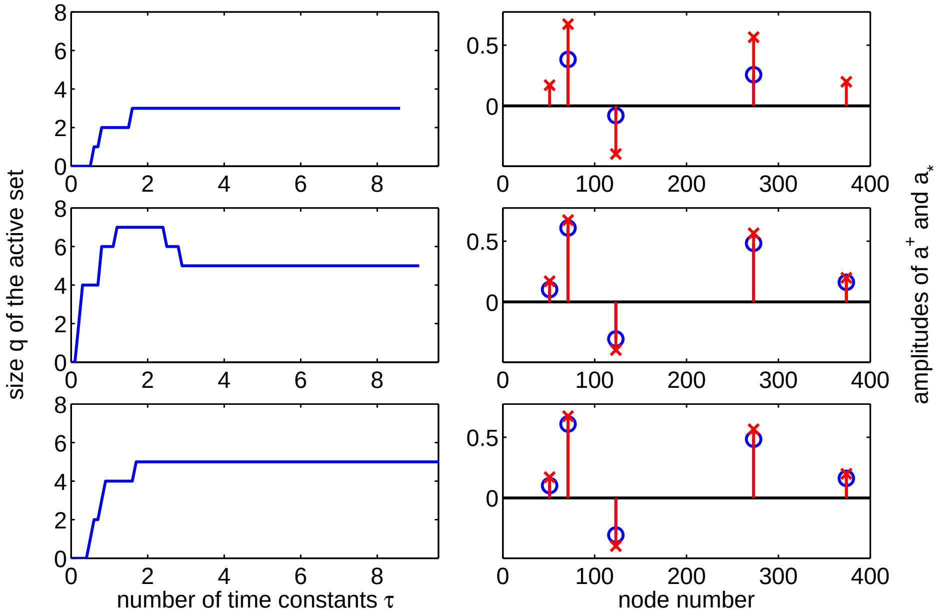
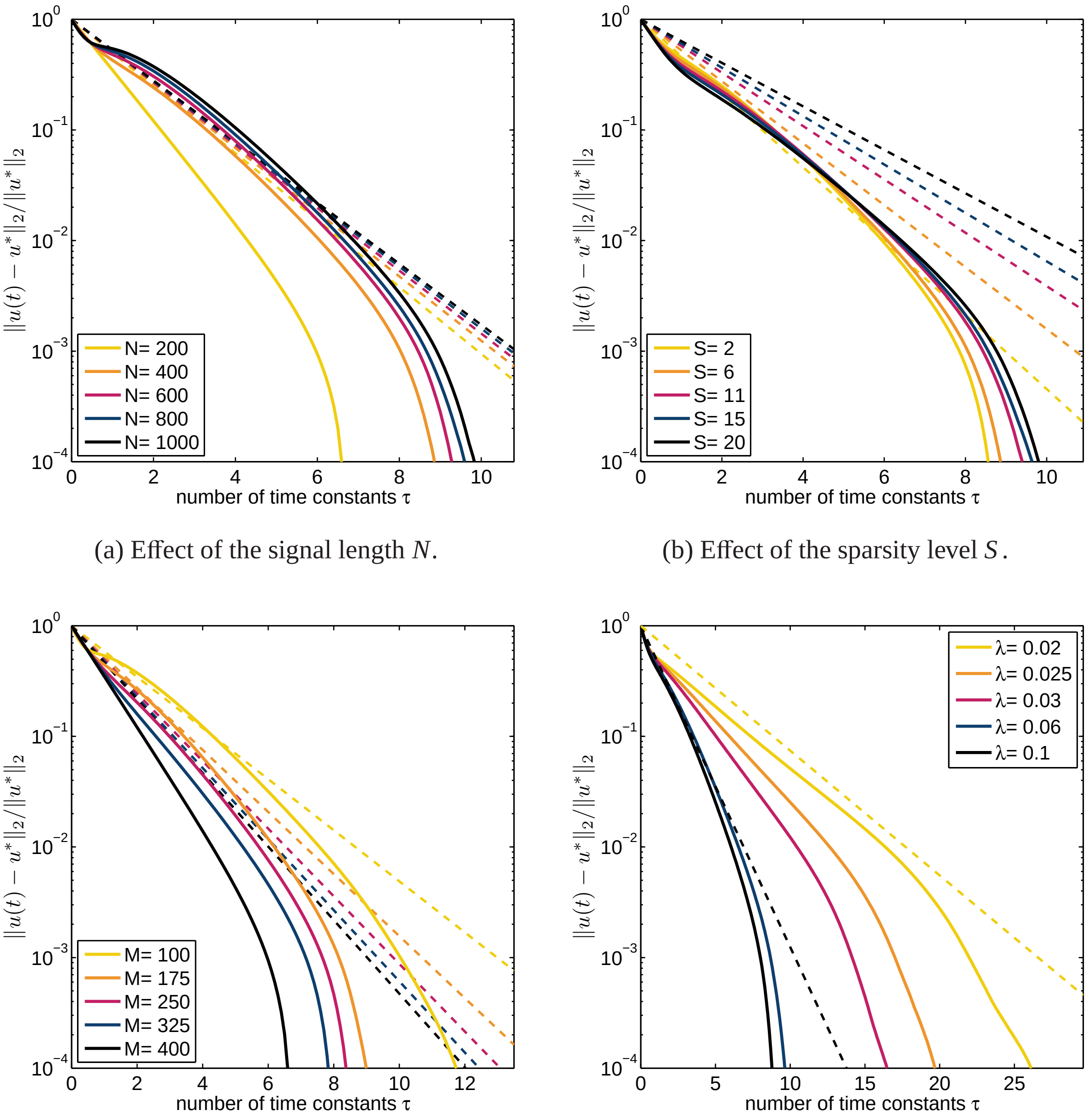


Figure 15: Number of active nodes (left column) and fixed point a_* reached by the LCA (right column) for different choices of the threshold. The red crosses represent the original signal $a^\dagger$ and the blue rounds represent the solutions a^*. A fixed threshold $\lambda = 0.3$ was used in the first row, $\lambda = 0.08$ in the second row, and the threshold was decreased from 0.3 to 0.08 according to an exponential decay in the third row.

faster convergence. To illustrate that this fact is confirmed in practice, the LCA is first run with a high threshold value of $\lambda = 0.3$. As shown in the first row of Figure 15, in this case the active set never contains more than three nodes that are part of the optimal support, but the final solution is missing two nodes from the target signal $a^\dagger$. In the second row, λ is fixed to a low value of 0.08. The final solution recovers all the nodes from $a^\dagger$. However, the largest active set visited during convergence now contains $q = 7$ nodes and the convergence is slower. Finally, in the last row, the threshold is started at 0.3 and decreased to the value 0.08 according to an exponential decay. As expected, the final solution is the same as in row 2. However, in this case the active set never contains more than the five nodes from the optimal support. Moreover, the support is recovered faster, in less than 2τ compared to 3τ in row 2.



(a) Effect of the signal length N.

(b) Effect of the sparsity level S.

(c) Effect of the number of measurements M.

(d) Effect of the threshold λ.

Figure 16: Evolution of the experimental mean-squared error $\|u(t) - u^*\|_2$ (plain line) and theoretical decay (dashed line) as the problem parameters are varied.

4.5.3 Estimate of the convergence speed

Finally, the experimental results in this section are used to test that the bound on the normalized ℓ_2-distance predicted by Theorem 6

$$\frac{\|u(t) - u^*\|_2}{\|u^*\|_2} \leq e^{-(1-d)t/\tau} \tag{46}$$

holds in practice. Both sides of the above expression are equal to 1 at $t = 0$. Since the matrix Φ is random Gaussian, it was shown in Section 4.4 that the constant d can be approximated by the RIP constant $\delta \sim \sqrt{S\log(N/S)/M}$. In Figure 16, the theoretical decay on the right-hand side of (46) (dashed lines) is plotted, along with the experimental normalized ℓ_2-distance on the left-hand side (solid lines) averaged over 100 trials. When they are not varying, the threshold is fixed to $\lambda = 0.1$, the number of measurements to $M = 200$, the sparsity to $S = 5$, and the signal length to $N = 400$.

As expected, the theoretical curves approximate the experimental decay. These upper bounds are not strict in practice since they rely on an estimate for the RIP constant δ, which cannot be exactly determined. However, these curves illustrate that the experimental curves qualitatively follow the theoretical predictions as the parameters N, M or S are varied in Figure 16a, 16c, and 16b, respectively.

In Figure 16d, the effect of the threshold λ on the experimental decay is explored. For values of λ larger than 0.06, the bound (46) with $d = \sqrt{S\log(N/S)/M}$ (dark blue dashed line) is valid even though more than S nodes may become active. Indeed, according to Figure 14, for $\lambda = 0.06$, the maximum size of the active set averaged over 100 trials is $q = 23 = 4.6S$, which is larger than S. As λ becomes smaller, more nodes are likely to enter the active set. To reflect this, the theoretical decay on the right-hand side of (46) is plotted again with $d = \sqrt{5S\log(N/S)/M}$ (yellow dashed line). The resulting curve is an upper bound even for very small values of the threshold, for which much more than $5S$ nodes become active during convergence. For instance, the maximum size of the active set averaged over 100 trials for $\lambda = 0.02$ is $180 = 36S$, which is much larger than $5S$. Consequently, the size of the largest active set during convergence seems to yield too conservative of a bound for most of the examples in practice.

4.6 *Proofs*

The proofs of the two main theorems of this section are given below. Both proofs rely on several lemmas and observations derived in the appendices.

4.6.1 Proof of Theorem 7

The proof that the active set Γ is a subset of $\Gamma_\dagger$ for all time $t \geq 0$ is done by induction over the switching times t_k.

Proof. The first induction hypothesis is that for all switching times $t_k \leq t_K$, the following holds:

$$\left|u_j(t)\right| \leq \lambda, \qquad \forall j \in \Gamma_\dagger^c \text{ and } \forall t \in [t_k, t_{k+1}). \tag{47}$$

If this condition is satisfied for all time $t \geq 0$, then nodes $j \in \Gamma_\dagger^c$ stay below threshold and the next active set $\Gamma_{k+1} = \Gamma(t_{k+1})$ is a subset of $\Gamma_\dagger$, so the theorem holds. An additional necessary induction hypothesis is

$$\left\|a_{\Gamma_k}(t_k) - a^\dagger\right\|_2 \leq B_\delta(S). \tag{48}$$

By the theorem hypotheses, the initial active set is a subset of $\Gamma_\dagger$ and (40) holds, proving readily that (47) and (48) hold at $t = 0$. Next, assume that the two induction hypotheses hold for a particular switching time t_k. If there is no more switching after t_k, then the theorem is proven. Otherwise, using the dynamics in (84), it follows that for all $j \in \Gamma_\dagger^c \subset \Gamma_k^c$ and for all $t \in [t_k, t_{k+1}]$

$$u_j(t) = e^{-(t-t_k)} u_j^{t_k} + e^{-t} \int_{t_k}^{t} e^{\nu} \rho_j(\nu) d\nu,$$

with $\rho_j(\nu) = \Phi_j^T \left(y - \Phi_{\Gamma_k} a_{\Gamma_k}(\nu)\right)$. The absolute value of the expression above can be bounded by

$$\begin{aligned}\left|u_j(t)\right| &= \left|e^{-(t-t_k)}u_j^{t_k} + e^{-t}\int_{t_k}^{t} e^{\nu}\rho_j(\nu)d\nu\right| \\ &\leq e^{-(t-t_k)}\left|u_j^{t_k}\right| + e^{-t}\int_{t_k}^{t} e^{\nu}\left|\rho_j(\nu)\right| d\nu \\ &\leq e^{-(t-t_k)}\left|u_j^{t_k}\right| + \left(1 - e^{-(t-t_k)}\right)\sup_{\nu'\in[t_k,t_{k+1}]}\left|\rho_j(\nu')\right|.\end{aligned}$$

Since, at time t_k, node $j \in \Gamma_\dagger^c$ is inactive, then $\left|u_j^{t_k}\right| \leq \lambda$. As a consequence, condition (47) is satisfied if

$$\sup_{\nu'\in[t_k,t_{k+1}]}\left|\rho_j(\nu')\right| \leq \lambda. \tag{49}$$

Since the matrix $\Phi_j^T\Phi_{\Gamma_\dagger}$ is a submatrix of $\Phi^T\Phi - I$ with $(S+1)$ distinct columns, and since Φ satisfies the RIP of order $(S+1)$, Lemma 15 yields that $\left\|\Phi_j^T\Phi_{\Gamma_\dagger}\right\| \leq \delta$. Then, for all time $t \in [t_k, t_{k+1}]$ and for all nodes $j \in \Gamma_\dagger^c$,

$$\begin{aligned}\left|\rho_j(t)\right| &= \left|\Phi_j^T\left(y - \Phi_{\Gamma_k}a_{\Gamma_k}(t)\right)\right| \\ &= \left|\Phi_j^T\left(\Phi_{\Gamma_\dagger}a^\dagger + \epsilon - \Phi_{\Gamma_k}a_{\Gamma_k}(t)\right)\right| && (y = \Phi_{\Gamma_\dagger}a^\dagger + \epsilon) \\ &= \left|\Phi_j^T\Phi_{\Gamma_\dagger}\left(a^\dagger - a_{\Gamma_k}(t)\right) + \Phi_j^T\epsilon\right| && (\text{since } \Gamma_k \subset \Gamma_\dagger) \\ &\leq \left|\Phi_j^T\Phi_{\Gamma_\dagger}\left(a^\dagger - a_{\Gamma_k}(t)\right)\right| + \left|\Phi_j^T\epsilon\right| \\ &\leq \left\|\Phi_j^T\Phi_{\Gamma_\dagger}\right\|\left\|a^\dagger - a_{\Gamma_k}(t)\right\|_2 + \left\|\Phi_{\Gamma_\dagger^c}^T\epsilon\right\|_\infty \\ &\leq \delta\left\|a^\dagger - a_{\Gamma_k}(t)\right\|_2 + \left\|\Phi_{\Gamma_\dagger^c}^T\epsilon\right\|_\infty.\end{aligned}$$

Lemma 12 is now applied to obtain a bound that holds uniformly across time:

$$\left\|a^\dagger - a(t)\right\|_2 \leq B_\delta(S), \qquad \forall t \in [t_k, t_{k+1}].$$

In particular, $\left\|a_{\Gamma_{k+1}}(t_{k+1}) - a^\dagger\right\|_2 \leq B_\delta(S)$ and the induction hypothesis (48) remains true at time t_{k+1}. Putting the pieces together and using condition (41), for all time $t \in [t_k, t_{k+1}]$ and for all nodes $j \in \Gamma_\dagger^c$, the following holds:

$$\begin{aligned}
\left|\rho_j(t)\right| &\leq \delta B_\delta(S) + \left\|\Phi_{\Gamma_\dagger^c}^T \epsilon\right\|_\infty \\
&\leq \lambda\left(1 - b\delta\sqrt{S} + b\delta\sqrt{S}\right) \\
&= \lambda.
\end{aligned}$$

This computation shows that (47) holds for all time $t \in [t_k, t_{k+1}]$. Since (47) holds at time t_{k+1}, it necessarily holds until the next switching time t_{k+2} (since, by definition, a switch occurs if a node crosses threshold), then the induction hypothesis (47) must hold for all $t \in [t_{k+1}, t_{k+2})$, and the proof by induction is complete. □

4.6.2 Proof of Theorem 8

The following presents a proof by induction on the switching times t_k that no more than q nodes are active during convergence, *i.e.*, $|\Gamma(t)| \leq q$ for all $t \geq 0$. The proof uses the set $\Delta(t)$ containing the q largest nodes in $u(t)$. While this set depends on time, the time index is removed in the notation for readability.

Proof. By Lemma 13, if

$$\left\|u_{\Delta(t)}(t)\right\|_2 \leq \lambda\sqrt{q} \tag{50}$$

for all time $t \geq 0$, then the theorem holds. The two induction hypotheses used to prove this result are that (50) and

$$\left\|a(t) - a^\dagger\right\|_2 \leq B_{\bar{\delta}}(q) \tag{51}$$

hold for all $t \leq t_k$.

By (42), the first condition (50) holds at $t = 0$. Moreover,

$$\begin{aligned}
\left\|a(0) - a^\dagger\right\|_2 &\leq \|a(0)\|_2 + \left\|a^\dagger\right\|_2 \\
&\leq \|u(0)\|_2 + \left\|a^\dagger\right\|_2 \\
&\leq \lambda\sqrt{q} + \left\|a^\dagger\right\|_2 \\
&\leq B_{\bar{\delta}}(q),
\end{aligned}$$

so (51) also holds at $t = 0$.

Next, assume that for some switching time t_k, (50) and (51) hold. If there is no more switching, the theorem is proven. Otherwise, since (51) holds at time t_k, Lemma 12 can be applied readily to prove that the induction condition (51) is true at time t_{k+1}. To prove that (50) holds, the dynamics on Δ for $t \in [t_k, t_{k+1}]$ are written as in (80):

$$u_\Delta(t) = e^{-(t-t_k)} u_\Delta(t_k) + e^{-t} \int_{t_k}^{t} e^{\nu} \rho_\Delta(\nu) d\nu,$$

where $\rho_\Delta(\nu) = a_\Delta(\nu) - \Phi_\Delta^T \Phi a(\nu) + \Phi_\Delta^T y$. The ℓ_2-norm of this quantity can be bounded as follows:

$$\begin{aligned} \|u_\Delta(t)\|_2 &\leq e^{-(t-t_k)} \|u_\Delta(t_k)\|_2 \quad + \quad e^{-t} \int_{t_k}^{t} e^{\nu} \sup_{\nu' \in t_k, t_{k+1}} \|\rho_\Delta(\nu')\|_2 \, d\nu \\ &\leq e^{-(t-t_k)} \|u_\Delta(t_k)\|_2 + \left(1 - e^{-(t-t_k)}\right) \sup_{\nu' \in t_k, t_{k+1}} \|\rho_\Delta(\nu')\|_2 \, . \end{aligned} \tag{52}$$

By the induction hypothesis (50), the following is true

$$\|u_\Delta(t_k)\|_2 \leq \lambda \sqrt{q},$$

and Lemma 13 implies that $\Gamma(t_k) = \Gamma_k \subset \Delta$ and Γ_k contains fewer than q nodes. The last step is to obtain a bound for all $t \in [t_k, t_{k+1}]$ for

$$\begin{aligned} \|\rho_\Delta(t)\|_2 &= \left\| a_\Delta(t) - \Phi_\Delta^T \Phi a(t) + \Phi_\Delta^T y \right\|_2 \\ &= \left\| a_\Delta^\dagger + \left(I_\Delta - \Phi_\Delta^T \Phi\right)(a(t) - a^\dagger) + \Phi_\Delta^T \epsilon \right\|_2 \\ &\leq \left\| a_\Delta^\dagger \right\|_2 + \left\| I_\Delta - \Phi_\Delta^T \Phi_{\Gamma_\dagger \cup \Gamma_k} \right\| \left\| a(t) - a^\dagger \right\|_2 + \left\| \Phi_\Delta^T \epsilon \right\|_2 . \end{aligned}$$

Since Φ satisfies the RIP with parameters $(S + p, \overline{\delta})$ and $\Gamma_k \subset \Delta$, Lemma 15 can be applied to the matrix $I_\Delta - \Phi_\Delta^T \Phi_{\Gamma_\dagger \cup \Delta}$ with $\Gamma_1 = \Delta$ and $\Gamma_2 = \Gamma_\dagger$, and Lemma 14 can be applied to Φ_Δ.

This results in a uniform bound on the quantity

$$\begin{aligned}
\|\rho_\Delta(t)\|_2 &\le \left\|a^\dagger_\Delta\right\|_2 + \left\|I_\Delta - \Phi_\Delta^T\Phi_{\Gamma_\dagger\cup\Delta}\right\| \left\|a(t) - a^\dagger\right\|_2 + \left\|\Phi_\Delta^T\epsilon\right\|_2 \\
&\le \left\|a^\dagger\right\|_2 + \bar{\delta}B_{\bar{\delta}}(q) + (1+\bar{\delta})\,\|\epsilon\|_2 \\
&= \left(1 + \bar{\delta}(1+\bar{\delta})(1-\bar{\delta})^{-2}\right)\left\|a^\dagger\right\|_2 \\
&\qquad + (1+\bar{\delta})\left(\bar{\delta}(1-\bar{\delta})^{-2}\sqrt{1-\bar{\delta}} + 1\right)\|\epsilon\|_2 \\
&\qquad + \bar{\delta}(1+\bar{\delta})(1-\bar{\delta})^{-2}\lambda\sqrt{q} \\
&\le (1+\bar{\delta})(1-\bar{\delta})^{-2}\left(\left\|a^\dagger\right\|_2 + \sqrt{1-\bar{\delta}}\,\|\epsilon\|_2 + \bar{\delta}\lambda\sqrt{q}\right).
\end{aligned}$$

Applying the theorem hypothesis (43) yields

$$\begin{aligned}
\|\rho_\Delta(t)\|_2 &< (1-\bar{\delta})^{-2}\left(1 - 3\bar{\delta} + \bar{\delta}(1+\bar{\delta})\right)\lambda\sqrt{q} \\
&= \lambda\sqrt{q}.
\end{aligned}$$

Plugging this result into (52) shows that $\|u_\Delta(t)\|_2 \le \lambda\sqrt{q}$ for all $t \in [t_k, t_{k+1}]$. In particular, the induction condition (50) holds at t_{k+1}, which finishes the proof. □

4.7 Summary

In this chapter and Chapter 4, the mathematical analysis of the LCA was carried out. Despite a nonsmooth activation function and possibly singular interconnection matrix that prevented the application of existing analytic results, the network was shown to converge exponentially fast from any initial point to the optimal solution. Prior to this study, algorithms for sparse recovery had been exclusively studied in the digital domain. The ISTA provides a useful reference as it is also designed to solve the ℓ_1-minimization program by taking a discrete step in the direction of the negative gradient and thresholding. This discrete-time algorithm was shown to converge with a linear rate in [60] to the solution of the ℓ_1-minimization program, for which an accuracy analysis was carried out in [9]. These two results are combined in the summary below.

ISTA for static recovery: *If Φ satisfies the RIP*[2] *, the threshold satisfies*

$$\lambda\sqrt{q} \gtrsim c_1 \left\|a^\dagger\right\|_2 + c_2\sigma$$

for some constants $c_0, c_1 \geq 0$, and the step size η is in the interval $\left(0, 2\left\|\Phi^T\Phi\right\|^{-1}\right)$, then ISTA converges with a linear rate; i.e., there exist $\bar{c} \in (0,1)$ and two constants $C_0, C_1 \geq 0$ such that for all iterations $l \geq 0$

$$\left\|a(l) - a^\dagger\right\|_2 \leq C_0\bar{c}^l + C_1.$$

The constant C_1 represents the optimal error $\left\|\hat{a}^\dagger - a^\dagger\right\|_2$ when solving (5) *and satisfies*

$$C_1 \leq C_2\lambda\sqrt{q} + C_3\sigma$$

for some $q \geq 0$ (which is typically on the order of S) and some constants $C_2, C_3 \geq 0$.

The results of Chapters 3 and 4 have provided similar convergence and accuracy guarantees for the LCA. Since the fixed points of the networks were shown to correspond to the solution to the ℓ_1-minimization program when the activation function is the soft-thresholding function, the accuracy result of [9] holds for the LCA as well. The analytic findings obtained for the LCA are summarized below.

LCA for static recovery: *If Φ satisfies the RIP*[2] *and the threshold satisfies*

$$\lambda\sqrt{q} \gtrsim c_2 \left\|a^\dagger\right\|_2 + c_3\sigma$$

for some constants $c_2, c_3 \geq 0$, then the LCA converges with an exponential rate; i.e., there exist $v \in (0,1)$ and two constants $C_4, C_5 \geq 0$ such that for all time $t \geq 0$

$$\left\|a(t) - a^\dagger\right\|_2 \leq C_4e^{-vt} + C_1,$$

where C_1 is again the optimal error achieved when solving (5). *In addition, the output of the LCA never contains more than q non-zero coefficients.*

[2] In [9], the author actually uses a slightly more general notion than the RIP, but the quantities used can be related to the classic RIP.

For the continuous-time algorithm, the linear rate of convergence becomes an exponential rate. The analysis in this thesis has also shown that the output of the LCA remains sparse, similar to the analysis carried out for certain greedy solvers such as OMP, ROMP, *etc*. This collection of results demonstrates that the LCA is a reasonable solution for sparse recovery that is worth implementing in analog VLSI for engineering applications. Eventually, a dedicated analog chip will have the potential to significantly improve the speed and power consumption necessary for real-time signal processing applications.

CHAPTER V

TRACKING OF TIME-VARYING SIGNALS

While there exist many well-established techniques with known performance guarantees to recover sparse signals from compressed measurements in the static case, only a few methods have been proposed to tackle the recovery of time-varying signals, and even fewer benefit from a theoretical analysis. In this chapter, the capacity to perform this tracking in real time is studied for both the LCA and ISTA, its discrete-time analogue. ISTA is a well-known digital solver for static sparse recovery, whose iteration is a first-order discretization of the LCA differential equation. The results of this chapter show that the outputs of both algorithms can track a time-varying signal while compressed measurements are streaming, even when no convergence criterion is imposed [61]. The ℓ_2-distance between the target signal and the outputs of both discrete- and continuous-time solvers is shown to decay exponentially fast to a bound that is essentially optimal.

5.1 Background and related work

First, the ISTA discrete iteration is reviewed, and a summary of results obtained in the static case is given along with several approaches that have been proposed in the literature to perform dynamic recovery.

5.1.1 The ISTA

The ISTA is one of the earliest digital algorithms developed for sparse recovery [19], and although it tends to converge slowly, many state-of-the-art solvers are only slight variations of its simple update rule [15, 52, 62, 63]. The ISTA is defined by a discrete update rule that can be seen as a generalized gradient step for the ℓ_1-minimization problem in (5). The l^{th}

iterate $a(l) \in \mathbb{R}^N$ is defined by[1]

$$a(l+1) = T_\lambda \left(a(l) + \eta\left(\Phi^T\left(y - \Phi a(l)\right)\right)\right). \tag{53}$$

The activation function $T_\lambda(\cdot)$ is the soft-thresholding function in (9). The constant η represents the size of the gradient step, which is usually required to be contained in the interval $\left(0, 2\left\|\Phi^T\Phi\right\|^{-1}\right)$ to ensure convergence. Several papers have shown that ISTA converges to the solution of (5) from any initial point $a(0)$ with a linear rate [58, 60].

To match the LCA equation, the extra variable $u(l)$ is introduced in the ISTA update iteration:

$$\begin{cases} u(l+1) & = a(l) + \eta\Phi^T\left(y - \Phi a(l)\right) \\ a(l+1) & = T_\lambda(u(l+1)) \end{cases}, \qquad \forall l \geq 0.$$

With this formulation, it is easy to see that ISTA is a first-order (or Euler method) discretization of the LCA dynamics. Using a step size dl for the discretization equal to the LCA time-constant $dl = t_{l+1} - t_l = \tau$, the LCA ODE (8) becomes

$$\begin{cases} \tau\dfrac{u(l+1) - u(l)}{\tau} & = -u(l) + a(l) + \Phi^T(y - \Phi a(l)) \\ a(l+1) & = T_\lambda(u(l+1)), \end{cases}$$

which can be written as

$$\begin{cases} u(l+1) & = a(l) + \Phi^T(y - \Phi a(l)) \\ a(l+1) & = T_\lambda(u(l+1)) \end{cases}.$$

This formulation matches the ISTA iteration when $\eta = 1$. As a consequence, simulating the ISTA on a digital computer with the appropriate parameter choice puts the LCA in the same framework as existing digital algorithms and facilitates the comparison of convergence time and computational complexity carried in Section 5.3.

[1] The iterate number l is in parenthesis, analogous to the continuous time index, and the n^{th} entry of the vector is put in subscript: $a_n(l)$.

5.1.2 Related work

Several approaches have been proposed to tackle the tracking of a high-dimensional sparse signal evolving with time from a set of undersampled streaming measurements. Classical methods for tracking signals include Kalman filtering and particle filtering [64]. These methods require knowledge of the underlying dynamics of the target and exploit no sparsity information. Some recent papers have built on these methods by incorporating a sparsity-aware criteria, either via convex relaxation [65,66] or greedy methods [67], and still require *a priori* knowledge of the target dynamics.

Another class of methods relies on building a probabilistic model for the evolution of the target's support and amplitudes, and uses Bayesian inference techniques to estimate the next time sample [68–70]. These methods also necessitate *a priori* knowledge of the target's behavior to adjust many parameters, and the recovery can be sensitive to inaccuracies in the model. While [70] proposes estimating the model parameters online, it only does so in the non-causal smoothing case, which can become computationally expensive as the number of parameters is large.

Finally, the last class of methods is based on optimization. For instance, in [71, 72], an optimization program is set up to account for the temporal correlation in the target, and the recovery is performed in batches. In [73], the best dynamical model is chosen among a family of possible dynamics or parameters. The performance of this technique is limited by the resolution and accuracy of the available dynamical models. In [22], a continuation approach is used to update the estimate of the target using the solution from the previous time-step. In [74–77], the optimization is solved using low-complexity iterative schemes. Unfortunately, these methods lack theoretical guarantees or at best provide convergence and accuracy results in the static case. Finally, in [78], a very general projection-based approach is studied. A convergence result is given, but it is not clear how the necessary assumptions apply in the time-varying setting and it does not come with an accuracy result.

The ISTA and LCA belong to the class of optimization-based schemes. The two algorithms do not rely on any model of the underlying dynamics, and a minimal number of parameters need to be adjusted that are already present in the static case. In the following analysis, convergence to the minimum of the objective in (5) or to a stopping criterion is not required. Rather, the LCA output evolves continuously with time as the input is streaming, while the standard ISTA iteration is performed as new measurements become available. This setting is particularly useful when signals are streaming at very high rates or computational resources are limited. Despite this simple setting, the analysis shows that the LCA and ISTA can both track a moving target accurately and provides an analytic expression for the evolution of the ℓ_2-distance between the output of both algorithms and the target for all time t. The techniques developed in this section provide a good foundation for the analysis of other algorithms that currently lack theoretical analysis, in particular iterative-thresholding schemes that extend the classic ISTA.

5.2 Tracking a time-varying input

In this section, the model used for the target signal and the two main theorems are presented. The resulting analysis provides an explicit expression for the tracking abilities of the ISTA and LCA when recovering a time-varying input $a^{\dagger}(t)$.

5.2.1 Signal model

The underlying target signal $a^{\dagger}(t)$ and the noise vector $\epsilon(t)$ are assumed to evolve continuously with time. As a consequence, the input $y(t)$ is

$$y(t) = \Phi a^{\dagger}(t) + \epsilon(t) \tag{54}$$

and is also continuous with time. The following analysis considers the general case where the measurements are corrupted by noise, but it remains valid in the noise-free case where $\epsilon = 0$. The target signal $a^{\dagger}(t)$ is assumed to remain S-sparse (*i.e.*, $\left|\Gamma_{\dagger}(t)\right| \leq S$ for all $t \geq 0$). Finally, the energy in the time-derivative of the target is required to satisfy the following bound for all time t:

$$\left\|\dot{a}^{\dagger}(t)\right\|_2 \leq -\frac{1}{\tau}\left\|a^{\dagger}(t)\right\|_2 + \mu, \qquad \forall t \geq 0. \tag{55}$$

This condition ensures that the energy in both the time-derivative $\left\|\dot{a}^{\dagger}(t)\right\|_2$ and the target itself $\left\|a^{\dagger}(t)\right\|_2$ remains bounded (see Lemma 2). Intuitively, the more energy is present in the target, the slower the variations must be for the algorithms to track them. The smaller the time constant τ of the solver is, the slower the target needs to vary to be tractable. Note that only the following condition is actually necessary in the proof of Theorem 10:

$$\left\|\dot{a}^{\dagger}(t)\right\|_2 \leq -\frac{1}{\tau}\left\|a^{\dagger}_{\Gamma^c}(t)\right\|_2 + \mu, \qquad \forall t \geq 0.$$

This condition is less restrictive than (55), since, as the LCA evolves, the output gets closer to the target signal and the energy in $\left\|a^{\dagger}_{\Gamma^c}(t)\right\|_2 = \left\|a^{\dagger}_{\Gamma^c}(t) - a_{\Gamma^c}(t)\right\|_2$ decreases. However, because the set Γ^c changes with time and the sequence of active sets is not known in advance, this condition is difficult to verify in practice.

The columns of Φ are assumed to have unit norm $\|\Phi_n\|_2 = 1$ and Φ to satisfy the RIP with parameters $(S + q, \delta)$ for some $q \geq 0$. Finally, the energy of the noise vector remains bounded and the constant σ is defined as

$$\|\epsilon(t)\|_2 \leq \frac{\sigma}{\sqrt{1+\delta}}, \qquad \forall t \geq 0. \tag{56}$$

5.2.2 Tracking abilities of ISTA

This section concerns the tracking abilities of the ISTA in a general setting, where a new measurement is received every P^{th} iteration:

$$y(kP) = \Phi a^{\dagger}(kP) + \epsilon(kP), \qquad \forall k \geq 0. \tag{57}$$

In this setting, the ISTA l^{th} iterate simply becomes

$$\begin{cases} u(l+1) & = a(l) + \eta\left(\Phi^T\left(y(l) - \Phi a(l)\right)\right) \\ a(l+1) & = T_\lambda(u(l+1)) \end{cases}, \qquad \forall l \geq 0. \tag{58}$$

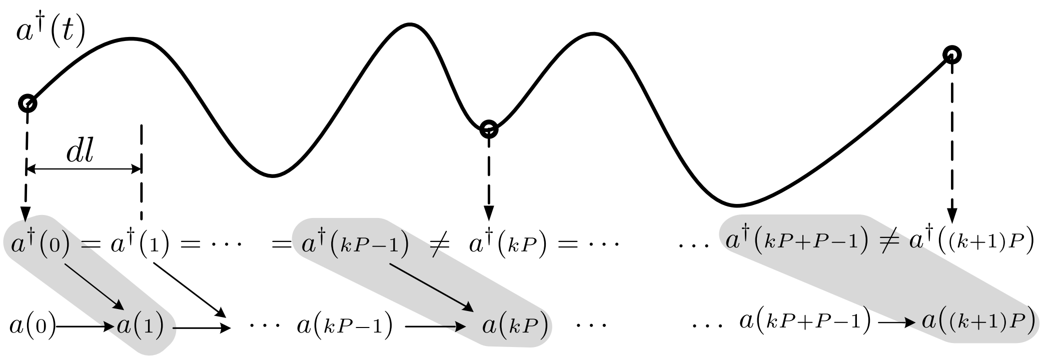

$a^\dagger(t)$
dl
$a^\dagger(0) = a^\dagger(1) = \cdots \quad = a^\dagger(kP-1) \neq a^\dagger(kP) = \cdots \quad \ldots \; a^\dagger(kP+P-1) \neq a^\dagger((k+1)P)$
$a(0) \to a(1) \to \cdots \; a(kP-1) \to a(kP) \; \cdots \quad \ldots \; a(kP+P-1) \to a((k+1)P)$

Figure 17: A new measurement of the underlying continuous-time signal $a^\dagger(t)$ is received every P^{th} ISTA iterate. During the subsequent $P - 1$ iterations, the target is treated as constant in the ISTA update rule. The quantity of interest is $\left\|a(kP) - a^\dagger(kP-1)\right\|_2$, which represents the last error before a new measurement is received.

For iterates l of the form $l = kP + i$, with $i = 0, \ldots, P-1$, since no new measurement has been received, the target signal $a^\dagger(kP+i)$ and the measurements $y(kP+i)$ are treated as constant signals (in other words, the algorithm does not assume a model on the dynamics of the target signal):

$$a^\dagger(kP+i) = a^\dagger(kP), \qquad \forall k \geq 0,\ \forall i = 0, \ldots, P-1. \tag{59}$$

This approach is illustrated in Figure 17. The step size for the discretization is $dl = t_{l+1} - t_l$. As a consequence, property (55) yields the following bound $\forall k \geq 0$:

$$\begin{aligned} \left\|a^\dagger(kP) - a^\dagger(kP-1)\right\|_2 &= \left\|\int_{t_{kP-1}}^{t_{kP}} \dot{a}^\dagger(t)dt\right\|_2 \\ &\leq \int_{t_{kP-1}}^{t_{kP}} \left\|\dot{a}^\dagger(t)\right\|_2 dt \\ &\leq \int_{t_{kP-1}}^{t_{kP}} \mu\, dt \\ &= \mu\, dl. \end{aligned} \tag{60}$$

Note that because the measurement vector $y(l)$ changes every P^{th} iteration, ISTA never converges to the optimum of (5) if P is small. This approach is of great interest for scenarios where the measurements are streaming at very high rates.

Theorem 9. *Assume that the dictionary* Φ *satisfies the RIP with parameters* $(S+2q,\delta)$ *for some* $q \geq 0$ *and that the gradient step* η *in* (58) *satisfies*

$$0 < \eta < \frac{2}{1+\delta}. \tag{61}$$

Define $c = |\eta - 1| + \delta\eta < 1$*. If the target signal satisfies condition* (60)*, the initial point* $a(0)$ *contains less than q active nodes and the following two conditions hold:*

$$\left\|u_{\Delta(0)}(0)\right\|_2 \leq \lambda\sqrt{q}, \tag{62}$$

$$\eta(1+\delta)\max\left\{\left\|a^{\dagger}(0)\right\|_2, \tau\mu\right\} + \eta\sigma \leq (1-c)\lambda\sqrt{q}, \tag{63}$$

then

1. *the output* $a(l)$ *never contains more than q active nodes for all* $l \geq 0$*; and*

2. *letting* $i = (l \mod P)$ *(i.e.,* $\exists! k \geq 0$ *such that* $l = kP + i$*, with* $0 \leq i \leq P-1$*), the* ℓ_2*-distance between the output and the target signal satisfies* $\forall l \geq 0$

   $$\left\|a(l+1) - a^{\dagger}(l)\right\|_2 \leq c^l\left(\left\|a(1) - a^{\dagger}(0)\right\|_2 - W\right) + \frac{c^{i+1}}{1-c^P}\mu\, dl + V, \tag{64}$$

   where

   $$V = (1-c)^{-1}\left(\eta\sigma + \lambda\sqrt{q}\right), \tag{65}$$

   $$W = \frac{c}{1-c^P}\mu\, dl + V. \tag{66}$$

This theorem shows that at every P^{th} iteration, the ℓ_2-distance between the output $a(kP)$ and the target signal $a^{\dagger}(kP-1)$ remains bounded and converges as $k \to \infty$ toward

$$V + \frac{c^P}{1-c^P}\mu\, dl = (1-c)^{-1}\left(\lambda\sqrt{q} + \eta\sigma\right) + \frac{c^P}{1-c^P}\mu\, dl$$

with a linear rate of convergence. This final value is essentially optimal, with the first term $(1-c)^{-1}\left(\lambda\sqrt{q} + \eta\sigma\right)$ corresponding to the error involved with solving (5). Together with the bound (63), they resemble the terms of Corollary 5.1 in [9] obtained for the static case.

The additional term $c^P(1-c^P)^{-1}\mu\, dl$ behaves like $\mu\, dl/P$ and corresponds to the error that is expected from having a time-varying input. The larger the variations in the target, the larger μ will be, which corresponds to a more difficult signal to track and a larger error. Conversely, the slower the target varies, the larger the value of P, and as expected, the smaller the final error is. When $P \to \infty$, this additional term disappears.

When the ISTA is considered as the discretization of the LCA, the discretization step dl for the Euler method is equal to τ, $\eta = 1$ in (58), and $P = 1$, so $c = \delta$ as discussed in Section 5.1.1. Then, the asymptotic final value becomes

$$\lesssim (1-\delta)^{-1}\left(\tau\mu + \sigma + \lambda\sqrt{q}\right).$$

5.2.3 Tracking abilities of the LCA

The following theorem shows that the number of non-zero elements in the LCA output remains bounded. It also provides an expression for the evolution over time of the ℓ_2-distance between the LCA output and the target signal.

Theorem 10. *Assume that the dictionary Φ satisfies the RIP with parameters $(S+q,\delta)$ for some $q \geq 0$. The following quantity depends on the threshold λ, the noise energy bound σ, the energy bound on the target signal μ, the parameter q and the RIP constant δ:*

$$D = (1-\delta)^{-1}\left(\tau\mu + \sigma + \lambda\sqrt{q}\right). \tag{67}$$

If the initial active set $\Gamma(0)$ contains less than q active nodes and the following two conditions hold:

$$\left\|u_{\Delta(0)}(0)\right\|_2 \leq \lambda\sqrt{q}, \tag{68}$$

$$\delta \cdot \max\left\{\left\|a(0) - a^{\dagger}(0)\right\|_2, D\right\} + \max\left\{\left\|a^{\dagger}(0)\right\|_2, \tau\mu\right\} + \sigma \leq \lambda\sqrt{q}, \tag{69}$$

then

1. *the active set never contains more than q active nodes (i.e., $|\Gamma(t)| \leq q$);*

2. *the energy of the q largest entries in absolute value $\Delta(t)$ in the state satisfies $\forall t \geq 0$*

$$\left\|u_{\Delta(t)}(t)\right\|_2 \leq e^{-t/\tau} \left\|u_{\Delta(0)}(0)\right\|_2 + \left(1 - e^{-t/\tau}\right) \lambda \sqrt{q}; \tag{70}$$

3. *the ℓ_2-distance between the LCA output and the target signal satisfies $\forall t \geq 0$*

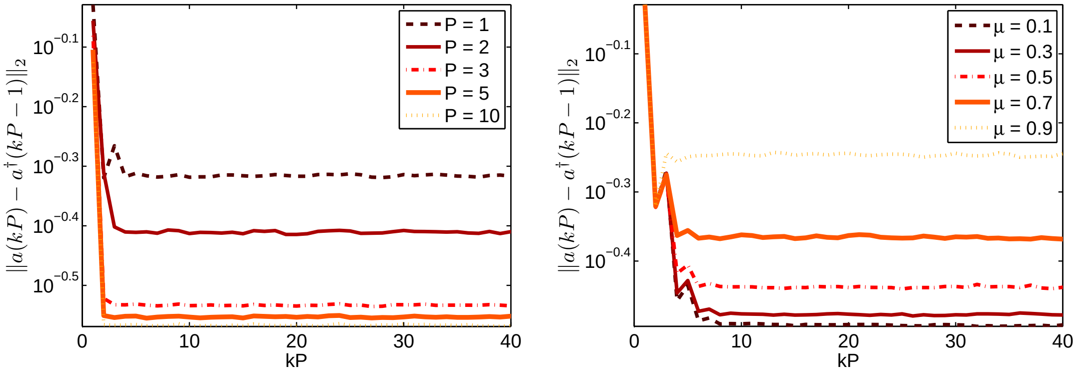
$$\left\|a(t) - a^{\dagger}(t)\right\|_2 \leq e^{-(1-\delta)t/\tau} \left\|a(0) - a^{\dagger}(0)\right\|_2 + \left(1 - e^{-(1-\delta)t/\tau}\right) D. \tag{71}$$

This theorem shows that the ℓ_2-distance between the LCA output and the target signal converges exponentially fast towards its final value

$$D = (1 - \delta)^{-1} \left(\tau\mu + \sigma + \lambda \sqrt{q}\right).$$

This quantity is equal to the final value obtained for ISTA when it corresponds to the first-order approximation of the LCA ODE. This bound is again essentially optimal for the problem. The first term $(1-\delta)^{-1}\left(\lambda\sqrt{q} + \sigma\right)$ corresponds to the expected error when solving (5), while the additional term $(1 - \delta)^{-1}\tau\mu$ corresponds to the error associated with recovering a time-varying signal. The error increases with μ, which corresponds to the energy of the variations in the target. Conversely, the error decreases with decreasing τ, corresponding to a faster solver. It is interesting to note that the initial conditions (68) and (69) are similar to the initial conditions of Theorem 3 in [55]. In particular, the analysis in [55] shows that for classic CS matrices for which $\delta \sim \sqrt{S/M \log(N/S)}$, the number of measurements required for (69) to hold is $O(S \log(N/S))$.

As a final remark, the convergence rate of both continuous and discrete algorithms depend on the RIP constant of the matrix Φ. However, the condition on the RIP constant is stronger in Theorem 9. This discrepancy can be explained by the fact that the ISTA is a discrete-time algorithm and, as a consequence, the set of active elements $\Gamma(l+1)$ may differ by as much as q elements from the previous active set $\Gamma(l)$. By contrast, the changes are continuous in the case of the LCA.



(a) Effect of the parameter P.

(b) Effect of the parameter μ.

Figure 18: Evolution of the ℓ_2-distance between the target and the output after every P^{th} ISTA iteration.

5.3 Simulations

The simulations in this section illustrate the previous theoretical results[2].

5.3.1 Synthetic data

A synthetic sparse vector $a^\dagger(k)$ for $k = 1, \ldots, 40$ of length $N = 512$ with sparsity $S = 40$ is generated as follows. For $k = 1$, $S = 40$ random amplitudes are drawn from a standard normal distribution and normalized to have norm e. Then, 39 consecutive time samples are obtained as follows

$$\alpha(k+1) = \sqrt{\frac{e^2 - \mu^2}{e^2}}\alpha(k) + \frac{\mu}{\sqrt{S}}v(k),$$

where $v(k)$ is a vector in $\mathbb{R}^S$ with amplitudes drawn from a standard normal distribution. Each sample in the sequence $\{\alpha(k)\}_{k=1,\ldots,40}$ has energy equal to e in expectation, and differences between consecutive samples have energy proportional to μ. To model support changes, a set of 10 sinusoids with frequencies drawn uniformly at random from $[0, 3]$ and random phases are generated. For 10 randomly selected indices, the corresponding target $a_n^\dagger(k)$ is set to the product of $\alpha_n(k)$ with the positive part of the sinusoids. For 10 different indices, the corresponding target $a_n^\dagger(k)$ are set to the product of $\alpha_n(k)$ with the negative

[2] Matlab code for running the experiments in this section can be downloaded from `http://users.ece.gatech.edu/~abalavoine3/code/LCA_ISTA_exp.zip`

part of the sinusoids. The remaining $S - 10$ nodes in the support $a_n^\dagger(k)$ are assigned to the remaining amplitudes $\alpha_n(k)$. This setup ensures that the sparsity of $a^\dagger$ is always $S = 40$ while letting 20 nodes switch between active and inactive. When they are not varied, the following values are used: $\eta = 1$, $\mu = 0.8$ and $P = 1$. The measurement matrix Φ is 256×512 with entries drawn from a standard normal distribution and columns normalized to 1. A Gaussian white noise vector with standard deviation $0.3 \left\|\Phi a^\dagger(0)\right\|_2 / \sqrt{M}$ is added to the measurements, which corresponds to a moderate level of noise. In Figure 18, the average over 1000 such trials of the ℓ_2-error $\left\|a(kP) - a^\dagger(kP-1)\right\|_2$ is plotted. The curves tend to a final value that matches the behavior predicted by Theorem 9 as $k \to \infty$. A higher value of P decreases the quantity $c^P \left(1 - c^P\right)^{-1} \mu\, dl$ and yields a lower final value, while a larger value of μ yields a larger final value.

Next, the threshold λ and the sparsity level S are varied, and for each pair 10 time samples of $a^\dagger(k)$ and associated measurements $y(k)$ are generated in the same fashion as before. The ISTA is run for $P = 5$ iterations per measurement. In Figure 19, the average over 100 such trials of the ratio of the maximum number of non-zero elements q in $a(l)$ over the sparsity level S is plotted. The figure shows that the maximum number of non-zero elements remains small (q is mostly contained between $1S$ and $10S$), which matches the two theorems' prediction.

5.3.2 Real data

Finally, the performance of ISTA in the streaming setting is tested on real data and compared against SpaRSA, a state-of-the-art LASSO solver [52], BPDN-DF (which adds a time-dependent regularization between frames), RWL1-DF (which additionally performs reweighting at each iteration) [66] and DCS-AMP (which uses a probabilistic model to describe the target's evolution) [70]. A total of 13 videos representing natural scenes are used to get 100 random sequences of 40 consecutive frames [3]. Since natural images are sparse in the wavelet domain, following the work in [79], the measurement matrix is taken

[3] The videos used can be downloaded at `http://trace.eas.asu.edu/yuv/`

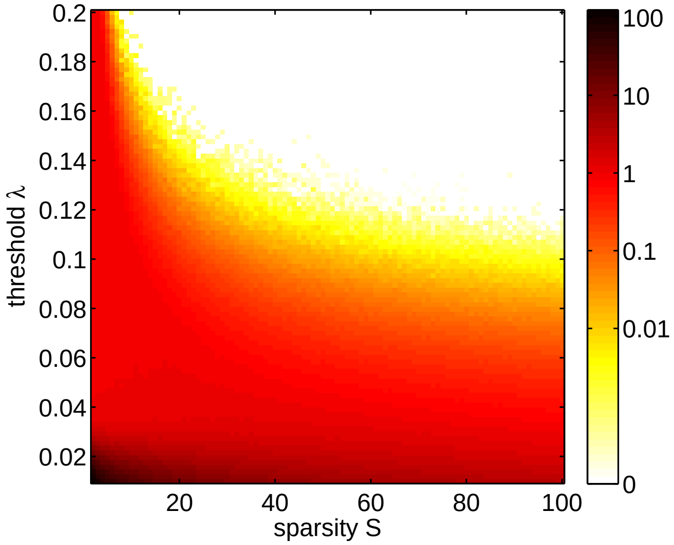


Figure 19: Ratio of the maximum number of non-zero elements q over the sparsity level S for several values of λ and S averaged over 100 trials.

to be $\Phi = AB$, where A consists of $M = 0.25N$ random rows of a noiselet matrix and B is a dual-tree discrete wavelet transform (DT-DWT) [80]. SpaRSA and BPDN-DF are given the estimate at the previous frame as a warm start for the following frame. The results obtained for the ISTA with $P = 1$ and $\eta = 1$ simulate the LCA ODEs. The regularized mean-squared error, defined by

$$rMSE(k) = \frac{\left\|a(k) - a^{\dagger}(k)\right\|_2}{\left\|a^{\dagger}(k)\right\|_2},$$

is plotted in Figure 20a, and the number of products involving the matrix Φ or its transpose is plotted in Figure 20b, both averaged over the 100 trials. The number of multiplications by Φ or Φ^T is preferred to the CPU time because it is a less arbitrary measure of the computational complexity for each algorithm. In figure 20a, the final rMSE reached by ISTA with $P = 3$ and $P = 10$ is contained between those of SpaRSA and BPDN-DF after about 19 and 6 frames, respectively. The average rMSE for ISTA with $P = 1$ converges much slower. However, one would expect that an analog implementation of the LCA would result in a faster time constant and would more closely match the performance of ISTA for $P = 10$ (assuming an analog constant 10 times smaller than the digital equivalent). The rMSE for RWL1-DF is much lower due to the additional reweighting steps. However, the complexity for this method is much larger than the other four. While the complexity of ISTA for $P = 10$ is similar to that of DCS-AMP, the complexity for $P = 3$ and $P = 1$ can be much smaller than any of the other approaches. In addition, the ISTA only involves a single parameter to adjust, while DCS-AMP has around 10. When its parameters are optimized non-causally for a specific video sequence, DCS-AMP performs only slightly worse than RWL1-DF. However, its performance degrades greatly when the parameters are not optimized for each individual video as in Figure 20a, which is a drawback for real-world applications.

5.4 Proofs

5.4.1 Lemma 2

The following lemma gives a bound on the energy of the target when its time-derivative satisfies (55).

Lemma 2. *If the target signal $a^{\dagger}(t)$ is continuous and satisfies* (55) *for all $t \geq 0$ then, $\forall t \geq 0$*

$$\begin{aligned}\left\|a^{\dagger}(t)\right\|_2 &\leq e^{-t/\tau}\left(\left\|a^{\dagger}(0)\right\|_2 - \tau\mu\right) + \tau\mu \\ &\leq \max\left\{\left\|a^{\dagger}(0)\right\|_2, \tau\mu\right\}.\end{aligned}$$

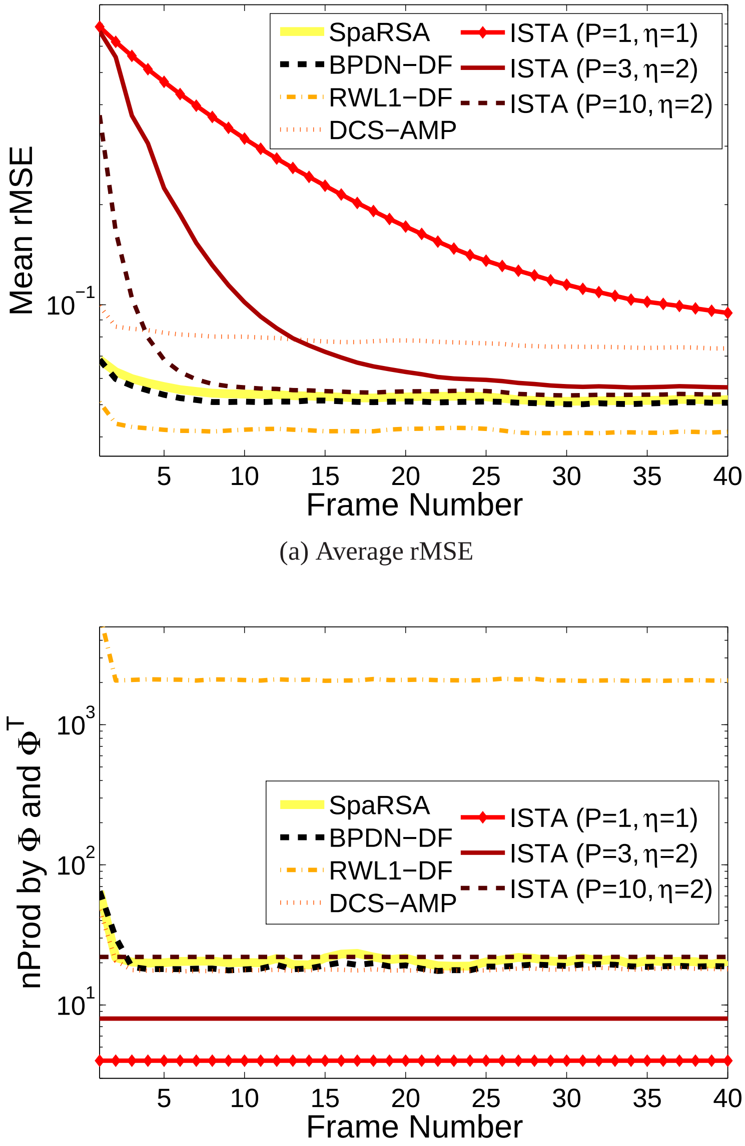
Proof. It suffices to notice that

$$\frac{d}{dt}\left(\|x(t)\|_2\right) = \frac{\frac{d}{dt}\left(\|x(t)\|_2^2\right)}{2\,\|x(t)\|_2} = \frac{x(t)^T\dot{x}(t)}{\|x(t)\|_2} \leq \|\dot{x}(t)\|_2\,,$$

where the last inequality comes from the Cauchy-Schwartz inequality. Thus, (55) implies

$$\frac{d}{dt}\left(\left\|a^{\dagger}(t)\right\|_2\right) \leq -\frac{1}{\tau}\left\|a^{\dagger}(t)\right\|_2 + \mu.$$

Since $a^{\dagger}(t)$ is continuous, the first inequality can be deduced from Lemma 9. The second inequality immediately follows from the monotonicity of the exponential. □



(a) Average rMSE

(b) Average number of products by Φ and Φ^T

Figure 20: Results of the experiment to recover the wavelet coefficients averaged over 100 random video sequences of 40 consecutive frames.

5.4.2 Proof of Theorem 9

Proof. To check that $c < 1$, the terms in hypothesis (61) can be reorganized to yield

$$\begin{aligned} \eta(1+\delta) < 2 \text{ and } \delta < 1 &\Rightarrow \eta\delta < \eta < 2 - \eta\delta \\ &\Rightarrow -1 + \eta\delta < \eta - 1 < 1 - \eta\delta \\ &\Rightarrow |\eta - 1| < 1 - \eta\delta \\ &\Rightarrow c < 1. \end{aligned}$$

In a first step, it is shown that $\left\|u(l)_{\Delta(l)}\right\|_2 \leq \lambda\sqrt{q}$ holds $\forall l \geq 0$ by induction on l. If so, by Lemma 13, the active set contains less than q elements and part 1) of the theorem is proven. By (62), this inequality holds for $l = 0$. Next, assume that $\left\|u(l)_{\Delta(l)}\right\|_2 \leq \lambda\sqrt{q}$ for some $l \geq 0$. By Lemma 13, it can be concluded that $\Gamma(l) \subset \Delta(l)$ and $|\Gamma(l)| \leq q$. As a consequence, the set

$$J = J(l+1) := \Delta(l+1) \cup \Gamma(l) \cup \Gamma_{\dagger}(l)$$

contains less than $S + 2q$ indices. Using the RIP of Φ, the eigenvalues of the matrix $\Phi_J^T\Phi_J$ are contained between $(1-\delta)$ and $(1+\delta)$ and

$$\begin{aligned} \left\|\eta\Phi_J^T\Phi_J - I_J\right\| &= \max\left\{|\eta(1+\delta) - 1|\,, |\eta(1-\delta) - 1|\right\} \\ &= |\eta - 1| + \eta\delta \quad = c. \end{aligned}$$

In addition, the form of the activation function (23) implies that

$$\|a(l)\|_2 \leq \left\|u_{\Gamma(l)}(l)\right\|_2 \leq \left\|u_{\Delta(l)}(l)\right\|_2 \leq \lambda\sqrt{q}.$$

Combining Lemma 2 with hypothesis (63) yields the following bound

$$\begin{aligned} \|u_J(l+1)\|_2 &= \left\|\eta\Phi_J^T\Phi\left(a^{\dagger}(l) - a(l)\right) + a_J(l) + \eta\Phi_J^T\epsilon(l)\right\|_2 \\ &\leq \left\|\left[\eta\Phi_J^T\Phi_J - I_J\right]a(l)\right\|_2 \; + \; \left\|\eta\Phi_J^T\Phi_J a^{\dagger}(l)\right\|_2 + \eta\left\|\Phi_J^T\epsilon(l)\right\|_2 \\ &\leq c\,\|a(l)\|_2 \; + \; \eta(1+\delta)\left\|a^{\dagger}(l)\right\|_2 \; + \; \eta\sqrt{1+\delta}\,\|\epsilon(l)\|_2 \\ &\leq c\lambda\sqrt{q} \; + \; \eta(1+\delta)\max\left\{\left\|a^{\dagger}(0)\right\|_2, \tau\mu\right\} \; + \; \eta\sigma \\ &\leq \lambda\sqrt{q}. \end{aligned}$$

Since $\Delta(l+1) \subset J(l+1)$, the induction hypothesis holds at $l+1$. As a consequence, part 1) of the theorem is proven, as well as the stronger result $\left\|u(l)_{J(l)}\right\|_2 \leq \lambda\sqrt{q}\ \ \forall l \geq 1$, which will be used in the remaining of the proof.

An induction on l is used to show that (64) holds $\forall l \geq 0$. It obviously holds for $l = 0$. Next, assume that (64) holds for some $l \geq 0$. There exist a unique $k \geq 0$ and a unique $0 \leq i \leq P-1$ such that $l = kP + i$. In the previous part of the proof, it was shown that $\|u_{J'}(l+2)\|_2 \leq \lambda\sqrt{q}$, where $J' = J(l+2) = \Delta(l+2) \cup \Gamma(l+1) \cup \Gamma_{\dagger}(l+1)$ and that J' contains less than $S + 2q$ indices. As a consequence, the RIP of $\Phi_{J'}^T\Phi_{J'}$ can be used to obtain the inequality

$$\begin{aligned}
\left\|a(l+2) - a^{\dagger}(l+1)\right\|_2 &\leq \|a(l+2) - u_{J'}(l+2)\|_2 \ + \ \left\|u_{J'}(l+2) - a^{\dagger}(l+1)\right\|_2 \\
&\leq \|u_{J'}(l+2)\|_2 \ + \ \left\|u_{J'}(l+2) - a^{\dagger}(l+1)\right\|_2 \\
&\leq \lambda\sqrt{q} + \left\|\eta\Phi_{J'}^T\epsilon(l+1) \ + \ \left(\eta\Phi_{J'}^T\Phi_{J'} - I_{J'}\right)\left(a^{\dagger}(l+1) - a(l+1)\right)\right\|_2 \\
&\leq \lambda\sqrt{q} \ + \ \eta\sigma \ + \ c\left\|a^{\dagger}(l+1) - a(l+1)\right\|_2 \\
&\leq \lambda\sqrt{q} \ + \ \eta\sigma \ + \ c\left\|a(l+1) - a^{\dagger}(l)\right\|_2 \ + \ c\left\|a^{\dagger}(l) - a^{\dagger}(l+1)\right\|_2 .
\end{aligned}$$

Using the induction hypothesis (64) at l, the analysis can be split into two cases.

First case: When $i = P-1$, $l = (k+1)P - 1$ and (60) yields $\left\|a^{\dagger}(l+1) - a^{\dagger}(l)\right\|_2 \leq \mu\, dl$. Thus,

$$\begin{aligned}
\left\|a(l+2) - a^{\dagger}(l+1)\right\|_2 &\leq c\left(c^l\left[\left\|a(1) - a^{\dagger}(0)\right\|_2 - W\right] + \frac{c^P}{1-c^P}\mu\, dl + V\right) + c\mu\, dl + \lambda\sqrt{q} + \eta\sigma \\
&\leq c^{l+1}\left[\left\|a(1) - a^{\dagger}(0)\right\|_2 - W\right] + \frac{c^{P+1}}{1-c^P}\mu\, dl + cV + c\mu\, dl + \lambda\sqrt{q} + \eta\sigma \\
&\leq c^{l+1}\left[\left\|a(1) - a^{\dagger}(0)\right\|_2 - W\right] + \frac{c}{1-c^P}\mu\, dl + V.
\end{aligned}$$

Therefore, the induction hypothesis (64) holds for $l+1 = (k+1)P$.

Second case: When $0 \le i \le P-2$, (59) yields $\left\|a^\dagger(l+1) - a^\dagger(l)\right\|_2 = 0$ and so

$$\begin{aligned}\left\|a(l+2) - a^\dagger(l+1)\right\|_2 &\le c\Big(c^l\left(\left\|a(1) - a^\dagger(0)\right\|_2 - W\right) + \frac{c^{i+1}}{1-c^P}\mu\, dl + V\Big) + \lambda\sqrt{q} + \eta\sigma \\ &\le c^{l+1}\left[\left\|a(1) - a^\dagger(0)\right\|_2 - W\right] + \frac{c^{i+2}}{1-c^P}\mu\, dl + cV + \lambda\sqrt{q} + \eta\sigma \\ &\le c^{l+1}\left[\left\|a(1) - a^\dagger(0)\right\|_2 - W\right] + \frac{c^{i+2}}{1-c^P}\mu\, dl + V.\end{aligned}$$

Since $l+1 = kP + (i+1)$, with $1 \le i+1 \le P-1$, this inequality proves the induction hypothesis (64) in the second case and finishes the proof. □

5.4.3 Proof of Theorem 10

Proof. The proof is done by induction on the switching time t_k. The induction hypothesis is that the active set Γ_k contains less than q active elements and that (70) and (71) hold $\forall t \le t_k$.

At time $t_0 = 0$, the theorem hypotheses imply that Γ_0 contains less than q active elements, and that (70) and (71) hold.

Next, assume that $\forall t \le t_k$ the active set contains less than q active elements and that (70) and (71) hold. In a first step, it is shown that (71) holds $\forall t \le t_{k+1}$. By the induction hypothesis, the active set Γ contains less than q active nodes for all $t \le t_k$, including the current active set Γ_k for $t \in [t_k, t_{k+1})$. As a consequence, the inequalities in Lemma 15 hold with $\Gamma_1 = \Gamma$ and $\Gamma_2 = \Gamma_\dagger$ for all $t \le t_{k+1}$.

The following time derivative can be computed $\forall t \le t_{k+1}$:

$$\begin{aligned}\tau\frac{d}{dt}\left(\frac{1}{2}\left\|a(t) - a^\dagger(t)\right\|_2^2\right) &= \tau\left(a(t) - a^\dagger(t)\right)^T\left(\dot{a}(t) - \dot{a}^\dagger(t)\right) \\ &= \left(a(t) - a^\dagger(t)\right)^T\left(-\Phi_\Gamma^T\Phi_\Gamma a(t) + \Phi_\Gamma^T y(t) - \lambda \mathsf{s}_\Gamma - \tau\dot{a}^\dagger(t)\right) \\ &= -\left(a(t) - a^\dagger(t)\right)^T\Phi_\Gamma^T\Phi_{(\Gamma\cup\Gamma_\dagger)}\left(a(t) - a^\dagger(t)\right) \\ &\quad + \left(a(t) - a^\dagger(t)\right)^T\left(\Phi_\Gamma^T\epsilon(t) - \lambda \mathsf{s}_\Gamma - \tau\dot{a}^\dagger(t)\right).\end{aligned}$$

Note that

$$
\begin{aligned}
&-\left(a(t)-a^\dagger(t)\right)^T \Phi_\Gamma^T\Phi_{(\Gamma\cup\Gamma_\dagger)}\left(a(t)-a^\dagger(t)\right)\\
&\qquad= -\left\|\Phi_\Gamma\left(a(t)-a^\dagger(t)\right)\right\|_2^2 + \left(a(t)-a^\dagger(t)\right)^T \Phi_\Gamma^T\Phi_{(\Gamma^c\cap\Gamma_\dagger)}a^\dagger_{\Gamma^c}(t)\\
&\qquad\leq -\left\|\Phi_\Gamma\left(a(t)-a^\dagger(t)\right)\right\|_2^2 + \left\|a_\Gamma(t)-a^\dagger_\Gamma(t)\right\|_2\left\|\Phi_\Gamma^T\Phi_{(\Gamma^c\cap\Gamma_\dagger)}a^\dagger_{\Gamma^c}(t)\right\|_2\\
&\qquad\leq -(1-\delta)\left\|a_\Gamma(t)-a^\dagger_\Gamma(t)\right\|_2^2 + \delta\left\|a_\Gamma(t)-a^\dagger_\Gamma(t)\right\|_2\left\|a^\dagger_{\Gamma^c}(t)\right\|_2\\
&\qquad= -(1-\delta)\left\|a(t)-a^\dagger(t)\right\|_2^2 + (1-\delta)\left\|a^\dagger_{\Gamma^c}(t)\right\|_2^2\\
&\qquad\qquad\qquad\qquad\qquad\qquad+\delta\left\|a_\Gamma(t)-a^\dagger_\Gamma(t)\right\|_2\left\|a^\dagger_{\Gamma^c}(t)\right\|_2\\
&\qquad\leq -(1-\delta)\left\|a(t)-a^\dagger(t)\right\|_2^2 + \left\|a^\dagger_{\Gamma^c}(t)\right\|_2\times\\
&\qquad\qquad\qquad\left((1-\delta)\left\|a(t)-a^\dagger(t)\right\|_2+\delta\left\|a(t)-a^\dagger(t)\right\|_2\right)\\
&\qquad= -(1-\delta)\left\|a(t)-a^\dagger(t)\right\|_2^2 + \left\|a^\dagger_{\Gamma^c}(t)\right\|_2\left\|a(t)-a^\dagger(t)\right\|_2.
\end{aligned}
$$

Plugging this inequality into the expression for the time derivative,

$$
\begin{aligned}
&\tau\frac{d}{dt}\left(\frac{1}{2}\left\|a(t)-a^\dagger(t)\right\|_2^2\right)+(1-\delta)\left\|a(t)-a^\dagger(t)\right\|_2^2\\
&\quad\leq \left\|a(t)-a^\dagger(t)\right\|_2\left\|a^\dagger_{\Gamma^c}(t)\right\|_2 + \left(a(t)-a^\dagger(t)\right)^T\left(\Phi_\Gamma^T\epsilon(t)-\lambda\mathsf{s}_\Gamma-\tau\dot{a}^\dagger(t)\right)\\
&\quad\leq \left\|a(t)-a^\dagger(t)\right\|_2\left\|a^\dagger_{\Gamma^c}(t)\right\|_2 + \left\|a(t)-a^\dagger(t)\right\|_2\left\|\Phi_\Gamma^T\epsilon(t)-\lambda\mathsf{s}_\Gamma-\tau\dot{a}^\dagger(t)\right\|_2\\
&\quad\leq \left\|a(t)-a^\dagger(t)\right\|_2\left\|a^\dagger_{\Gamma^c}(t)\right\|_2 + \left\|a(t)-a^\dagger(t)\right\|_2\left(\left\|\Phi_\Gamma^T\epsilon(t)\right\|_2+\lambda\left\|\mathsf{s}_\Gamma\right\|_2+\tau\left\|\dot{a}^\dagger(t)\right\|_2\right)\\
&\quad\leq \left\|a(t)-a^\dagger(t)\right\|_2\left(\left\|a^\dagger_{\Gamma^c}(t)\right\|_2+\tau\left\|\dot{a}^\dagger(t)\right\|_2\right) + \left\|a(t)-a^\dagger(t)\right\|_2\left(\sqrt{1+\delta}\left\|\epsilon(t)\right\|_2+\lambda\sqrt{q}\right)\\
&\quad\leq \left\|a(t)-a^\dagger(t)\right\|_2\left(\tau\mu+\sigma+\lambda\sqrt{q}\right).
\end{aligned}
$$

The bound on the energy in the target's derivative (55) was used to obtain the last inequality. Noting that

$$
\frac{d}{dt}\left(\left\|x(t)\right\|_2\right)=\frac{\frac{d}{dt}\left(\left\|x(t)\right\|_2^2\right)}{2\left\|x(t)\right\|_2},
$$

the following inequality holds

$$
\frac{d}{dt}\left(\left\|a(t)-a^\dagger(t)\right\|_2\right)\leq -(1-\delta)/\tau\left\|a(t)-a^\dagger(t)\right\|_2+1/\tau\left(\sigma+\lambda\sqrt{q}+\tau\mu\right).
$$

Since $\left\|a(t) - a^{\dagger}(t)\right\|_2$ is continuous, Lemma 9 can be applied to obtain, $\forall t \leq t_{k+1}$,

$$\left\|a(t) - a^{\dagger}(t)\right\|_2 \leq e^{-(1-\delta)t/\tau} \left\|a(0) - a^{\dagger}(0)\right\|_2 + \left(1 - e^{-(1-\delta)t/\tau}\right) \frac{\sigma + \lambda\sqrt{q} + \tau\mu}{1-\delta}.$$

This inequality shows that (71) holds for all $t \leq t_{k+1}$.

Next, the hypothesis (70) is shown to hold for all $t \leq t_{k+1}$. Though the set $\Delta(t)$ varies with time, $\left\|u_{\Delta(t)}(t)\right\|_2$ is continuous for all $t \geq 0$ (as a continuous function of the supremum of the continuous functions $|u_i(t)|$). Moreover, the following time derivative can be computed $\forall t \leq t_{k+1}$:

$$\begin{aligned}\tau\frac{d}{dt}\left(\frac{1}{2}\left\|u_{\Delta}(t)\right\|_2^2\right) &= \tau u_{\Delta}(t)^T \dot{u}_{\Delta}(t)\\ &= u_{\Delta}(t)^T\Big(- u_{\Delta}(t) + a_{\Delta}(t) - \Phi_{\Delta}^T\Phi a(t) + \Phi_{\Delta}^T y(t)\Big)\\ &\leq - \left\|u_{\Delta}(t)\right\|_2^2 + \left\|u_{\Delta}(t)\right\|_2 \left\|\rho_{\Delta}(t)\right\|_2,\end{aligned}$$

where $\rho_{\Delta}(t) = a_{\Delta}(t) - \Phi_{\Delta}^T\Phi a(t) + \Phi_{\Delta}^T y(t)$. This quantity can be bounded $\forall t \leq t_{k+1}$ by

$$\begin{aligned}\left\|\rho_{\Delta}(t)\right\|_2 &= \left\|a_{\Delta}(t) - \Phi_{\Delta}^T\Phi a(t) + \Phi_{\Delta}^T y(t)\right\|_2\\ &= \left\|a_{\Delta}^{\dagger}(t) + \left(I_{\Delta} - \Phi_{\Delta}^T\Phi\right)(a(t) - a^{\dagger}(t)) + \Phi_{\Delta}^T\epsilon(t)\right\|_2\\ &\leq \left\|a^{\dagger}(t)\right\|_2 + \left\|\Phi_{\Delta}^T\epsilon(t)\right\|_2 + \left\|I_{\Delta} - \Phi_{\Delta}^T\Phi_{(\Delta\cup\Gamma_k)}\right\| \left\|a(t) - a^{\dagger}(t)\right\|_2\\ &\leq \max\left\{\left\|a^{\dagger}(0)\right\|_2, \tau\mu\right\} + \sigma + \delta\left\|a(t) - a^{\dagger}(t)\right\|_2,\end{aligned}$$

where Lemma 15 with $\Gamma_1 = \Delta$ and $\Gamma_2 = \Gamma_{\dagger}$ and Lemma 2 were applied. Finally, the bound (71) obtained for $\left\|a(t) - a^{\dagger}(t)\right\|_2$ for all $t \leq t_{k+1}$ and the monotonicity of the exponential yield

$$\begin{aligned}\left\|\rho_{\Delta}(t)\right\|_2 &\leq \max\left\{\left\|a^{\dagger}(0)\right\|_2, \tau\mu\right\} + \sigma + \delta\left[e^{-(1-\delta)t/\tau}\left\|a(0) - a^{\dagger}(0)\right\|_2 + \left(1 - e^{-(1-\delta)t/\tau}\right)D\right]\\ &\leq \max\left\{\left\|a^{\dagger}(0)\right\|_2, \tau\mu\right\} + \sigma + \delta\max\left\{\left\|a(0) - a^{\dagger}(0)\right\|_2, D\right\}\\ &\leq \lambda\sqrt{q},\end{aligned}$$

where the last inequality comes from the theorem's hypothesis (69). As a consequence, the following inequality holds $\forall t \leq t_{k+1}$:

$$\tau\frac{d}{dt}\left(\left\|u_{\Delta}(t)\right\|_2\right) \leq - \left\|u_{\Delta}(t)\right\|_2 + \lambda\sqrt{q}.$$

Using Lemma 9 again yields, $\forall t \leq t_{k+1}$,

$$\begin{aligned}\|u_\Delta(t)\|_2 &\leq e^{-t/\tau}\,\|u_\Delta(0)\|_2 + e^{-t/\tau}\int_0^t e^{\nu/\tau}\lambda\sqrt{q}d\nu\\ &\leq e^{-t/\tau}\,\|u_\Delta(0)\|_2 + \left(1-e^{-t/\tau}\right)\lambda\sqrt{q},\end{aligned}$$

which shows that (70) holds for all $t \leq t_{k+1}$.

Finally, the last induction hypothesis is proven to hold; *i.e.*, the next active set Γ_{k+1} is shown to contain less than q indices. Since (70) holds $\forall t \leq t_{k+1}$, together with (68) this inequality implies that

$$\begin{aligned}\|u_\Delta(t_{k+1})\|_2 &\leq e^{-t_{k+1}/\tau}\,\|u_\Delta(0)\|_2 + \left(1-e^{-t_{k+1}/\tau}\right)\lambda\sqrt{q}\\ &\leq \lambda\sqrt{q}.\end{aligned}$$

Applying Lemma 13 shows that the active set Γ_{k+1} contains less than q indices and finishes the proof. □

5.5 Summary

Previous analysis had shown that the ISTA converges to the solution of the ℓ_1-minimization problem with a linear rate, and the analysis in Chapters 3 and 4 showed that the LCA converges to the same solution with an exponential rate when these algorithms are recovering a static signal (*cf*, Section 4.7). In this chapter, an analysis for both the continuous-time LCA and discrete-time ISTA was given for the online recovery of a time-varying signal from streaming compressed measurements. In this setting, no convergence criterion is necessary before proceeding to the next frame, and a new measurement is fed in input as soon as it becomes available. The analysis showed that the convergence rate of the ℓ_2-distance between the target signal and the output of the ISTA is still linear, and that it is still exponential for the output of the LCA. In addition, an expression for the best possible error achievable (corresponding to achieving convergence for each frame) was given. These results are simplified and summarized below.

ISTA for dynamic recovery: *If Φ satisfies the RIP with parameters $(s + 2q, \delta)$ for some $q \geq 0$, the threshold λ satisfies*

$$\lambda \sqrt{q} \gtrsim c_6 \beta + c_7 \sigma$$

for some constants $c_6, c_7, \beta \geq 0$ such that $\left\| a^{\dagger}(t) \right\|_2 \leq \beta$ for all $t \geq 0$, and the step size η is in the interval $\left(0, 2(1 - \delta)^{-1}\right)$, then the ISTA converges with a linear rate; i.e., there exist $c \in (0, 1)$ and two constants $C_5, C_6 \geq 0$ such that, for all iterations $l \geq 0$,

$$\left\| a(l) - a^{\dagger} \right\|_2 \leq C_5 c^l + C_6.$$

The constant C_6 represents the optimal error if the ISTA had infinite iterations per frame to converge and satisfies

$$C_6 \leq C_7 \lambda \sqrt{q} + C_8 \sigma + C_9 \mu \; dl$$

for some constants $C_7, C_8, C_9, \mu \geq 0$ such that $\left\| \dot{a}^{\dagger}(t) \right\|_2 \leq \mu$ for all $t \geq 0$, and where dl is the time between two iterates. In addition, the output never contains more than q non-zero coefficients.

LCA for dynamic recovery: *If Φ satisfies the RIP with parameters $(s + q, \delta)$ for some $q \geq 0$, and the threshold satisfies*

$$\lambda \sqrt{q} \gtrsim c_8 \beta + c_9 \sigma$$

for some constants $c_8, c_9, \beta \geq 0$ such that $\left\| a^{\dagger}(t) \right\|_2 \leq \beta$ for all $t \geq 0$, then the LCA converges with an exponential rate; i.e., there exist $v \in (0, 1)$ and two constants $C_7, C_8 \geq 0$ such that, for all time $t \geq 0$,

$$\left\| a(t) - a^{\dagger} \right\|_2 \leq C_7 e^{-vt} + C_8.$$

The constant C_8 represents the steady state error if the LCA converged for each frame and satisfies

$$C_8 \leq C_9 \lambda \sqrt{q} + C_{10} \sigma + C_{11} \tau \mu$$

for some constants $C_9, C_{10}, C_{11}, \mu \geq 0$ *such that* $\left\|\dot{a}^{\dagger}(t)\right\|_2 \leq \mu$ *for all* $t \geq 0$*, and where* τ *is the time constant of the LCA. In addition, the output of the LCA never contains more than q non-zero coefficients.*

The links between the static setting in Section 4.7 and the dynamic setting above appear explicitly in this summary. The convergence rates for both the discrete- and continuous-time algorithms remain the same between the two settings. The optimal error in the dynamic case is composed of the static error $C'\lambda\sqrt{q} + C''\sigma$ plus a term $\mu\ dl$ or $\tau\mu$, which reflects how much energy is in the derivative of the target signal *vs* how fast each algorithm completes one ‘iteration’. Thus, the results of this chapter naturally extend those obtained in the static setting.

It had been previously observed in literature that limiting the number of iterations in the streaming setting could yield good convergence results. For instance, in [67] the authors mention that a simplified version of their algorithm that only executes one iteration per measurement still performs well. However, no analysis had been previously provided for such iteration-limited settings. The analysis presented in this thesis could potentially be applied to obtain similar convergence and accuracy results for these algorithms that currently lack analysis. Moreover, while the simulations of the LCA ODEs with the appropriate parameters (in particular $P = 1$) suggest that its convergence is slow, one would expect its time constant τ to be much faster than the time dl for a digital algorithm to complete one iteration. If so, the actual behavior of an analog implementation of the LCA would be closer to the ISTA simulated with $P = 10$ (assuming an analog constant 10 times smaller than its digital equivalent). Consequently, the results of this chapter support the idea that an analog implementation of the LCA has the potential to lead to a low-power solver for the real-time recovery of time-varying signals.

CHAPTER VI

SUMMARY AND FUTURE DIRECTIONS

The focus of this thesis was to determine what type of continuous-time systems can be used to solve nonsmooth optimization problems, with specific application to CS. The analysis developed has shown that a class of recurrent neural networks can be used to perform a wide class of complex optimization programs. Recurrent neural networks are characterized by distributed information-processing units and a matrix of feedback interconnections. The highly parallel structure of these networks makes them amenable to analog implementation, which designates them as a promising approach for real-time applications. While significant research has been put into providing performance guarantees for several classes of neural networks, this thesis has provided new results that broaden previous guarantees to a larger class. The neural networks in this extended class can be used to solve sparse recovery problems that arise in CS. In addition, this thesis has presented convergence and accuracy results for the recovery of time-varying sparse signals in both discrete and continuous time.

6.1 Summary

6.1.1 General performance guarantees

The first contribution of this thesis was the mathematical analysis of the class of LCA neural networks. The LCA is characterized by an interconnection matrix with a nontrivial nullspace and an activation function that can have flat regions and may be unbounded. It was shown under what conditions the fixed points of these networks correspond to critical points of the desired optimization problem. The convergence of these networks to their set of fixed points was proven by taking a Lyapunov-type approach. The support of the solution was shown to be recovered in finite time under a condition that is expected to hold with near certainty. A stronger convergence result was then established using a recently developed analytic tool called the nonsmooth Łojasiewicz inequality. With this approach, it was shown that under some mild assumptions on the activation function, which are often true in practice, the trajectories of both the internal states and outputs converge toward a unique fixed point, even when there exists a continuous subset of solutions to the optimization problem. Finally, the convergence rate of the LCA networks was shown to be exponential, and an analytic expression for the convergence speed was derived. All of these findings have expanded the state of knowledge in neural network analysis. In addition, they have shown that the LCA can be used to solve a wide class of optimization programs.

6.1.2 Application to CS

The second contribution of this thesis was to specialize the previous results to CS recovery. When the activation function is the soft-thresholding function, the LCA solves the ℓ_1-minimization program, which is the most famous objective function for sparse recovery. Some strong guarantees are associated with the solution of this optimization program when it is used to recover a sparse signal. Unfortunately, even the most efficient digital solvers cannot achieve real-time recovery for problems of large sizes. The analysis presented has shown that the LCA takes an efficient path towards recovering the sparse solution. In addition, an estimate for the convergence speed that only depends on the problem parameters and is independent of the input signal has been derived. The analysis uses the RIP and has yielded interesting parallels to existing digital algorithms. In particular, an analog to the S-step property and a less restrictive condition were shown to hold for the LCA for a number of measurements equivalent to those obtained for standard digital solvers. These findings have demonstrated that the LCA has the potential to be used as a real-time solver with potentially better scaling properties than its digital equivalents.

6.1.3 Tracking of time-varying signals

The last contribution of this thesis was to predict the convergence behavior of the LCA and ISTA, its discrete-time equivalent, when they are recovering a time-varying signal from streaming measurements. In this study, the measurements are continuously fed in input and the solvers are constrained to operate in real-time to avoid delays. This situation is of particular interest when the measurements are streaming at high rates or the computational resources are limited. While guarantees have been obtained for many solvers in the static case, few approaches have been developed for the dynamic case, and most lack performance guarantees. The findings of this thesis have provided upper bounds for the evolution of the error for both solvers over time. These bounds are essentially optimal and prove that the LCA and ISTA can be used to track time-varying signals from streaming measurements. Such theoretical results provide a solid foundation for the analysis of the many solvers that extend the classic ISTA. In addition, they show the potential of extending CS theory to a wider range of applications that involve dynamically evolving signals.

6.2 Comparative overview of results

In this section, a synthesis of some selected results from literature and from this thesis are presented in parallel to put the contributions of this thesis in perspective.

Many digital algorithms for sparse recovery have been studied for convergence and accuracy in the static case. For instance, combining the convergence result in [60] for the ISTA with the accuracy result for the ℓ_1-minimization program in [9] yields the following.

ISTA for static recovery: *If Φ satisfies the RIP, the threshold satisfies*

$$\lambda\sqrt{q} \gtrsim c_1 \left\|a^{\dagger}\right\|_2 + c_2\sigma$$

for some constants $c_0, c_1 \geq 0$, and the step size η is in the interval $\left(0, 2\left\|\Phi^T\Phi\right\|^{-1}\right)$, then ISTA converges with a linear rate; i.e., there exist $\overline{c} \in (0,1)$ and two constants $C_0, C_1 \geq 0$ such that, for all iterations $l \geq 0$,

$$\left\|a(l) - a^{\dagger}\right\|_2 \leq C_0\overline{c}^l + C_1.$$

The constant C_1 represents the optimal error $\left\|\hat{a}^\dagger - a^\dagger\right\|_2$ when solving (5) *and satisfies*

$$C_1 \leq C_2 \lambda \sqrt{q} + C_3 \sigma$$

for some $q \geq 0$ (which is typically on the order of S) and constants $C_2, C_3 \geq 0$.

This thesis has extended the previous result to the continuous-time LCA algorithm for ℓ_1-minimization. The combination of Theorems 6 and 8 with the accuracy result for ℓ_1-minimization in [9] implies the following result.

LCA for static recovery: *If Φ satisfies the RIP and the threshold satisfies*

$$\lambda \sqrt{q} \gtrsim c_2 \left\|a^\dagger\right\|_2 + c_3 \sigma$$

for some constants $c_2, c_3 \geq 0$, then the LCA converges with an exponential rate; i.e., there exist $\nu \in (0, 1)$ and two constants $C_4, C_5 \geq 0$ such that, for all time $t \geq 0$,

$$\left\|a(t) - a^\dagger\right\|_2 \leq C_4 e^{-\nu t} + C_1,$$

where C_1 is again the optimal error achieved when solving (5)*. In addition, the output of the LCA never contains more than q non-zero coefficients.*

Finally, Theorems 9 and 10 have provided similar guarantees in the case where the ISTA and LCA are driven by a time-varying signal.

ISTA for dynamic recovery: *If Φ satisfies the RIP with parameters $(s + 2q, \delta)$ for some $q \geq 0$, the threshold λ satisfies*

$$\lambda \sqrt{q} \gtrsim c_6 \beta + c_7 \sigma$$

for some constants $c_6, c_7, \beta \geq 0$ such that $\left\|a^\dagger(t)\right\|_2 \leq \beta$ for all $t \geq 0$, and the step size η is in the interval $\left(0, 2(1-\delta)^{-1}\right)$, then the ISTA converges with a linear rate; i.e., there exist $c \in (0, 1)$ and two constants $C_5, C_6 \geq 0$ such that, for all iterations $l \geq 0$,

$$\left\|a(l) - a^\dagger\right\|_2 \leq C_5 c^l + C_6.$$

The constant C_6 represents the optimal error if the ISTA had infinite iterations per frame to converge and satisfies

$$C_6 \leq C_7 \lambda \sqrt{q} + C_8 \sigma + C_9 \mu \, dl$$

for some constants $C_7, C_8, C_9, \mu \geq 0$ such that $\left\|\dot{a}^\dagger(t)\right\|_2 \leq \mu$ for all $t \geq 0$, and where dl is the time between two iterates. In addition, the output never contains more than q non-zero coefficients.

LCA for dynamic recovery: *If Φ satisfies the RIP with parameters $(s+q, \delta)$ for some $q \geq 0$ and the threshold satisfies*

$$\lambda \sqrt{q} \gtrsim c_8 \beta + c_9 \sigma$$

for some constants $c_8, c_9, \beta \geq 0$ such that $\left\|a^\dagger(t)\right\|_2 \leq \beta$ for all $t \geq 0$, then the LCA converges with an exponential rate; i.e., there exist $\nu \in (0,1)$ and two constants $C_7, C_8 \geq 0$ such that, for all time $t \geq 0$,

$$\left\|a(t) - a^\dagger\right\|_2 \leq C_7 e^{-\nu t} + C_8.$$

The constant C_8 represents the steady state error if the LCA converged for each frame and satisfies

$$C_8 \leq C_9 \lambda \sqrt{q} + C_{10} \sigma + C_{11} \tau \mu$$

for some constants $C_9, C_{10}, C_{11}, \mu \geq 0$ such that $\left\|\dot{a}^\dagger(t)\right\|_2 \leq \mu$ for all $t \geq 0$, and where τ is the time constant of the LCA. In addition, the output of the LCA never contains more than q non-zero coefficients.

The four results synthesized above bring to light the three areas where this thesis has provided significant contributions: the continuous-time recovery of static signals, and the discrete- and continuous-time recovery of dynamic signals. These results also show the parallels that exist between the discrete- and continuous-time algorithms, for which similar optimal errors are achieved but with linear and exponential rates of convergence, respectively. Finally, these results highlight the links between static and dynamic recovery, where an additional $\tau\mu$ or $\mu\ dl$ term appears to capture the tradeoff between the energy in the derivative of the target and the time constant of the corresponding solver.

6.3 Future directions

6.3.1 Discontinuous activation functions

The convergence results obtained in this thesis have assumed that the LCA activation function is continuous. However, several programs that arise in CS recovery necessitate a discontinuous activation function, including the ideal ℓ_0-minimization problem. Extending the results of this thesis to discontinuous activation functions would further broaden the tools available for neural network analysis and demonstrate the ability of the LCA to solve these complex optimization programs. A first step in this direction has been presented in Appendix D, which shows that the convergence result when the fixed points are not isolated still holds in the discontinuous case. In a similar way, it seems possible to extend most of the results in this thesis using Filippov's approach to approximate solutions of ODEs with discontinuous right-hand sides with absolutely continuous functions.

6.3.2 Matrix uncertainty

The results presented in thesis have assumed that the LCA ODE can be implemented accurately. Unfortunately, it is well-known that analog circuitry inevitably introduces errors in its various parameters. For instance, floating-gate transistors may be used to implement the weights of the matrices, but will likely suffer from inaccuracies due to the manufacturing or programming processes [59]. In addition, the sharp transition necessary in the soft-thresholding function may not be realizable in practice. Consequently, a study of the effect of errors in the various parameters would be valuable to understand the level of accuracy achievable in practical applications. Unfortunately, several difficulties arise when modeling these inaccuracies. In particular, if the interconnection matrix is no longer symmetric due to inaccuracies, an energy function for the network cannot be written, and new analytic tools must be developed to study the network convergence.

APPENDIX A

PROPERTIES OF THE LCA

This appendix provides several useful properties of the activation function $T_\lambda(\cdot)$, cost function $C(\cdot)$, and objective function $V(\cdot)$ under Assumptions 1 and/or 2.

A.1 Properties of the cost and activation functions

The lemma below presents two relationships satisfied by the state and output variables and a bound on the subgradient of the activation function when it satisfies Assumption 1.

Lemma 3. *If the activation function $T_\lambda(\cdot)$ satisfies Assumption 1, then for all $u_n \in \mathbb{R}$ and $a_n = T_\lambda(u_n)$ the following properties hold:*

$$\operatorname{sign}(u_n) = \operatorname{sign}(a_n), \tag{72}$$

$$|a_n|^2 \leq u_n a_n \leq |u_n|^2 . \tag{73}$$

Moreover, There exists $0 < \alpha$ such that for all non-constant nodes $n \in Z^c$ and for all $\zeta_n \in \partial T_\lambda(u_n)$

$$|\zeta_n| \leq \alpha. \tag{74}$$

Proof. Since $T_\lambda(\cdot)$ is locally Lipschitz on $\mathbb{R}$, Proposition 2.1.2 of [36] implies that there exists $\alpha > 0$ such that $|\zeta| \leq \alpha$ for all $\zeta \in \partial T_\lambda(u_n)$, so (74) holds.

Since $T_\lambda(0) = 0$ and $T_\lambda(\cdot)$ is nondecreasing on $\mathbb{R}$, $a_n = T_\lambda(u_n) \geq 0$ for all $u_n \geq 0$, and $a_n = T_\lambda(u_n) \leq 0$ for all $u_n \leq 0$, which proves (72). This fact also implies that

$$a_n u_n = \operatorname{sign}(a_n)\,|a_n|\,\operatorname{sign}(u_n)\,|u_n| = |a_n|\,|u_n|$$

for all $u_n \in \mathbb{R}$. Finally, condition (24) yields that, for all $u_n \in \mathbb{R}$,

$$|a_n|^2 \leq |a_n|\,|u_n| = a_n u_n \leq |u_n|^2 ,$$

which proves (73). □

Next, a method for building a cost function $C(\cdot)$ with useful properties and that satisfies the relationship (19) is presented. This lemma is a foundation for many results in Chapter 3.

Lemma 4. *If the activation function $T_\lambda(\cdot)$ satisfies Assumption 1, there exists a cost function $C(\cdot)$ that satisfies the relationship* (19) *and obeys*

1. *$C(\cdot)$ is locally Lipschitz continuous on $\mathbb{R}$,*

2. *$C(\cdot)$ is even on $\mathbb{R}$,*

3. *$C(\cdot)$ is nondecreasing on $\mathbb{R}^+$,*

4. *$C(0) = 0$,*

5. *$C(\cdot)$ is regular on $\mathbb{R}$.*

Proof. Since the activation function $T_\lambda(\cdot)$ is continuous and increasing on $\mathbb{R}$, it is surjective on $[0, a]$ for all $a \in T_\lambda(\mathbb{R})$ (where $T_\lambda(\mathbb{R})$ is the image of $\mathbb{R}$ by $T_\lambda(\cdot)$). In other words, for all $v \in (0, a)$ there exists $u \in \mathbb{R}$ such that $v = T_\lambda(u)$. As a consequence, a function $z^{-1}(\cdot)$ can be defined on $T_\lambda(\mathbb{R})$ as follows:
$\forall v \in T_\lambda(\mathbb{R})$, let $u \in \mathbb{R}$ such that $v = T_\lambda(u)$.

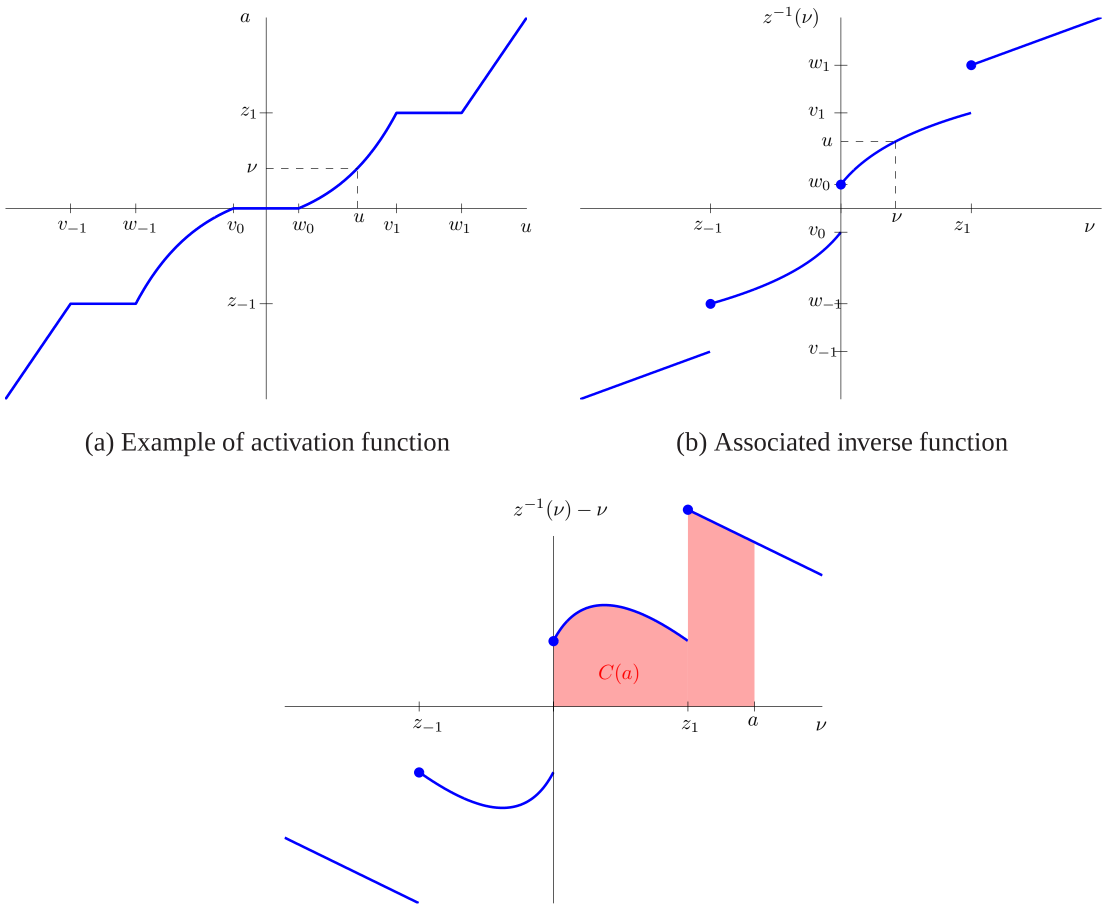
1. if $u \in \mathrm{Z}^c$ (which is the set of nodes that do not yield a constant output), then u is the unique point in $\mathbb{R}^N$ satisfying $v = T_\lambda(u)$, and $z^{-1}(v)$ is defined as $z^{-1}(v) = u$,

2. if $u \in \mathrm{Z}$, there exists $k \in \mathcal{K}$, such that $v = T_\lambda(u_k)$ for all $u_k \in [v_k, w_k]$. In that case, $z^{-1}(v)$ can be chosen to be $z^{-1}(v) = w_k$.

Figure 21b shows a visual example of how to construct $z^{-1}(\cdot)$ for the particular activation function plotted in Figure 21a.

Using this definition for $z^{-1}(\cdot)$, the following quantity is well-defined on $T_\lambda(\mathbb{R})$:

$$C(a) = \int_0^a z^{-1}(v) - v \, dv. \tag{75}$$



(a) Example of activation function

(b) Associated inverse function

(c) Associated cost function

Figure 21: Example of a generic activation function $T_\lambda(\cdot)$ satisfying Assumption 1, associated inverse function $z^{-1}(\cdot)$ and associated cost function $C(\cdot)$.

The function $C(\cdot)$ defined this way is locally Lipschitz on $T_\lambda(\mathbb{R})$ and differentiable for a.a. $a \in T_\lambda(\mathbb{R})$. Figure 21c shows the cost function associated with the activation function plotted in Figure 21a.

The following derivation shows that $C(\cdot)$ indeed satisfies (19). There are two cases.

1. At points a where $C(\cdot)$ is differentiable, the subgradient reduces to $\partial C(a) = \{C'(a)\}$, and the fundamental theorem of calculus applied to (75) yields

$$C'(a) = z^{-1}(a) - a = u - a.$$

   As a consequence, for such a, (19) holds.

2. For a point a where $C(\cdot)$ is not differentiable,

$$\exists k \in \mathcal{K}\ s.t.\ \forall u_k \in [v_k, w_k] \qquad a = T_\lambda(u_k).$$

Since $T_\lambda(\cdot)$ is continuous and strictly increasing on the intervals immediately adjacent to $[v_k, w_k]$, there exist two constants $\delta_1 > 0$ and $\delta_2 > 0$ such that $[w_k, w_k + \delta_1] \subset \mathbb{R}\backslash \bigcup_{k' \in \mathcal{K}} [v_{k'}, w_{k'}]$ and $[v_k - \delta_2, v_k] \subset \mathbb{R}\backslash \bigcup_{k' \in \mathcal{K}} [v_{k'}, w_{k'}]$. In other words, for δ_1 and δ_2 sufficiently small, the activation function is not constant on the intervals $[w_k, w_k + \delta_1]$ and $[v_k - \delta_2, v_k]$.

Letting $\{w_m^+\}_{m \geq 0}$ be a sequence of points in $[w_k, w_k + \delta]$ that converges to w_k, the sequence $\{a_m^+\}_{m \geq 0} = \{T_\lambda(w_m^+)\}_{m \geq 0}$ converges to $a = T_\lambda(w_k)$ by continuity of $T_\lambda(\cdot)$. Similarly, letting $\{v_m^-\}_{m \geq 0}$ be a sequence of points in $[v_k - \delta, v_k]$ that converges to v_k, the sequence $\{a_m^-\}_{m \geq 0} = \{T_\lambda(v_m^-)\}_{m \geq 0}$ converges to $a = T_\lambda(v_k)$. Using the fundamental theorem of calculus and the fact that $C(\cdot)$ is differentiable at a_m^+ and a_m^- for all $m \geq 0$ (by construction) yields, $\forall t \geq 0$ sufficiently small,

$$\begin{aligned} \frac{C(a_m^+ + t) - C(a_m^+)}{t} &= \frac{1}{t} \int_{a_m^+}^{a_m^+ + t} z^{-1}(v) - v \, dv \\ &\xrightarrow[t \to 0]{} z^{-1}(a_m^+) - a_m^+ = w_m^+ - a_m^+ \\ &\xrightarrow[\substack{t \to 0 \\ m \to \infty}]{} w_k - a \end{aligned}$$

and

$$\begin{aligned} \frac{C(a_m^- - t) - C(a_m^-)}{t} &= -\frac{1}{t} \int_{a_m^- - t}^{a_m^-} z^{-1}(v) - v \, dv \\ &\xrightarrow[t \to 0]{} -z^{-1}(a_m^-) + a_m^- = -v_m^- + a_m^- \\ &\xrightarrow[\substack{t \to 0 \\ m \to \infty}]{} -v_k + a. \end{aligned}$$

Using the definition of the subgradient in Section 2.3.1, it can be easily seen that for all $\xi \in \partial C(a)$, the generalized directional derivative satisfies $C^\circ(a; 1) = w_k - a \geq 1\xi$ and $C^\circ(a; -1) = -v_k + a \geq -1\xi$. As a consequence, $\xi \in [v_k - a, w_k - a]$ and thus

$$\partial C(a) = [v_k - a, w_k - a].$$

This equality proves that indeed $u_k - a \in \partial C(a)$ for all $u_k \in [v_k, w_k]$.

As a consequence, the derivation above shows that in every case $u - a \in \partial C(a)$ holds for all $a \in T_\lambda(\mathbb{R})$, which proves that (19) holds.

By inspection of (75), it is immediate that $C(0) = 0$.

It is also easy to check that, since $T_\lambda(\cdot)$ is nondecreasing and odd, the function $z^{-1}(\cdot)$ defined above is nondecreasing and odd. As a consequence, for all $a \in T_\lambda(\mathbb{R})$, the following holds:

$$\begin{aligned} C(-a) &= \int_0^{-a} \left(z^{-1}(v) - v\right)\, dv \\ &= \int_0^{a} \left(z^{-1}(-v) + v\right)\, (-dv) \\ &= \int_0^{a} \left(z^{-1}(v) - v\right)\, dv = C(a). \end{aligned}$$

This computation proves that $C(\cdot)$ is even.

Finally, for all $v \in T_\lambda(\mathbb{R})$ such that $v \geq 0$, letting $u = z^{-1}(v)$, condition (24) implies that $z^{-1}(v) - v = u - z(u) \geq 0$. This fact proves that $C(\cdot)$ is nondecreasing on $\mathbb{R}^+$ by the positivity of the integral. As a consequence, $C(a) \geq C(0) = 0$ for all $a \in T_\lambda(\mathbb{R}^+)$ and, by symmetry, for all $a \in T_\lambda(\mathbb{R})$.

To show that $C(\cdot)$ is regular, it suffices to notice that the usual one-sided derivative exists for all $a \in T_\lambda(\mathbb{R})$. There are two cases.

1. For point a where $C(\cdot)$ is differentiable, the result is obvious.

2. For points a where $C(\cdot)$ is not differentiable, it follows by construction of $z^{-1}(\cdot)$ that $z^{-1}(a) = w_k$ for some $k \in \mathcal{K}$. As a consequence, the right-sided derivative exists and is the limit for $t > 0$ sufficiently small of

   $$\frac{C(a+t) - C(a)}{t} = \frac{1}{t} \int_a^{a+t} z^{-1}(v) - v\, dv.$$

   By the fundamental theorem of calculus, this quantity converges to $z^{-1}(a) - a = u - a$ as $t \to 0$. So $C'(a; 1) = C^\circ(a; 1)$. For the left-sided derivative, taking $t > 0$

sufficiently small, the following integral can be computed

$$\begin{aligned}\frac{C(a-t)-C(a)}{t} &= -\frac{1}{t}\int_{a-t}^{a} z^{-1}(v) - v\, dv \\ &= -\int_{a-t}^{a} z_0^{-1}(v) - v\, dv,\end{aligned}$$

where $z_0^{-1}(v) = z^{-1}(v)$ everywhere except at a where the function is defined to be $z_0^{-1}(a) = v_k$. The two integrals are equal because they only differ at one point. The function $z_0^{-1}(\cdot)$ is now continuous on $[a-t, a]$ and thus, the one-sided integral exists and can be computed as $t \to 0$ to get $C'(a; -1) = -v_n + a_n = C^\circ(a; -1)$.

As a consequence, $C(\cdot)$ is regular on $T_\lambda(\mathbb{R})$. □

The final lemma below gives a set of properties for variables that are useful in the study of the LCA dynamics. These variables were defined in (36), and their definition is given again below:

$$\widetilde{u}_n(t) = u_n(t) - u_n^*,$$

$$\widetilde{a}_n(t) = a_n(t) - a_n^* = T_\lambda(\widetilde{u}_n(t) + u_n^*) - T_\lambda(u_n^*).$$

Intuitively, these variables measure the distance of the states and outputs from any arbitrary fixed points u^* and $a^* = T_\lambda(u^*)$ of (8).

Lemma 5. *If the activation function $T_\lambda(\cdot)$ satisfies Assumption 1, then the set of variables $\widetilde{u}$ and $\widetilde{a}$ defined in* (36) *satisfies the following properties:*

(i) $\text{sign}(\widetilde{a}_n) = \text{sign}(\widetilde{u}_n)$,

(ii) $|\widetilde{a}_n| \leq \alpha\, |\widetilde{u}_n|$,

(iii) $\widetilde{a}_{\mathcal{T}}^T \widetilde{a}_{\mathcal{T}} \leq \alpha \widetilde{u}_{\mathcal{T}}^T \widetilde{a}_{\mathcal{T}} \leq \alpha^2 \widetilde{u}_{\mathcal{T}}^T \widetilde{u}_{\mathcal{T}}$ *for any $\mathcal{T}$ (in particular for $\mathcal{T} = \widetilde{\Gamma}$).*

Proof. Each of the three properties will be treated separately.

(i) For any $\widetilde{u}_n \in \mathbb{R}$, let $\mathsf{s}_n = \operatorname{sign}(\widetilde{u}_n)$. Since the activation function is nondecreasing and odd (*i.e.*, $T_\lambda(-u_n) = -T_\lambda(u_n)$),

$$\begin{aligned}
\mathsf{s}_n = \operatorname{sign}(\widetilde{u}_n) &\Rightarrow 0 \leq \mathsf{s}_n\widetilde{u}_n \\
&\Rightarrow \mathsf{s}_n u_n^* \leq \mathsf{s}_n\widetilde{u}_n + \mathsf{s}_n u_n^* \\
&\Rightarrow T_\lambda\left(\mathsf{s}_n u_n^*\right) \leq T_\lambda\left(\mathsf{s}_n\widetilde{u}_n + \mathsf{s}_n u_n^*\right) \qquad \text{(since } T_\lambda(\cdot) \text{ is nondecreasing)} \\
&\Rightarrow \mathsf{s}_n T_\lambda\left(u_n^*\right) \leq \mathsf{s}_n T_\lambda\left(\widetilde{u}_n + u_n^*\right) \qquad \text{(since } T_\lambda(\cdot) \text{ is odd)} \\
&\Rightarrow 0 \leq \mathsf{s}_n\left[T_\lambda\left(\widetilde{u}_n + u_n^*\right) - T_\lambda\left(u_n^*\right)\right] \\
&\Rightarrow 0 \leq \mathsf{s}_n\widetilde{a}_n \\
&\Rightarrow \operatorname{sign}(\widetilde{a}_n) = \mathsf{s}_n = \operatorname{sign}(\widetilde{u}_n).
\end{aligned}$$

(ii) Since $T_\lambda(\cdot)$ is locally Lipschitz on $\mathbb{R}$, the mean-value theorem for nonsmooth functions (Theorem 2.3.7 in [36]) applies and states that there exist $\overline{\overline{u}}_n \in (\widetilde{u}_n + u_n^*, u_n^*)$ and $\overline{\overline{\zeta}}_n \in \partial T_\lambda\left(\overline{\overline{u}}_n\right)$ such that

$$T_\lambda(\widetilde{u}_n + u_n^*) - T_\lambda(u_n^*) = \overline{\overline{\zeta}}_n\left(\widetilde{u}_n + u_n^* - u_n^*\right) = \overline{\overline{\zeta}}_n\widetilde{u}_n.$$

Applying the bound (74) on the subgradients of $T_\lambda(\cdot)$ on $\mathbb{R}$ yields

$$\left|\widetilde{a}_n\right| = \left|T_\lambda(\widetilde{u}_n + u_n^*) - T_\lambda(u_n^*)\right| = \left|\overline{\overline{\zeta}}_n\widetilde{u}_n\right| \leq \alpha\left|\widetilde{u}_n\right|.$$

(iii) Properties (i) and (ii) imply the final property:

$$\begin{aligned}
\widetilde{a}_{\mathcal{T}}^T\widetilde{a}_{\mathcal{T}} &= \sum_{n\in\mathcal{T}}\widetilde{a}_n\widetilde{a}_n = \sum_{n\in\mathcal{T}}\left|\widetilde{a}_n\right|\left|\widetilde{a}_n\right| \\
&\leq \sum_{n\in\mathcal{T}}\alpha\left|\widetilde{u}_n\right|\left|\widetilde{a}_n\right| = \alpha\sum_{n\in\mathcal{T}}\widetilde{u}_n\widetilde{a}_n = \alpha\widetilde{u}_{\mathcal{T}}^T\widetilde{a}_{\mathcal{T}} \\
&\leq \sum_{n\in\mathcal{T}}\alpha\left|\widetilde{u}_n\right|\alpha\left|\widetilde{u}_n\right| = \alpha^2\sum_{n\in\mathcal{T}}\widetilde{u}_n\widetilde{u}_n = \alpha^2\widetilde{u}_{\mathcal{T}}^T\widetilde{u}_{\mathcal{T}}.
\end{aligned}$$

□

A.2 Time derivative of the objective function

This section contains results on the evolution of the objective function with respect to time.

Lemma 6. *For an activation function satisfying Assumption 1 and a cost function constructed as in Lemma 4, the objective $V(a(\cdot))$ in* (6) *is continuous and regular on $\mathbb{R}^+$. In addition, its time derivative satisfies for a.a. $t \geq 0$ and for any $\zeta_n \in \partial T_\lambda(u_n(t))$ the two equalities*

$$\dot{V}(a(t)) = -\sum_{n \notin Z} \zeta_n \left|\dot{u}_n(t)\right|^2 , \tag{76}$$

$$\dot{V}(a(t)) = -\sum_{n \notin Z} \frac{1}{\zeta_n} \left|\dot{a}_n(t)\right|^2 . \tag{77}$$

Proof. Since the activation function $T_\lambda(\cdot)$ is locally Lipschitz on $\mathbb{R}$, it is differentiable almost everywhere. For constant nodes $n \in \mathrm{Z}$, the output is constant and thus $\dot{a}_n(t) = 0$. Using the chain rule (12), non-constant nodes for $n \notin \mathrm{Z}$ satisfy $\dot{a}_n(t) = \zeta_n \dot{u}_n(t)$ for any $\zeta_n \in \partial T_\lambda(u_n(t))$. Since $C(\cdot)$ is regular on $T_\lambda(\mathbb{R})$, $V(a(\cdot))$ is regular for all $t \geq 0$ and by the chain rule (14), any element in $\partial V(a(t))$ can be used to compute the time derivative $\dot{V}(a(\cdot))$ along the LCA trajectories. In particular, by Lemma 1, using $-\dot{u}(t) \in \partial V(a(t))$ yields, for a.a. $t \geq 0$,

$$\begin{aligned}
\dot{V}(a(t)) &= -\dot{u}(t)^T \dot{a}(t) \\
&= -\sum_{n=1}^{N} \dot{u}_n(t) \dot{a}_n(t) \\
&= -\sum_{n \notin \mathrm{Z}} \zeta_n \left|\dot{u}_n(t)\right|^2 \\
&= -\sum_{n \notin \mathrm{Z}} \frac{1}{\zeta_n} \left|\dot{a}_n(t)\right|^2 ,
\end{aligned}$$

where the last inequality holds since $\zeta_n > 0$ for all $n \notin \mathrm{Z}$. □

Using the expression for the derivative of $V(\cdot)$ with respect to time, it is straightforward to show that the objective function $V(\cdot)$ is decreasing and converges to a positive value.

Corollary 2. *For an activation function satisfying Assumption 1 and a cost function constructed as in Lemma 4, the objective $V(a(\cdot))$ in* (6) *is decreasing for all $t \geq 0$ and converges to a limit $V^* \geq 0$ as t goes to infinity.*

Proof. Equation (77) in Lemma 6 states that a.a. $t \geq 0$, for any $\zeta_n \in \partial T_\lambda(u_n(t))$

$$\frac{dV(a(t))}{dt} = -\sum_{n \notin Z} \frac{1}{\zeta_n} |\dot{a}_n(t)|^2 ,$$

with $\zeta_n > 0$ for all $n \notin Z$ (corresponding to non-constant outputs by definition). As a consequence

$$\frac{dV(a(t))}{dt} \leq 0, \qquad \text{for a.a. } t \geq 0$$

This inequality shows that since $C(\cdot)$ in continuous on $T_\lambda(\mathbb{R})$ and lower-bounded by zero by Lemma 4, the objective function $V(a(t))$ is continuous, bounded below by zero, and nonincreasing for all $t \geq 0$. Thus, $V(a(t))$ converges to a constant value $V^* \geq 0$ as t goes to infinity (note: the continuity of $V(a(t))$ is essential for this result to hold). □

A.3 Boundedness of the objective, the states and outputs

The following result proves that, while the activation function may be constant on many intervals and unbounded, the state and output are guaranteed to remain bounded throughout convergence.

Lemma 7. *For an activation function satisfying Assumption 1 and a cost function constructed as in Lemma 4, the objective $V(a(\cdot))$ in* (6) *satisfies for all $t \geq 0$*

$$0 \leq V(a(t)) \leq V(a(0)).$$

In addition, the output $a(t)$ and state variables $u(t)$ of the system (8) *are bounded $\forall t \geq 0$.*

Proof. From (76) and the fact that $\zeta_n > 0$ for all $n \notin Z$, it can be concluded that $\dot{V}(a(t)) \leq 0$ for a.a. $t \geq 0$. As a consequence, $\forall t > 0$

$$V(a(t)) - V(a(0)) = \int_0^t \dot{V}(a(s))ds,$$

and since $0 < t$ and $\dot{V}(a(s)) \leq 0$ for a.a. $s \in (0, t)$, by the positivity of the integral, it can be seen that $V(a(t)) \leq V(a(0))$ for all $t \geq 0$.

In the following, the boundedness of the state $u(t)$ is shown. For this proof, it is first shown that both $||\Phi a(t)||_2$ and $||\Phi u(t)||_2$ are bounded $\forall t \geq 0$. By Lemma 4, $C(a_n) \geq 0$ for all $a_n \in \mathbb{R}$. Thus, for all $t \geq 0$,

$$0 \leq \frac{1}{2} ||y - \Phi a(t)||_2^2 \leq V(a(t)) \leq V(a(0)).$$

The triangle inequality yields

$$||\Phi a(t)||_2 - ||y||_2 \leq \sqrt{2V(a(0))}.$$

This inequality shows that $||\Phi a(t)||_2$ is bounded $\forall t \geq 0$. As a consequence, there must exist a constant $C_1 \geq 0$ such that, $\forall t \geq 0$,

$$\left\|(I - \Phi\Phi^T)\Phi a(t) + \Phi^T y\right\|_2 \leq \sigma_1 ||\Phi a(t)||_2 + \left\|\Phi^T y\right\|_2 \leq C_1,$$

where $\sigma_1 \geq 0$ is the largest eigenvalue of the interconnection matrix $W = \Phi\Phi^T - I$. This inequality implies that $||\Phi u(t)||_2$ is also bounded for $t \geq 0$. Indeed, using the Cauchy-Schwartz inequality, the time-derivative of $1/2\, ||\Phi u(t)||_2^2$ satisfies

$$\begin{aligned} \frac{d}{dt}\frac{1}{2} ||\Phi u(t)||_2^2 &= u(t)\Phi^T\Phi\dot{u}(t) \\ &= u^T\Phi^T\Phi(-u(t) + a(t) - \Phi^T\Phi a(t) + \Phi^T y) \\ &\leq -\, ||\Phi u(t)||_2^2 + ||\Phi u(t)||_2\, C_1 \\ &\leq -\, ||\Phi u(t)||_2 \left(||\Phi u(t)||_2 - C_1\right). \end{aligned}$$

As a consequence, the set $\left\{u \in \mathbb{R}^N\ s.t.\ ||\Phi u||_2 \leq C_1\right\}$ is attractive, and by continuity, $||\Phi u(t)||_2$ is bounded $\forall t \geq 0$. It is not possible to conclude directly that $||u(t)||_2$ is bounded because the matrix Φ may be singular. Any vector u in its nullspace can grow unbounded while $||\Phi u||_2$ remains bounded. However, $u(t)$ can be decomposed into its component $u_1(t)$ that lies in the nullspace of Φ and its component $u_2(t)$ that lies in the range of Φ^T. These two vectors are orthogonal (this property comes from the singular value decomposition of Φ), and the following shows that each of them is bounded. Since $u_1(t)$ is in the nullspace of

Φ, $\Phi u(t) = \Phi u_2(t)$. Since $u_2(t)$ is in the range of Φ^T, there exists $x_2(t) \in \mathbb{R}^M$ such that $u_2(t) = \Phi^T x_2(t)$. Using the Cauchy-Schwartz inequality yields

$$\begin{aligned}
\|x_2(t)\|_2 \, \|\Phi u(t)\|_2 &\geq x_2(t)^T \Phi u(t) \\
&= x_2(t)^T \Phi u_2(t) \\
&= x_2(t)^T \Phi \Phi^T x_2(t) \\
&\geq \sigma_2^2 \, \|x_2(t)\|_2^2 \, ,
\end{aligned}$$

where $\sigma_2 > 0$ is the smallest singular value of Φ^T restricted to its range (so it is strictly positive). Letting σ_3 be the largest singular value of Φ^T,

$$\|u_2(t)\|_2 = \left\|\Phi^T x_2(t)\right\|_2 \leq \sigma_3 \, \|x_2\|_2 \leq \sigma_3 \sigma_2^{-2} \, \|\Phi u(t)\|_2 \, .$$

The inequality above shows that $\|u_2(t)\|_2$ is bounded, since $\|\Phi u(t)\|_2$ is bounded. Moreover, using the fact that $\Phi u_1(t) = 0$, the time-derivative of $1/2 \, \|u_1(t)\|_2^2$ can be computed as follows:

$$\begin{aligned}
\frac{d}{dt} \frac{1}{2} \, \|u_1(t)\|_2^2 &= u_1(t)^T \dot{u}_1(t) \\
&= u_1(t)^T \left(-u(t) + a(t) + \Phi^T y - \Phi^T \Phi a(t)\right)_1 \\
&= -u_1(t)^T u_1(t) + u_1(t)^T a_1(t) \leq 0,
\end{aligned}$$

where the last inequality follows from (73). As a consequence, $\|u_1(t)\|_2$ is also bounded $\forall t \geq 0$. The two bounds obtained prove that $\|u(t)\|_2 \leq \|u_1(t)\|_2 + \|u_2(t)\|_2$ is bounded $\forall t \geq 0$.

Finally, since $T_\lambda(\cdot)$ is continuous on $\mathbb{R}$ and $\|u(t)\|_2$ is bounded $\forall t \geq 0$, $\|T_\lambda(u(t))\|_2$ is bounded $\forall t \geq 0$, which means that $\|a(t)\|_2$ is bounded $\forall t \geq 0$. □

The following corollary demonstrates that under certain conditions of Assumptions 1 and 2, the subgradients of the activation function at non-constant nodes are lower-bounded by a strictly positive constant.

Corollary 3. *If the activation function $T_\lambda(\cdot)$ in* (23) *satisfies conditions* (24) *and* (26)*, then there exist two constants $0 < \beta \leq \alpha$ such that for all non-constant nodes $n \in Z^c$, $\forall t \geq 0$, and $\forall \zeta_n \in \partial T_\lambda(u_n(t))$, the following holds:*

$$\beta \leq \zeta_n \leq \alpha.$$

Proof. The proof that (24) implies the existence of α was done in Lemma 3. By Lemma 7, there exists a bound $\mu > 0$ such that $\|u_n(t)\|_2 \leq \mu$ for all $t \geq 0$. Since by Assumption 1, there exists only a finite number of intervals in $\mathcal{Z}$ (where $T_\lambda(\cdot)$ is constant) on any bounded interval of $\mathbb{R}$, there must also exist only a finite number of open intervals in $[0,\mu]\backslash\mathcal{Z}$. Each of these open intervals $\mathcal{U} \subset [0,\mu]\backslash\mathcal{Z}$ is obviously bounded, and thus (26) guarantees the existence of a constant $\beta_U > 0$ such that $\zeta_n \geq \beta_U$ for all $u_n \in \mathcal{U}$ and all $\zeta_n \in \partial T_\lambda(u_n)$. As a consequence, since there is only a finite number of these intervals $\mathcal{U}$, the minimum β over all the constants β_U exists, and it is guaranteed that $\beta > 0$. Thus, the first part of the corollary's inequality holds for any $u_n \in [0,\mu]\backslash\mathcal{Z}$, and as a result for all $t \geq 0$. □

A.4 Subanalicity of the objective

The final part of this appendix proves that if the activation function $T_\lambda(\cdot)$ is subanalytic, then the associated cost function and objective function are also subanalytic. The proof only uses the facts that $T_\lambda(\cdot)$ is subanalytic and bounded on bounded intervals, but is stated under the stronger conditions in Assumptions 1 and 2.

Lemma 8. *If the activation function $T_\lambda(\cdot)$ satisfies Assumptions 1 and 2, then the associated cost function $C(\cdot)$ constructed as in Lemma 4 and the objective function $V(a(\cdot))$ in* (6) *are subanalytic.*

Proof. From Assumption 2, $T_\lambda(\cdot)$ is subanalytic. As a consequence, by the definition of a subanalytic function in Section 2.4.1, every point $(u,a) \in \mathbb{R} \times \mathbb{R}$ admits a neighborhood $\mathcal{B}_{\delta_1}(u) \times \mathcal{B}_{\delta_2}(a)$ for some $\delta_1, \delta_2 > 0$, such that

$$(u,a) \in \mathrm{Graf}\, T_\lambda \cap (\mathcal{B}_{\delta_1}(u) \times \mathcal{B}_{\delta_2}(a)) \quad \Leftrightarrow \quad (u,a) \in A,$$

where A is a bounded semianalytic subset of $\mathbb{R} \times \mathbb{R}$. Furthermore, since $T_\lambda(\cdot)$ is locally Lipschitz, it is locally bounded, and so for any $(u, a) \in \operatorname{Graf} T_\lambda(\mathbb{R})$, there exists a bounded semianalytic set $B \subset \mathbb{R}^m$, with $m \geq 1$, a finite stratification $\{I_i, J_i, B_i\}_{i=1,\dots,p} \subset [0, u] \times [0, a] \times B$ and analytic functions $f_i(\cdot, \cdot, \cdot) : I_i \times J_i \times B_i \to \mathbb{R}$ for $i = 1, \dots, p$ such that

$$f_i(u, a, y) = 0, \qquad \forall (u, a, y) \in (\operatorname{Graf} T_\lambda \times B) \cap (I_i \times J_i \times B_i)\,.$$

Since each $f_i(\cdot)$ is analytic, by the implicit function theorem, there exist a finite number of subsets $\left\{J'_{ij} \times B'_{ij} \to I'_{ij}\right\}_{j=1,\dots,q} \subset J_i \times B_i \to I_i$ and analytic functions $g_{ij}(\cdot, \cdot) : J'_{ij} \times B'_{ij} \to I'_{ij}$ that satisfy

$$u = g_{ij}(a, y) \;\;\Leftrightarrow\;\; f_i(u, a, y) = 0, \qquad \forall (u, a, y) \in (\operatorname{Graf} T_\lambda \times B) \cap \left(I'_{ij} \times J'_{ij} \times B'_{ij}\right).$$

As a consequence, for any $(a, c) \in \operatorname{Graf} c$ with $|a| \leq a_0$,

$$\begin{aligned} c = C(a) &= \int_0^a z^{-1}(v) - v \, dv \\ &= \sum_{i=1}^{p} \int_{J_i} z^{-1}(v) - v \, dv \\ &= \sum_{i=1}^{p} \sum_{j=1}^{q} \int_{J'_{ij} \times B'_{ij}} g_{ij}(v, y) - v \, dv \, dy. \end{aligned}$$

The derivation above uses the fact that $(z^{-1}(v), v) \in \operatorname{Graf} T_\lambda$ by construction of $z^{-1}(\cdot)$. Since each $g_{ij}(J'_{ij} \times B'_{ij}) = I'_{ij}$ is bounded for any $|a| < a_0$, then the above expression is also bounded by some constant $c_0 > 0$. As a consequence,

$$\begin{aligned} &(a, c) \in \operatorname{Graf} c \cup [-a_0, a_0] \times [0, c_0] \qquad \Leftrightarrow \\ &\qquad \left(a \;\; , \quad \sum_{i=1}^{p} \sum_{j=1}^{q} \int_{J'_{ij} \times B'_{ij}} g_{ij}(v, y) - v \, dv \, dy\right) \cap [-a_0, a_0] \times [0, c_0]. \end{aligned}$$

Since the above expression only contains bounded semianalytic sets and analytic functions, the set $\operatorname{Graf} c$ is subanalytic, and as a result $C(\cdot)$ is subanalytic.

To show that $V(\cdot)$ is subanalytic, its graph is expressed as the projection onto the first and last component of the set

$$\Big\{(a, v_1, b, v_2, v) \in \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N \times \mathbb{R} \times \mathbb{R} \;\; s.t. \\ \frac{1}{2}\,\|y - \Phi a\|_2 = v_1,\; \lambda \sum_{n=1}^{N} C(b_n) = v_2,\; a = a_2,\; v = v_1 + v_2\Big\} \\ = (\mathrm{Graf} F_1 \times \mathrm{Graf} F_2 \times R) \bigcap \Big\{a, a_2, v_1, v_2, v \in \mathbb{R}^{2N+3} \;\; s.t.\; a = a_2,\; v = v_1 + v_2\Big\},$$

where $F_1(a) = \dfrac{1}{2}\,\|y - \Phi a\|_2^2$ and $F_2(b) = \sum_{n=1}^{N} C(b_n)$ are subanalytic and locally bounded. The projection theorem of [44, Th 2.3] implies that $\mathrm{Graf} V$ is a subanalytic set. As a consequence, $V(\cdot)$ is subanalytic on $\mathbb{R}^N$ □

APPENDIX B

DIFFERENTIAL EQUATIONS

In this appendix, several properties that apply to solutions of differential equations are presented. They are then applied to the LCA trajectories, which are solutions of (8).

B.1 Gronwall's Lemma

The following is a fundamental result for solutions of differential equations, known as Gronwall's Lemma. It can take two forms (one with an equality and one with an inequality) that have been combined in the lemma below. This result is fundamental for the analyses of the LCA and is applied on multiple occasions in the proofs.

Lemma 9 (Gronwall's Lemma)**.** *Let $a \in \mathbb{R}$. If $x(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ satisfies*

$$\begin{aligned} \frac{dx(t)}{dt} &\leqq -ax(t) + F(x(t), t), \\ x(0) &= x_0, \end{aligned}$$

then the following holds $\forall t \geq 0$:

$$x(t) \leqq e^{-at} x_0 + e^{-at} \int_0^t e^{as} F(x(s), s) ds. \tag{78}$$

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

$$\begin{aligned} \frac{d}{dt}\left(e^{at} x(t)\right) &= a e^{at} x(t) + e^{at} \frac{d}{dt}\left(x(t)\right) \\ &\leqq a e^{at} x(t) + e^{at}\left(-ax(t) + F(x(t), t)\right) \\ &\leqq e^{at} F(x(t), t). \end{aligned}$$

Integrating on both sides from 0 to t (using the positivity of the integral) yields

$$e^{at} x(t) - x(0) \leqq \int_0^t e^{as} F(x(s), s) ds.$$

As a consequence,

$$x(t) \leqq e^{-at}x_0 + e^{-at}\int_0^t e^{as}F(x(s), s)ds. \qquad \square$$

A proof similar to that of Gronwall's Lemma yields the following result, which applies to a linear system of ODEs with a constant matrix.

Lemma 10. *Let $x(\cdot) : \mathbb{R}^+ \to \mathbb{R}^N$, A a symmetric matrix in $\mathbb{R}^{N\times N}$ and $b(\cdot,\cdot) : \mathbb{R}^N \times \mathbb{R}^+ \to \mathbb{R}^N$. The solution to the system of ODE*

$$\begin{cases} \dot{x}(t) = Ax(t) + b(x(t), t) \\ x(t_k) = x^{t_k} \end{cases} \tag{79}$$

is $\forall t \geq t_k$

$$x(t) = e^{A(t-t_k)}x^{t_k} + e^{At}\int_{t_k}^t e^{-Av}b(v)dv. \tag{80}$$

In the case where $b(\cdot,\cdot) = b$ is a constant vector in $\mathbb{R}^N$, the solution can also be written as

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

In the above expression, $\left(I - e^{At}\right)A^{-1}$ is well-defined even when the matrix A is singular. To illustrate this fact, the matrix A can be expressed as $A = P\Lambda P^{-1}$, where Λ is a diagonal matrix with diagonal elements λ_i; *i.e.* $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Using this decomposition yields

$$\begin{aligned} \left(I - e^{At}\right)A^{-1} &= P\left(I - e^{\Lambda t}\right)\Lambda^{-1}P^{-1} \\ &= P\operatorname{diag}\!\left(\left(1 - e^{\lambda_1 t}\right)\lambda_1^{-1}, \ldots, \left(1 - e^{\lambda_n t}\right)\lambda_n^{-1}\right)P^{-1} \end{aligned}$$

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

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

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

B.2 LCA trajectories for ℓ_1-minimization

In this section, the activation function is assumed to be the soft-thresholding function, so that the LCA solves the ℓ_1-minimization program (5). In this case, the LCA is a type of switched linear system [81], where the dynamics are governed by a linear ODE that changes every time a node crosses threshold (*i.e.*, moves into or out of the active set). Between switching times, the active set Γ is fixed, $\dot{a}_\Gamma(t) = \dot{u}_\Gamma(t)$, and the ODE (8) can be partially decoupled as follows:

$$\dot{a}_\Gamma(t) = -\Phi_\Gamma^T \Phi_\Gamma a_\Gamma(t) + \Phi_\Gamma^T y - \lambda \mathsf{s}_\Gamma(t), \tag{81}$$

$$\dot{u}_{\Gamma^c}(t) = -u_{\Gamma^c}(t) - \Phi_{\Gamma^c}^T \Phi_\Gamma a_\Gamma(t) + \Phi_{\Gamma^c}^T y. \tag{82}$$

Applying the results in Lemma 10, the solution to (81) on the active set Γ between switching times t_k and t_{k+1} is given by

$$a_\Gamma(t) = e^{-A(t-t_k)} a_\Gamma^{t_k} + \left(I - e^{-A(t-t_k)}\right) A^{-1} \left(\Phi_\Gamma^T y - \lambda \mathsf{s}_\Gamma\right), \tag{83}$$

where $A = \Phi_\Gamma^T \Phi_\Gamma$ and $a_\Gamma^{t_k} = a_\Gamma(t_k)$. In the case where $\Phi_\Gamma^T \Phi_\Gamma$ is nonsingular, the point

$$a_\Gamma^\infty = A^{-1} \left(\Phi_\Gamma^T y - \lambda \mathsf{s}_\Gamma\right)$$

can be viewed as the steady state of (81) if the active set and sign vector s_Γ remained unchanged until convergence. The points a_Γ^∞ play a key role in the analysis of the LCA in Chapter 4 and in the following (see Lemma 11).

The solution to the linear ODE (82) on the inactive set Γ^c between switching times t_k and t_{k+1} is given by

$$u_{\Gamma^c}(t) = e^{-(t-t_k)} u_{\Gamma^c}^{t_k} + e^{-t} \int_{t_k}^{t} e^{\nu} \rho_{\Gamma^c}(\nu) d\nu, \tag{84}$$

where $\rho_{\Gamma^c}(\nu) = \Phi_{\Gamma^c}^T \left(y - \Phi_\Gamma a_\Gamma(\nu)\right)$ and $u_{\Gamma^c}^{t_k} = u_{\Gamma^c}(t_k)$. Letting t go to infinity in equations (83) and (84), the fixed point a^*, which is supported on the final active set Γ_*, must satisfy

$$a_{\Gamma_*}^* = \left(\Phi_{\Gamma_*}^T \Phi_{\Gamma_*}\right)^{-1} \left(\Phi_{\Gamma_*}^T y - \lambda \mathsf{s}_{\Gamma_*}\right),$$

$$u_{\Gamma_*^c}^* = \Phi_{\Gamma_*^c}^T \left(y - \Phi_{\Gamma_*} a_{\Gamma_*}^*\right).$$

Since a node j is in the inactive set Γ_*^c if and only if $\left|u_j\right| \leq \lambda$, the two equations above translate immediately to

$$\begin{aligned} a^*_{\Gamma_*} &= \left(\Phi^T_{\Gamma_*}\Phi_{\Gamma_*}\right)^{-1}\left(\Phi^T_{\Gamma_*}y - \lambda \mathsf{s}_{\Gamma_*}\right), \\ \left\|\Phi^T_{\Gamma^c_*}\left(y - \Phi_{\Gamma_*}a^*_{\Gamma_*}\right)\right\|_\infty &\leq \lambda, \end{aligned} \tag{85}$$

which are the two well-known optimality conditions for a^* to be the solution of (5) [82].

B.3 LCA inequalities

The proofs of Theorems 7 and 8 make use of the following two lemmas, which provide bounds on some relevant quantities. The first lemma is stated below and bounds the ℓ_2-distance between the points a^∞_Γ and the target signal $a^\dagger$.

Lemma 11. *Let a^∞ be a vector supported on a set Γ that contains less than p indices, $\mathsf{s}_\Gamma = \operatorname{sign}\left(a^\infty_\Gamma\right)$, and assume*

$$\Phi^T_\Gamma \Phi_\Gamma a^\infty = \Phi^T_\Gamma y - \lambda \mathsf{s}_\Gamma.$$

Let $R = \left|\Gamma \cup \Gamma_\dagger\right|$ be the number of elements in the support of $\left(a^\infty - a^\dagger\right)$. If Φ satisfies the RIP with parameters (R, δ), then the following holds:

$$\left\|a^\infty - a^\dagger\right\|_2 \leq \underbrace{(1-\delta)^{-1}\left(\left\|a^\dagger\right\|_2 + \sqrt{1-\delta}\,\|\epsilon\|_2 + \lambda\sqrt{p}\right)}_{=(1-\delta)(1+\delta)^{-1}B_\delta(p)}.$$

Proof. Since Φ satisfies the RIP with parameters (S, R), using the results in Lemma 14 and 15 with $|\Gamma| = p \leq R$, $\Gamma_1 = \Gamma$ and $\Gamma_2 = \Gamma_\dagger$, then

$$\begin{aligned} \left\|\left(\Phi^T_\Gamma\Phi_\Gamma\right)^{-1}\right\| &\leq (1-\delta)^{-1}, \\ \left\|\Phi^T_\Gamma\Phi_{\Gamma_\dagger \cap \Gamma^c}\right\| &\leq \delta, \\ \left\|\left(\Phi^T_\Gamma\Phi_\Gamma\right)^{-1}\Phi^T_\Gamma\right\|^2 &\leq (1-\delta)^{-1}. \end{aligned}$$

Splitting $a^\dagger$ into its components on Γ and Γ^c yields the two equalities

$$\begin{aligned} a^\dagger_\Gamma &= \left(\Phi^T_\Gamma\Phi_\Gamma\right)^{-1}\Phi^T_\Gamma\Phi_\Gamma a^\dagger_\Gamma, \\ \Phi\left(a^\dagger - a^\dagger_\Gamma\right) &= \Phi_{\Gamma^c}a^\dagger_{\Gamma^c}. \end{aligned}$$

Applying these facts to finish the proof results in

$$\begin{aligned}
\left\|a^{\infty} - a^{\dagger}\right\|_2 &= \left\|\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\left(\Phi_\Gamma^T y - \lambda \mathsf{s}_\Gamma\right) - a^{\dagger}\right\|_2 \\
&= \left\|\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\left(\Phi_\Gamma^T\left(\Phi a^{\dagger} + \epsilon\right) - \lambda \mathsf{s}_\Gamma\right) - a^{\dagger}_{\Gamma} - a^{\dagger}_{\Gamma^c}\right\|_2 \\
&= \left\|\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\Phi_\Gamma^T\Phi_{\Gamma^c}a^{\dagger}_{\Gamma^c} + a^{\dagger}_{\Gamma} + \left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\Phi_\Gamma^T\epsilon - \lambda\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\mathsf{s}_\Gamma - a^{\dagger}_{\Gamma} - a^{\dagger}_{\Gamma^c}\right\|_2 \\
&\leq \left\|\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\right\|\left\|\Phi_\Gamma^T\Phi_{\Gamma_\dagger\cap\Gamma^c}\right\|\left\|a^{\dagger}_{\Gamma^c}\right\|_2 + \left\|a^{\dagger}_{\Gamma^c}\right\|_2 \\
&\qquad\qquad + \left\|\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\Phi_\Gamma^T\right\|\|\epsilon\|_2 + \lambda\left\|\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\right\|\|\mathsf{s}_\Gamma\|_2 \\
&\leq (1-\delta)^{-1}\delta\left\|a^{\dagger}_{\Gamma^c}\right\|_2 + \left\|a^{\dagger}_{\Gamma^c}\right\|_2 + \sqrt{1-\delta}^{-1}\|\epsilon\|_2 + \lambda(1-\delta)^{-1}\sqrt{p} \\
&\leq (1-\delta)^{-1}\left(\left\|a^{\dagger}\right\|_2 + \sqrt{1-\delta}\,\|\epsilon\|_2 + \lambda\sqrt{p}\right). \qquad \square
\end{aligned}$$

The lemma below states that the ℓ_2-distance of the output $a(t)$ to the target signal $a^{\dagger}$ remains bounded for all time $t \geq 0$.

Lemma 12. *Assume that, at switching time t_k, the current active set Γ_k contains less than p indices, Φ satisfies the RIP with parameters (R_k, δ), where $R_k = \left|\Gamma_k \cup \Gamma_\dagger\right|$, and that*

$$\left\|a(t_k) - a^{\dagger}\right\|_2 \leq B_\delta(p).$$

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

$$\left\|a(t) - a^{\dagger}\right\|_2 \leq B_\delta(p).$$

Proof. Define $a^{\infty}_{\Gamma_k}$ such that $\Phi_{\Gamma_k}^T\Phi_{\Gamma_k}a^{\infty}_{\Gamma_k} = \Phi_{\Gamma_k}^T y - \lambda\mathsf{s}_{\Gamma_k}$. Lemma 11 implies that

$$\left\|a^{\infty}_{\Gamma_k} - a^{\dagger}\right\|_2 \leq (1-\delta)(1+\delta)^{-1}B_\delta(p).$$

Using the dynamics in (83) shows that, $\forall t \in [t_k, t_{k+1})$,

$$\begin{aligned}
\left\|a(t) - a^\dagger\right\|_2 &= \left\|a_{\Gamma_k}(t) - a^\dagger\right\|_2 \\
&= \left\|e^{-A(t-t_k)} a_{\Gamma_k}(t_k) + \left(I - e^{-A(t-t_k)}\right) a^\infty_{\Gamma_k} - a^\dagger\right\|_2 \\
&\le \left\|\left(e^{-A(t-t_k)} + e^{-(1-\delta)(t-t_k)} I_{\Gamma_k^c}\right)\left(a_{\Gamma_k}(t_k) - a^\dagger\right)\right\|_2 \\
&\qquad\qquad + \left\|\left(I - e^{-A(t-t_k)} - e^{-(1-\delta)(t-t_k)} I_{\Gamma_k^c}\right)\left(a^\infty_{\Gamma_k} - a^\dagger\right)\right\|_2 \\
&\le \left\|e^{-A(t-t_k)} + e^{-(1-\delta)(t-t_k)} I_{\Gamma_k^c}\right\| \left\|a_{\Gamma_k}(t_k) - a^\dagger\right\|_2 \\
&\qquad\qquad + \left\|I - e^{-A(t-t_k)} - e^{-(1-\delta)(t-t_k)} I_{\Gamma_k^c}\right\| \left\|a^\infty_{\Gamma_k} - a^\dagger\right\|_2 \\
&\le e^{-(1-\delta)(t-t_k)} \left\|a_{\Gamma_k}(t_k) - a^\dagger\right\|_2 \;+\; \left(1 - e^{-(1+\delta)(t-t_k)}\right) \left\|a^\infty_{\Gamma_k} - a^\dagger\right\|_2 \\
&\le e^{-(1-\delta)(t-t_k)} B_\delta(p) \;+\; \left(1 - e^{-(1+\delta)(t-t_k)}\right) \frac{1-\delta}{1+\delta} B_\delta(p) \\
&\overset{(i)}{\le} B_\delta(p).
\end{aligned}$$

The function $h(\cdot)$ below is used to prove the last inequality in the above derivation:

$$h(t) = \left(1 - e^{-(1-\delta)t}\right) B - \left(1 - e^{-(1+\delta)t}\right) \frac{1-\delta}{1+\delta} B.$$

The derivative of $h(t)$ is

$$\begin{aligned}
h'(t) &= (1-\delta)\, e^{-(1-\delta)t} B - (1-\delta)\, e^{-(1+\delta)t} B \\
&= (1-\delta)\, B\left(e^{-(1-\delta)t} - e^{-(1+\delta)t}\right) \ge 0.
\end{aligned}$$

Since $h(0) = 0$, and $h'(t) \ge 0$ for all $t \ge 0$, then $h(t) \ge 0$ for all $t \ge 0$, and the inequality (i) holds.

Finally, since the vector $a(t) - a^\dagger$ is continuous with time:

$$\left\|a_{\Gamma_{k+1}}(t_{k+1}) - a^\dagger\right\|_2 = \left\|a_{\Gamma_k}(t_{k+1}) - a^\dagger\right\|_2 \le B_\delta(p). \qquad \square$$

Finally, the lemma below is used repeatedly in the proofs and states that if the energy in the q nodes with largest magnitude in $u(t)$ satisfy a certain inequality, then there is no more than q active nodes at time t.

Lemma 13. *If* Δ *contains the indices of the* q *entries with largest absolute values in* $u(t)$ *and*

$$||u_\Delta(t)||_2 \leq \lambda\sqrt{q},$$

then the active set Γ *corresponding to the non-zero elements in* $a(t) = T_\lambda(u(t))$ *is a subset of* Δ *and contains less than* q *indices; i.e.,* $\Gamma \subset \Delta$, *and as a result* $|\Gamma| \leq q$.

Proof. Since Δ contains the q nodes with largest absolute values in $u(t)$, then $\forall j \in \Delta^c$,

$$\left|u_j(t)\right| \leq \frac{||u_\Delta(t)||_2}{\sqrt{q}} \leq \lambda.$$

As a consequence, nodes in Δ^c are below threshold, which proves that only the nodes in Δ can be non-zero in $a(t)$. Thus, $\Gamma \subset \Delta$ and $|\Gamma| \leq |\Delta| = q$. □

APPENDIX C

MATRIX PROPERTIES

The following lemmas are consequences of the RIP, defined in (2), that are used repeatedly in the proofs in this thesis. The first lemma has been proven several times in literature (for instance [28, Prop. 3.1, 3.2]), but it is repeated below for completeness.

Lemma 14. *If* Φ *satisfies the RIP with parameters* (S, δ) *and the set* Γ *contains less than* S *indices, then* $\forall x \in \mathbb{R}^N$ *supported on* Γ *and* $\forall y \in \mathbb{R}^M$*, the following holds:*

$$\begin{aligned}
\left\|\Phi_\Gamma^T x\right\|_2 &\leq \sqrt{1+\delta}\,\|x\|_2\,, \\
(1-\delta)\,\|x\|_2 \leq \left\|\Phi_\Gamma^T \Phi_\Gamma x\right\|_2 &\leq (1+\delta)\,\|x\|_2\,, \\
\frac{1}{1+\delta}\,\|x\|_2 \leq \left\|\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1} x\right\|_2 &\leq \frac{1}{1-\delta}\,\|x\|_2\,, \\
\frac{1}{\sqrt{1+\delta}}\,\|x\|_2 \leq \left\|\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1} \Phi_\Gamma^T x\right\|_2 &\leq \frac{1}{\sqrt{1-\delta}}\,\|x\|_2\,.
\end{aligned}$$

Proof. The RIP implies that the S non-zero singular values of Φ_Γ are contained between $\sqrt{1-\delta}$ and $\sqrt{1+\delta}$, which entails the first inequality. Taking the singular value decomposition of $\Phi_\Gamma = U\Sigma V^T$, where U is a $M\times S$ matrix with orthogonal columns, Σ is a $S\times S$ matrix with the S non-zero singular values of Φ_Γ on the diagonal, and V is a $S\times N$ unitary matrix, it is easy to check that the singular values of $\Phi_\Gamma^T\Phi_\Gamma = V\Sigma^2V^T$ are contained between $(1-\delta)$ and $(1+\delta)$, which proves the second inequality. This fact also implies that the singular values of $\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1} = V\left(\Sigma^2\right)^{-1}V^T$ are contained between $(1+\delta)^{-1}$ and $(1-\delta)^{-1}$, yielding the third inequality. Finally, the singular value decomposition of $\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\Phi_\Gamma^T$ is equal to $\left(\Phi_\Gamma^T\Phi_\Gamma\right)^{-1}\Phi_\Gamma^T = \left(V\left(\Sigma^2\right)^{-1}V^T\right)V\Sigma U^T = V\Sigma^{-1}U^T$, which shows that its singular values are contained between $\sqrt{1+\delta}^{-1}$ and $\sqrt{1-\delta}^{-1}$ and proves the last inequality. □

The following lemma provides slightly more complicated consequences of the RIP that involve two (not necessarily disjoint) subsets of indices.

Lemma 15. *If Φ satisfies the RIP with parameters $(S+q,\delta)$, the set Γ_1 contains less than q indices, and the set Γ_2 contains less than S indices, then $\forall x \in \mathbb{R}^N$ supported on $\Gamma_1 \cup \Gamma_2$, the following holds:*

$$\left\|\Phi_{\Gamma_1}^T \Phi_{\left(\Gamma_1^c \cap \Gamma_2\right)} x\right\|_2 \leq \delta \left\|x\right\|_2 ,$$
$$\left\|\left(I_{\Gamma_1} - \Phi_{\Gamma_1}^T \Phi_{(\Gamma_1 \cup \Gamma_2)}\right) x\right\|_2 \leq \delta \left\|x\right\|_2 .$$

Proof. Since the set $\Gamma_1 \cup \Gamma_2$ contains less that $S+q$ indices, the RIP implies that the eigenvalues of $\Phi_{(\Gamma_1\cup\Gamma_2)}^T \Phi_{(\Gamma_1\cup\Gamma_2)}$ are contained between $(1-\delta)$ and $(1+\delta)$. As a consequence, the eigenvalues denoted by $\text{sp}(\cdot)$ (for spectrum) of the following matrix

$$G_{(\Gamma_1\cup\Gamma_2)} := I_{(\Gamma_1\cup\Gamma_2)} - \Phi_{(\Gamma_1\cup\Gamma_2)}^T \Phi_{(\Gamma_1\cup\Gamma_2)}$$

can be deduced:

$$\text{sp}\left(G_{(\Gamma_1\cup\Gamma_2)}\right) \leq \max\left\{1-(1-\delta), -1+(1+\delta)\right\} = \delta,$$

and

$$\text{sp}\left(G_{(\Gamma_1\cup\Gamma_2)}\right) \geq \min\left\{1-(1+\delta), -1+(1-\delta)\right\} = -\delta.$$

The matrices $\Phi_{\Gamma_1}^T \Phi_{\left(\Gamma_1^c\cap\Gamma_2\right)}$ and $\left(I_{\Gamma_1} - \Phi_{\Gamma_1}^T \Phi_{(\Gamma_1\cup\Gamma_2)}\right)$ are submatrices of the matrix $G_{(\Gamma_1\cup\Gamma_2)}$, in the sense that

$$\Phi_{\Gamma_1}^T \Phi_{\left(\Gamma_1^c\cap\Gamma_2\right)} = \Pi_{\Gamma_1} G_{(\Gamma_1\cup\Gamma_2)} \Pi_{\Gamma_1^c}$$

and

$$\left(I_{\Gamma_1} - \Phi_{\Gamma_1}^T \Phi_{(\Gamma_1\cup\Gamma_2)}\right) = \Pi_{\Gamma_1} G_{(\Gamma_1\cup\Gamma_2)},$$

where Π_{Γ_i} denotes the projection onto the set of indices Γ_i. The operator norm of the projection operator Π_{Γ_i} is 1. As a consequence, the operator norms of the two matrices $\Phi_{\Gamma_1}^T \Phi_{\left(\Gamma_1^c\cap\Gamma_2\right)}$ and $\left(I_{\Gamma_1} - \Phi_{\Gamma_1}^T \Phi_{(\Gamma_1\cup\Gamma_2)}\right)$ are bounded by the operator norm of the larger matrix $G_{(\Gamma_1\cup\Gamma_2)}$, which its largest eigenvalue and is equal to δ. □

APPENDIX D

DISCONTINUOUS ACTIVATION FUNCTION

In CS and other applications, it can be useful to consider a neural network with a discontinuous activation function. This appendix provides some preliminary results for the study of such networks.

D.1 Cost function

In a first step, the method developed in Lemma 4 for building a cost function $C(\cdot)$ that satisfies the relationship (19) is extended to the case where the activation function may contain discontinuities.

Assumption 4. *The activation function $T_\lambda(\cdot)$ is locally bounded, admits directional derivatives, is odd and nondecreasing on $\mathbb{R}$. In addition, there exist $\lambda \geq 0$, and locally finitely many $\{(v_k, w_k, z_k, d_k)\}_{k\in\mathcal{K}}$ in $\mathbb{R}\times\mathbb{R}\times\mathbb{R}$, with $u_k < w_k$, such that $T_\lambda(\cdot)$ has a jump discontinuity at d_k and has the form*

$$a_n = T_\lambda(u_n) = \begin{cases} 0, & |u_n| \leq \lambda \\ z_k, & |u_n| \in \bigcup_{k\in\mathcal{K}} [v_k, w_k] := \mathcal{Z} \\ \multicolumn{2}{l}{\textit{is strictly increasing otherwise with } \zeta_n > 0,\ \forall \zeta_n \in \partial T_\lambda(u_n)} \end{cases} \tag{86}$$

and satisfies

$$|T_\lambda(u_n)| \leq |u_n|\,, \qquad \forall u_n \in \mathbb{R}. \tag{87}$$

Explicitly, the form in (86) *means that $T_\lambda(\cdot)$*

- *is exactly zero on the interval $[-\lambda, \lambda]$,*
- *is constant on a countable and locally finite number of intervals denoted by $\mathcal{Z}$ (which include the interval $[-\lambda, \lambda]$ and potentially the case where w_k is equal to infinity for some k),*

- *has a locally finite number of points d_k at which is it discontinuous, and*

- *is otherwise strictly increasing on any open interval $\mathcal{U}$ in $\mathbb{R}\backslash\mathcal{Z}$ (where $T_\lambda(\cdot)$ is not constant) with strictly positive subgradients.*

Lemma 16. *If the activation function $T_\lambda(\cdot)$ satisfies Assumption 4, there exists a cost function $C(\cdot)$ that satisfies the relationship* (19) *and obeys*

1. *$C(\cdot)$ is locally Lipschitz continuous on $\mathbb{R}$,*

2. *$C(\cdot)$ is even on $\mathbb{R}$,*

3. *$C(\cdot)$ is nondecreasing on $\mathbb{R}^+$,*

4. *$C(0) = 0$,*

5. *$C(\cdot)$ is regular on $\mathbb{R}$.*

Proof. Similar to the construction of the inverse function in Lemma 4, the first step is to construct an inverse function $z^{-1}(\cdot)$ for the activation function $T_\lambda(\cdot)$. For points that belong to the image $T_\lambda(\mathbb{R})$ of the activation function, the inverse function can be defined as in Lemma 4. To reiterate, for these points $\nu \in T_\lambda(\mathbb{R})$, there exists $u \in \mathbb{R}$ such that $\nu = T_\lambda(u)$. As a consequence, the function $z^{-1}(\cdot)$ can be defined on $T_\lambda(\mathbb{R})$ as follows:
$\forall \nu \in T_\lambda(\mathbb{R})$, let $u \in \mathbb{R}$ such that $\nu = T_\lambda(u)$.

1. if $u \in \mathrm{Z}^c$ (which is the set of nodes that do not yield a constant output), then u is the unique point in $\mathbb{R}^N$ satisfying $\nu = T_\lambda(u)$, and $z^{-1}(\nu)$ is defined as $z^{-1}(\nu) = u$,

2. if $u \in \mathrm{Z}$, there exists $k \in \mathcal{K}$ such that $\nu = T_\lambda(u_k)$ for all $u_k \in [v_k, w_k]$. In that case, $z^{-1}(\nu)$ is chosen to be $z^{-1}(\nu) = w_k$.

For points that do not belong to the image of the activation function $\nu \notin T_\lambda(\mathbb{R})$, there exists a point of discontinuity d_k such that

$$\lim_{\substack{u \to d_k \\ u < d_k}} T_\lambda(u) \le \nu \le \lim_{\substack{u \to d_k \\ u > d_k}} T_\lambda(u).$$

The two points

$$l_k = \lim_{\substack{u \to d_k \\ u < d_k}} T_\lambda(u),$$

$$m_k = \lim_{\substack{u \to d_k \\ u > d_k}} T_\lambda(u),$$

exist since $T_\lambda(\cdot)$ admits directional derivatives by assumption. Then, the function $z^{-1}(v)$ can be chosen as $z^{-1}(v) = d_k$ for all $v \in [l_k, m_k)$. Figure 22b illustrates how to construct $z^{-1}(\cdot)$ for a particular activation function plotted in Figure 22a.

Using this definition, the following quantity is well-defined on $T_\lambda(\mathbb{R})$:

$$C(a) = \int_0^a z^{-1}(v) - v \, dv. \tag{88}$$

The function $C(\cdot)$ defined this way is locally Lipschitz on $T_\lambda(\mathbb{R})$ and differentiable for a.a. $a \in T_\lambda(\mathbb{R})$. Figure 22c shows the cost function associated with the activation function plotted in Figure 22a.

The following derivation shows that $C(\cdot)$ is regular and satisfies (19). There are two cases.

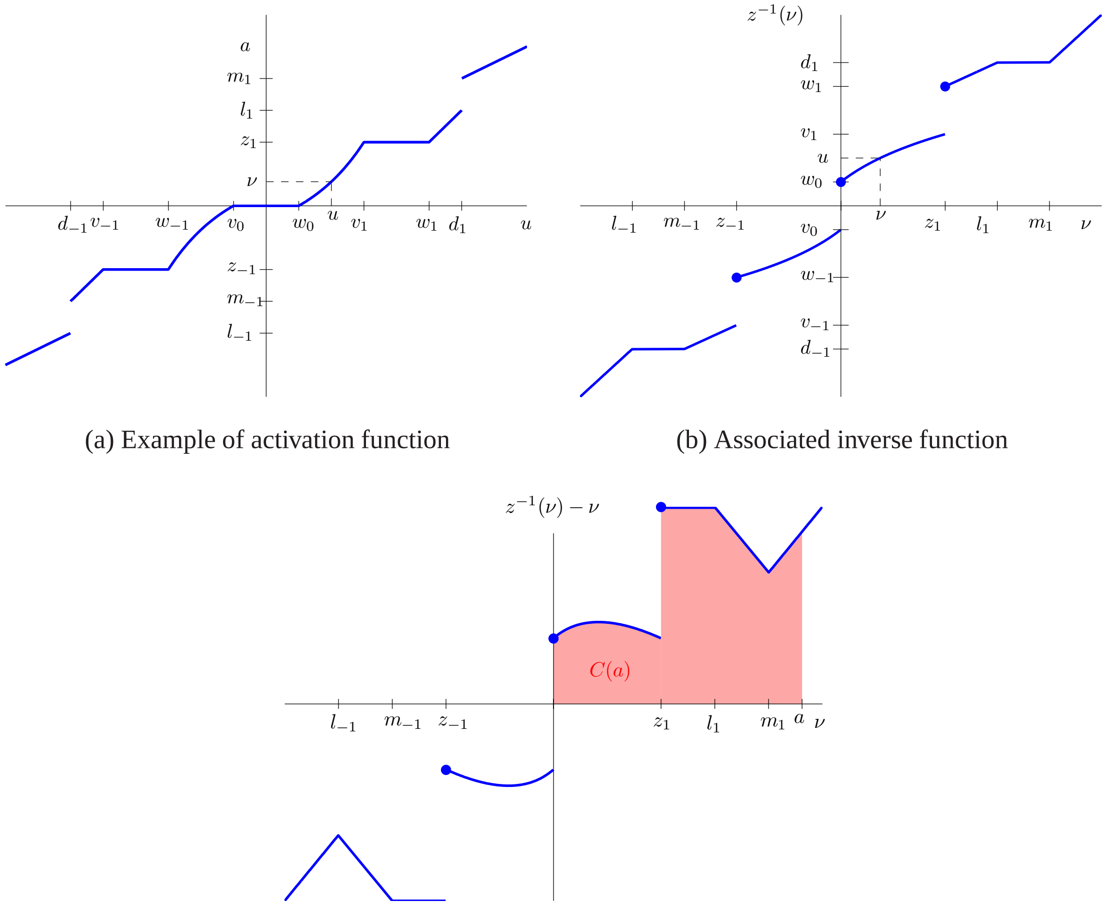
1. At points a where $C(\cdot)$ is differentiable, the subgradient reduces to $\partial C(a) = \{C'(a)\}$, and the fundamental theorem of calculus yields

   $$C'(a) = z^{-1}(a) - a = u - a.$$

   As a consequence, for such a, (19) holds.

2. For points a where $C(\cdot)$ is not differentiable, it follows by construction of $z^{-1}(\cdot)$ that $z^{-1}(a) = w_k$ for some $k \in \mathcal{K}$. As a consequence, a similar analysis to the proof of Lemma 4 shows that the right-sided derivative of $C(\cdot)$ exists and is the limit for $t > 0$ sufficiently small of

   $$\frac{C(a+t) - C(a)}{t} = \frac{1}{t} \int_a^{a+t} z^{-1}(v) - v \, dv.$$



(a) Example of activation function

(b) Associated inverse function

(c) Associated cost function

Figure 22: Example of a generic activation function $T_\lambda(\cdot)$ satisfying Assumption 4, associated inverse function $z^{-1}(\cdot)$ and associated cost function $C(\cdot)$.

By the fundamental theorem of calculus, the above quantity converges to $z^{-1}(a)-a = u-a$ as $t \to 0$, and as a result $C'(a;1) = C^\circ(a;1)$. For the left-sided derivative, the following integral can be computed for all $t > 0$ sufficiently small:

$$\begin{aligned}\frac{C(a-t)-C(a)}{t} &= -\frac{1}{t}\int_{a-t}^{a} z^{-1}(v) - v\, dv \\ &= -\int_{a-t}^{a} z_0^{-1}(v) - v\, dv,\end{aligned}$$

where $z_0^{-1}(v) = z^{-1}(v)$ everywhere except at a where the function is defined to be $z_0^{-1}(a) = v_k$. The two integrals are equal because they only differ at one point. The function $z_0^{-1}(\cdot)$ is now continuous on $[a-t, a]$, and thus, the one-sided integral exists and can be computed as $t \to 0$ to get $C'(a;-1) = -v_n + a_n = C^\circ(a;-1)$. Using the definition of the subgradient in Section 2.3.1, for all $\xi \in \partial C(a)$, the generalized directional derivative satisfies $C^\circ(a;1) = w_k - a \geq 1\xi$ and $C^\circ(a;-1) = -v_k + a \geq -1\xi$. As a consequence $\xi \in [v_k - a, w_k - a]$, which means that

$$\partial C(a) = [v_k - a, w_k - a].$$

This equality proves that indeed $u_k - a \in \partial C(a)$ for all $u_k \in [v_k, w_k]$.

As a consequence, $C(\cdot)$ is regular on $T_\lambda(\mathbb{R})$, and $u - a \in \partial C(a)$ holds for all $a \in T_\lambda(\mathbb{R})$, which proves that (19) holds.

By inspection of (75), it is immediate that $C(0) = 0$.

It is easy to check that, since $T_\lambda(\cdot)$ is nondecreasing and odd, $z^{-1}(\cdot)$ defined above is also nondecreasing and odd. As a consequence, for all $a \in T_\lambda(\mathbb{R})$, the following holds:

$$\begin{aligned}C(-a) &= \int_0^{-a} \left(z^{-1}(v) - v\right)\, dv \\ &= \int_0^{a} \left(z^{-1}(-v) + v\right)\, (-dv) \\ &= \int_0^{a} \left(z^{-1}(v) - v\right)\, dv = C(a).\end{aligned}$$

This computation proves that $C(\cdot)$ is even.

Finally, for all $\nu \in T_\lambda(\mathbb{R})$ such that $\nu \geq 0$, letting $u = z^{-1}(\nu)$, condition (87) implies that $z^{-1}(\nu) - \nu = u - z(u) \geq 0$. This bound proves that $C(\cdot)$ is nondecreasing on $\mathbb{R}^+$ by the positivity of the integral. As a consequence, $C(a) \geq C(0) = 0$ for all $a \in T_\lambda(\mathbb{R}^+)$, and by symmetry for all $a \in T_\lambda(\mathbb{R})$. □

Hard-thresholding function

Using this technique, it is possible to build a cost function satisfying (19) for the hard-thresholding activation function, defined by

$$T_\lambda(u) = \begin{cases} 0, & \text{if } |u| \leq \lambda \\ u, & \text{if } |u| > \lambda \end{cases}.$$

The associated inverse function $z^{-1}(\cdot)$ constructed as in Lemma 16 is

$$z^{-1}(\nu) = \begin{cases} \lambda, & \text{if } \nu \in [0, \lambda) \\ -\lambda, & \text{if } \nu \in [-\lambda, 0) \\ \nu, & \text{if } |\nu| > \lambda \end{cases} \cdot$$

Integrating the function $z^{-1}(\nu) - \nu$ between 0 and a yields the cost function:

$$C(a) = \begin{cases} \lambda\,|a| - \dfrac{a^2}{2}, & \text{if } |a| \leq \lambda \\ \dfrac{\lambda^2}{2}, & \text{if } |a| > \lambda \end{cases}. \tag{89}$$

For large values of a, this function behaves like the ideal ℓ_0-pseudo norm scaled by $\lambda^2/2$. This observation matches the result in [20], which states that the hard-thresholding function can be used to approximately solve the ideal ℓ_0-minimization problem (4) with a tradeoff parameter of $\lambda^2/2$.

D.2 Solutions to ODEs with a discontinuous right-hand side

Because the activation function $T_\lambda(\cdot)$ may now have points of discontinuity, the theory developed in Chapter 3 does not apply. In a first step, it is necessary to define what a solution of the ODE (8) might be when the right-hand side is discontinuous. A well-established approach in mechanics and nonlinear neural networks consists in approximating the trajectory $u(\cdot)$ by the solution of (8) in the sense of Filippov [37]. Using the theory of Filippov, a function $u(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ is a solution of (8) if it is absolutely continuous on $\mathbb{R}^+$ and satisfies the differential inclusion:

$$\dot{u}(t) \in -u(t) + (I - \Phi^T \Phi) co \{T_\lambda(u(t))\} + \Phi^T y, \qquad \text{for a.a. } t \geq 0,$$

where $co\{T_\lambda(u)\} = (co\{T_\lambda(u_1)\}, \ldots, co\{T_\lambda(u_N)\})^T$ and

$$co\{T_\lambda(u_n)\} = [T_\lambda(u_n^-), T_\lambda(u_n^+)]$$

is the traditional convex hull with

$$T_\lambda(u_n^-) = \lim_{\substack{v \to u \\ v < u}} T_\lambda(v),$$
$$T_\lambda(u_n^+) = \lim_{\substack{v \to u \\ v > u}} T_\lambda(v).$$

At points d_k where $T_\lambda(\cdot)$ is discontinuous, $co\{T_\lambda(d_k)\}$ is an interval while it is a singleton at points where $T_\lambda(\cdot)$ is continuous. A function $F(\cdot) : \mathbb{R} \to \mathbb{R}$ is called *absolutely continuous* on $\mathbb{R}$ if for any $\varepsilon > 0$, there exists $R > 0$, such that for all finite sequences of intervals $\{(x_k, y_k)\}_{k \geq 0}$ disjoint in $\mathbb{R}$

$$\sum_k (y_k - x_k) \leq R \qquad \Rightarrow \qquad \sum_k |F(y_k) - F(x_k)| \leq \varepsilon.$$

It is well know that for an absolutely continuous function $F(\cdot)$, the theorem of calculus yields that $F(\cdot)$ is differentiable a.e. and the following holds:

$$F(y) = F(x) + \int_x^y G(v) dv,$$

where $G(v) = F'(v)$ for a.a. $v \in \mathbb{R}$. Moreover, it was shown in [83] that the chain rule (14) holds when $x(\cdot)$ is only absolutely continuous on any bounded interval of $\mathbb{R}^+$.

D.3 Convergence result

Unfortunately, the proof of convergence of the LCA with a discontinuous activation function does not derive readily from previous analysis obtained for neural networks with discontinuities. For the analysis in [83], the activation function needs to be bounded, and the interconnection matrix $(\Phi^T\Phi - I)$ needs to be Lyapunov diagonally stable, which requires that it is nonsingular. Similarly, in [84], the interconnection matrix is nonsingular and the activation function is bounded. In [85, 86], the boundedness assumption on the activation function is dropped, but the interconnection matrix is assumed to be nonsingular in [85] and to be Lyapunov diagonally stable in [86], which also implies nonsingularity.

Nevertheless, the proof of Theorem 4 and 5 in Section 3.5 is based on the Łojasiewicz inequality and only requires the cost function to be continuous. As seen in Lemma 16, it is possible to create a cost function $C(\cdot)$ that is locally Lipschitz continuous and satisfies (19), even when the activation function has discontinuities. As a consequence, Theorems 4 and 5 still hold without modifications.

REFERENCES

[1] A. Cichocki and R. Unbehauen, *Neural Networks for Optimization and Signal Processing*. Wiley, 1993.

[2] E. J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” *IEEE Trans. Inf. Theory*, vol. 52, pp. 489– 509, Feb. 2006.

[3] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 52, pp. 1289–1306, Apr. 2006.

[4] E. J. Candès and T. Tao, “Decoding by linear programming,” *IEEE Trans. Inf. Theory*, vol. 51, pp. 4203–4215, Dec. 2005.

[5] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *arXiv:1011.3027*, Nov. 2011.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by Basis Pursuit,” *SIAM Rev.*, vol. 43, pp. 129–159, Mar. 2001.

[7] R. Gribonval and M. Nielsen, “Highly sparse representations from dictionaries are unique and independent of the sparseness measure,” *Appl. Comput. Harmon. Anal.*, vol. 22, pp. 335–355, May 2007.

[8] E. J. Candès, “The restricted isometry property and its implications for compressed sensing,” *Comptes Rendus Acad. Sci., Serie I*, vol. 346, pp. 589–592, May 2008.

[9] T. Zhang, “Some sharp performance bounds for least squares regression with L1 regularization,” *Ann. Statist.*, vol. 37, pp. 2109–2144, Oct. 2009.

[10] 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.*, vol. 20, pp. 2526–2563, Oct. 2008.

[11] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proc. Natl. Acad. Sci.*, vol. 79, pp. 2554 –2558, Apr. 1982.

[12] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain,” *Psychol. Rev.*, vol. 65, pp. 386–408, Nov. 1958.

[13] J. J. Hopfield, “Neurons with graded response have collective computational properties like those of two-state neurons,” *Proc. Natl. Acad. Sci.*, vol. 81, pp. 3088–3092, May 1984.

[14] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, “An interior-point method for large-scale L1-regularized least squares,” *IEEE J. Sel. Topics Signal Process.*, vol. 1, pp. 606–617, Dec. 2007.

[15] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE J. Sel. Topics Signal Process.*, vol. 1, pp. 586–597, Dec. 2007.

[16] E. van den Berg and M. P. Friedlander, “Probing the Pareto frontier for basis pursuit solutions,” *SIAM J. Sci. Comput.*, vol. 31, pp. 890–912, Nov. 2008.

[17] S. Becker, J. Bobin, and E. J. Candès, “NESTA: A fast and accurate first-order method for sparse recovery,” *SIAM J. Imaging Sci.*, vol. 4, no. 1, pp. 1–39, 2011.

[18] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, “Bregman iterative algorithms for L1-minimization with applications to compressed sensing,” *SIAM J. Imaging Sci.*, vol. 1, no. 1, pp. 143–168, 2008.

[19] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–1457, Aug. 2004.

[20] T. Blumensath and M. E. Davies, “Iterative thresholding for sparse approximations,” *J. Fourier Anal. Appl.*, vol. 14, no. 5-6, pp. 629–654, 2008.

[21] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *Ann. Stat.*, vol. 32, pp. 407–499, Apr. 2004.

[22] M. S. Asif and J. Romberg, “Dynamic updating for L1 minimization,” *IEEE J. Sel. Topics Signal Process.*, vol. 4, pp. 421–434, Apr. 2010.

[23] D. Donoho and Y. Tsaig, “Fast solution for L1-norm minimization problems when the solution may be sparse,” *IEEE Trans. Inf. Theory*, vol. 54, pp. 4789–4812, Nov. 2008.

[24] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via Orthogonal Matching Pursuit,” *IEEE Trans. Inf. Theory*, vol. 53, pp. 4655–4666, Dec. 2007.

[25] T. Zhang, “Sparse recovery with Orthogonal Matching Pursuit under RIP,” *IEEE Trans. Inf. Theory*, vol. 57, pp. 6215–6221, Sept. 2011.

[26] M. A. Davenport and M. B. Wakin, “Analysis of Orthogonal Matching Pursuit using the Restricted Isometry Property,” *IEEE Trans. Inf. Theory*, vol. 56, pp. 4395–4401, Sept. 2010.

[27] D. Needell and R. Vershynin, “Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit,” *Found. Comput. Math.*, vol. 9, p. 317–334, Apr. 2009.

[28] D. Needell and J. A. Tropp, “CoSaMP: Iterative signal recovery from incomplete and inaccurate samples,” *Appl. Comput. Harmon. Anal.*, vol. 26, pp. 301–321, Mar. 2008.

[29] A. M. Lyapunov, *The General Problem of the Stability of Motion*. Taylor & Francis, 1892.

[30] J. J. Hopfield and D. W. Tank, “”Neural” computation of decisions in optimization problems,” *Biol. Cybern.*, vol. 52, no. 3, pp. 141–152, 1985.

[31] D. W. Tank and J. J. Hopfield, “Simple ’neural’ optimization networks: An A/D converter, signal decision circuit, and a linear programming circuit,” *IEEE Trans. Circuits Syst.*, vol. 33, pp. 533 – 541, May 1986.

[32] A. Michel and D. Gray, “Analysis and synthesis of neural networks with lower block triangular interconnecting structure,” *IEEE Trans. Circuits Syst.*, vol. 37, pp. 1267 –1283, Oct. 1990.

[33] M. Forti, A. Liberatore, S. Maneti, and M. Marini, “Global asymptotic stability for a class of nonsymmetric neural networks,” in *IEEE Int. Symp. Circuits, Syst.*, pp. 2580 –2583, May 1993.

[34] H. Yang and T. S. Dillon, “Exponential stability and oscillation of Hopfield graded response neural network,” *IEEE Trans. Neural Netw.*, vol. 5, pp. 719–729, Sept. 1994.

[35] M. Forti and A. Tesi, “New conditions for global stability of neural networks with application to linear and quadratic programming problems,” *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 42, pp. 354–366, July 1995.

[36] F. H. Clarke, *Optimization and Nonsmooth Analysis*. Society for Industrial Mathematics, Jan. 1987.

[37] A. F. Filippov, *Differential equations with discontinuous righthand sides: Control systems*. Springer, Sept. 1988.

[38] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317. Springer, 2011.

[39] R. T. Rockafellar, “Directionally Lipschitzian functions and subdifferential calculus,” *London Math. Soc.*, vol. s3-39, pp. 331–355, Sept. 1979.

[40] J. P. Hespanha, “Uniform stability of switched linear systems: Extensions of LaSalle’s invariance principle,” *IEEE Trans. Autom. Control*, vol. 49, pp. 470– 482, Apr. 2004.

[41] S. Liu and J. Wang, “A simplified dual neural network for quadratic programming with its KWTA application,” *IEEE Trans. Neural Netw.*, vol. 17, pp. 1500–1510, Nov. 2006.

[42] M. Forti and A. Tesi, “Absolute stability of analytic neural networks: an approach based on finite trajectory length,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, pp. 2460 – 2469, Dec. 2004.

[43] W. Lu and J. Wang, “Convergence analysis of a class of nonsmooth gradient systems,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, pp. 3514–3527, Dec. 2008.

[44] J. Bolte, A. Daniilidis, and A. Lewis, “The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems,” *SIAM J. Optim.*, vol. 17, pp. 1205–1223, Jan. 2007.

[45] M. Forti, P. Nistri, and M. Quincampoix, “Convergence of neural networks for programming problems via a nonsmooth Łojasiewicz inequality,” *IEEE Trans. Neural Netw.*, vol. 17, pp. 1471–1486, Nov. 2006.

[46] X. Xue and W. Bian, “Subgradient-based neural networks for nonsmooth convex optimization problems,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, pp. 2378–2391, Sept. 2008.

[47] S. Łojasiewicz, “Une propriété topologique des sous-ensembles analytiques réels,” *Colloques internationaux du C.N.R.S. Les équations aux dériveés partielles*, vol. 117, pp. 87–89, 1963.

[48] A. Balavoine, C. J. Rozell, and J. Romberg, “Global convergence of the Locally Competitive Algorithm,” in *IEEE DSP/SPE*, pp. 431–436, Jan. 2011.

[49] A. Balavoine, J. Romberg, and C. J. Rozell, “Convergence and rate analysis of neural networks for sparse approximation,” *IEEE Trans. Neural Netw.*, vol. 23, pp. 1377–1389, Sept. 2012.

[50] A. Balavoine, C. J. Rozell, and J. Romberg, “Convergence of a neural network for sparse approximation using the nonsmooth Łojasiewicz inequality,” in *Int. Joint Conf. Neural Netw.*, (Dallas, TX), pp. 1–8, Aug. 2013.

[51] A. S. Charles, P. Garrigues, and C. J. Rozell, “A common network architecture efficiently implements a variety of sparsity-based inference problems,” *Neural Comput.*, vol. 24, pp. 3317–3339, Dec. 2012.

[52] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, “Sparse reconstruction by separable approximation,” *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, 2009.

[53] M. S. Asif and J. Romberg, “On the LASSO and Dantzig selector equivalence,” *Conf. Inf. Sci. Syst. (CISS)*, pp. 1–6, Mar. 2010.

[54] E. J. Candès, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Comm. Pure Appl. Math.*, vol. 59, pp. 1207–1223, Aug. 2006.

[55] A. Balavoine, C. Rozell, and J. Romberg, “Convergence speed of a dynamical system for sparse recovery,” *IEEE Trans. Signal Process.*, vol. 61, pp. 4259–4269, Jan. 2013.

[56] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, pp. 2845–2862, Nov. 2001.

[57] D. Needell and R. Vershynin, "Signal recovery from incomplete and inaccurate measurements via Regularized Orthogonal Matching Pursuit," *IEEE J. Sel. Topics Signal Process.*, vol. 4, pp. 310 –316, Apr. 2010.

[58] E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for L1-minimization: Methodology and convergence," *SIAM J. Optim.*, vol. 19, pp. 1107–1130, Jan. 2008.

[59] S. Shapero, A. S. Charles, C. J. Rozell, and P. Hasler, "Low power sparse approximation on reconfigurable analog hardware," *IEEE J. Emerg. Sel. Topic Circuits Syst.*, vol. 2, pp. 530 –541, Sept. 2012.

[60] K. Bredies and D. A. Lorenz, "Linear convergence of iterative soft-thresholding," *J. Fourier Anal. Appl.*, vol. 14, pp. 813–837, Dec. 2008.

[61] A. Balavoine, C. Rozell, and J. Romberg, "Iterative and continuous-time soft-thresholding with a dynamic input," 2014. in preparation.

[62] A. Beck and M. Teboulle, "A fast Iterative Shrinkage-Thresholding Algorithm with application to wavelet-based image deblurring," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 693–696, Apr. 2009.

[63] J. M. Bioucas-Dias and M. A. Figueiredo, "A new TwIST: Two-Step Iterative Shrinkage/Thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, pp. 2992–3004, Dec. 2007.

[64] Z. Chen, "Bayesian filtering: From Kalman filters to particle filters, and beyond," *[Online]*, 2003. Available: http://soma.crl.mcmaster.ca/˜zhechen/download/ieee_bayesian.ps.

[65] N. Vaswani, "Kalman filtered Compressed Sensing," in *IEEE Int. Conf. Image Process.*, pp. 893–896, Oct. 2008.

[66] A. S. Charles and C. J. Rozell, "Dynamic filtering of sparse signals using reweighted L1," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 6451–6455, May 2013.

[67] W. Dai, D. Sejdinovic, and O. Milenkovic, "Gaussian dynamic compressive sensing," in *Int. Conf. Sampling Theory and Appl.*, (Singapore), May 2011.

[68] D. Sejdinovic, C. Andrieu, and R. Piechocki, "Bayesian sequential compressed sensing in sparse dynamical systems," in *48th Annu. Allerton Conf. Commun., Control, Comput.*, pp. 1730–1736, Sept. 2010.

[69] B. Shahrasbi, A. Talari, and N. Rahnavard, "TC-CSBP: Compressive sensing for time-correlated data based on belief propagation," in *45th Annu. Conf. Inf. Sci. Syst.*, pp. 1–6, 2011.

[70] J. Ziniel, L. Potter, and P. Schniter, "Tracking and smoothing of time-varying sparse signals via approximate belief propagation," in *Asilomar Conf. Signals, Syst., Comput.*, pp. 808–812, Nov. 2010.

[71] D. Angelosante, G. Giannakis, and E. Grossi, "Compressed sensing of time-varying signals," in *16th Int. Conf. Digit. Signal Process.*, pp. 1–8, July 2009.

[72] D. Angelosante, J. Bazerque, and G. Giannakis, "Online adaptive estimation of sparse signals: Where RLS meets the L1-Norm," *IEEE Trans. Signal Process.*, vol. 58, pp. 3436–3447, July 2010.

[73] E. C. Hall and R. M. Willett, "Dynamical models and tracking regret in online convex programming," in *Int. Conf. Mach. Learn.*, vol. 28, (Atlanta), 2013.

[74] Y. Zakharov and V. Nascimento, "DCD-RLS adaptive filters with penalties for sparse identification," *IEEE Trans. Signal Process.*, vol. 61, pp. 3198–3213, June 2013.

[75] K. Slavakis, Y. Kopsinis, and S. Theodoridis, "Adaptive algorithm for sparse system identification using projections onto weighted L1 balls," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3742–3745, Mar. 2010.

[76] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: the sparse RLS algorithm," *IEEE Trans. Signal Process.*, vol. 58, pp. 4013–4025, Aug. 2010.

[77] Y. Chen, Y. Gu, and A. Hero, "Sparse LMS for system identification," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3125–3128, Apr. 2009.

[78] K. Slavakis, Y. Kopsinis, S. Theodoridis, and S. McLaughlin, "Generalized thresholding and online sparsity-aware learning in a union of subspaces," *IEEE Trans. Signal Process.*, vol. 61, pp. 3760–3773, Aug. 2013.

[79] J. Romberg, "Imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, pp. 14–20, Mar. 2008.

[80] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process. Mag.*, vol. 22, pp. 123–151, Nov. 2005.

[81] R. A. DeCarlo, M. S. Branicky, S. Pettersson, and B. Lennartson, "Perspectives and results on the stability and stabilizability of hybrid systems," *Proc. IEEE*, vol. 88, pp. 1069–1082, July 2000.

[82] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, pp. 1341–1344, June 2004.

[83] M. Forti, M. Grazzini, P. Nistri, and L. Pancioni, "Generalized Lyapunov approach for convergence of neural networks with discontinuous or non-Lipschitz activations," *Physica D: Nonlinear Phenomena*, vol. 214, pp. 88–99, Feb. 2006.

[84] M. Forti, “M-matrices and global convergence of discontinuous neural networks,” *International Journal of Circuit Theory and Applications*, vol. 35, no. 2, p. 105–130, 2007.

[85] L. Li and L. Huang, “Dynamical behaviors of a class of recurrent neural networks with discontinuous neuron activations,” *Applied Mathematical Modelling*, vol. 33, pp. 4326–4336, Dec. 2009.

[86] H. Wu, “Global stability analysis of a general class of discontinuous neural networks with linear growth activation functions,” *Information Sciences*, vol. 179, pp. 3432–3441, Sept. 2009.