A Linearly Convergent Algorithm for Solving a Class of Nonconvex/Affine Feasibility Problems

Amir Beck and Marc Teboulle

July 30, 2010

Abstract

We introduce a class of nonconvex/affine feasibility problems, called (NCF), that consists of finding a point in the intersection of affine constraints with a nonconvex closed set. This class captures some interesting fundamental and NP hard problems arising in various application areas such as sparse recovery of signals and affine rank minimization that we briefly review. Exploiting the special structure of (NCF), we present a simple gradient projection scheme which is proven to converge to a unique solution of (NCF) at a linear rate under a natural assumption explicitly given defined in terms of the problem's data.

Keywords: Nonconvex affine feasibility, inverse problems, gradient projection algorithm, linear rate of convergence, scalable restricted isometry, mutual coherence of a matrix, sparse signal recovery, compressive sensing, affine rank minimization.

AMS 2010 Subject Classification: 90C30, 90C26

1 Introduction

Let $\mathbb{E}$ and $\mathbb{V}$ be finite dimensional Euclidean spaces, $\mathcal{A}:\mathbb{E}\to\mathbb{V}$ a given linear mapping and $\mathbf{b}\in\mathbb{V}$ a vector of observations. Consider the feasibility problem defined by

$$(NCF)\ \text{Find } \mathbf{x}\in\mathcal{C} \text{ such that } \mathcal{A}(\mathbf{x})=\mathbf{b},$$

where $\mathcal{C}\subseteq\mathbb{E}$ is a set which describes some a priori information on the unknown element $\mathbf{x}$. One natural approach for tackling the feasibility problem (NCF) is via the associated minimization problem

$$(NC)\ \ \min\left\{\frac{1}{2}\|\mathcal{A}(\mathbf{x})-\mathbf{b}\|^2:\ \mathbf{x}\in\mathcal{C}\right\} \tag{1.1}$$

for some given norm in $\mathbb{V}$.[1]

[1] Throughout the paper $\|\cdot\|$ will denote the endowed norm of the relevant Euclidean space (either $\mathbb{E}$ or $\mathbb{V}$).

The above problem formulations are very well known and have been extensively studied over the last several decades, in particular when $\mathcal{C}$ is a closed convex subset of $\mathbb{E}$, giving rise to the so-called convex feasibility problems, see the comprehensive review paper [1] and references therein. Problems of this kind naturally arise in the area of linear inverse problems which covers a wide range of data processing problems such as imaging sciences, optics, astrophysics to name just few, see e.g., the classical monograph [17] and references therein. In such situations, one has to derive an estimate of some physical quantity of interest (e.g., a signal or an image) from given measurements and some a priori information described through the set $\mathcal{C}$, see for instance the in-depth review paper [12]. Furthermore, it should be noted that nonconvex feasibility problems have also been studied in the literature, see for instance, [13],[7]. In particular, the method of successive projections for closed convex sets was extended in [13] to a class of nonconvex compact sets satisfying some hypothesis.

A current trend of research in the data processing areas (e.g., signal processing, machine learning etc..) which has recently attracted a lot of attention focuses on solving problems that can recover *sparse* objects. Finding the sparsest solution of a linear system or the more general problem that consists of finding a low rank matrix satisfying linear matrix equations are at the heart of these current activities. These problems being generally NP hard are often solved by their convex relaxations. The current algorithmic, theoretical and applications literature is vast, and we refer the reader to the excellent very recent survey papers [6] and [21] and references therein.

In this paper we depart from the convex relaxation approach. We focus on the class of problems (NCF) where the constraint set $\mathcal{C}$ is a closed and *nonconvex* subset of $\mathbb{E}$, which will be defined to naturally captures sparsity features, and we propose to solve the nonconvex feasibility problem (NCF) via a very simple gradient projection algorithm which under a natural assumption on the problem's data is proven to converge linearly to a global optimal solution of (NC).

The paper is organized as follows. In Section2 we define the problem and give some examples arising in fundamental applications that naturally fit our formalism. Section 3 first gives some background on the so-called Restricted Isometry Property (RIP) which has been central in the analysis of sparse recovery problems via their *convex* relaxations. This leads us to introduce a natural extension of RIP, called *Scalable Restricted Isometry Property* (SRIP) for the class of problems under study, and that will play a key role in the analysis of the proposed gradient projection scheme and which here solved directly the nonconvex problem. The analysis is developed in Section 4, where we prove that despite the nonconvex nature of the problem, if (SRIP) is satisfied, the gradient projection method converges at a linear rate to a global optimal solution of (NC) also shown to be unique. The convergence is established both for a constant and backtracking stepsize rules, the later being particularly useful in applications as it does not require the knowledge of any unknown parameter. The algorithm is useful and efficient whenever the projection map onto the nonconvex set is easy to compute, this is shown to be the case in the context of sparse recovery problems, for which we also derive a further interesting consequence from our main convergence result.

2 Problem Statement, Motivation and Examples

2.1 General Problem Statement

In most practical applications, prior knowledge on some desired features of the unknown $\mathbf{x} \in \mathbb{E}$ is available, and can be quantified by some given function, e.g., a norm like function. The motivation for the proposed definition will be described below.

Definition 2.1. $\mathcal{S}$ is the set of all functions $\varphi : \mathbb{E} \to \mathbb{R}_+$ which are lower semi-continuous (lsc) function satisfying the following properties:

$$(i) \qquad \varphi(\mathbf{0}) = 0, \tag{2.1}$$

$$(ii) \qquad \varphi(\mathbf{x}) = \varphi(-\mathbf{x}) \text{ (symmetry)}, \tag{2.2}$$

$$(iii) \qquad \varphi(\mathbf{x}+\mathbf{y}) \leq \varphi(\mathbf{x}) + \varphi(\mathbf{y}) \text{ (subadditivity)}. \tag{2.3}$$

We are interested in the situation where $\varphi \in \mathcal{S}$ is *nonconvex* and we want to solve the nonconvex feasibility problem:

$$(NCF) \;\; \text{Find } \mathbf{x} \in \mathcal{C}_s \text{ such that } \mathcal{A}(\mathbf{x}) = \mathbf{b},$$

where the admissible constraint is defined by the closed *nonconvex* set

$$\mathcal{C}_s := \{\mathbf{x} \in \mathbb{E} : \; \varphi(\mathbf{x}) \leq s\} \tag{2.4}$$

for some fixed given $s > 0$.

To solve (NCF), we consider the related nonconvex minimization problem

$$(NC) \;\; \min \left\{ f(\mathbf{x}) \equiv \frac{1}{2}\|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|^2 : \mathbf{x} \in \mathcal{C}_s \right\}, \tag{2.5}$$

where $\|\cdot\|$ is the underlying norm of the Euclidean space $\mathbb{V}$. For example $\|\cdot\|_2$ when $\mathbb{E} = \mathbb{R}^n$ and $\|\cdot\|_F$ (the Frobenius norm) when $\mathbb{E} = \mathbb{R}^{m\times n}$. Given that (NCF) has a solution, the optimal value of (NC) is zero and $\bar{\mathbf{x}}$ is an optimal solution of (NC) if and only if $\bar{\mathbf{x}}$ is a solution to (NCF).

This formalism encompasses a wide class of problems which has attracted considerable interest in the recent literature and which has triggered the motivation of this work, this will be briefly discussed below. We end by noting that well known alternative ways to tackle (NCF) include the following three closely related problems:

$$\begin{array}{ll} \min & \{\varphi(\mathbf{x}) : \; \|\mathcal{A}(\mathbf{x}) - \mathbf{b}\| \leq \eta, \; \mathbf{x} \in \mathbb{E}\} \; (\eta > 0, \text{ perturbed case}), \\ \min & \{\varphi(\mathbf{x}) : \mathcal{A}(\mathbf{x}) = \mathbf{b}, \; \in \mathbb{E}\}, \\ \min & \{\|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|^2 + \tau\varphi(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}, \end{array} \tag{2.6}$$

where the last formulation corresponds to a penalty approach, with a penalty parameter $\tau > 0$ which measures the tradeoff between the error in the approximation measured by $\|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|^2$ and the desired property of the unknown $\mathbf{x}$ quantified by the function $\varphi(\mathbf{x})$. Note that all these formulations remain essentially NP hard for the choices of the nonconvex function $\varphi \in \mathcal{S}$ which are described next.

2.2 Motivation and Examples

We briefly describe three models of interest in applications that naturally fit as special cases of the proposed formalism of this paper.

Example 2.1. [*Compressive Sensing*] Roughly speaking, in the new emerging compressed sensing technology we are interested in recording as much information as possible in a signal or image $\mathbf{x}$ in the "cheapest" way. In other words, under suitable condition on the problem's data, few measurements are enough to correctly recover a signal, see [14] for more details.

Let $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^m$. Here the mapping $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ can be represented by an $m \times n$ matrix $\mathbf{A}$ satisfying $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for every $\mathbf{x} \in \mathbb{R}^n$ (for the sake of notation consistency with the other examples, we will often not use the "matrix" notation). A typical approach is to select a sparse vector, namely with many zero components, that solves a linear system of equations $\mathcal{A}(\mathbf{x}) = \mathbf{b}$ for $\mathbf{x} \in \mathbb{R}^n$. Let $\|\mathbf{x}\|_0$ be the l_0-norm[2] of $\mathbf{x}$ which counts the number of nonzero components of $\mathbf{x}$. Given that the observed vector $\mathbf{b} \in \mathbb{R}^m$ and that the number of measurements is smaller than the size of the vector $\mathbf{x}$, i.e., that $m < n$, the sparse reconstruction problem amounts to finding an s-sparse solution (with $s \ll n$) of a nonempty linear system, i.e.,

$$\text{find } \mathbf{x} \in \mathbb{R}^n \text{ with } \|\mathbf{x}\|_0 \leq s \text{ such that } \mathcal{A}(\mathbf{x}) = \mathbf{b}.$$

Clearly, this problem is a special case of our model (NCF) with $\mathcal{S} \ni \varphi(\mathbf{x}) := \|\mathbf{x}\|_0$, since the l_0-norm satisfies all the premises of Definition 2.1.

Example 2.2. [*Affine Rank Minimization*]. Let $\mathbb{E} = \mathbb{R}^{m\times n}, \mathbb{V} = \mathbb{R}^p$ and $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ a linear map. The problem consists of finding a matrix $\mathbf{x} \in \mathbb{R}^{m\times n}$ of minimal rank, that satisfies a given system of linear matrix equations $\mathcal{A}(\mathbf{x}) = \mathbf{b}$. This is a fundamental problem in many diverse areas, see [21]. Recall that the rank of a matrix is the number of its positive singular values. Thus, when $\mathbf{x}$ is a square diagonal matrix with diagonal elements x_j, the rank function coincides with the l_0-norm of $\mathbf{x}$, and the affine rank minimization problem can be viewed as a natural extension of the previous compressive sensing example.

Now, with $\varphi(\mathbf{x}) := \text{rank}(\mathbf{x})$, one has $\varphi \in \mathcal{S}$ since $\text{rank}(\mathbf{0}) = 0$, $\text{rank}(-\mathbf{x}) = \text{rank}(\mathbf{x})$ and $\text{rank}(\mathbf{x} + \mathbf{y}) \leq \text{rank}(\mathbf{x}) + \text{rank}(\mathbf{y})$ and the conditions in Definition 2.1 are thus satisfied, so that the problem of finding a matrix of rank at most s satisfying $\mathcal{A}(\mathbf{x}) = \mathbf{b}$ fits our model (NCF).

Note that both problems described in Examples 2.1 and 2.2 are also often tackled through either one of the corresponding three related optimization problems described via (2.6).

Example 2.3. [l_p*-pseudo norm minimization,* $0 < p < 1$].

Let $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^m$ and $\varphi_p(\mathbf{x}) := \|\mathbf{x}\|_p^p = \sum_{j=1}^n |x_j|^p$ $(0 < p < 1)$. The l_p pseudo-norms are connected to the l_0-norm via the relation $\|\mathbf{x}\|_0 = \lim_{p\to 0^+} \|\mathbf{x}\|_p^p$ (with the convention $0^0 = 0$). Thus, for instance, one could try to solve an approximation of the sparse recovery problem by solving the resulting nonconvex minimization models with $\varphi_p(\cdot)$ for small p. This

[2]This is by some abuse of terminology, since $\|\mathbf{x}\|_0$ is not a norm, as it clearly does not satisfy the the homogeneity property of a norm.

approach is well known, and it has been recently considered by several authors, see e.g., [10] and references therein.

We now verify that $\varphi_p \in \mathcal{S}$. Clearly, we have $\varphi_p(\mathbf{0}) = 0, \varphi_p(-\mathbf{x}) = \varphi_p(\mathbf{x})$. Moreover, it is easy to see that for any $p \in (0,1)$ one has $(u+v)^p \leq u^p + v^p$ for all $u, v \geq 0$, from which it follows that $\|\mathbf{x}+\mathbf{y}\|_p^p \leq \|\mathbf{x}\|_p^p + \|\mathbf{y}\|_p^p$ so that, as in the previous two examples, the conditions of Definition 2.1 are satisfied and thus this problem fits our formalism.

The last example, but now with $p = 1$, that results in the l_1-norm $\varphi_1(\mathbf{x}) = \|\mathbf{x}\|_1 := \sum_{j=1}^n |x_j|$ of $\mathbf{x} \in \mathbb{R}^n$, and which is a *convex* relaxation of the l_0-norm[3], is of particular interest. It leads us in the next section to first review some of the recent interesting results in sparse recovery problems which rely on the so-called *restricted isometry property* and also provide the motivation for introducing a natural extension of this notion within our formalism, and that will play an essential role in our analysis.

3 A Scalable Restricted Isometry Property – SRIP

3.1 Convex Relaxation and Restricted Isometry

In sparse solutions of linear systems and affine rank minimization, one faces to solve two computationally intractable combinatorial problems [20, 21]:

$$\begin{aligned} &(CS) && \min\{\|\mathbf{x}\|_0 : \ \mathcal{A}(\mathbf{x}) = \mathbf{b}, \mathbf{x} \in \mathbb{R}^n\}, \\ &(AR) && \min\{\text{rank}(\mathbf{x}) : \mathcal{A}(\mathbf{x}) = \mathbf{b}, \mathbf{x} \in \mathbb{R}^{m\times n}\}. \end{aligned}$$

Recent and extensive studies (see [6, 21] and their references) have shown that under appropriate assumptions on the data, that will be discussed shortly, it is possible to solve these problems via their *convex* relaxations. More precisely, we replace the l_0-norm and the rank function by their tractable convex counterparts, namely the l_1-norm in (CS) and the Ky-Fan (nuclear) norm in (AR). The nuclear norm of a matrix $\mathbf{x} \in \mathbb{R}^{m\times n}$ is denoted by $\|\mathbf{x}\|_*$ and is defined as the sum of the nonzero singular values of $\mathbf{x}$. It is the convex envelope of the rank function over the set $\{\mathbf{x} \in \mathbb{R}^{m\times n} : \ \|\mathbf{x}\|_F \leq 1\}$, see [18]. It should be noted that the idea of using the l_1-norm in the context of sparsity is not a new idea, and goes back to some works in geophysics, see [23, 22].

The convex relaxed problems for (CS) and (AR) which provide lower bounds to the original problems then read as two well-known problems:

$$\begin{aligned} &(ConvCS) && \min\{\|\mathbf{x}\|_1 : \ \mathcal{A}(\mathbf{x}) = \mathbf{b}, \mathbf{x} \in \mathbb{R}^n\} \ \text{(Basis Pursuit, [11])}, \\ &(ConvAR) && \min\{\|\mathbf{x}\|_* : \mathcal{A}(\mathbf{x}) = \mathbf{b}, \mathbf{x} \in \mathbb{R}^{m\times n}\} \ \text{(Trace minimization, [18])}. \end{aligned}$$

Both problems above are tractable convex optimization problems that can be efficiently solved by many convex minimization schemes, see for instance the fast and simple optimal gradient based scheme recently developed in [2], and also the recent review [3] and references therein.

[3]The l_1-norm of $\mathbf{x} \in \mathbb{R}^n$ is the lowest convex envelope of $\|\mathbf{x}\|_0$ over the l_∞ unit ball.

The main question that has been extensively investigated in the literature is then

Main question: For which $\mathcal{A}$, a sparse solution (a low rank matrix) can be recovered? That is to say, under which conditions an optimal solution of the original nonconvex problems (CS) and (AR) can be obtained by solving their convex counterparts (ConvCS) and (ConvAR) respectively?

One of the first results to answer that question was for the compressed sensing l_0-minimization problem (CS) and was obtained via the concept of *mutual coherence of a matrix*, which is also related the forthcoming property. For the interested reader, we have briefly summarized some of these pertinent results in the appendix.

Another concept which plays a fundamental role in answering the main stated question is the so-called *Restricted Isometry Property*, for short (RIP). Below we state the definition of RIP for the general matrix rank minimization recently introduced in [21] as a natural generalization of the vector case which is recalled after the definition (and which can be recovered by setting $\mathbf{x}$ to be a diagonal matrix).

Definition 3.1. The linear map $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ with $m < n$ and $1 \leq d \leq m$ is said to satisfy the Restricted Isometry Property (RIP) with the isometry constant δ_d associated to $\mathcal{A}$, if δ_d is the smallest number such that the following holds:

$$(1-\delta_d)\|\mathbf{x}\|^2 \leq \|\mathcal{A}(\mathbf{x})\|^2 \leq (1+\delta_d)\|\mathbf{x}\|^2 \text{ for all } \mathbf{x} \in \mathbb{R}^{m\times n} \text{ s.t. } \operatorname{rank}(\mathbf{x}) \leq d,$$

where $\|\cdot\|$ stands here for the Frobenius norm.

In the vector case (originally developed for compressed sensing problems, see e.g., [9]), the linear mapping $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ reads $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, and the RIP condition reduces to:

$$(1-\delta_d)\|\mathbf{x}\|^2 \leq \|\mathcal{A}(\mathbf{x})\|^2 \leq (1+\delta_d)\|\mathbf{x}\|^2 \text{ for all } \mathbf{x} \in \mathbb{R}^{n} \text{ s.t. } \|\mathbf{x}\|_0 \leq d.$$

The following two results answer the main question stated above. In the sequel we use the terminology "$\mathbf{x}$ is s-sparse" for all vectors such that $\|\mathbf{x}\|_0 \leq s$.

In [9], the following result has recently been proven for problem (CS).

Theorem 3.1 ([9])**.** *Consider problem (CS). Let* $\mathbf{b} = \mathcal{A}(\bar{\mathbf{x}})$ *for some s-sparse vector* $\bar{\mathbf{x}} \in \mathbb{R}^n$ *with* $s \geq 1$*. Then,*
(i) if $\delta_{2s} < 1$*, the l_0 problem (CS) has a unique s-sparse solution;*
(ii) if $\delta_{2s} < \sqrt{2} - 1$*, the optimal solution of the l_1-problem (ConvCS) is the same as of the l_0 problem.*

In a similar vein, in [21], the previous result has been extended for the rank minimization problem.

Theorem 3.2 ([21])**.** *Consider problem (AR). Let* $\mathbf{b} = \mathcal{A}(\bar{\mathbf{x}})$ *for some matrix* $\bar{\mathbf{x}} \in \mathbb{R}^{m\times n}$ *of rank* $s \geq 1$*. Then,*
(i) if $\delta_{2s} < 1$*, then* $\bar{\mathbf{x}}$ *is the unique matrix of rank at most s.*
(ii) if $\delta_{5s} < 1/10$*, then the optimal solution of the convex problem (ConvAR) coincides with the minimum rank solution of problem (AR).*

If either of the above RIP assumptions are satisfied for $\mathcal{A}$, for some given d, and with the requested upper bound on δ_d, we will simply write that RIP(d, δ_d) holds. Also, it is useful to note that if $s \leq t$, then $\delta_s \leq \delta_t$, i.e, $RIP(s, \delta_s) \implies RIP(t, \delta_t)$.

The good news: For both the vector and matrix cases, it has been proven that for some classes of random matrices (e.g., with i.i.d gaussian entries), the corresponding RIP can be proven to be satisfied with overwhelming probability. Details on these probabilistic analysis can be found for instance in [9, 14, 21]. However, not much is known for arbitrary *deterministic* matrices. .

The bad news: The RIP suffers from two major drawbacks:
(i) The RIP is lacking scalability.
(ii) Finding/computing the isometry parameter δ_d can be as difficult as solving the original NP hard problems (CS) and (AR).

Both issues will be addressed in this paper within our general model and the proposed algorithm. The first issue is addressed next, by introducing a natural modification of RIP.

3.2 SRIP: Scalable Restricted Isometry Property

As just mentioned, an evident drawback of the RIP assumption is its lack of scalability. For example, if a linear operator $\mathcal{A}$ satisfies the RIP with some parameters (s, δ_s), then surely $2\mathcal{A}$ will *not* satisfy the RIP with the same parameters. This is not the case for the notion introduced below which remedies this drawback by considering a straightforward and natural generalization of the RIP for our general problem (NCF), and which we call the scalable restricted isometry property (SRIP).

Let $\varphi \in \mathcal{S},\ d > 0$ and $\mathcal{A} : \mathbb{E} \to \mathbb{V}$. We write SRIP$(d, \alpha)$ if the following holds:

SRIP(d, α): There exist $\nu_d, \mu_d > 0$ satisfying $\frac{\mu_d}{\nu_d} < \alpha$ such that

$$\nu_d \|\mathbf{x}\| \leq \|\mathcal{A}(\mathbf{x})\| \leq \mu_d \|\mathbf{x}\| \quad \text{for every } \mathbf{x} \in \mathcal{C}_d.$$

By its definition, if SRIP(d, α) holds for some (d, α), then $\alpha > 1$. Of course, SRIP(d, α) might hold true for certain values of d, α and fail for others. The assumption is restrictive when d is "large" and α is "small" and loose when d is "small" and α is "large". This is reflected in the following lemma whose simple proof is omitted.

Lemma 3.1. *Suppose that $d_1 \leq d_2$ and $\alpha_1 \geq \alpha_2$. If SRIP(d_1, α_1) is satisfied, then SRIP(d_2, α_2) is also satisfied.*

Plugging

$$\mu_d^2 = 1 + \delta_d, \nu_d^2 = 1 - \delta_d. \tag{3.1}$$

In SRIP, the relationship between RIP and SRIP (in the settings of Examples 2.1, 2.2) is revealed through the following obvious result.

Lemma 3.2. *Let $\beta \in (0, 1)$. If RIP(d, δ_d) is satisfied for $\delta_d < \beta$, then SRIP$\Big(d, \sqrt{\frac{1+\beta}{1-\beta}}\Big)$ holds true.*

We re-emphasize that here we are concerned with solving the nonconvex model (NCF) directly rather than relaxing it. Much like the second drawback of the RIP alluded above (i.e., the necessity of knowing δ_{2s}), the determination of the unknown parameters (ν_d, μ_d) of SRIP appears as equally difficult. However, thanks to the proposed algorithmic framework which is developed next, we will show that finding/approximating these parameters is not an issue.

4 A Linearly Convergent Gradient Projection Method

4.1 The Gradient Projection Method for Solving Problem (NCF)

The gradient projection algorithm for minimizing a smooth function over some closed set is very well known and due to its simplicity is particularly adequate for solving large scale problems. However, even for convex problems, it suffers from a slow (e.g., sublinear) rate of convergence, see [4], and references therein.

We will prove that if SRIP$(2s, \sqrt{2})$ holds, the gradient projection method actually converges linearly to the solution of the nonconvex problem (NCF) which is also shown to be unique.

Before proceeding, we recall the notion of orthogonal projection. For a nonempty closed possibly nonconvex set $C \subseteq \mathbb{E}$, the projection of $\mathbf{y} \in \mathbb{E}$ onto C, written $P_C(\mathbf{y})$ is a multivalued map (as opposed to the convex case in which orthogonal projections are guaranteed to be single-valued operators) defined by

$$P_C(\mathbf{y}) := \operatorname{argmin}\{\|\mathbf{x} - \mathbf{y}\|^2 : \mathbf{x} \in C\}.$$

Consider the basic gradient projection method for solving problem (NC):

$$\min\{f(\mathbf{x}) : \ \mathbf{x} \in \mathcal{C}_s\}, \text{ where } f(\mathbf{x}) := \frac{1}{2}\|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|^2.$$

The gradient of f is simply given by $\nabla f(\mathbf{x}) = \mathcal{A}^*(\mathcal{A}(\mathbf{x}) - \mathbf{b})$, where $\mathcal{A}^*$ stands for the adjoint map to $\mathcal{A}$. The gradient projection method generates a sequence $\mathbf{x}_k$ via:

$$\text{(GP)} \quad \mathbf{x}_{k+1} \in P_{\mathcal{C}_s}\left(\mathbf{x}_k - \frac{1}{T_k}\nabla f(\mathbf{x}_k)\right), k = 0, 1, 2, \ldots$$

where T_k is an appropriately chosen (inverse) stepsize and $\mathbf{x}_0 \in \mathbb{E}$ is arbitrary.

Note that applying (GP) requires to compute an orthogonal projection onto the set $\mathcal{C}_s$ defined in (2.4). This set is nonempty and closed by the lower semi-continuity of φ. Finding an orthogonal projection onto a nonconvex set is by itself a nonconvex optimization problem, and as such is not necessarily an easy one. However, as seen below, it can be efficiently computed for the sets involved in sparse recovery problems. Note that in both cases below the resulting projections are in general not single valued, and when applying (GP) we can select any element of the resulting multivalued projection in an arbitrary fashion.

- **Case A.** Let $\mathbf{x} \in \mathbb{R}^n$ and $\varphi(\mathbf{x}) = \|\mathbf{x}\|_0$. In this case, the orthogonal projection $P_{C_s}(\mathbf{x})$ of $\mathbf{x} \in \mathbb{R}^n$ onto the set $\mathcal{C}_s$ is simply a vector consisting of the s components of $\mathbf{x}$ with the largest absolute values and zeros otherwise.

- **Case B.** Let $\mathbf{x} \in \mathbb{R}^{m\times n}$ and $\varphi(\mathbf{x}) = \text{rank}(\mathbf{x})$. The set of orthogonal projections $P_{\mathcal{C}_s}(\mathbf{x})$ is computed via a truncated singular value decomposition ([19]) as follows: if $\mathbf{x} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is a singular value decomposition of $\mathbf{x}$, then $P_{\mathcal{C}_s}(\mathbf{x})$ consists of matrices of the form $\mathbf{x} = \mathbf{U}\mathbf{\Sigma}_s\mathbf{V}^T$ where the diagonal $\mathbf{\Sigma}_s$ includes the s singular values with largest absolute value (otherwise zero).

4.2 Linear Rate of Convergence Analysis for GP

We assume that SRIP$(2s, \sqrt{2})$ holds. We will consider two versions of algorithm (GP). The first one is with a constant stepsize where we assume that

$$T_k = \bar{T} \in [\mu_{2s}^2, 2\nu_{2s}^2),$$

where μ_{2s}, ν_{2s} are as in the definition of SRIP. An evident drawback of the fixed steppsize setting is the requirement that at least μ_{2s} should be known. In order to avoid the need for knowing this parameter, we also introduce a variant of the method with a backtracking stepsize rule that *does not* require the knowledge of μ_{2s} for computational implementation, see Remark 4.1 below. This backtracking procedure requires that SRIP should hold with a parameter α which is smaller than $\sqrt{2}$ (but on the other hand, can be arbitrary close to $\sqrt{2}$).

Gradient Projection with backtracking:
Input: $\varphi \in \mathcal{S},\ s > 0,\ \mathbf{x}_0 \in \mathcal{C}_s$ arbitrary,
$\eta > 1$ - backtracking parameter,
$T_0 \in (0, \mu_{2s})$ initial stepsize.
Step $k (k \geq 0)$**:**
(a) Compute $\mathbf{x}_{k+1} \in P_{\mathcal{C}_s}\left(\mathbf{x}_k - \frac{1}{T_k}\nabla f(\mathbf{x}_k)\right)$.
(b) If $\|\mathcal{A}(\mathbf{x}_{k+1} - \mathbf{x}_k)\| > \sqrt{T_k}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$, set $T_k \leftarrow \eta T_k$ and go back to (a).
(c) Set $T_{k+1} \leftarrow T_k$.
(d) Set $k \leftarrow k + 1$.

Remark 4.1. It is very easy to find a $T_0 \in (0, \mu_{2s}^2)$ without actually knowing μ_{2s}. For example, by taking an arbitrary $\mathbf{v} \in \mathcal{C}_{2s}$, we get that $\frac{\|A(\mathbf{v})\|}{\|\mathbf{v}\|} \leq \mu_{2s}$, so we can pick $T_0 \in \left(0, \frac{\|A(\mathbf{v})\|^2}{\|\mathbf{v}\|^2}\right)$.

From the definition of the backtracking procedure, we first establish the following useful fact on the inverse step size T_k.

Proposition 4.1. *For all $k \geq 0$,*

$$T_k \leq \eta\mu_{2s}^2. \tag{4.1}$$

Proof. This is proved by induction on k. For $k = 0$ the claim is valid by the choice of T_0. Suppose that the claim is true for k and we will prove it for $k + 1$. If no backtracking steps were done in step (b), then $T_{k+1} = T_k$ and the claim is correct by the induction assumption. Otherwise, if backtracking steps were performed during step (b), then, in particular, $\gamma = \frac{T_{k+1}}{\eta}$ satisfies $\|\mathcal{A}(\mathbf{x}_{k+2} - \mathbf{x}_{k+1})\| > \sqrt{\gamma}\|\mathbf{x}_{k+2} - \mathbf{x}_{k+1}\|$, which together with

the fact that $\mathbf{x}_{k+2} - \mathbf{x}_{k+1} \in \mathcal{C}_{2s}$ and the SRIP assumption imply that $\sqrt{\gamma} \leq \mu_{2s}$ and hence $T_{k+1} \leq \eta\mu_{2s}^2$. □

We are now ready to prove our main result which shows that if SRIP$(2s, \sqrt{2}/\eta)$ is satisfied, then the function values of the sequence generated by the above algorithm converges linearly to zero. For example, if $\eta = 1.1$, then SRIP$(2s, 1.285...)$ is required to hold true instead of SRIP$(2s, 1.414...)$.

Theorem 4.1. *Consider the GP method with either a constant stepsize $T_k = \bar{T} \in [\mu_{2s}^2, 2\nu_{2s}^2)$ or with a backtracking stepsize rule with parameter η and suppose that SRIP$(2s, \sqrt{2}/\xi)$ is satisfied where $\xi = 1$ for the constant stepsize setting and $\xi = \eta > 1$ for the backtracking scenario. Then*

$$f(\mathbf{x}_{k+1}) \leq (\rho - 1) f(\mathbf{x}_k), \ \forall k \geq 0$$

with $\rho < 2$ given by

$$\rho = \begin{cases} \frac{\bar{T}}{\nu_{2s}^2} & \text{constant stepsize} \\ \\ \frac{\eta\mu_{2s}^2}{\nu_{2s}^2} & \text{backtracking.} \end{cases}$$

As a consequence,

$$f(\mathbf{x}_{k+1}) \leq (\rho - 1)^k f(\mathbf{x}_0), \text{ for every } k \geq 0$$

and $f(\mathbf{x}_k) \to 0$ as $k \to \infty$.

Proof. Let

$$q_k(\mathbf{x}, \mathbf{x}_k) := f(\mathbf{x}_k) + \langle \mathbf{x} - \mathbf{x}_k, \nabla f(\mathbf{x}_k) \rangle + \frac{T_k}{2} \|\mathbf{x} - \mathbf{x}_k\|^2. \tag{4.2}$$

Then the GP method can be equivalently rewritten as

$$\mathbf{x}_{k+1} \in \operatorname{argmin}\{q_k(\mathbf{x}, \mathbf{x}_k) : \mathbf{x} \in \mathcal{C}_s\},$$

and hence, in particular, for a solution $\bar{\mathbf{x}}$ of (NCF) it holds that

$$q_k(\mathbf{x}_{k+1}, \mathbf{x}_k) \leq q_k(\bar{\mathbf{x}}, \mathbf{x}_k). \tag{4.3}$$

Now, since $f(\mathbf{x}) = \frac{1}{2}\|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|^2$, it follows that

$$\begin{aligned} f(\mathbf{x}_{k+1}) &= f(\mathbf{x}_k) + \langle \mathbf{x}_{k+1} - \mathbf{x}_k, \nabla f(\mathbf{x}_k) \rangle + \frac{1}{2}\|\mathcal{A}(\mathbf{x}_{k+1} - \mathbf{x}_k)\|^2 \\ &\leq f(\mathbf{x}_k) + \langle \mathbf{x}_{k+1} - \mathbf{x}_k, \nabla f(\mathbf{x}_k) \rangle + \frac{T_k}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2, \end{aligned}$$

where the last inequality follows from the fact that $\mathbf{x}_k - \mathbf{x}_{k+1} \in \mathcal{C}_{2s}$ (by the subadditivity and symmetry of the function $\varphi \in \mathcal{S}$) and from the fact that the definition of the stepsize (in the constant or backtracking settings) implies that $\|\mathcal{A}(\mathbf{x}_{k+1} - \mathbf{x}_k)\| \leq \sqrt{T_k}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$. Therefore, we have shown that $f(\mathbf{x}_{k+1}) \leq q_k(\mathbf{x}_{k+1}, \mathbf{x}_k)$ so that

$$f(\mathbf{x}_{k+1}) = q_k(\mathbf{x}_{k+1}, \mathbf{x}_k) \overset{(4.3)}{\leq} q_k(\bar{\mathbf{x}}, \mathbf{x}_k). \tag{4.4}$$

On the other hand,

$$\begin{aligned} q_k(\bar{\mathbf{x}}, \mathbf{x}_k) &= f(\mathbf{x}_k) + \langle \bar{\mathbf{x}} - \mathbf{x}_k, \nabla f(\mathbf{x}_k) \rangle + \frac{T_k}{2} \|\bar{\mathbf{x}} - \mathbf{x}_k\|^2 \\ &\leq f(\mathbf{x}_k) + \langle \bar{\mathbf{x}} - \mathbf{x}_k, \nabla f(\mathbf{x}_k) \rangle + \frac{T_k}{2\nu_{2s}^2} \|\mathcal{A}(\bar{\mathbf{x}} - \mathbf{x}_k)\|^2 \\ &\overset{\mathcal{A}(\bar{\mathbf{x}})=\mathbf{b}}{=} f(\mathbf{x}_k) + \langle \bar{\mathbf{x}} - \mathbf{x}_k, \nabla f(\mathbf{x}_k) \rangle + \frac{T_k}{2\nu_{2s}^2} \|\mathbf{b} - \mathcal{A}(\mathbf{x}_k)\|^2 \\ &= \left(1 + \frac{T_k}{\nu_{2s}^2}\right) f(\mathbf{x}_k) + \langle \bar{\mathbf{x}} - \mathbf{x}_k, \nabla f(\mathbf{x}_k) \rangle \\ &= \left(1 + \frac{T_k}{\nu_{2s}^2}\right) f(\mathbf{x}_k) - 2f(\mathbf{x}_k) \\ &= \left(\frac{T_k}{\nu_{2s}^2} - 1\right) f(\mathbf{x}_k), \end{aligned}$$

which along with (4.1) (in the backtracking setting), implies the result. □

Corollary 4.1. *Suppose that SRIP*$(2s, \sqrt{2})$ *holds true. Then the sequence* $\{\mathbf{x}_k\}$ *generated by GP converges to the unique optimal solution of (NC), and hence of (NCF).*

Proof. Let $\{\mathbf{x}_k\}$ be the sequence generated by the GP method with a constant stepsize $\bar{T} = \mu_{2s}^2$ and let $\bar{\mathbf{x}}$ be a solution of (NCF). Then by Theorem 4.1 we have that

$$f(\mathbf{x}_k) \leq (\rho - 1)^{k-1} f(\mathbf{x}_0)$$

On the other hand,

$$f(\mathbf{x}_k) = \frac{1}{2}\|\mathcal{A}(\mathbf{x}_k) - \mathbf{b}\|^2 = \frac{1}{2}\|\mathcal{A}(\mathbf{x}_k) - \mathcal{A}(\bar{\mathbf{x}})\|^2 \geq \frac{1}{2\nu_{2s}^2}\|\mathbf{x}_k - \bar{\mathbf{x}}\|^2.$$

Therefore, $\mathbf{x}_k \to \bar{\mathbf{x}}$ and since $\bar{\mathbf{x}}$ was chosen arbitrarily its uniqueness follows. □

The next corollary, follows immediately from Theorem 4.1, and bounds the number of iterations required to obtain an ε-optimal solution of (NC).

Corollary 4.2. *Consider the setting of Theorem 4.1. Then for* $k \geq 1 + \frac{\log(1/\varepsilon)+C}{D}$*, the (GP) algorithm produces an* $\mathbf{x}$ *such that*

$$\|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|^2 \leq \varepsilon,$$

where $C := \log(2f(\mathbf{x}_0))$, $D := \log\left(\frac{1}{\rho-1}\right)$.

Remark 4.2. Recently, an algorithm called *"the iterative M-sparse algorithm"* was analyzed in [5] for solving the l_0 problem ($\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^m, \varphi(\mathbf{x}) = \|\mathbf{x}\|_0$). This method is in fact nothing else but the gradient projection algorithm with a *constant step size fixed and equal to 1.* It was proved in [5] that if the columns of the matrix are normalized, and $\|\mathbf{A}\|_2 < 1$, this algorithm converges to a *local* minimum of (NC).

Our approach has focused on using SRIP to solve directly the nonconvex feasibility problem (NCF) via a simple gradient projection method. On the other hand, RIP was used to determine conditions that warrant recovery of solutions for the nonconvex optimization problems such as (CS) and (AR) by solving their convex relaxations (ConvCS) and (ConvAR) respectively, namely, it is also needed to apply convex minimization schemes to solve these relaxed problems and achieve the same goals. While a direct comparison of these results is not fully transparent (e.g., in terms of complexity, the parameters involved etc..), it is nevertheless worthwhile to make the following remarks.

Remark 4.3. In the (CS) case, if RIP$(2s, \delta_{2s})$ is satisfied with $\delta_{2s} < \sqrt{2}-1$, then this implies (by Lemma 3.2) that SRIP $(2s, \alpha)$ holds true with $\alpha = \sqrt{\frac{1+\sqrt{2}-1}{1-(\sqrt{2}-1)}} = \sqrt{\frac{1}{\sqrt{2}-1}} = 1.5538...$ which is less restrictive than the assumption $\alpha = \sqrt{2}$ used in Theorem 4.1. On the other hand, note that the later condition does not imply the condition on RIP of Theorem 3.1(ii).

Remark 4.4. In the (AR) case, we can be more precise. The condition in Theorem 3.2(ii) requires that RIP should hold with $\delta_{5s} < 0.1$. This condition is worse than the assumption of Theorem 4.1. Indeed, by Lemma 3.2 it implies that SRIP$(5s, \alpha)$ is satisfied with $\alpha = \sqrt{\frac{1.1}{0.9}} = 1.105...$, which is a more restrictive than the condition SRIP$(2s, \sqrt{2})$, see Lemma 3.1.

We end by showing another interesting consequence of Theorem 4.1 which is particularly relevant to sparse recovery problems. Let us focus again on the setting of problem (CS) in Example 2.1, that is, $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^m$ and $\varphi(\mathbf{x}) = \|\mathbf{x}\|_0$. The support of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined to be the set of indices of the nonzero components:

$$\text{supp}(\mathbf{x}) = \{i \in \{1, 2, \ldots, n\} : x_i \neq 0\}.$$

Our final result shows stabilization of the support in the sense that the support of $\mathbf{x}_k$ is contained in the support of the unique solution of (NCF) from a certain iteration of (GP).

Corollary 4.3. *Consider the setting of Theorem 4.1 and let $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^m$ and $\varphi(\mathbf{x}) = \|\mathbf{x}\|_0$. Let $\bar{\mathbf{x}}$ be the unique solution of (NCF). Then there exists $\bar{k}$ such that for every $k \geq \bar{k}$ the inclusion*

$$\text{supp}(\mathbf{x}_k) \subseteq \text{supp}(\bar{\mathbf{x}})$$

holds true.

Proof. For every set $S \subseteq \{1, 2, \ldots, n\}$ let us define:

$$f_S^* = \min \left\{ \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 : x_i = 0, i \notin S \right\}. \tag{4.5}$$

If $|S| \leq s$ and $\text{supp}(\bar{\mathbf{x}}) \nsubseteq S$ then $f_S^* > 0$ since otherwise, if $f_S^* = 0$, it would mean by the uniqueness of $\bar{\mathbf{x}}$ that the optimal solution of (4.5) is $\bar{\mathbf{x}}$ in contradiction to $\text{supp}(\bar{\mathbf{x}}) \nsubseteq S$. Let us now define the number

$$g = \min_S \{ f_S^* : |S| \leq s, S \subseteq \{1, 2, \ldots, n\}, \text{supp}(\bar{\mathbf{x}}) \nsubseteq S \}, \tag{4.6}$$

which is positive. Now, since $f(\mathbf{x}_k) \to 0$, it follows that there exists $\bar{k}$ such that

$$f(\mathbf{x}_k) < g \tag{4.7}$$

for all $k \geq \bar{k}$. Let $k \geq \bar{k}$ and let us assume in contradiction that $\text{supp}(\mathbf{x}_k) \not\subseteq \text{supp}(\bar{\mathbf{x}})$. Then

$$f(\mathbf{x}_k) \geq f^*_{\text{supp}(\mathbf{x}_k)} \geq g,$$

where the last inequality follows from the definition of g, which is a contradiction to (4.7). □

Appendix

We briefly summarize some of the first results providing sufficient conditions warranting recovery of sparse vectors for the compressed sensing l_0-minimization problem via the convex l_1-norm problem (ConvCS). These were obtained via the concept of *mutual coherence of a matrix*, see [6] for more details and references.

Definition 4.1. [11] Let $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{m \times n}$ with $m \leq n$ and with normalized columns $\|\mathbf{a}_i\| = 1$ for all $i = 1, \ldots, n$. Then the mutual coherence $M(\mathbf{A})$ of the matrix $\mathbf{A}$ is defined by

$$M(\mathbf{A}) := \max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle| = \max_{i \neq j} |(\mathbf{A}^T \mathbf{A})_{ij}|.$$

Clearly, $0 \leq M(\mathbf{A}) \leq 1$. Furthermore, it has been shown that

$$M(\mathbf{A}) \geq \sqrt{\frac{n-m}{m(n-1)}}.$$

Note that the mutual coherence of a matrix is generally easy to compute even for large matrices.

Using the mutual coherence of a matrix given in Definition 4.1, the following sufficient condition relating (CS) to its convex relaxation (ConvCS) was proven in [15].

Theorem 4.2. *[15] Consider problem (CS) with $\mathcal{A}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x}$. If a solution $\mathbf{x} \in \mathbb{R}^n$ of problem (CS) satisfies*

$$\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{M(\mathbf{A})}\right),$$

then it is unique and coincides with the optimal solution of the convex problem (ConvCS).

We note that for a special class of matrices which are the concatenation of two orthogonal square matrices U, V, i.e., with $\mathbf{A} := [\mathbf{U}, \mathbf{V}]$, the above result has been improved in [16] by requiring the weaker condition:

$$\|\mathbf{x}\|_0 < \frac{(\sqrt{2} - \frac{1}{2})}{M(\mathbf{A})}.$$

Further results in the same spirit have been derived for the noisy compressed sensing, see e.g., [8].

Finally we note that the mutual coherence of a matrix $M(\mathbf{A})$ given in Definition 4.1 is closely related to RIP as shown in the following result.

Lemma 4.1. . *Let* $\mathbf{A} = [\mathbf{a}_1, \ldots, a_n] \in \mathbb{R}^{m \times n}$ *with* $m \leq n$ *and with normalized columns* $\|\mathbf{a}_i\| = 1$ *for all* $i = 1, \ldots, n$. *Then, with* $\delta_s \leq (s-1)M(\mathbf{A})$, *the matrix* $\mathbf{A}$ *with mutual coherence* $M(\mathbf{A})$ *satisfies RIP*(s, δ_s).

Proof. . This follows immediately from the definition of RIP and using the Gershgorin circles theorem (see e.g., [19]). □

Acknowledgement. We thank two anonymous referees for their useful comments and suggestions. This research was partially supported by the Israel Science Foundation under ISF Grant 489-06.

References

[1] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, **38**, 367–426, (1996).

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, **2**, 183–202, (2009).

[3] A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery problems, pp. 42–88. In *Convex Optimization in Signal Processing and Communications*. Editors: Daniel Palomar and Yonina Eldar. Cambridge University Press, (2010).

[4] D. Bertsekas. *Non-Linear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[5] T. Blumensath and M. E. Davies. Iterative hard thresholding for Sparse Approximations, *The Journal of Fourier Analysis and Applications* **14**, no. 4, 629–654, (2008).

[6] Bruckstein, A. M.; Donoho, D. L., and Elad, M. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, **51**, 34–81, (2009).

[7] J. A. Cadzow. Signal enhancement – a composite property mapping algorithm. *IEEE Trans. Acoustics, Speech, Signal Process.*, **36**, 49–62, 1988.

[8] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489.509, 2006.

[9] E. J. Candes, The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, 346 589-592, 2008.

[10] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization *IEEE Signal Process. Letters*,14, 707–710, 2007.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review* ,**43**, 129–159, 2001.

[12] P. L. Combettes. The foundations of set theoeretic estimation. *Proc. IEEE*, **81**, 182–208, (1993).

[13] P. L. Combettes, and H. J. Trussell. Method of successive projections for finding a common point of sets in metric spaces. *J. Optim. Theory Appl.*, **67**, 487–507, 1990.

[14] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, **52**, 1289–1306, 2006.

[15] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, **47**, 2845–2862, 2001.

[16] M. Elad, and A.M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Theory* **48**, 2558–2567, 2002.

[17] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht, 1996.

[18] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *American Control Conference*, 3272–3278, 2004.

[19] G. Golub, and C. V. Loan, *Matrix computations*, 3rd ed. Johns Hopkins University, Press, 1996.

[20] B. K. Natarajan. Sparse approximation solutions to linear systems. *SIAM J. Computing*, **24**, 227–234, 1995.

[21] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. Preprint (2007). To appear in SIAM Review.

[22] F. Santosa, and W.W. Symes. Linear inversion of band-limited reflection seismograms, *SIAM J. Sci. Statist. Comput.*, **7**, 1307–1330, (1986).

[23] H. L. Taylor, S. C. Banks, and J. F. McCoy. Deconvolution with the l_1 norm. *Geophysics*, **44**, 39–52, (1979).