A CONVEX OPTIMIZATION ALGORITHM FOR SPARSE REPRESENTATION AND APPLICATIONS IN CLASSIFICATION PROBLEMS

REINALDO SANCHEZ ARIAS

Computational Science Program



APPROVED:

Miguel Argáez, Ph.D., Chair

Rodrigo Romero, Ph.D.

Martine Ceberio, Ph.D.

Benjamin C. Flores, Ph.D.
Dean of the Graduate School

*A mi amada madre Alid,*

*mi padre Reinaldo, y mi hermano Juan Camilo*

*que son la luz de mi vida.*

# A CONVEX OPTIMIZATION ALGORITHM FOR SPARSE REPRESENTATION AND APPLICATIONS IN CLASSIFICATION PROBLEMS

by

REINALDO SANCHEZ ARIAS

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Computational Science Program

THE UNIVERSITY OF TEXAS AT EL PASO

May 2013

# Acknowledgements

I would like to express my gratitude to my advisor Dr. Miguel Argáez for his guidance and encouragement. I am thankful for the opportunity to work with him and for his support and advice. I also want to acknowledge Dr. Rodrigo Romero and Dr. Martine Ceberio for kindly accepting being part of my thesis committee, and for their valuable observations and help.

I would like to thank the Computational Science Program and Department of Mathematical Sciences professors and staff for all their hard work and dedication. Special thanks to Cindy Aguilar and Maria Salayandia for their always kind assistance. My gratitude to Dr. Argáez and Dr. Velazquez who gave me the opportunity to start my doctoral studies at UTEP. Also thanks to Dr. Sewell, Dr. Yi, Dr. Kreinovich, Dr. Wojciechowski, Dr. Martinez, and Dr. Duque, for their help through my career.

Infinitely many thanks to my family for their endless support. To my dear mother Alid for her unlimited love and encouragement in every moment; to my brother Juan Camilo for his motivation and kindness; and to my dad Reinaldo for always lighting my way and for taking care of us from Heaven. This work is dedicated to them. I thank God every day for having such an amazing family. Nothing in a simple paragraph can express the love I have for the three of you. During all these years, you were my inspiration, and your support and love are irreplaceable.

Thanks to my dear love Mirel for always putting a smile on my face, and for helping me to stay positive every time. It is a blessing to have her in my life. Also thanks to her family for their affection and trust.

During this experience away from home I have been blessed with the company of great people and friends. I am forever thankful to my fellow classmates and friends Paula (mi hermanita), Anibal, Carlos, Ron, Javier, Clemente, and Quintero for their affection and the great memories. Also, thanks to The Khamsis, Samer, Abe, Josh, Mich, John and all the

other ones that have made me feel as part of their family. I am glad I had the chance to share this adventure with them. Many thanks also to Rev. Wilson Cuevas and Rev. Henry Beck for their prayers and counseling.

This work was supported by the Department of the Army ARL Grant No. W911NF-07-2-0027 and the College of Science at UTEP.

# Abstract

In pattern recognition and machine learning, a classification problem refers to finding an algorithm for assigning a given input data into one of several categories. Many natural signals are sparse or compressible in the sense that they have short representations when expressed in a suitable basis. Motivated by the recent successful development of algorithms for sparse signal recovery, we apply the selective nature of sparse representation to perform classification. Any test sample is represented in an overcomplete dictionary with the training sample as base elements. A given test sample can be expressed as a linear combination of only those training samples belonging to the same class, therefore providing a naturally sparse representation. Finding the correct coefficients in a given basis or training dataset, allows us to identify the correct category or class of any given input that needs to be categorized. In order to find such sparse linear representation, we implement an $\ell_1$-minimization algorithm. This methodology overcomes the lack of robustness with respect to outliers, and in contrast to other classification algorithms, no model selection dependence is involved in the optimization method. The minimization algorithm is a convex relaxation-like algorithm that has been proven to efficiently recover sparse signals. To study its performance, the proposed method is applied to several test datasets with different number of features and samples. A dimensionality reduction technique is also proposed and implemented as part of the classification process.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Many areas of engineering and applied mathematics benefit greatly from the results and methods developed in numerical optimization. Applications in signal processing, machine learning, economics, geophysics, biosciences, physical phenomena simulations, among others are enhanced and improved with the different achievements in numerical optimization.

Several engineering and science applications involve solving linear inverse problems that are usually ill-conditioned and for which the use of regularization techniques is required to be able to propose useful solutions. Recently, regularization via *sparsity* constraints has become very popular, where we look for an approximate solution to a linear system of equations, with the requirement that it has as few nonzero components as possible. This kind of problems can be found in several applications in machine learning, image and signal processing, and coding and information theory among others. Many natural signals are sparse or compressible in the sense that they have short representations when expressed in a suitable basis. Moreover, it has been proven that sparse signals can effectively approximate compressible signals [6], [9],[14].

The theory of *compressed sensing* (compressive sampling) has been a "hot" research topic of interest for many applied mathematics and engineering researchers in recent years. The work in this area was initiated in late 2004 by Emmanuel Cands, Justin Romberg, and Terence Tao [9], and independently by David Donoho [19]. The general theme aims to answer the question of how much information is necessary to accurately reconstruct a signal. Several encouraging numerical results followed by theoretical conditions and characterizations, showed that one can reconstruct sparse or compressible signals accurately from a very limited number of measurements. This motivated the rise of new techniques imaging sciences and

signal processing, based on compressed sensing and sparse representation methods, with a broad set of applications in engineering and sciences.

In machine learning and pattern recognition, the term "classification" refers to the result of an algorithm/technique for assigning a given set of input data into one of a given number of categories. An example would be assigning a given email into "spam" or "non-spam" classes, or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). An algorithm that implements classification is referred to as a classifier. Inspired by the recent successful development of algorithms for sparse signal recovery [17, 23, 28, 35], we apply the *selective nature* of sparse representation to perform classification. Any test sample is represented in an overcomplete dictionary with the training sample as base elements. In such way, test samples can be expressed as a linear combination of only those training samples belonging to the same class, therefore providing a naturally sparse representation. In order to find the sparsest linear representation we propose an optimization algorithm based on $\ell_1$-minimization that allows us to overcome the lack of robustness to outliers [8]. Convex relaxation techniques based on $\ell_1$ optimization have shown great results in compressed sensing and sparse representation problems. The use and development of $\ell_1$ optimization approaches present several advantages over nonconvex optimization and greedy pursuit methodologies also proposed to attack the sparse signal reconstruction problem.

*Sparse representations* of signals have received a great deal of attention in recent years. The sparse representation problem consists in searching for the most compact representation of a signal in terms of a linear combination of *atoms* in an overcomplete *dictionary.*

Research has focused on *pursuit methods* for solving the optimization problem, such as matching pursuit [37], orthogonal matching pursuit [40], basis pursuit [17], and also on the *applications* of a sparse representation for different tasks, such as signal separation, denoising, and coding.

This dissertation is organized as follows:

**Chapter 2** presents the mathematical background in the theory of compressed sensing that

gave rise to the development of efficient optimization algorithms for sparse signal recovery. We explain the formulation for the sparse representation problem, the ideas guaranteeing the recovery of sparse signals via $\ell_1$ minimization, and some of the strategies to solve this problem.

**Chapter 3** includes a description of the novel $\ell_1$-minimization algorithm we propose to use for solving the classification problem. Convergence results for our algorithm and comparisons with other state-of-the-art solvers are also presented.

**Chapter 4** explains the classification problem we aim to solve using a sparse representation approach. We motivate the use of the sparse representation in classifications tasks, and show how to use the $\ell_1$ minimization algorithm developed in this work for solving classification problems. A description of the mathematical formulation and the strategies used are presented.

**Chapter 5** presents numerical results of the classification technique we propose in this work for different datasets. We explain the experiment design, describe the datasets used, and present a comparison of our results with commonly used algorithms for classification.

**Chapter 6** includes a discussion of techniques to enhance the classification algorithm. A novel dimensionality reduction approach is also proposed.

**Chapter 7** includes the conclusions of our work and the future research directions we have in mind to improve our technique. We describe some strategies that can be used to enhance our methods and discuss their viability.

# Chapter 2

# Sparse Solution of Linear Inverse Problems

The problem of sparse representation consists in representing a given signal as a linear combination of as few "base" elements as possible from a fixed collection. That is, we aim to identify a sparse vector $x \in \mathbb{R}^n$ such that the target signal $b \in \mathbb{R}^m$ can be represented by $Ax \approx b$, where $A$ is a known $m \times n$ matrix. In this chapter we formulate the problem that must be solved in order to obtain approximate sparse solutions to linear systems of equations, and discuss the strategies that have been proposed in recent years.



## 2.1 Problem Formulation

Consider a real matrix $A \in \mathbb{R}^{m\times n}$ whose columns $a_j$ have unit Euclidean norm, that is $\|a_j\|_2 = 1$, for $j = 1, ..., n$. We will often refer to this type of matrix as the *dictionary*. We say that a vector (signal) $x \in \mathbb{R}^n$ is *k-sparse* if $\|x\|_0 \leq k$, where the counting function $\|\cdot\|_0 \colon \mathbb{R}^n \to \mathbb{R}$, known as the $\ell_0$ "norm" [22], gives the number of nonzero elements in its argument. In other words,

$$\|x\|_0 = \text{card}\,\{i \colon x_i \neq 0\}\,. \tag{2.1}$$

Even though we call it the $\ell_0$-norm, one can easily verify that it does not satisfy the positive homogeneity (positive scalability) property in the definition of a norm. Namely we have that $\|\lambda x\|_0 \neq |\lambda|\|x\|_0$, for any given nonzero scalar $\lambda$.

A signal $x$ is said to be *nearly sparse* if the rearranged entries of $x$ decay exponentially

when sorted in decreasing order of magnitude [9]. Since compressible signals are well approximated by sparse ones, the framework of sparse approximation applies to this class too.

Given that we are looking for the sparsest vector $x$ satisfying the linear system of equations $Ax = b$, we are interested in solving the following optimization problem:

$$\begin{aligned} \min \quad & \|x\|_0 \\ \text{subject to} \quad & Ax = b, \end{aligned} \tag{2.2}$$

assuming that the matrix $A \in \mathbb{R}^{m \times n}$ is short and wide, that is $m \ll n$. Unfortunately, Problem (2.2) is a combinatorial minimization problem and NP-hard (non-deterministic polynomial-time) [39]. Therefore any algorithm that is intended to solve (2.2) given the matrix $A$ and the vector $b$ will be computationally intractable. Thus, strategies to overcome this difficulty had to be developed, which gave rise to different algorithmic approaches with remarkable results in different applications.

## 2.2 Algorithmic Approaches

During the last decade, several strategies have been proposed to find approximate solutions to problem (2.2). These different approaches include:

***Convex Relaxation.*** In this case, the objective function in Problem (2.2) is replaced by a convex function (as the $\ell_1$ norm), overcoming the combinatorial nature of the problem [17].

***Nonconvex Optimization.*** The idea consists in relaxing the $\ell_0$ norm with a related nonconvex function, and attack the problem by identifying the corresponding stationary points. The use of $\ell_q$ quasi-norms ($0 < q < 1$) has been studied in [2, 14, 38].

***Greedy Pursuit.*** Iterative refinement of a sparse solution is proposed, by successively identifying those entries in the vector producing the greatest improvement [37].

In this work, we focus on developing a ***Convex Relaxation*** technique for finding an approximate solution to the sparse representation problem. The strategy uses an $\ell_1$ relaxation of the $\ell_0$ norm, through which successful recovering of sparse signals has been shown. We solve the $\ell_1$ optimization problem by iteratively solving a sequence of convex subproblems that depend on a regularization parameter. The methodology falls in a path-following framework, and convergence results along with comparison with other methodologies show its efficiency in finding sparse solutions of linear systems of equations.

## 2.3 $\ell_1$-minimization problem

A practical alternative to Problem (2.2) is the $\ell_1$ minimization approach, which consists in finding the solution to the problem

$$\begin{aligned} \min \quad & \|x\|_1 \\ \text{subject to} \quad & Ax = b, \end{aligned} \tag{2.3}$$

where $\|x\|_1 = \sum_{i=1}^{n} |x_i|$. We now have an optimization problem whose objective function is convex, unlike the $\ell_0$-norm in Problem (2.2). However, we must have special conditions on the matrix $A$ and on the sparsity of $x$ in order to guarantee that the solution of Problem (2.3) will lead us to find the solution of the original problem.

The motivation for this approach comes from studying the theory of ***Compressed Sensing*** (compressive sampling) which has been a research topic of interest in the last years. The work in this area initiated in late 2004 by Emmanuel Candès, Justin Romberg and Terence Tao [9], and independently by David Donoho [19]. The general theme aims to answer the question: *How much information is necessary to accurately reconstruct a signal?*. It turns out that one can reconstruct *sparse* or *compressible* signals accurately from a very limited number of measurements. We wish to recover an object $x \in \mathbb{R}^n$, using information from a collection of $m$ linear measurements $b_i = \langle a_i, x \rangle$ for $i = 1, ..., m$. In matrix notation, we can

write this as $b = Ax$, where $A \in \mathbb{R}^{m \times n}$ has the vectors $a_i$ as rows. We will assume that $m < n$ and the *measurement matrix* $A$ has full rank [22].

### Restricted Isometry Property

We will say that a matrix $A$ satisfies the *restricted isometry property* (RIP) with parameters $(r, \delta)$ if (see [10]):

$$(1-\delta)\|x\|_2^2 \quad \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2, \quad \text{for all } r \text{ sparse vectors } x. \tag{2.4}$$

The restricted isometry constant $\delta_r$ of a matrix $A$ is the smallest number satisfying (2.4). This property essentially requires that every set of columns with cardinality less than $r$ approximately behaves like an orthonormal system [9]. In the following, let us present a mathematical result that shows the effectiveness of the $\ell_1$ optimization approach when finding sparse solutions to linear systems of equations.

### The Null Space Property

A matrix $A \in \mathbb{R}^{m \times n}$ satisfies the *null space property* (NSP) of order $r$ with constant $\gamma \in (0, 1)$ if (see [15])

$$\|v_S\|_1 \quad \leq \gamma \|v_{S^c}\|_1, \tag{2.5}$$

for all sets $S \subset \{1, ..., n\}$ with $\#S \leq r$, and $v \in \ker(A)$. Here $S^c$ is the complement of $S$ in the set $\{1, ..., n\}$. It can be shown that if a matrix $A$ satisfies the *restricted isometry property* (2.4) then it also satisfies the *null space property* (see [15]).

## Sparse Recovery Result

Let $A \in \mathbb{R}^{m\times n}$ be a matrix satisfying the NSP of order $r$ with constant $\gamma \in (0,1)$. Let $x^*$ be the solution of the $\ell_1$-minimization Problem (2.3). If $x \in \mathbb{R}^n$ and $Ax = b$, then

$$\|x - x^*\|_1 \leq \tfrac{2(1+\gamma)}{(1-\gamma)}\sigma_x, \tag{2.6}$$

where $\sigma_x$ is a quantity that depends on the sparsity of $x$. If the vector $x$ is $r$-sparse then $x = x^*$.

**Proof** Since $Ax = Ax^* = b$, then the vector $v = x - x^*$ is in $\ker(A)$. Also, since $x^*$ solves (2.3), then $\|x^*\|_1 \leq \|x\|_1$. Let $S$ be the set of the $r$ largest components of $x$ in absolute value. We have

$$\|x^*_S\|_1 + \|x^*_{S^c}\|_1 \leq \|x_S\|_1 + \|x_{S^c}\|_1.$$

Notice also that (use triangle inequality)

$$\|x_S\|_1 - \|v_S\|_1 + \|v_{S^c}\|_1 - \|x_{S^c}\|_1 \leq \|x_S\|_1 + \|x_{S^c}\|_1,$$

so we get

$$\begin{aligned} \|v_{S^c}\|_1 &\leq \|v_S\|_1 + 2\|x_{S^c}\|_1, \\ &\leq \gamma\|v_{S^c}\|_1 + 2\sigma_x. \end{aligned}$$

Therefore, $\|v_{S^c}\|_1 \leq \frac{2}{(1-\gamma)}\sigma_x$. Since $v = x - x^*$, then

$$\begin{aligned} \|x - x^*\| &= \|v_S\| + \|v_{S^c}\| \\ &\leq (\gamma+1)\|v_{S^c}\| \\ &\leq \tfrac{2(1+\gamma)}{(1-\gamma)}\sigma_x. \end{aligned}$$

In case the vector $x$ is $r$ sparse, then $\|x_{S^c}\|_1 = \sigma_x = 0$, so we get $x = x^*$.

We have shown with the results presented above, that the notion of $\ell_1$ minimization is indeed an effective technique for finding the sparsest solution $x^*$ of a linear system of equations $Ax = b$.

## 2.4 Convex Relaxation Strategies

The $\ell_1$ convex relaxation approach has been proven to sucessfully find sparse solutions to linear system of equations [22]. In the following, we briefly describe some state-of-the-art algorithms developed for finding approximate solutions of the sparse representation problem based on $\ell_1$ optimization techniques.

### Donoho, Saunders et al. - Basis Pursuit (BP)

In their work [17], Donoho et al. proposed to reformulate Problem (2.3) as a linear programing problem of the form

$$\begin{aligned} \min_x \quad \sum_{i=1}^{n} & \quad u_i \\ \text{s.t} & \quad -u_i \leq x_i \leq u_i, \\ & \quad Ax = b. \end{aligned} \tag{2.7}$$

They were able to solve linear programs of size 8192 by 212,992. In their work, reasonable success with a primal-dual logarithmic barrier method and a conjugate gradient solver was obtained. It is easy to check that Problem (2.3) is equivalent to

$$\begin{aligned} \min & \quad c^T z \\ \text{subject to} & \quad \Phi z = f, \; z \geq 0, \end{aligned} \tag{2.8}$$

by letting $\Phi = [A, -A]$, $f = b$, $c = (\mathbb{1}; \mathbb{1})$, $z = (u, v)$ and $x = u - v$. Here, $\mathbb{1} \in \mathbb{R}^n$ is a vector with all components equal to 1.

Even though the approach provides strong guarantees and stability, it relies on *linear programming*, whose methods do not yet have strong polynomially bounded runtimes. It is worthwhile to mention, that the work by the authors of [17] was done several years before the results Candès and Tao proved on the recovery of sparse signals via the $\ell_1$ minimization approach. In [11], Candès and Tao characterized the conditions that must be satisfied for finding the actual solution to the original problem (2.2), when using the $\ell_1$-minimization alternative.

A natural variation to the basis pursuit Problem (2.3) consists in relaxing the linear constraint in order to consider an error tolerance, say $\epsilon \geq 0$, for the situation when the signal is contaminated with some additive noise. More specifically, the following problem is considered:

$$\begin{aligned} \min \quad & \|x\|_1 \\ \text{subject to} \quad & \|Ax - b\|_2 \leq \epsilon. \end{aligned} \tag{2.9}$$

The work in [9] claims that the convex relaxation approach (2.9) is also effective in finding an approximated solution of the sparse problem (2.2) whenever the observations are contaminated with a bounded additive noise.

## Boyd, Lustig et al.

Boyd and his research group [35] proposed to solve a generalized version of (2.3) that allows certain degree of noise, given by the unconstrained minimization problem

$$\min_x \quad \lambda \|x\|_1 + \|Ax - b\|_2^2. \tag{2.10}$$

where the parameter $\lambda > 0$ is used as a penalization parameter balancing the tradeoff between error and sparsity.

First, Problem (2.10) is posed as the following constrained optimization problem

$$\begin{aligned} \min \quad & \lambda \sum_{i=1}^{n} u_i + \|Ax - b\|_2^2 \\ \text{s.t} \quad & -u_i \leq x_i \leq u_i. \end{aligned} \tag{2.11}$$

Secondly, using the notions of interior-point method (log-barrier method) they designed an algorithm to find a solution of the dual problem of (2.11). Their method makes use of a preconditioned conjugate gradient (PCG) to accelerate convergence and stabilize the algorithm. They also showed the application of their algorithm on a magnetic resonance imaging (MRI) data set. One drawback of their approach is that each step would require the solution of a Newton system of the form $H\Delta x = g$, where $H \in \mathbb{R}^{2n\times 2n}$ is the Hessian matrix and $g$ is the gradient at the current iterate. To overcome this difficulty, they compute a search direction of an approximate Newton system using a PCG. This alternative is commonly known as the Truncated Newton Method. The truncation rule for the PCG provides the condition for terminating the algorithm. The total number of PCG iterations required by the truncated Newton interior-point method depends on the value of the regularization parameter $\lambda$ and a given relative tolerance $\epsilon$. An implementation of their algorithm is available at `http://www.stanford.edu/~boyd/l1_ls/`.



## Figueiredo, Wright et al.

Figueiredo et al. [23] studied the unconstrained problem

$$\min_x \quad \lambda\|x\|_1 + \tfrac{1}{2}\|b - Ax\|_2^2, \tag{2.12}$$