Iteratively Re-weighted Schemes for Non-smooth Optimization

Daiwei He

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Aleksandr Aravkin, Chair

James V. Burke, Chair

Dmitriy Drusvyatskiy

Program Authorized to Offer Degree:
Mathematics

University of Washington

Abstract

Iteratively Re-weighted Schemes for Non-smooth Optimization

Daiwei He

Co-Chairs of the Supervisory Committee:
Professor Aleksandr Aravkin
Department of Applied Mathematics

Professor James V. Burke
Department of Mathematics

Iteratively Re-weighted Least Squares (IRLS) has long been used to solve both convex optimization problems, including ℓ_1 regression and compressed sensing, as well as non-convex optimization problems, including ℓ_p regression for $(0 < p < 1)$.

The thesis is organized as follows. Following the introduction in Chapter 1, in Chapter 2 we give a robust phase retrieval counterpart to the seminal paper by Candes and Tao on compressed sensing (ℓ_1 regression) [*Decoding by linear programming.* IEEE Transactions on Information Theory, 51(12):42034215, 2005]. Chapter 3 answers a question raised in [*Iteratively reweighted least squares minimization for sparse recovery*, Communications on Pure and Applied Mathematics, **63**(2010) 1–38]. In particular, we find examples where IRLS algorithm in the paper provably fails and provide a remedy. In Chapter 4 we show that under the assumption that the entries of A are i.i.d. standard normal, locally linear convergence rate is achieved for IRLS algorithm, when applied to robust phase retrieval problem $\min_x \||Ax| - b\|_1$. Furthermore, we provide several other IRLS variants which can be applied to phase retrieval problems. Both the noiseless case and the case with sparse noise are considered. In Chapter 5 we talk about the application of IRLS in general cases.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisors, Jim Burke and Sasha Aravkin. I am deeply grateful to Jim, who has spent enormous amount of time with me on my research. His knowledge in optimization and creativity is always very inspirational. The start of research topic is proposed by Jim. I also learned a lot from his elegant and organized way in presentation and mathematical writing. I would like to thank Sasha for taking me as a student and helping me with the applied aspects related to my thesis. None of my thesis would exists without their help.

I would like for Dmitriy Drusvyatskiy and Mehran Mesbahi for serving in my committee. I would like to thank Dima for advice on my research.

I would like to thank my officemates for bringing happiness and laughter to our office. Many thanks to all of my friends in Seattle, either working in industry, doing research in academia or studying at University of Washington. You make my harsh life towards a Ph.D. more fun.

At last I want to thank my parents for their love and support.

DEDICATION

to my parents

Chapter 1

INTRODUCTION

In the thesis we mainly consider the non-smooth problem

$$\min_x J(x) := \|A_0x - b\|^2 + \sum_{i=1}^{\ell} \text{dist}(A_ix|C_i) + \sum_{i=\ell+1}^{\ell+h} \text{dist}(A_ix|C_i)\text{dist}(A_{i+h}x|C_{i+h}), \quad (1.0.1)$$

where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^{n_0}$, A_i is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^{n_i}$ and C_i's are closed sets. The norm $\|\cdot\|$ represents Euclidean norm and the distance function is induced by Euclidean norm.

When $A_0 = 0_{n_0\times n}$, $b = 0_{n_0}$ and $h = 0$, for each $1 \le i \le \ell$, $A_i \in \mathbb{R}^{1\times n}$ and $C_i = \{b_i\}$ for $b_i \in \mathbb{R}$, problem (1.0.1) is the ℓ_1-regression problem

$$\min_x \|Ax - b\|_1, \quad (1.0.2)$$

where $A = (A_1^T, A_2^T, ..., A_\ell^T)^t$ and $b := (b_1, b_2, ..., b_\ell)^T$. Here for $v = (v_1, v_2, ..., v_n)^T \in \mathbb{R}^n$, $|v| := (|v_1|, |v_2|, ..., |v_n|)^T$. In a seminal paper of Candes and Tao [22], ℓ_1-regression is used to solve a classical error correction problem. In particular, one hopes to recover a vector from $b = Ax_* + e \in \mathbb{R}^\ell$, where the measurement matrix $A \in \mathbb{R}^{\ell\times n}$ and the error vector $e \in \mathbb{R}^\ell$ is a sparse. The author of [22] proved that under certain condition of measurement matrix A, if the size of the support of the error vector e is small, $\|e\|_0 := |\{i|e_i \neq 0\}| \le sm$ for some constant s, then the vector x_* is the unique solution to problem (1.0.2). We say matrix $\Phi \in \mathbb{R}^{(\ell-n)\times\ell}$, with full row rank, is an annihilator matrix of A if $\Phi A = 0$. The condition used in the proof of [22] is called Restricted Isometry Property (RIP) (See Section 2.2 of Chapter 2 for details). It is a property of the annihilator matrix Φ. In particular, they proved that if matrix A with entries being independent and identically distributed random variables, then it's annihilator matrix Φ satisfies RIP with high probability.

Two other interesting problems of form (1.0.1) are real robust phase retrieval problem.

$$\min_x \left\|(Ax)^2 - b\right\|_1 = \sum_{i=1}^{\ell} \operatorname{dist}(b_i A_i x | \{-b_i^2, b_i^2\}) + \sum_{i=1}^{\ell} \operatorname{dist}(A_i x | \{0\}) \operatorname{dist}(A_i x | \{-b_i, b_i\}) \quad (1.0.3)$$

and

$$\min_x \left\||Ax| - b\right\|_1 = \sum_{i=1}^{\ell} \operatorname{dist}(A_i x | \{-b_i, b_i\}) \quad (1.0.4)$$

Likewise, the square and magnitude of a vector means component-wise square and magnitude, respectively. These two problems are non-convex. Problem (1.0.3) is weakly convex [33, 41] and is widely studied in several literatures [33, 41, 24]. Problem is relatively harder since it's neither weakly convex nor subdifferentially regular [57]. Since the forms of (1.0.2), (1.0.3) and (1.0.4) looks similarly, we raise a naturally question, despite the intrinsic non-convexity of (1.0.3) and (1.0.4).

Question 1.0.1. *Are there any robust phase retrieval counterparts to the result of Candes and Tao [22]?*

Chapter 2 gives a positive answer to Question 1.0.1. In particular, we prove that under the assumption that the entries of measurement matrix A are of i.i.d. $N(0,1)$ random variables, for $p \in \{1,2\}$, there exists universal constants $c_0, c_1, c_2 > 0$ and $s \in (0,1)$, if $\ell \geq c_0 n$ and $x_* \in \mathbb{R}^n$ is such $|\{i | |A_i x_*|^p \neq b_i\}| \leq sm$, then x_* is a global minimizer of the robust phase retrieval problems; If it is further assumed that $\ell \geq 2n - 1$, then, x_* is the unique global minimizer of the robust phase retrieval problems, up to multiplication by -1. All the events above hold with probability at least $1 - c_1 \exp(-c_2 \ell)$.

Next we come to the algorithmic aspects of problem (1.0.1). It is known that ℓ_1-regression problem is equivalent to compressed sensing (sparse recovery) problem

$$\min_y \|y\|_1 \text{ such that } \Phi y = c, \quad (1.0.5)$$

where $\Phi \in \mathbb{R}^{(\ell-n)\times \ell}$ and $c \in \mathbb{R}^{\ell - n}$, $\Phi b = -c$ and $\operatorname{range}(A) = \operatorname{Null}(\Phi)$. In [32], Iteratively Re-weighted Least Squares (IRLS) algorithm is used to solve sparse recovery problem (1.0.5).

It is called DDFG-IRLS in the thesis. Intuitively, IRLS algorithm solves a constrained re-weighted least square problem and update weights afterwards in each sub-step. In [32], the authors show that if the matrix Φ satisfies the the *null space property of order* K *for* $0 < \gamma < 1$ (see Section 3.3 of Chapter 3 for details), then the DDFG-IRLS algorithm converges to the unique k-sparse solution when $k < K - 2\gamma(1-\gamma)^{-1}$, and this k-sparse solution coincides with the unique ℓ_1 solution, where a vector is k-sparse if it has k nonzero components. In addition, the authors also establish the local linear convergence of the DDFG-IRLS algorithm when $0 < \gamma < 1 - 2/(K+2)$. On the other hand, it is known that for $k \leq K$ the k-sparse and ℓ_1 solutions are unique and coincide [49, 39, 32]. In [32, Remark 5.4], the authors note that their proof method does not apply for $K - 2\gamma(1-\gamma)^{-1} \leq k \leq K$, and state that they were unsuccessful in finding an example where the algorithm fails when k falls in this range. A natural question is

Question 1.0.2. *Is the failure of the analysis in [32] when* $K - 2\gamma(1-\gamma)^{-1} \leq k \leq K$ *due to the theoretical analysis technique or the algorithm itself? More precisely, does there exist an example where DDFG-IRLS fails when* $K - 2\gamma(1-\gamma)^{-1} \leq k \leq K$*? If yes, can one improve the algorithm for all* $k \leq K$ *theoretically?*

We answer Question 1.0.2 in Chapter 3. We construct a family of examples where the DDFG-IRLS algorithm fails when $k = K$, and provide a modification to their algorithm that provably converges to the unique k-sparse solution for $k \leq K$. In addition, we show that this modification is locally linearly convergent for all $k \leq K$ and $\gamma \in (0, 1)$ which increases the range of γ values for which linear convergence is assured. For our modification of DDFG-IRLS, we change the ϵ updating strategy.

In Chapter 4 we consider the local linear convergence of IRLS for Problem (1.0.4), when the measurement matrix A is i.i.d. standard normal. We show both in the noiseless case and the in the case with sparse noise, with number of non-zero entries no larger than a constant fraction of total number of measurements, IRLS converges locally linearly, with high probability. Also we provide a wide variety of IRLS algorithm applicable to phase

retrieval problems.

In Chapter 5 we discuss the applications of IRLS in the general setting (1.0.1). Historically the method of iteratively re-weighted least squares method (IRLS) is introduced to solve l_p $(0 < p < \infty)$ regression. It becomes popular for solving compress sensing ([25, 32]) and matrix recover problem ([59, 44, 55, 51]) recently. In Chapter 5, we first show there exist step size in each sub-step of IRLS. For each sub-step of IRLS, sub-step with a step size of α for $0 < \alpha < 2$ will also lead to convergence of the algorithm ($\alpha = 1$ corresponds to the classical IRLS). A good choice of step size α can result in less number of iterations, less running time in practice, and higher accuracy, though the theoretic time complexity remains the same as the classical IRLS. Then we show that IRLS can be applied to non-convex settings. Problems such as real robust phase retrieval and Nesterov's Chebyshev-Rosenbrock functions can be solved by IRLS. We give the iterates complexity in this case. we also discuss the applications of IRLS in the matrix case.

Chapter 2

ON THE GLOBAL MINIMIZERS OF REAL ROBUST PHASE RETRIEVAL WITH SPARSE NOISE

Abstract

We study a class of real robust phase retrieval problems under a Gaussian assumption on the coding matrix when the received signal is sparsely corrupted by noise. The goal is to establish conditions on the sparsity under which the input vector can be exactly recovered. The recovery problem is formulated as the minimization of the ℓ_1 norm of the residual. The main contribution is a robust phase retrieval counterpart to the seminal paper by Candes and Tao on compressed sensing (ℓ_1 regression) [*Decoding by linear programming.* IEEE Transactions on Information Theory, 51(12):42034215, 2005]. Our analysis depends on a key new property on the coding matrix which we call the *Absolute Range Property* (ARP). This property is an analogue to the Null Space Property (NSP) in compressed sensing. When the residuals are computed using squared magnitudes, we show that ARP follows from a standard Restricted Isometry Property (RIP). However, when the residuals are computed using absolute magnitudes, a new and very different kind of RIP or growth property is required. We conclude by showing that the robust phase retrieval objectives are sharp with respect to their minimizers with high probability.

2.1 Introduction

Phase retrieval has been widely studied in machine learning, signal processing and optimization. The goal of phase retrieval is to recover a signal x provided the observations of the amplitude of its linear measurements:

$$|\langle a_i,\, x\rangle| = b_i, \quad 1 \le i \le m \tag{2.1.1}$$

where $a_i \in \mathbb{C}^n$ or $\mathbb{R}^n$, $b_i \in \mathbb{R}$ are observations, and x is an unknown variable we wish to recover (e.g. see [57]). A well studied form of the phase retrieval problem is

$$|\langle a_i, x\rangle|^2 = b_i, \quad 1 \leq i \leq m, \tag{2.1.2}$$

where b_i now represent the squared magnitudes of the observations. It is shown in [64] that the phase retrieval problem is NP-hard. Recent work on the phase retrieval problem [24, 41, 33, 20, 19] focuses on the real phase retrieval problem where it is assumed that $a_i \in \mathbb{R}^n$ for each $i = 1, 2, ..., m$. This is the line of inquiry we follow. In the following discussion the m rows of the matrix $A \in \mathbb{R}^{m\times n}$ are the vectors $a_i \in \mathbb{R}^n$.

The two most popular approaches to the real phase retrieval problem are through semidefinite programming relaxations [5, 19, 21, 28, 34, 53, 75] and convex-composite optimization [10, 33, 41]. These approaches formulate real phase retrieval problem as an optimization problem of the form

$$\min_x \rho(|Ax|^2 - b), \tag{2.1.3}$$

where ρ is chosen to be either the ℓ_1 or the square of ℓ_2 norm, and, for any vector $z \in \mathbb{R}^m$, $|z|$ and z^2 are vectors in $\mathbb{R}^m$ whose components are the absolute value and squares of those in z. The objective in (2.1.3) is a composition of a convex and a smooth function, and is called convex-composite. This structure plays a key role in both optimality conditions and algorithm development for (2.1.3) [10].

In the noiseless case, when there exists a vector $x_* \in \mathbb{R}^n$ such that $|Ax_*|^2 = b$ (or, $|Ax_*| = b$), a gradient based method called Wirtinger Flow (WF) was introduced by [20] to solve the smooth problem

$$\min_x \left\| |Ax|^2 - b \right\|_2^2.$$

WF admits a linear convergence rate when properly initialized. Further work along this line includes the Truncated Wirtinger Flow (TWF), e.g., see [28]. Truncated Wirtinger Flow requires $m \geq Cn$ measurements as opposed to the $m \geq Cn \log n$ measurements in WF to obtain a linear rate. A similar approach using sub-gradient is used to minimize $\min_x \||Ax| - b\|_2^2$ in [76] for the noiseless case.

Contributions. In this paper we address two forms of the *robust* phase retrieval problem, where the optimization objective takes the form

$$\min_x f_p(x) := \| |Ax|^p - b \|_1 \quad \text{for } p = 1, 2, \tag{2.1.4}$$

and it is assumed that the matrix A satisfies the following Gaussian assumption:

$$\text{G}: \qquad \text{The entries of } A \text{ are i.i.d. standard Gaussians } N(0,1).$$

Our goal is to establish a robust phase retrieval counterpart to the seminal paper by Candes and Tao on compressed sensing (ℓ_1 regression) [22].

Compressed sensing problems [36] take the form

$$\min_y \|y\|_1 \text{ such that } \Phi y = c, \tag{2.1.5}$$

where $\Phi \in \mathbb{R}^{n \times N}, y \in \mathbb{R}^N, c \in \mathbb{R}^n$. This problem is known to be equivalent to the ℓ_1 linear regression problem

$$\min_x \|Ax - b\|_1 , \tag{2.1.6}$$

where $\Phi b = -c$ and $A \in \mathbb{R}^{N \times (N-n)}$ (e.g., the columns of A form basis of Null(Φ)). In [22] it is shown that there is a universal constant $s \in (0,1)$ such that, under suitable conditions on A (e.g., Assumption G), if x^* satisfies $\|Ax_* - b\|_0 \leq sm$, then x^* is the unique solution to (2.1.6), with high probability. We prove similar exact recovery results for the two robust phase retrieval problems (2.1.4). In particular, we show that $\{x_*, -x_*\} = \operatorname{argmin} f_p$ with high probability, when $m \geq 2n - 1$ (Theorem 4.8.2). In this situation, the solution set to $\min f_p$ and the ℓ_0 phase retrieval problem coincide, that is,

$$\{x_*, -x_*\} = \operatorname*{argmin}_x \| |Ax|^p - b \|_0 . \tag{2.1.7}$$

Thus, the ℓ_0 phase retrieval problem can be solved by the ℓ_1 phase retrieval problem $\min f_p$, when there exists an x_* with sufficiently sparse noise.

A key underlying structural requirement used by [22] is the Restricted Isometry Property (RIP). We also make use of an RIP property in the $p = 2$ case. However, in the $p = 1$ case a

new property, which we call the p-Absolute Growth Property (p-AGP) (see Definition 2.2.3), is required. When $p = 2$, RIP implies 2-AGP. The p-AGP holds under Assumption G, with high probability (see Lemmas 2.4.1 and 2.4.7). A second key property, which mimics the so-called Null Space Property (NSP) in compressed sensing [31, 32, 37, 48], is also introduced. We call this the p-Absolute Range Property (p-ARP) (see Definition 2.2.1), and show that p-AGP implies p-ARP under Assumption G with high probablility. In [18], it is shown that, for problem (2.1.5), if Φ satisfies RIP with parameter $\delta_{2s} < \sqrt{2} - 1$, then Φ satisfies NSP of order s. Correspondingly, we show that the p-AGP implies the p-ARP with high probability under Assumption G. (see Lemmas 2.4.2 and 2.4.8).

There are separate classes of methods for solving (2.1.4) for $p = 2$ and $p = 1$. When $p = 1$, one can apply a smoothing method to the absolute value function [3, 57], or use other relaxation techniques that preserve the nonsmooth objective but introduce auxiliary variables [81]. When $p = 2$, the solution methods typically exploit the convex-composite structure of the objective f_2. These methods rely on two key conditions on the function f_2: weak convexity (i.e., $f + \frac{\rho}{2}\|\cdot\|^2$ is convex for some $\rho > 0$) and sharpness (i.e., $f(x) - \min f \geq c \cdot \operatorname{dist}(x, \mathcal{X})$ for some $c > 0$ where $\mathcal{X}$ is the set of minimizers of f). Under these two properties, Duchi and Ruan [41], Drusvyatskiy, Davis and Paquette [33] and Charisopoulos, et al.[24] establish convergence and iteration complexity results for prox-linear and subgradient algorithms. Recently [78] and [26] considered gradient-based methods for the problem $\min_x f_2(x)$ when the noise is sm sparse for some $s < 1$. To establish locally linear convergence of their algorithms the authors of [78] require that the measurements satisfy $m \geq cn \log n$ for $c > 0$, while the authors of [26] require that $s < c/\log m$ for some $c > 0$. The results in [41] and [24] require $m \geq cn$ for some $c > 0$ and for some $s \in [0, \frac{1}{2})$ sufficiently small.

Conditions for the weak convexity of f_2 follow from results in [41, 33] under assumptions weaker than Assumption G. In the noiseless case, the sharpness of f_2 also follows from results in [41, 33]. In the noisy case, sharpness is established in [41, 33] under same assumptions on the sparsity of the noise.

We establish sharpness for both f_1 and f_2 under Assumption G uniformly for all possible

supports of the sparse noise. Our result for $p = 2$ case has a similar flavor to those in [41, 24], but more closely parallels the result of Candes and Tao in the compressed sensing case. When $p = 1$, our result has no precedence in the literature and requires a new approach. The function f_1 is not weakly convex since it is not even subdifferentially regular [57].

This paper is organized as follows. In section 2, we introduce the new properties p-ARP and p-AGP and provide a detailed description of how our program of proof parallels the program used in compressed sensing. In Section 3, we show that if A satisfies p-ARP and the residual $||Ax_*|^p - b|$ is sufficiently sparse, then $\{\pm x_*\} \subset \operatorname{argmin} f_p$ with equality under Assumption G. In section 4, we show that Assumption G implies that p-AGP implies p-ARP with high probability. In the last section we show that f_p is sharp with respect to $\operatorname{argmin} f_p$, with high probability.

2.1.1 Notation

Lower case letters (i.e. x, y) denote vectors, while x_i denotes the ith component of the x. $c_0, c_1, c_2, \tilde{c_0}, \tilde{c_1}, C$ denote universal constants. $\|x\|$, $\|x\|_1$ denote the Euclidean and ℓ_1 norms of vector x, while $\|x\|_0$ denotes the ℓ_0 'norm' $|\{i|x_i \neq 0\}|$. For a matrix X, $\|X\|_F$ denotes the Frobenius and $\|X\|$ denotes the ℓ_2 operator norm. When $x = (x_i)_{1\leq i\leq n}$ is a vector, $|x| := (|x_i|)_{1\leq i\leq n}$ and $x^p := (x_i^p)_{1\leq i\leq m}$. For a vector $v \in \mathbb{R}^m$, and $T \in [m] := \{1, 2, ..., m\}$, v_T is defined to be a vector in $\mathbb{R}^m$ where the ith entry is v_i if $i \in T$ and 0 else where. $\operatorname{supp}(x) := \{i|x_i \neq 0\}$. We say a vector x is L sparse if $\|x\|_0 := |\operatorname{supp}(x)| \leq L$.

2.2 The Roadmap

Recall from the compressed sensing literature [31, 32] that a matrix $\Phi \in \mathbb{R}^{m\times n}$ satisfies Null Space Property (NSP) of order L at $\psi \in (0, 1)$ if

$$\|y_T\|_1 \leq \psi \|y_{T^c}\|_1 \quad \forall\, y \in \operatorname{Null}(\Phi) \text{ and } |T| \leq L. \tag{2.2.1}$$

It is shown in [37, 48] that every L-sparse signal $y_* \in \mathbb{R}^m$ is the unique minimizer of the compressed sensing problem (2.1.5) with $b = \Phi y_*$ if and only if $\Phi \in \mathbb{R}^{p\times m}$ satisfies NSP of

order L for some $\psi \in (0,1)$. NSP of order L is implied by the Restricted Isometry Property (RIP) for a sufficiently small RIP parameter δ_{2L} [18], where a matrix $\Phi \in \mathbb{R}^{p\times m}$ is said to satisfy RIP with constant δ_L if [22]

$$(1-\delta_L)\,\|y\|_2^2 \le \|\Phi y\|_2^2 \le (1+\delta_L)\,\|y\|_2^2 \quad \forall\, L\text{-sparse vectors } y \in \mathbb{R}^m. \tag{2.2.2}$$

It is known that RIP is satisfied under many distributional hypothesis on the matrix Φ, for example, random matrices Φ with entries i.i.d. Gaussian or Bernoulli random variables are known to satisfies RIP with high probability for $L \le Cm/\log m$ for constant C [6, 22, 23, 70]. Recapping, the general pattern of the proof for establishing that sufficiently sparse y_* is the unique minimizer of problem (2.1.5) using distributional assumptions on Φ is given in the following program:

$$\text{(CS)} \qquad \begin{matrix}\text{Distributional}\\ \text{Assumptions}\end{matrix} \xRightarrow{[22]} \text{RIP} \xRightarrow{[18]} \text{NSP} \xLeftrightarrow{[37,\ 48]} y_* \text{ minimizes (2.1.5)}.$$

We extend this program to the class of robust phase retrieval problems

$$\min_x f_p(x) := \||Ax|^p - b\|_1 \tag{2.2.3}$$

for $p \in \{1,2\}$, to show that, under Assumption G, and when the residuals $|Ax_*|^p - b$ are sufficiently sparse, the vectors $\pm x_*$ are the global minimizers of the real robust phase retrieval problems (2.2.3) with high probability. In our program, we substitute NSP and RIP with new properties called the p-Absolute Range Property (p-ARP) and the p-Absolute Growth Property (p-AGP), respectively.

Definition 2.2.1 (p-Absolute Range Property (p-ARP))**.** *For $p \in \{1,2\}$, we say $A \in \mathbb{R}^{m\times n}$ satisfies the p-Absolute Range Property of order L_p for $\psi_p \in (0,1)$ if, for any $x, y \in \mathbb{R}^n$ and for any $T \subseteq [m]$ with $|T| \le L_p$,*

$$\|(|Ax|^p - |Ay|^p)_T\|_1 \le \psi_p\,\|(|Ax|^p - |Ay|^p)_{T^c}\|_1 \;\; \forall\, x, y \in \mathbb{R}^n \text{ and } T \subseteq [m] \text{ with } |T| \le L_p. \tag{2.2.4}$$

In order for Definition 2.2.1 to make sense, m must be significantly larger than n. This is illustrated by the following example.

Example 2.2.2. *For $p \in \{1,2\}$, an example in which ARP does not hold for any order L is $A = I_n$ for any $\psi \in (0,1)$. An example in which ARP of order $L = 1/3$ holds is $A = (I_n, I_n, I_n)^T$ for any $\psi \in [\frac{1}{2}, 1)$.*

The connection between p-ARP and NSP is seen by observing the parallels between (2.2.4) the fact that Φ satisfies NSP of order L for $\psi \in (0,1)$ (2.2.1) if

$$\|(Ax - Ay)_T\|_1 \leq \psi \, \|(Ax - Ay)_{T^c}\| \quad \forall\, x, y \in \mathbb{R}^n \text{ and } T \subseteq [m] \text{ with } |T| \leq L,$$

where the columns of A form a basis of Null(Φ).

Definition 2.2.3 (p-Absolute Growth Property (p-AGP))**.** *For $p \in \{1,2\}$, we say that the matrix $A \in \mathbb{R}^{m \times n}$ satisfies the p-Absolute Growth Property if there exists constants $0 < \mu_1 < \mu_2 < 2\mu_1$ and a mapping $\phi_p : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ such that*

$$\mu_1 \phi_p(x,y) \leq \frac{1}{m} \left\| |Ax|^p - |Ay|^p \right\|_1 \leq \mu_2 \phi_p(x,y) \quad \forall\, x, y \in \mathbb{R}^n. \tag{2.2.5}$$

The mapping ϕ_p is introduced to accommodate the fact that the robust phase retrieval problem cannot have unique solutions since if x_* solves (2.2.3) then so does $-x_*$. For this reason, (2.2.5) implies that if $x = \pm y$, then $\phi_p(x,y) = 0$. In what follows, we take

$$\phi_2(x,y) := \left\| xx^T - yy^T \right\|_F \quad \text{and} \quad \phi_1(x,y) := \min\{\|x+y\|, \|x-y\|\} \quad \forall\, x, y \in \mathbb{R}^n. \tag{2.2.6}$$

The relationship between RIP and p-AGP is now seen by comparing (2.2.2) with (2.2.5). A fundamental (and essential) difference is that RIP for compressed sensing applies to any selection of L columns from Φ where L is considered to be small since it determines the sparsity of the solution. On the other hand, our p-AGP applies to the rows of A corresponding to the zero entries in the sparse residual vector $|Ax_*|^p - b$.

We can now more precisely describe how our program of proof parallels the one used for compressed sensing.

$$p=2: \quad \mathrm{G} \xrightarrow{\text{Lem } 2.4.1} \text{RIP} \Rightarrow \text{2-AGP} \xrightarrow{\text{Lem } 2.4.2} \text{2-ARP} \xrightarrow{\text{Thm } 4.8.2} \begin{matrix} x_* \text{minimizes} \\ f_2(x) \end{matrix}$$

$$p=1: \quad \mathrm{G} \xrightarrow{\text{Lem } 2.4.7} \text{1-AGP} \xrightarrow{\text{Lem } 2.4.8} \text{1-ARP} \xrightarrow{\text{Thm } 4.8.2} \begin{matrix} x_* \text{minimizes} \\ f_1(x) \end{matrix}$$

2.3 Global minimization under p-ARP

In this section we parallel the discussion given in [32] with NSP replaced by p-ARP. We begin by introducing a measure of residual sparsity. For a vector $y \in \mathbb{R}^n$, let $T \subseteq [m]$ be the set of indices corresponding to the L largest entries in the residual vector $||Ax|^p - b|$ and define

$$\sigma_L^p(x) := \|(|Ax|^p - b)_{T^c}\|_1 .$$

Note that $\sigma_L^p(x) = 0$ if and only if $\||Ax|^p - b\|_0 \leq L$.

Lemma 2.3.1. *Let $A \in \mathbb{R}^{m\times n}$, $p \in \{1,2\}$ and $L \in (0,m)$. If the matrix A satisfies p-ARP of order L for $\psi \in (0,1)$, then*

$$\||Ax|^p - |Ay|^p\|_1 \leq \frac{1+\psi}{1-\psi}\left(\||Ax|^p - b\|_1 - \||Ay|^p - b\|_1 + 2\sigma_L^p(y)\right), \tag{2.3.1}$$

for all $x, y \in \mathbb{R}^n$.

Proof. In either case 1 or 2 above, let T be the set of indices of the L largest entries in $||Ay|^p - b|$. Then

$$\begin{aligned}
\|(|Ax|^p - |Ay|^p)_{T^c}\|_1 &\leq \|(|Ax|^p - b)_{T^c}\|_1 + \|(|Ay|^p - b)_{T^c}\|_1 \\
&= \||Ax|^p - b\|_1 - \|(|Ax|^p - b)_T\|_1 + \sigma_L^p(y) \\
&= \|(|Ay|^p - b)_T\|_1 - \|(|Ax|^p - b)_T\|_1 \\
&\qquad + \||Ax|^p - b\|_1 - \||Ay|^p - b\|_1 + 2\sigma_L^p(y) \\
&\leq \|(|Ax|^p - |Ay|^p)_T\|_1 + \||Ax|^p - b\|_1 - \||Ay|^p - b\|_1 + 2\sigma_L^p(y).
\end{aligned} \tag{2.3.2}$$

By p-ARP,

$$\|(|Ax|^p - |Ay|^p)_T\|_1 \leq \psi \|(|Ax|^p - |Ay|^p)_{T^c}\|_1 . \tag{2.3.3}$$

Consequently, by (2.3.2) and (2.3.3),

$$\|(|Ax|^p - |Ay|^p)_{T^c}\| \leq \frac{1}{1-\psi}(\||Ax|^p - b\|_1 - \||Ay|^p - b\|_1 + 2\sigma_L^p(y)). \tag{2.3.4}$$

By (2.3.3), we know

$$\begin{aligned} \||Ax|^p - |Ay|^p\| &= \|(|Ax|^p - |Ay|^p)_T\|_1 + \|(|Ax|^p - |Ay|^p)_{T^c}\|_1 \\ &\leq (1+\psi) \|(|Ax|^p - |Ay|^p)_{T^c}\|_1 . \end{aligned}$$

By combining this with (2.3.4), we obtain (4.8.5) which holds true for all $x, y \in \mathbb{R}^n$. □

The main result of this section now follows.

Theorem 2.3.2. *Let $L \in (0, m)$, $p \in \{1, 2\}$, and suppose $x_* \in \mathbb{R}^n$ is such that $(|Ax_*|^p - b)$ is L sparse. Let the assumptions of Lemma 2.3.1 holds. Then x_* is a global minimizer of the robust phase retrieval problem (2.2.3). Moreover, for any x,*

$$\||Ax|^p - |Ax_*|^p\|_1 \leq \frac{2(1+\psi)}{1-\psi}\sigma_L^p(x).$$

If $\tilde{x}$ is another global minimizer, then $|A\tilde{x}| = |Ax_*|$. If it is further assumed that the entries of A are i.i.d. standard Gaussians and $m \geq 2n-1$, then, with probability 1, x_* is the unique solution of (2.2.3) up to multiplication by -1.

Proof. By lemma 2.3.1, since $\sigma_L^p(x_*) = 0$,

$$\||Ax|^p - |Ax_*|^p\|_1 \leq \frac{1+\psi}{1-\psi}(\||Ax|^p - b\|_1 - \||Ax_*|^p - b\|_1) \quad \forall\, x \in \mathbb{R}^n, \tag{2.3.5}$$

and so $\||Ax|^p - b\|_1 \geq \||Ax_*|^p - b\|_1$ for all x, i.e., x_* is a global minimizer. Again by Lemma 2.3.1,

$$\||Ax|^p - |Ax_*|^p\|_1 \leq \frac{1+\psi}{1-\psi}(\||Ax_*|^p - b\|_1 - \||Ax|^p - b\|_1 + 2\sigma_L^p(x)) \tag{2.3.6}$$

$$\leq \frac{2(1+\psi)}{1-\psi}\sigma_L^p(x) \tag{2.3.7}$$

Inequality (2.3.5) also implies that if there is another minimizer $\tilde{x}$, then $|Ax_*| = |A\tilde{x}|$. The final statement on the uniqueness of x_* is established in [5, Corollary 2.6]. □

In the next section we show that under Assumption G, p-ARP of order $L = sm$ holds for a sufficiently small constant s, with high probability.

2.4 Assumption G $\implies$ p-AGP $\implies$ p-ARP

In this section we use of the Gaussian Assumption G on the matrix A to show that p-AGP holds for A with high probability, and that p-AGP implies p-ARP of order $L := sm$ with high probability for a constant $s \in (0,1)$. The cases $p = 2$ and $p = 1$ are treated separately since different techniques are required.

2.4.1 $\mathbf{p = 2}$

We begin by re-stating [28, Lemma 1] in our notation, where the conclusion of [28, Lemma 1] is called RIP in [24].

Lemma 2.4.1 (Assumption G $\implies$ 2-AGP(RIP))**.** *[28, Lemma 1] Under Assumption G, there exists universal constants c_0, c_1, C such that for $\epsilon \in (0,1)$, if $m > c_0 n \epsilon^{-2} \log \frac{1}{\epsilon}$, then with probability at least $1 - C\exp(-c_1\epsilon^2 m)$,*

$$0.9(1-\epsilon)\left\|M\right\|_F \leq \frac{1}{m}\sum_{i=1}^{m}|A_i M A_i^T| \leq \sqrt{2}(1+\epsilon)\left\|M\right\|_F \tag{2.4.1}$$

for all symmetric rank-2 matrices M which implies 2-AGP with $M = xx^T - yy^T$, $\mu_1 = 0.9(1-\epsilon)$ and $\mu_2 = \sqrt{2}(1+\epsilon)$.

Lemma 2.4.2 (Assumption G $\implies$ 2-AGP $\implies$ 2-ARP)**.** *Under assumption G, there exist universal constants $c_0, c_1, C > 0, s \in (0,1), \psi \in (0,1)$ such that if $m > c_0 n$ and $A \in \mathbb{R}^{m\times n}$ satisfies G, then*

$$\left\|(|Ax|^2 - |Ay|^2)_T\right\|_1 \leq \psi \left\|(|Ax|^2 - |Ay|^2)_{T^c}\right\|_1 \;\; \forall\, x, y \in \mathbb{R}^n \text{ and } T \subseteq [m] \text{ with } |T| \leq sm$$

with probability at least $1 - C\exp(-c_1 m)$. Consequently, 2-ARP holds for m with high probability for m sufficiently large.

Proof. We first derive conditions on ϵ, $s \in (0,1)$ so that $\psi \in (0,1)$ exists. To this end let ϵ, $s \in (0,1)$ be given. Let $T \subset [m]$ be any subset of sm indices and denote by A_{T^c} the $(1-s)m \times n$ sub-matrix of A whose rows correspond to the indices in T^c. With this notation, we have $|A_{T^c}x| = |Ax|_{T^c}$. Also note that the entries of the matrix A_{T^c} satisfy G. By Lemma 2.4.1, there exist universal constants c_0, c_1, C such that if $m > c_0 n\epsilon^{-2}\log\frac{1}{\epsilon}$, then, for $M = xx^T - yy^T$ and each subset $T \subseteq |m|$ with $|T| = sm$,

$$0.9(1-\epsilon)\left\|xx^T - yy^T\right\|_F \leq \frac{1}{(1-s)m}\left\|(|Ax|^2 - |Ay|^2)_{T^c}\right\|_1 \leq \sqrt{2}(1+\epsilon)\left\|xx^T - yy^T\right\|_F \quad (2.4.2)$$

fails to hold with probability no greater than $C\exp(-c_1\epsilon^2(1-s)m)$, that is, 2-AGP holds for A_{T^c}. Since there are

$$\binom{m}{(1-s)m} = \binom{m}{sm} \leq \left(e\frac{m}{sm}\right)^{sm} = \left(\frac{e}{s}\right)^{sm}$$

such T's, the event

$$B := \{(2.4.2) \text{ holds for every } T \subseteq [m] \text{ with } |T| = sm\} \cap \{(2.4.1) \text{ holds}\},$$

satisfies

$$\begin{aligned}\mathbb{P}(B) &\geq 1 - C(e/s)^{sm}\exp(-c_1\epsilon^2(1-s)m) - C\exp(-c_1\epsilon^2 m)\\ &= 1 - C\exp((1+c_1\epsilon^2)sm + sm\log(\frac{1}{s}) - c_1\epsilon^2 m) - C\exp(-c_1\epsilon^2 m).\end{aligned}$$

Choose $\hat{s} > 0$ so that $(1+c_1\epsilon^2)\hat{s} + \hat{s}\log(\frac{1}{\hat{s}}) < \frac{c_1}{2}\epsilon^2$. Then, for all $s \in (0,\hat{s})$, $\mathbb{P}(B) \geq 1 - 2C\exp(-(c_1/2)\epsilon^2 m)$. Thus, if event B occurs, we have

$$\begin{aligned}\left\|(|Ax|^2 - |Ay|^2)_T\right\|_1 &= \left\||Ax|^2 - |Ay|^2\right\|_1 - \left\|(|Ax|^2 - |Ay|^2)_{T^c}\right\|_1\\ &\leq \sqrt{2}(1+\epsilon)m\left\|xx^T - yy^T\right\|_F - 0.9(1-\epsilon)(1-s)m\left\|xx^T - yy^T\right\|_F\\ &\leq \frac{\sqrt{2}(1+\epsilon) - 0.9(1-\epsilon)(1-s)}{0.9(1-\epsilon)(1-s)}\left\|(|Ax|^2 - |Ay|^2)_{T^c}\right\|_1,\end{aligned} \quad (2.4.3)$$

where the first inequality follows from (2.4.1) applied to the first term and (2.4.2) applied to the second, and the second inequality follows by (2.4.2). Consequently, as long as $s \in (0,\hat{s})$ is

chosen so that $\psi := \frac{\sqrt{2}(1+\epsilon)-0.9(1-\epsilon)(1-s)}{0.9(1-\epsilon)(1-s)} < 1$, the conclusion follows. This can be accomplished by choosing ϵ so that $\frac{\sqrt{2}(1+\epsilon)}{1.8(1-\epsilon)} < 1$ (or equivalently, $0 < \epsilon < \frac{1.8-\sqrt{2}}{1.8+\sqrt{2}}$) and then choosing $s \in (0, \min\{\hat{s}, 1 - \frac{\sqrt{2}(1+\epsilon)}{1.8(1-\epsilon)}\})$. □

2.4.2 $\mathbf{p = 1}$

This case requires a series of four technical lemmas in order to establish the main results. We list these lemmas below, and their proofs are in the appendix (Section 2.7).

Lemma 2.4.3. *Under assumption G, there exist universal constants C_0, C_1, C_2 such that for $\tilde{\epsilon} > 0$ sufficiently small, if $m > C_0 n \tilde{\epsilon}^{-4} \log \tilde{\epsilon}^{-1}$, then with probability at least $1 - C_1 \exp(-C_2 \tilde{\epsilon}^4 m)$,*

$$(1-\tilde{\epsilon})\sqrt{\frac{2}{\pi}}\,\|h\| \le \frac{1}{m}\sum_{i=1}^{m}|A_i h| \le (1+\tilde{\epsilon})\sqrt{\frac{2}{\pi}}\,\|h\| \quad \forall\, h \in \mathbb{R}^n. \tag{2.4.4}$$

Lemma 2.4.4. *Under assumption G, there exists universal constants $\tilde{c}_0, \tilde{c}_1, \tilde{C}$ such that for $\tilde{\epsilon}$ sufficiently small, if $m > \tilde{c}_0 n \tilde{\epsilon}^{-2} \log \frac{1}{\tilde{\epsilon}}$, then with probability at least $1 - \tilde{C}\exp(-\tilde{c}_1 \tilde{\epsilon}^2 m)$,*

$$\frac{1}{m}\sum_{i=1}^{m}\left||A_i x|^2 - |A_i y|^2\right|^{\frac{1}{2}} \ge 0.77(1-\tilde{\epsilon})\left\|xx^T - yy^T\right\|_F^{\frac{1}{2}} \quad \forall\, x, y \in \mathbb{R}^n. \tag{2.4.5}$$

Lemma 2.4.5. *For $x,\ y \in \mathbb{R}^n$, if $x^T y \ge 0$ (i.e. $\|x-y\| \le \|x+y\|$), then*

$$\|x+y\| + (\sqrt{2}-1)\|x-y\| \ge \|x\| + \|y\| \tag{2.4.6}$$

Lemma 2.4.6. *For $x, y \in \mathbb{R}^n$,*

$$\sqrt{2}\left\|xx^T - yy^T\right\|_F \ge \|x+y\|\,\|x-y\| \tag{2.4.7}$$

We first show that if the matrix A satisfies Assumption G, then it satisfies 1−AGP with high probability.

Lemma 2.4.7 (Assumption G $\implies$ 1-AGP)**.** *Under assumption G, there exist universal constants $\tilde{C}_0, \tilde{C}_1, \tilde{C}_2 > 0$ such that for $\tilde{\epsilon} > 0$ sufficiently small, if $m > \tilde{C}_0 n \tilde{\epsilon}^{-4} \log \frac{1}{\tilde{\epsilon}}$, then with probability at least $1 - \tilde{C}_1 \exp(-\tilde{C}_2 \tilde{\epsilon}^4 m)$,*

$$\mu_1 \phi_1(x,y) \le \frac{1}{m}\left\||Ax| - |Ay|\right\|_1 \le \mu_2 \phi_1(x,y) \quad \forall\, x, y \in \mathbb{R}^n, \tag{2.4.8}$$

where $\phi_1(x,y)$ *is defined in* (2.2.6), $\mu_1 = \sqrt{\frac{2}{\pi}}(2-\sqrt{2}-\tilde{\epsilon})$ *and* $\mu_2 = \sqrt{\frac{2}{\pi}}(1+\tilde{\epsilon})$. *Consequently, 1-AGP holds with high probability for* m *sufficiently large.*

Proof. By Lemma 4.4.4 and Lemma 2.4.4, there exist universal constant c_0, c_1, c_2 such that for ϵ sufficiently small, if $m > c_0 n\epsilon^{-4}\log\frac{1}{\epsilon}$, then with probability at least $1-c_1\exp(-c_2\epsilon^4 m)$, (4.4.9) and (2.7.6) hold. Since we can substitute y by $-y$ if necessary, without loss of generality, we assume $\|x-y\| \le \|x+y\|$.

The right hand inequality in (4.4.10) easily follows by (4.4.9) and triangle inequality

$$\||Ax|-|Ay|\|_1 \le \|A(x-y)\|_1 .$$

For the left hand inequality of (4.4.10), we consider two cases: (1) $\|x-y\| \le \|x+y\| \le 10\|x-y\|$, and (2) $\|x+y\| \ge 10\|x-y\|$.

(1) Assume $\|x-y\| \le \|x+y\| \le 10\|x-y\|$. By (4.4.9), we know

$$\begin{aligned}
&\frac{1}{m}\||Ax|-|Ay|\|_1 = \frac{1}{m}\sum_{i=1}^m ||A_ix|-|A_iy|| \\
&= \frac{1}{m}\sum_{i=1}^m |A_i(x+y)| + \frac{1}{m}\sum_{i=1}^m |A_i(x-y)| - \frac{1}{m}\sum_{i=1}^m |A_ix| - \frac{1}{m}\sum_{i=1}^m |A_iy| \\
&\ge \sqrt{\frac{2}{\pi}}\left((1-\epsilon)\|x+y\| + (1-\epsilon)\|x-y\| - (1+\epsilon)\|x\| - (1+\epsilon)\|y\|\right) \\
&\ge \sqrt{\frac{2}{\pi}}((2-\sqrt{2}-\sqrt{2}\epsilon)\|x-y\| - 2\epsilon\|x+y\|) \\
&\ge \sqrt{\frac{2}{\pi}}(2-\sqrt{2}-(\sqrt{2}+20)\epsilon)\|x-y\| ,
\end{aligned} \tag{2.4.9}$$

where the second equality is from $||a|-|b|| = |a+b|+|a-b|-|a|-|b|$ for $a,b\in\mathbb{R}$(since if $ab \ge 0$, then $||a|-|b|| = |a-b|$ and $|a+b| = |a|+|b|$ and if $ab<0$, then $||a|-|b|| = |a+b|$ and $|a-b| = |a|+|b|$), the first inequality is from Lemma 4.4.4 (with $=h$ successively set to $x+y$, $x-y$, x, and y), the second inequality uses Lemma 2.4.5 to replace $\|x\|+\|y\|$, and the last inequality follows from our assumption that $\|x+y\| \le 10\|x-y\|$.

(2) Assume $\|x+y\| \geq 10\,\|x-y\|$. We have

$$\begin{aligned}
\frac{1}{m}\,\||Ax|-|Ay|\|_1 &= \frac{1}{m}\sum_{i=1}^{m}||A_ix|-|A_iy|| \\
&\geq \left(\frac{1}{m}\sum_{i=1}^{m}||A_ix|^2-|A_iy|^2|^{\frac{1}{2}}\right)^2 \Bigg/ \left(\frac{1}{m}\sum_{i=1}^{m}(|A_ix|+|A_iy|)\right) \\
&\geq \sqrt{\frac{\pi}{2}}\frac{0.77^2(1-\epsilon)^2\left\|xx^T-yy^T\right\|_F}{(1+\epsilon)(\|x\|+\|y\|)} \\
&\geq \frac{0.77^2\sqrt{\pi}(1-\epsilon)^2\,\|x+y\|\,\|x-y\|}{2(1+\epsilon)(\|x\|+\|y\|)} \\
&\geq \frac{0.77^2\sqrt{\pi}(1-\epsilon)^2\,\|x+y\|\,\|x-y\|}{2(1+\epsilon)(\|x+y\|+(\sqrt{2}-1)\,\|x-y\|)} \\
&\geq \frac{5\cdot 0.77^2\sqrt{\pi}(1-\epsilon)^2}{(\sqrt{2}+9)(1+\epsilon)}\,\|x-y\|
\end{aligned} \tag{2.4.10}$$

where the first inequality is by Cauchy-Schwartz inequality applied to the vectors with $u_i = ||A_ix|-|A_iy||^{\frac{1}{2}}$ and $v_i = ||A_ix|+|A_iy||^{\frac{1}{2}}$, the second inequality is by Lemma 2.4.4 and Lemma 4.4.4, the third inequality is by Lemma 2.4.6, the fourth inequality is by Lemma (2.4.5) and the last inequality is by $\|x+y\| \geq 10\,\|x-y\|$. When $0<\epsilon<0.01$, one can show by direct computation that

$$\frac{5\cdot 0.77^2\sqrt{\pi}(1-\epsilon)^2}{(\sqrt{2}+9)(1+\epsilon)} > 0.02+\sqrt{\frac{2}{\pi}}(2-\sqrt{2}),$$

and so

$$\frac{1}{m}\,\||Ax|-|Ay|\|_1 \geq \sqrt{\frac{2}{\pi}}(2-\sqrt{2})\,\|x-y\| \tag{2.4.11}$$

Consequently,

$$\frac{1}{m}\,\||Ax|-|Ay|\|_1 \geq \sqrt{\frac{2}{\pi}}\left(2-\sqrt{2}-(20+\sqrt{2})\epsilon\right))\,\|x-y\|\,.$$

By substituting ϵ with $\tilde{\epsilon}/(\sqrt{2}+20)$ and adjusting c_0, c_1, c_2 we arrived at the desired result. □

Lemma 2.4.8 (Assumption G $\implies$ 1-AGP $\implies$ 1-ARP)**.** *Under Assumption G, there exist universal constants $c_0, c_1, C>0, s\in(0,1), \psi\in(0,1)$ such that if $m>c_0n$, then*

$$\|(|Ax|-|Ay|)_T\|_1 \leq \psi\,\|(|Ax|-|Ay|)_{T^c}\|_1 \;\; \forall\, x,y\in\mathbb{R}^n \text{ and } T\subseteq[m] \text{ with } |T|\leq sm$$

holds with probability at least $1 - C\exp(-c_1 m)$. *Consequently, 1-ARP holds with high probability for* m *sufficiently large.*

Proof. The proof strategy is similar to Lemma 2.4.2. Let $\phi_1(x, y)$ be as defined in (2.2.6). Again, we first derive conditions on ϵ, $s \in (0, 1)$ so that $\psi \in (0, 1)$ exists. To this end let ϵ, $s \in (0, 1)$ be given. By Lemma 2.4.7, there exist universal constants c_0, c_1, C such that if $m > c_0 n\epsilon^{-4} \log \frac{1}{\epsilon}$, then, for any $x, y \in \mathbb{R}^n$ and each subset $T \subseteq |m|$ with $|T| = sm$, the double sided inequality

$$\sqrt{\frac{2}{\pi}}(2 - \sqrt{2} - \epsilon)\phi_1(x, y) \leq \frac{1}{(1 - s)m} \|(|Ax| - |Ay|)_{T^c}\|_1 \leq \sqrt{\frac{2}{\pi}}(1 + \epsilon)\phi_1(x, y) \tag{2.4.12}$$

fails to hold with probability no larger than $C\exp(-c_1\epsilon^2(1 - s)m)$, that is, 1-AGP holds for A_{T^c}. We know for the event $B := \{(2.4.12) \text{ holds for every } \text{ with } |T| = sm\} \cap \{(4.4.10) \text{ holds}\}$, by taking s sufficient small, there exist positive constant constant $\tilde{c}$ and $\tilde{C}$ such that $\mathbb{P}(B) \geq 1 - \tilde{C}\exp(-\tilde{c}\epsilon^4 m)$. On the event B, we obtain

$$\begin{aligned}
\|(|Ax| - |Ay|)_T\|_1 &= \||Ax| - |Ay|\|_1 - \|(|Ax| - |Ay|)_{T^c}\|_1 \\
&\leq \sqrt{\frac{2}{\pi}}(1 + \epsilon)m\phi_1(x, y) - \sqrt{\frac{2}{\pi}}(2 - \sqrt{2})(1 - \epsilon)(1 - s)m\phi_1(x, y) \\
&\leq \frac{(1 + \epsilon) - (2 - \sqrt{2})(1 - \epsilon)(1 - s)}{(2 - \sqrt{2})(1 - \epsilon)(1 - s)} \|(|Ax| - |Ay|)_{T^c}\|_1
\end{aligned}$$

So as long as we choose $s \in (0, 1)$ such that $\psi := \frac{(1+\epsilon)-(2-\sqrt{2})(1-\epsilon)(1-s)}{(2-\sqrt{2})(1-\epsilon)(1-s)} < 1$, the conclusion follows. More precisely, $0 < s < 1 - \frac{1+\epsilon}{2(2-\sqrt{2})(1-\epsilon)}$ (Note ϵ must be chosen such that $\frac{1+\epsilon}{2(2-\sqrt{2})(1-\epsilon)} < 1$ in advance, which is possible since $2(2 - \sqrt{2}) > 1$).

□

By combining the results of this section with those of Section 2.3 we show under Assumption G that the solutions to the ℓ_0 optimization problem (2.1.7) and ℓ_1 optimization problem (2.2.3) coincide with high probability when the residuals are sufficiently sparse. Methods for solving (2.2.3) often require that the objective function f_p satisfies a sharpness condition. In the next section, we consider this sharpness condition.

2.5 Sharpness

In this section we show that, under assumption G, if $|Ax_*|^p - b$ is sufficiently sparse, then the function

$$f_p(x) := \frac{1}{m} \| \, |Ax|^p - b\|_1$$

is sharp with respect to the solution set $\{x_*, -x_*\}$ with high probability, for $p = 1, 2$. Sharpness is an extremely useful tool for analyzing the convergence and the rate of convergence of optimization algorithms [9, 12, 13, 14, 24, 33, 41].

Definition 2.5.1. *[14] Let $f : \mathbb{R}^n \to \mathbb{R}$ and set $\mathcal{X} := \operatorname{argmin} f$. Then f is said to be sharp with respect to $\mathcal{X}$ if*

$$f(x) \geq \min_x f \ + \ \mu dist(x.\mathcal{X}) \quad \forall x \in \mathbb{R}^n,$$

where $dist(x.\mathcal{X}) := \inf_{y \in \mathcal{X}} \|x - y\|$.

Theorem 2.5.2. *Let Assumption G hold and let $p \in \{1, 2\}$. Then there exist constants $C_p, c_{p0}, c_{p1} > 0$ and $s_p \in (0, 1)$, such that if $\| \, |Ax_*|^p - b\|_0 \leq s_p m$, then, for $m \geq c_{p0} n$, f_p is sharp with probability at least $1 - C_p \exp(-c_{p1} m)$.*

Proof. Let $C_p, c_{p0}, c_{p1} > 0$ and $s_p \in (0, 1)$ be as in Lemma 2.4.2 for $p = 2$ and as in Lemma 2.4.8 for $p = 1$. By either Lemma 2.4.2 ($p = 2$) or Lemma 2.4.8 ($p = 1$), A satisfies p-ARP of order $s_p m$ for $\psi_p \in (0, 1)$ for $p = 1, 2$, where s_p and ψ_p are constants depending on p. Hence, by (2.3.5),

$$f_p(x) - f_p(x_*) \geq \frac{1 - \psi_p}{m(1 + \psi_p)} \left\| |Ax|^2 - |Ax_*|^2 \right\|_1 \tag{2.5.1}$$

For $p = 2$, Lemma 2.4.1 tells us that if $m \geq c_{p0} \epsilon^{-2} \log(\frac{1}{\epsilon}) n$, then, with probability at least $1 - C_p \exp\left(c_{p1} \epsilon^{-2} \log \frac{1}{\epsilon} m\right)$,

$$\begin{aligned}
\frac{1}{m} \left\| |Ax|^2 - |Ax_*|^2 \right\|_1 &\geq 0.9(1 - \epsilon) \left\| xx^T - yy^T \right\|_F \\
&\geq 0.45\sqrt{2}(1 - \epsilon) \, \|x + x_*\| \, \|x - x_*\| \\
&= 0.45\sqrt{2}(1 - \epsilon)\phi_1(x, x_*) \max\{\|x - x_*\|, \|x + x_*\|\} \\
&\geq 0.45\sqrt{2}(1 - \epsilon) \, \|x_*\| \operatorname{dist}(x, \{x_*, -x_*\}),
\end{aligned} \tag{2.5.2}$$

where $\phi_1(x, x_*)$ is defined in (2.2.6). For $p = 1$, Lemma 2.4.7 tells us that, if $m \geq c_{p0}\epsilon^{-4}\log(\frac{1}{\epsilon})n$, then, with probability at least $1 - C_p \exp\left(c_{p1}\epsilon^{-4}\log\frac{1}{\epsilon}m\right)$,

$$\frac{1}{m}\||Ax| - |Ax_*|\|_1 \geq \sqrt{\frac{2}{\pi}}(2 - \sqrt{2} - \epsilon)\text{dist}(x, \{x_*, -x_*\}).$$

Thus, in either case, by taking an $0 < \epsilon < 1$ small enough and using (2.5.1), there is constant $\mu > 0$ such that

$$f_p(x) - f_p(x_*) \geq \mu\text{dist}(x, \mathcal{X}),$$

where $\mathcal{X}$ is argmin f_p. □

It is shown in [24, 41] that if f_2 is sharp and weakly convex at argmin f_2, then prox-linear method and subgradient descent method with geometrically decreasing stepsize converges locally quadratically and locally linearly, respectively. Since weak convexity of f_2 under assumption G is already shown in [41, 24, 33], sharpness in this regime guarantees these two algorithms converge with the specified rate. In both algorithms proper initialization is needed (e.g., Section 5 of [78]).

2.6 Concluding Remarks

There are a number of recent results discussing the nature of the solution set to the robust phase retrieval problem $\min_x f_2(x)$ with sparse noise under weaker distributional hypothesis than employed here [24, 41, 78, 26]. The focus of these works are algorithmic. Their goal is to show their methods are robust to outliers, and, in addition, some establish the sharpness of f_2 in order to prove rates of convergence [24, 41]. Although these works use weaker distributional hypothesis, the probability of successful recovery is an average over all possible subsets $T \subseteq [m]$ with $|T| = sm$ for some $s \in (0, \frac{1}{2})$. Consequently, the value of s in their results is larger than ours. The reason for this difference is that, in our result, successful recovery is valid for all possible subsets $T \subseteq [m]$ with $|T| = sm$ for some $s \in (0, 1)$, with uniformly high probability. A more precise description is this difference follows.

In [24, 41], the random matrix A and the random index set $T \subseteq [m]$, with $|T| = sm$ for $s \in (0, \frac{1}{2})$, are drawn independently of each other. Let $w \in \{0,1\}^m$ denote the random indicator vector of T, that is, $w_i = 1$ if $i \in T$ and $w_i = 0$ otherwise. Let $z \in \mathbb{R}^m$ be an arbitrary vector. The noisy model in [24, 41] has the form

$$\min_x \tilde{f}_2(x) := \left\| |Ax|^2 - (\vec{1} - w) \odot b - w \odot z \right\|_1,$$

where $b = |Ax_*|^2$, $\vec{1}$ represents the vector with 1 in each entry and $\odot$ represents the elementwise product of vectors. The authors in [24, 41] prove sharpness of $\tilde{f}_2$ with respect to x_* with high probability. Due to the independence of A and T, in fact, they show that the probabillity

$$\mathbb{P}(\tilde{f}_2 \text{ is sharp}) = \frac{1}{\binom{m}{sm}} \sum_{T_0 : |T_0| = sm} \mathbb{P}(\tilde{f}_2^{T_0} \text{ is sharp})$$

is high, where $\tilde{f}_2^{T_0}(x) := \left\| |Ax|^2 - (\vec{1} - w_0) \odot b - w_0 \odot z \right\|_1$ and w_0 is the indicator vector for a *fixed* index set T_0. On the other hand, we show that with high probability, $\tilde{f}_2^{T_0}$ is sharp for *all* possible T_0 with $|T_0| = sm$. Our result is a stronger implication, however, it comes at the expense of a smaller value for s. By design, this result closely parallels the result in [22] for compressed sensing.

2.7 Appendix

In this appendix we provide the proofs for Lemmas 4.4.4, 2.4.4, 2.4.5, and 2.4.6. These proofs make use of a Hoeffding-type inequality [74] explained below. A random variable X is said to be sub-gaussian [74, Definition 5.7] if

$$\|X\|_{\psi_2} := \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p} \tag{2.7.1}$$

is finite, and is said to be centered if it has zero expectation. By [74, Proposition 5.10], there is a universal constant $c > 0$ such that if $X_1, ..., X_N$ are independent centered sub-gaussian random variables, then, for every $a = \{a_1, ..., a_N\} \in \mathbb{R}^N$ and $t \geq 0$, we have

$$\mathbb{P}\left(|\sum_{i=1}^N a_i X_i| \geq t \right) \leq e \cdot \exp(-\frac{ct^2}{K^2 \|a\|^2}), \tag{2.7.2}$$

where $K := \max_i \|X_i\|_{\psi_2}$.

Proof of Lemma 4.4.4: First observe that the inequality (4.4.9) is trivially true for $h = 0$. Next, let $h \in \mathbb{R}^n \setminus \{0\}$ and $0 < \epsilon < \sqrt{2} - 1$. Observe that $\frac{|A_i h|}{\|h\|}$ are independent sub-gaussian random variables with mean $\sqrt{\frac{2}{\pi}}$. Therefore, $\frac{|A_i h|}{\|h\|} - \sqrt{\frac{2}{\pi}}$ is a centered sub-gaussian random variable. Hence, (2.7.2) tells us that there are universal constants $C > 0$ and $c_0 > 0$ such that

$$\mathbb{P}\left(\left|\sum_{i=1}^{m}\left(\frac{|A_i h|}{\|h\|} - \sqrt{\frac{2}{\pi}}\right)\right| > m\sqrt{\frac{2}{\pi}}\epsilon\right) \leq C\exp(-c_0 m\epsilon^2). \tag{2.7.3}$$

Therefore (4.4.9) holds for each fixed $h \in \mathbb{R}^n \setminus \{0\}$ with probability $1 - C\exp(-c_0 m\epsilon^2)$. We now show that there exist a universal event with large probability, in which (4.4.9) holds for every h. On the unit sphere $S := \{x | \|x\| = 1\}$ construct an ϵ-net $\mathcal{N}_\epsilon$ with $|\mathcal{N}_\epsilon| \leq (1 + \frac{2}{\epsilon})^n$ [74, Lemma 5.2], i.e., for any $h \in S$, there exists $h_0 \in \mathcal{N}_\epsilon \subseteq S$ such that $\|h - h_0\| \leq \epsilon$. Taking the probability of the union of the events in (2.7.3) for all the points $h_0 \in \mathcal{N}_\epsilon$, we obtain the bound $C(1 + \frac{2}{\epsilon})^n \exp(-c_0 m\epsilon^2)$. Hence, (4.4.9) holds for each $h_0 \in \mathcal{N}_\epsilon$ with probability at least $1 - C(1 + \frac{2}{\epsilon})^n \exp(-c_0 m\epsilon^2)$. On the intersection of these events and the event of Lemma 2.4.1, we deduce, for any h with $\|h\| = 1$,

$$\begin{aligned}
\frac{1}{m}\Big|\sum_{i=1}^{m}|A_i h| - \sum_{i=1}^{m}|A_i h_0|\Big| &\leq \frac{1}{m}\sum_{i=1}^{m}||A_i h| - |A_i h_0|| \\
&\leq \frac{1}{m}\sum_{i=1}^{m}||A_i h|^2 - |A_i h_0|^2|^{\frac{1}{2}} \\
&\leq \Big(\frac{1}{m}\sum_{i=1}^{m}||A_i h|^2 - |A_i h_0|^2|\Big)^{\frac{1}{2}} \\
&\leq 2^{1/4}(1+\epsilon)^{1/2}\left\|hh^T - h_0 h_0^T\right\|_F^{\frac{1}{2}} \\
&\leq 2^{1/4}(1+\epsilon)^{1/2}(\|h - h_0\|\,\|h\| + \|h - h_0\|\,\|h_0\|)^{\frac{1}{2}} \\
&\leq 2^{5/4}\epsilon^{1/2},
\end{aligned} \tag{2.7.4}$$

where the second inequality follows since $||a| - |b||^2 \leq (|a| + |b|)||a| - |b||$, the third from the concavity of $(\cdot)^2$, the fourth is by Lemma 2.4.1, the fifth is by triangle inequality and the

last inequality is from $\|h\| = \|h_0\| = 1$ and $\|h - h_0\| \le \epsilon$. Hence

$$(1 - \epsilon - 2^{3/4}\sqrt{\pi\epsilon})\sqrt{\frac{2}{\pi}} \le \frac{1}{m}\sum_{i=1}^{m} |A_i h| \le (1 + \epsilon + 2^{3/4}\sqrt{\pi\epsilon})\sqrt{\frac{2}{\pi}}$$

holds for all $\|h\| = 1$ with probability at least $1 - (1 + \frac{2}{\epsilon})^n \exp(-c_0 m\epsilon^2) - c_2 \exp(-c_3 m\epsilon^2)$, for $m \ge c_1 n\epsilon^{-2}\log(\frac{1}{\epsilon})$. For $c_1 > 0$ sufficiently large and ϵ small, the probability is at least

$$\begin{aligned} & 1 - c_2 \exp(-c_3 m\epsilon^2) - \exp(-c_0 m\epsilon^2 + 2n\log(\frac{1}{\epsilon})) \\ & \ge 1 - c_2 \exp(-c_3 m\epsilon^2) - \exp(-(c_0 - \frac{2}{c_1})m\epsilon^2) \\ & \ge 1 - \tilde{c}_2 \exp(-\tilde{c}_3 m\epsilon^2), \end{aligned} \tag{2.7.5}$$

for some $\tilde{c}_2$, $\tilde{c}_3 > 0$. By letting $\tilde{\epsilon} = \epsilon + 2^{3/4}\sqrt{\pi\epsilon} < (1 + 2^{3/4}\sqrt{\pi})\sqrt{\epsilon}$ so that $\epsilon \ge k\tilde{\epsilon}^2$ for $k > 0$, we arrive at the desired result. □

Proof of Lemma 2.4.4: We only need to prove

$$\frac{1}{m}\sum_{i=1}^{m} \left|A_i M A_i^T\right|^{\frac{1}{2}} \ge 0.77(1 - \epsilon)\|M\|_F^{\frac{1}{2}} \tag{2.7.6}$$

holds for all rank-2 matrix M with high probability. Clearly this inequality holds when $M = 0$. Assume $M \ne 0$. Furthermore, since we can divide (2.7.6) by $\|M\|^{\frac{1}{2}}$ on both sides, we can assume $\|M\| = 1$. Moreover, using the eigenvalue decomposition of M, we can assume that $M = z_1 z_1^T - s z_2 z_2^T$ where $z_1^T z_2 = 0$, $\|z_1\| = \|z_2\| = 1$ and $s \in [-1, 1]$. Since for each i, $A_i z_1$ and $A_i z_2$ are independent standard gaussians,

$$\left|A_i M A_i^T\right|^{\frac{1}{2}} = \left|(A_i z_1)^2 - s(A_i z_2)^2\right|^{\frac{1}{2}} \le \left((A_i z_1)^2 + (A_i z_2)^2\right)^{\frac{1}{2}} \le |A_i z_1| + |A_i z_2| \tag{2.7.7}$$

are sub-gaussian. Set $e(s) := \mathbb{E}\left|A_i M A_i^T\right|^{\frac{1}{2}} = \mathbb{E}\left|Z_1^2 - sZ_2^2\right|^{\frac{1}{2}}$ where Z_1 and Z_2 are independent standard gaussian scaler random variables. Notice $\|M\|_F = \left\|z_1 z_1^T - s z_2 z_2^T\right\|_F = \sqrt{1 + s^2}$ and

$$\begin{aligned} e(s) = \mathbb{E}\left|Z_1^2 - sZ_2^2\right|^{\frac{1}{2}} &= \frac{1}{2\pi}\int_0^\infty r^2 e^{-\frac{r^2}{2}} dr \int_0^{2\pi} \left|\cos^2\theta - s\sin^2\theta\right|^{-\frac{1}{2}} d\theta \\ &= \frac{1}{2\sqrt{2\pi}}\int_0^{2\pi} \left|\cos^2\theta - s\sin^2\theta\right|^{\frac{1}{2}} d\theta \end{aligned} \tag{2.7.8}$$

We draw a plot of $\frac{e(s)}{\|M\|_F} = \int_0^{2\pi} \left|\cos^2\theta - s\sin^2\theta\right|^{\frac{1}{2}} d\theta/(2\sqrt{2\pi(1+s^2)})$ when $s \in [-1,1]$ through a numerical experiment.

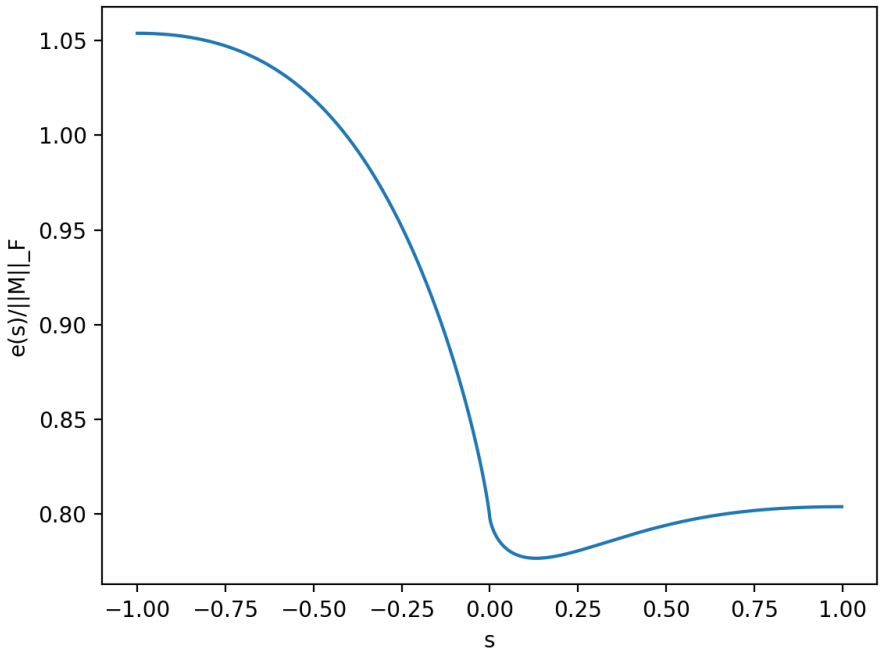

Figure 2.1: Values of $\frac{e(s)}{\|M\|_F}$ when $s \in [-1,1]$.

Numerical experiment above shows that $\frac{e(s)}{\|M\|_F} \geq 0.77$ (hence $e(s) \geq 0.77$) for all $a \in [-1,1]$. Note that for each i, $Y_i := \frac{\left|A_i M A_i^T\right|^{\frac{1}{2}}}{e(s)} - 1$ is a centered sub-gaussian random variable. Hence, by (2.7.1) and (2.7.7),

$$\|Y_i\|_{\psi_2} \leq \sup_{p\geq 1} p^{-\frac{1}{2}}\left(\frac{2(\mathbb{E}|Z|^p)^{\frac{1}{p}}}{e(s)} + 1\right) \leq \frac{2}{0.77}\|Z\|_{\psi_2} + 1 < +\infty.$$

where Z is a standard gaussian variable. Hence, (2.7.2) tells us that there exist universal constants $C > 0$ and $c_0 > 0$ such that

$$\mathbb{P}\left(\left|\sum_{i=1}^m \left(\frac{\left|A_i M A_i^T\right|^{\frac{1}{2}}}{e(s)} - 1\right)\right| > m\epsilon\right) \leq \hat{C}\exp(-\hat{c}_0 m\epsilon^2) \tag{2.7.9}$$

Consequently, for fixed M,

$$\frac{1}{m}\sum_{i=1}^{m}\left|A_iMA_i^T\right|^{\frac{1}{2}} \geq (1-\epsilon)e(s) \geq 0.77(1-\epsilon)\left\|M\right\|_F \tag{2.7.10}$$

holds with probability at least $1-\hat{C}\exp(-\hat{c}_0m\epsilon^2)$.

Next we generalize (2.7.10) to all rank-2 matrices M. Again, by scale invariance, we assume $\|M\|_F = 1$. Consequently, we only need to prove (2.4.4) holds with high probability for all $M \in \mathcal{M} := \{\beta uu^T + \gamma vv^T | \, \|u\| = \|v\| = 1, \; u^Tv = 0 \;\text{ and }\; \beta^2+\gamma^2 = 1\}$. Set $\mathcal{S}_{\epsilon^2} := \mathcal{T}_{\epsilon^2} \times \mathcal{N}_{\epsilon^2} \times \mathcal{N}_{\epsilon^2}$ where $\mathcal{T}_{\epsilon^2}$ is an ϵ^2-net of $[-1,1]$ and $\mathcal{N}_{\epsilon^2}$ is an ϵ^2-net of the unit sphere $\{x \in \mathbb{R}^n | \, \|x\| = 1\}$. Since $|\mathcal{T}_{\epsilon^2}| \leq \frac{2}{\epsilon^2}$ and $|\mathcal{N}_{\epsilon^2}| \leq \left(\frac{3}{\epsilon^2}\right)^n$, we know $|\mathcal{S}_{\epsilon^2}| \leq \left(\frac{3}{\epsilon}\right)^{4n+2}$. Let E denote the event that (2.7.10) holds for every $(\beta_0, u_0.v_0) \in \mathcal{S}_{\epsilon^2}$. Consequently,

$$\mathbb{P}(E) \geq 1 - 2\hat{C}\left(\tfrac{3}{\epsilon}\right)^{4n+2}\exp(-\hat{c}_0m\epsilon^2).$$

For $M \in \mathcal{M}$, we want to approximate $M = \beta uu^T + \gamma vv^T$ by an element $M_0 = \beta_0u_0u_0^T + \gamma_0v_0v_0^T \in \mathcal{M}$ with $(\beta_0, u_0, v_0) \in \mathcal{S}_{\epsilon^2}$. More precisely, let $(\beta_0, u_0, v_0) \in \mathcal{S}_{\epsilon^2}$ and $M_0 = \beta_0u_0u_0^T + \operatorname{sgn}(\gamma)\sqrt{1-\beta_0^2}v_0v_0^T$ be such that $|\beta-\beta_0| \leq \epsilon^2$, $\|u-u_0\| \leq \epsilon^2$ and $\|v-v_0\| \leq \epsilon^2$. Consequently, we have

$$|\gamma - \operatorname{sgn}(\gamma)\sqrt{1-\beta_0^2}| = |\sqrt{1-\beta^2} - \sqrt{1-\beta_0^2}| \leq |\beta^2-\beta_0^2|^{\frac{1}{2}} \leq \sqrt{2}\,|\beta-\beta_0|^{\frac{1}{2}} \leq \sqrt{2}\epsilon.$$

Also note that

$$\begin{aligned}\left\|\beta uu^T - \beta_0u_0u_0^T\right\|_F &\leq |\beta-\beta_0|\left\|uu^T\right\|_F + \left\|\beta_0u(u-u_0)^T\right\|_F + \left\|\beta_0(u-u_0)u_0^T\right\|_F \\ &= |\beta-\beta_0|\,\|u\|^2 + |\beta_0|\,\|u-u_0\|\,(\|u\|+\|u_0\|) \\ &\leq 3\epsilon^2 < 4\epsilon\end{aligned} \tag{2.7.11}$$

Similarly we can prove $\left\|\gamma vv^T - \operatorname{sgn}(\gamma)\sqrt{1-\beta_0^2}v_0v_0^T\right\| \leq 2\epsilon^2 + 2\epsilon < 4\epsilon$. On the intersection

of events where (2.4.1) holds and E, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}|A_iMA_i^T|^{\frac{1}{2}}-\frac{1}{m}\sum_{i=1}^{m}|A_iM_0A_i^T|^{\frac{1}{2}}\right|\leq\frac{1}{m}\sum_{i=1}^{m}\left|\left|A_iMA_i^T\right|^{\frac{1}{2}}-\left|A_iM_0A_i^T\right|^{\frac{1}{2}}\right|\\
&\leq\frac{1}{m}\sum_{i=1}^{m}\left|\left|A_iMA_i^T\right|-\left|A_iM_0A_i^T\right|\right|^{\frac{1}{2}}\\
&\leq\left(\frac{1}{m}\sum_{i=1}^{m}\left|\left|A_iMA_i^T\right|-\left|A_iM_0A_i^T\right|\right|\right)^{\frac{1}{2}}\\
&\leq\left(\frac{1}{m}\sum_{i=1}^{m}\left|A_i(M-M_0)A_i^T\right|\right)^{\frac{1}{2}}\\
&\leq\left(\frac{1}{m}\sum_{i=1}^{m}\left|A_i(\beta uu^T-\beta_0u_0u_0^T)A_i^T\right|+\left|A_i(\gamma vv^T-\gamma_0v_0v_0^T)A_i^T\right|\right)^{\frac{1}{2}}\\
&\leq\left(\frac{1}{m}\sum_{i=1}^{m}\left|A_i(\beta uu^T-\beta_0u_0u_0^T)A_i^T\right|\right)^{\frac{1}{2}}+\left(\frac{1}{m}\sum_{i=1}^{m}\left|A_i(\gamma vv^T-\gamma_0v_0v_0^T)A_i^T\right|\right)^{\frac{1}{2}}\\
&\leq 2^{\frac{1}{4}}(1+\epsilon)^{\frac{1}{2}}\left\|\beta uu^T-\beta_0u_0u_0^T\right\|_F^{\frac{1}{2}}+2^{\frac{1}{4}}(1+\epsilon)^{\frac{1}{2}}\left\|\gamma vv^T-\gamma_0v_0v_0^T\right\|_F^{\frac{1}{2}}\\
&\leq 2^{\frac{9}{4}}(1+\epsilon)^{\frac{1}{2}}\epsilon^{\frac{1}{2}},
\end{aligned}$$

where the second inequality is by $||a|-|b||^2\leq|a^2-b^2|$ for any $a,b\in\mathbb{R}$, the third inequality is by concavity of $(\cdot)^2$, the fourth and the fifth inequalities are by triangle inequality, the sixth inequality is by $a^2+b^2\leq(a+b)^2$ for any $a,b\in\mathbb{R}$ and the seventh inequality is by the right hand side of equation (2.4.1). Consequently, if $m>c_0n\epsilon^{-2}\log\frac{1}{\epsilon}$

$$\frac{1}{m}\sum_{i=1}^{m}\left|A_iMA_i^T\right|^{\frac{1}{2}}\geq 0.77(1-\epsilon-2^{\frac{9}{4}}(1+\epsilon)^{\frac{1}{2}}\epsilon^{\frac{1}{2}}) \tag{2.7.12}$$

holds with probability at least $1-2\hat{C}\left(\frac{3}{\epsilon}\right)^{4n+2}\exp(-\hat{c}_0m\epsilon^2)-C\exp(-c_1\epsilon^2m)$. As in (2.7.5), by making c_0 large, we are able to make the probability $\geq 1-\hat{\hat{C}}\exp(-\hat{\hat{c}}_0m\epsilon^2)$ for some constants $\hat{\hat{C}}$ and $\hat{\hat{c}}_0$. By letting $\tilde{\epsilon}:=\epsilon+2^{\frac{9}{4}}(1+\epsilon)^{\frac{1}{2}}\epsilon^{\frac{1}{2}}$ and adjust constants $\hat{\hat{C}},\hat{\hat{c}}_0,c_0$ we arrive at the desired result. □

Proof of Lemma 2.4.5: If $x=0$ or $y=0$ or $x=y$, the inequality holds. Thus, in particular, by the symmetry of (2.4.6) in x and y, we can assume that $\|x\|\geq\|y\|>0$. Dividing (2.4.6)

by $\|x\|$, tells us that we can assume $\|x\| = 1$ and $\|y\| = t$ for $t \in [0,1]$. Set $\rho := \frac{x^T y}{\|y\|} \in [0,1]$, and define $h(t,\rho) := \sqrt{t^2 - 2\rho t + 1} + \sqrt{t^2 + 2\rho t + 1} - 1 - t = \|x+y\| + \|x-y\| - \|x\| - \|y\|$. If $x = y$, we are done; otherwise, set $q(t,\rho) := \frac{h(t,\rho)}{\sqrt{t^2 - 2\rho t + 1}} = \frac{\|x+y\| + \|x-y\| - \|x\| - \|y\|}{\|x-y\|}$, for each $(t,\rho) \in [0,1] \times [0,1]$. We now show that the minimum value of q over $[0,1] \times [0,1]$ is $2 - \sqrt{2}$. For fixed $t \in [0,1]$,

$$\begin{aligned}
\frac{\partial q(t,\rho)}{\partial \rho} &= \frac{t(t^2+1)}{(t^2 - 2\rho t + 1)^{\frac{3}{2}}} \left[\frac{2}{(t^2 + 2\rho t + 1)^{\frac{1}{2}}} - \frac{t+1}{t^2+1} \right] \\
&\geq \frac{t(t^2+1)}{(t^2 - 2\rho t + 1)^{\frac{3}{2}}} \left[\frac{2}{t+1} - \frac{t+1}{t^2+1} \right] \\
&\geq 0,
\end{aligned}$$

where the first inequality follows since $t^2 + 2\rho t + 1 \leq (1+t)^2$ as $\rho \in [0,1]$, and the last inequality follows since $2(t^2+1) \geq (t+1)^2$. That is, $q(t,\rho)$ is increasing with respect to ρ when $\rho \in [0,1]$ for each fixed $t \in [0,1]$. Also

$$\frac{dq(t,0)}{dt} = -\frac{1-t}{(1+t^2)^{\frac{3}{2}}} \leq 0.$$

Hence $q(t,0)$ is decreasing for $t \in [0,1]$. We know for each $t \in [0,1]$, $\rho \in [0,1]$,

$$q(t,\rho) \geq q(t,0) \geq q(1,0) = 2 - \sqrt{2}$$

Thus $h(t,\rho) \geq (2 - \sqrt{2}) \|x - y\|$, which leads to the desired result. □

Proof of Lemma 2.4.6: If $x = y = 0$, we are done. Next assume at least one of x and y is non-zero. We assume $\|x\| = 1$ and $\|y\| = t \in [0,1]$ since we can divide (2.4.7) by $\max\{\|x\|, \|y\|\}$ on both sides. Set $\rho := \frac{x^T y}{\|y\|}$. We have

$$
\begin{aligned}
\sqrt{2}\left\|xx^T - yy^T\right\|_F &= \sqrt{2}\left(\sum_{i,j}(x_i x_j - y_i y_j)^2\right)^{\frac{1}{2}} \\
&= \sqrt{2}\left((\sum_i x_i^2)(\sum_j x_j^2)+(\sum_i y_i^2)(\sum_j y_j^2)-2(\sum_i x_i y_i)(\sum_j x_j y_j)\right)^{\frac{1}{2}} \\
&= \left(2(1+t^4) - 4\rho^2 t^2\right)^{\frac{1}{2}} \\
&\geq \left((1+t^2)^2 - 4\rho^2 t^2\right)^{\frac{1}{2}} \\
&= \sqrt{1+t^2+2\rho t}\sqrt{1+t^2-2\rho t} \\
&= \|x+y\|\,\|x-y\|\,,
\end{aligned}
$$

where the inequality follows by the algebraic geometric mean inequality. □

Chapter 3

IRLS FOR SPARSE RECOVERY REVISITED: EXAMPLES OF FAILURE AND A REMEDY

Abstract

Compressed sensing is a central topic in signal processing with myriad applications, where the goal is to recover a signal from as few observations as possible. Iterative re-weighting is one of the fundamental tools to achieve this goal. This paper re-examines the iteratively reweighted least squares (IRLS) algorithm for sparse recovery proposed by Daubechies, Devore, Fornasier, and Güntürk in *Iteratively reweighted least squares minimization for sparse recovery*, Communications on Pure and Applied Mathematics, **63**(2010) 1–38. Under the null space property of order K, the authors show that their algorithm converges to the unique k-sparse solution for k strictly bounded above by a value strictly less than K, and this k-sparse solution coincides with the unique ℓ_1 solution. On the other hand, it is known that, for k less than or equal to K, the k-sparse and ℓ_1 solutions are unique and coincide. The authors emphasize that their proof method does not apply for k sufficiently close to K, and remark that they were unsuccessful in finding an example where the algorithm fails for these values of k.

In this note we construct a family of examples where the Daubechies-Devore-Fornasier-Güntürk IRLS algorithm fails for $k = K$, and provide a modification to their algorithm that provably converges to the unique k-sparse solution for k less than or equal to K while preserving the local linear rate. The paper includes numerical studies of this family as well as the modified IRLS algorithm, testing their robustness under perturbations and to parameter selection.

3.1 Introduction

The fundamental problem in compressed sensing is to recover the sparsest solution x_* to a linear equation of the form $\Phi x = y$ for a given y, where $\Phi \in \mathbb{R}^{\ell \times N}$ is the measurement

matrix and $\ell < N$. We denote the set of solutions to the equation $\Phi x = y$ by $\Phi^{-1}(y)$ which is assumed to be non-empty throughout. The problem of obtaining the sparsest solution can be posed as the minimization of the so-called 0-norm, $\|x\|_0$, over $\Phi^{-1}(y)$, where $\|x\|_0$ is the number of non-zero components in the vector x. Since the 0-norm problem is NP hard, in practice [27] one replaces this problem with the ℓ_1 minimization (or basis pursuit) problem

$$\min_{x \in \Phi^{-1}(y)} \|x\|_1 . \tag{BP}$$

The relationship of BP to the 0-norm problem has been intensively studied over the past few years [17, 22, 36, 35]. Compressed sensing has applications to a range of signal processing areas, including image acquisition, sensor networks and image reconstruction [27, ?, 72].

Numerous algorithms have been proposed for solving BP and its various reformulations, which include the basis pursuit denoising (BPDN) problem:

$$\min_x \{\|x\|_1 \,|\, \|\Phi x - y\|_2 \leq \sigma\},$$

the LASSO problem: $\min_x \|x\|_1 + \frac{\mu}{2} \|\Phi x - y\|_2^2$, and the ℓ_1-regression problem:

$$\min_x \|Az - b\|_1 \tag{ℓ_1R}$$

under the correspondences $\operatorname{rge}(A) = \operatorname{Null}(\Phi)$ and $\Phi b = y$ [22] (see Section 3.5 for details). Algorithms designed to solve these problems include the iteratively reweighted least squares (IRLS) algorithms [11] which apply to ℓ_1R, the FISTA algorithm [7, 77] which applies to the LASSO, and the homotopy algorithm [61], the alternating direction method of multipliers (ADMM) [8, 45], and the level-set method described in [1] which all apply to BPDN. However, the focus of this paper is the IRLS algorithm described in [32] which we refer to as the DDFG-IRLS algorithm.

In [32], the authors show that if the matrix Φ satisfies the the *null space property of order* K *for* $0 < \gamma < 1$ (see Section 3.3 for details), then the DDFG-IRLS algorithm converges to the unique k-sparse solution when $k < K - 2\gamma(1-\gamma)^{-1}$, and this k-sparse solution coincides with the unique ℓ_1 solution, where a vector is k-sparse if it has k nonzero components. In

addition, the authors also establish the local linear convergence of the DDFG-IRLS algorithm when $0 < \gamma < 1 - 2/(K+2)$. On the other hand, it is known that for $k \leq K$ the k-sparse and ℓ_1 solutions are unique and coincide [49, 39, 32]. In [32, Remark 5.4], the authors note that their proof method does not apply for $K - 2\gamma(1-\gamma)^{-1} \leq k \leq K$, and state that they were unsuccessful in finding an example where the algorithm fails when k falls in this range. In this note we construct a family of examples where the DDFG-IRLS algorithm fails when $k = K$, and provide a modification to their algorithm that provably converges to the unique k-sparse solution for $k \leq K$. In addition, we show that this modification is locally linearly convergent for all $k \leq K$ and $\gamma \in (0, 1)$ which increases the range of γ values for which linear convergence is assured.

Iteratively re-weighted least squares algorithms (IRLS) for solving ℓ_p minimization problems for $1 \leq p \leq \infty$ have been in the literature for many years beginning with the Ph.D. thesis of Lawson [52]. For $0 < p \leq 1$, IRLS was used to solve sparse reconstruction in [47], and a theory for solving ℓ_p minimization problems in general can be found in [62]. We refer the reader to [63] for a survey on IRLS methods applied to robust regression. More recently, cluster point convergence of IRLS smoothing methods for problems of the form $\min f(x) + \lambda \|x\|_0$, where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, is given in [54]. In addition, an IRLS algorithm has been developed for convex inclusions of the form $A_i x + b_i \in C_i,\ i = 1, \ldots, n$ where the sets C_i are all assumed to be convex [16]. In this case, the authors establish the iteration complexity of their method. However, all of these methods focus on general linear systems and do not specifically address the problem of compressed sensing where the null space properties play a key role. Daubechies, Devore, Fornasier, and Güntürk [32] focus on the compressed sensing case where $\|x\|_0$ is approximated by a smoothing of the norms $\|x\|_p$ for $0 < p \leq 1$. We follow Daubechies, Devore, Fornasier, and Güntürk in the $p = 1$ case and suggest a simple modification to their method for updating the smoothing parameter. This modification allows us to obtain stronger convergence properties.

Our discussion proceeds as follows. In Section 2 we discuss the DDFG-IRLS algorithm and our modification to the smoothing parameter update procedure. In Section 3 we prove

the stronger convergence and rate of convergence properties for the modified algorithm. Our proofs closely parallel those given in [32] but contain some simplifications. In Section 4, we construct a family of examples where the DDFG-IRLS algorithm fails but our modifications succeed. These results are illustrated numerically in Section 5 where we also provide a few numerical experiments to illustrate the numerical stability of the modified algorithm. In particular, we show that on randomly chosen problems the two methods have virtually identical performance characteristics.

3.2 The Modified IRLS Algorithm

Our algorithm is similar to the IRLS algorithm given in [32]. The primary innovation is the manner in which the smoothing parameter ϵ_k is is updated. In [32], ϵ_k is updated by the rule

$$\epsilon_{k+1} = \min\left\{\epsilon_k, \frac{r_{K+1}(x^{k+1})}{N}\right\},$$

where, for $x = (x_1, ..., x_N)^T \in \mathbb{R}^N$,

$$r_i(x) \text{ is the } i\text{th largest element of } \{|x_j| | 1 \leq j \leq N\}.$$

On the other hand, the algorithm below employs the update rule

$$\epsilon_{k+1} = \min\left\{\epsilon_k, \frac{\eta(1-\gamma)\sigma_K(x^{k+1})}{N}\right\}, \tag{3.2.1}$$

where $\eta \in (0,1)$ is chosen and fixed at the beginning of the iteration, the parameters γ and K come from A2, and

$$\sigma_j(z) := \sum_{\nu > j} r_\nu(z), \quad j = 1, \ldots, N. \tag{3.2.2}$$

As stated, the algorithm is an iteratively re-weighted least squares algorithm where the weights at each iteration are given by

$$w_i^k := ((x_i^k)^2 + \epsilon_k^2)^{-1/2} \quad i = 1, \ldots, N. \tag{3.2.3}$$

Moreover, given a positive weight vector $w \in \mathbb{R}^N_{++}$, we define the associated inner product by

$$\langle u,\, v\rangle_w := \sum_{i=1}^N w_i u_i v_i \quad \forall\, u, v \in \mathbb{R}^N,$$

and the corresponding weighted 2-norm by $\|u\|_w := \sqrt{\langle u, u\rangle_w}$. With this notation, our algorithm can be stated as follows.

Algorithm 1: An IRLS algorithm for compressed sensing.

Input : $x^0 \in \mathbb{R}^N$

Initialize $\epsilon_0 = 1$ and $\eta \in (0, 1)$

while *not converge* **do**

  $w_i^k \leftarrow ((x_i^k)^2 + \epsilon_k^2)^{-1/2} \quad i = 1, \ldots, N.$

  $x^{k+1} \leftarrow \operatorname{argmin} \left\{ \|x\|_{w^k}^2 \mid x \in \Phi^{-1}(y) \right\}.$

  $\epsilon_{k+1} \leftarrow \min \left\{ \epsilon_k, \frac{\eta(1-\gamma)\sigma_K(x^{k+1})}{N} \right\}.$

  If $\epsilon_{k+1} = 0$, stop.

  $k \leftarrow k + 1.$

end

Output: x^{k+1}

In general, the null space parameters K and γ are unknown, however, we show in Section 3.5.2 that the performance of both algorithms is robust with respect to their choice. In particular, by taking $K = N/2$ and $\gamma = .9$, the algorithms DDFG-IRLS and Algorithm 1 perform essentially the same in successfully solving the BP problem.

3.3 Convergence

We follow the proof strategy given in [32] for establishing the convergence and rate of convergence of Algorithm 1. Given $\epsilon > 0$, consider the smoothed ℓ_1 objective

$$J(x, \epsilon) := \sum_{i=1}^{n} \sqrt{x_i^2 + \epsilon^2}.$$

Since $\epsilon > 0$, the function $J(x, \epsilon)$ is strictly convex in x. Hence, the minimizer in x over any convex set is unique if it exists. For each $\epsilon \geq 0$, set

$$x^\epsilon = \operatorname*{argmin}_{x \in \Phi^{-1}(y)} J(x, \epsilon).$$

The smoothing function $J(x, \epsilon)$ is used to measure the progress of the iteratively re-weighted iterates. For this we require that Φ satisfies the null space property NSP.

Assumption 3.3.1. *[31, Section 3]* ***Null Space Property (NSP)*** *A matrix $\Phi \in \mathbb{R}^{\ell \times N}$ satisfies NSP of order K for $\gamma \in (0,1)$ if and only if*

$$\|z_T\|_1 \leq \gamma \|z_{T^c}\|_1 \quad \forall\, z \in Null(\Phi) \tag{3.3.1}$$

and for all index sets $T \subset \{1, \ldots, N\}$ of cardinality not exceeding K.

Observe that since (3.3.1) holds for all index sets $T \subset \{1, \ldots, N\}$ of cardinality K, we must have $K < N/2$. The null space property is intimately connected to the k-sparsity of solutions to the basis pursuit problem BP.

Lemma 3.3.2 (NSP + K-sparsity imply uniqueness)**.** *[32, Lemma 4.3] Assume A2 holds and $\Phi^{-1}(y)$ contains an K-sparse vector x^*. Then x^* is the unique ℓ_1-minimizer in $\Phi^{-1}(y)$ and for all $v \in \Phi^{-1}(y)$,*

$$\|v - x^*\|_1 \leq 2\frac{1+\gamma}{1-\gamma}\sigma_L(v).$$

We now show that the null space property guarantees the boundedness of any sequence generated by Algorithm 1.

Lemma 3.3.3 (Boundness of $\{x^n\}$)**.** *Let Assumption 3.3.1 hold, and suppose $\{x^n\}$ is a sequence generated by Algorithm 1. Then the sequence $\{J(x^n, \epsilon_n)\}$ is non-increasing, $\|x^n\|_1 \leq J(x^0, \epsilon_0)$, for all $n \in \mathbb{N}$, and $\sum_{i=1}^{\infty} \|x^{n+1} - x^n\|_{w^n}^2 < \infty$.*

Proof. By concavity of the square root function $\sqrt{b} + \frac{1}{2\sqrt{b}}(a-b) \geq \sqrt{a}$ for $0 \leq a, b$, and so

$$J(x^{n+1}, \epsilon_n) - J(x^n, \epsilon_n) \leq \frac{1}{2}(\left\|x^{n+1}\right\|_{w^n}^2 - \|x^n\|_{w^n}^2). \tag{3.3.2}$$

By completing the square and rearranging terms, we have

$$\left\|x^{n+1}\right\|_{w^n}^2 - \|x^n\|_{w^n}^2 = -\left\|x^{n+1} - x^n\right\|_{w^n}^2 + 2\left\langle x^{n+1},\, x^{n+1} - x^n\right\rangle_{w^n}. \tag{3.3.3}$$

Since $x^{n+1} = \operatorname{argmin}_{x \in \Phi^{-1}(y)} \|x\|_{w^n}$, we know

$$\left\langle x^{n+1},\, x^{n+1} - x^n \right\rangle_{w^n} = 0. \tag{3.3.4}$$

By combining (3.3.2), (3.3.3) and (3.3.4) and using the fact that $\{\epsilon_n\}$ is non-increasing, we have

$$J(x^{n+1}, \epsilon_{n+1}) - J(x^n, \epsilon_n) \leq J(x^{n+1}, \epsilon_n) - J(x^n, \epsilon_n) \leq -\frac{1}{2} \left\| x^{n+1} - x^n \right\|_{w^n}^2 .$$

Hence $\|x^n\|_1 \leq J(x^n, \epsilon_n) \leq J(x^0, \epsilon_0)$. Moreover, by telescoping we know

$$\sum_{n=1}^{\infty} \left\| x^{n+1} - x^n \right\|_{w^n}^2 \leq 2J(x^0, \epsilon_0) < \infty.$$

□

Our convergence proof also relies on the following lemma.

Lemma 3.3.4. *[32, Lemma 4.2] Let Assumption 3.3.1 hold. Then, for any $z, z' \in \Phi^{-1}(y)$, we have*

$$\|z - z'\|_1 \leq \frac{1-\gamma}{1+\gamma} \left[\|z'\|_1 - \|z\|_1 + 2\sigma_K(z) \right], \tag{3.3.5}$$

where σ_K is defined in (3.2.2).

The main convergence result makes use of the following notation: for $S \subseteq [N] := \{1, 2, 3, ..., N\}$ and $x \in \mathbb{R}^N$, define $x_S \in \mathbb{R}^N$ componentwise by

$$(x_S)_i = \begin{cases} x_i, & i \in S, \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 3.3.5 (Convergence of Algorithm 1)**.** *Let Assumption 3.3.1 hold, and let $y \in \mathbb{R}^m$ and $x_0 \in \mathbb{R}^N$ be given. If $\{x_k\}$ is generated by Algorithm 1 initialized at x_0, then there is an $\bar{x} \in \mathbb{R}^N$ such that $x_k \to \bar{x}$. Moreover, the following hold.*

(1) If $\epsilon := \lim_{n\to\infty} \epsilon_n = 0$, then $\bar{x}$ is K-sparse in which case $\bar{x}$ is the unique ℓ_1 - minimizer.

(2) If there exists a K-sparse $x^* \in \Phi^{-1}(y)$, then $\bar{x} = x^*$ is the unique ℓ_1 - minimizer and $\lim_{n\to\infty} \epsilon_n = 0$.

Proof. Part (1): The proof the part (1) is similar to the proof of [32, Theorem 5.3(i)]. First observe that ϵ is well-defined since the sequence $\{\epsilon_n\}_{n=1}^{\infty}$ is non-increasing. Moreover, by definition, $\sigma_K(x) = 0$ if and only if x is K-sparse. Consequently if for any iteration n_0 we have $\epsilon_{n_0+1} = 0$, then Algorithm 1 terminates at x^{n_0} with x^{n_0} K-sparse, and so part (1) follows from Lemma 3.3.2. Therefore, we assume that the algorithm does not terminate and $0 < \epsilon_n \to 0$. In this case, there must be a subsequence $\mathcal{N} \subset \mathbb{N}$ such that $\sigma_K(x^n) \overset{\mathcal{N}}{\to} 0$. Since Lemma 3.3.3 tells us that the sequnce $\{x^n\}$ is bounded, there is a further subsequence $\mathcal{N}' \subset \mathcal{N}$ and a point $\bar{x} \in \Phi^{-1}(y)$ such that $x^n \overset{\mathcal{N}'}{\to} \bar{x}$ with $\sigma_K(\bar{x}) = 0$. Hence, by Lemma 3.3.2, $\bar{x}$ is the unique K-sparse ℓ_1-minimizer.

Next let $\mathcal{J} \subset \mathbb{N}$ be any subsequence. Again, by Lemma 3.3.3, there is a further subsequence $\mathcal{J}' \subset \mathcal{J}$ and a point x' such that $x^n \overset{\mathcal{J}'}{\to} x'$. Let $i \in \mathcal{N}'$ and $j \in \mathcal{J}'$ be such that $i < j$. Then

$$\begin{aligned}
\left\|x^i - x^j\right\|_1 &\leq \frac{1-\gamma}{1+\gamma}(\left\|x^j\right\|_1 - \left\|x^i\right\|_1 + 2\sigma_K(x^i)) && \text{(by (3.3.5))} \\
&\leq \frac{1-\gamma}{1+\gamma}(J(x^j, \epsilon_j) - J(x^i, \epsilon_i) + N\epsilon_i + 2\sigma_K(x^i)) \\
&\leq \frac{1-\gamma}{1+\gamma}(N\epsilon_i + 2\sigma_K(x^i)). && \text{(by Lemma 3.3.3)}
\end{aligned}$$

Consequently, $\bar{x} = x'$. Hence the entire sequence $\{x^n\}$ must converge to $\bar{x}$ since every subsequence has a further subsequence convergent to $\bar{x}$.

Part (2): First we assume $\epsilon = \inf_n \epsilon_n = \lim_{n\to\infty} \epsilon_n > 0$ and establish a contradiction. By Lemma 3.3.3, every subsequence $\mathcal{N} \subset \mathbb{N}$ has a further subsequence $\mathcal{N}' \subset \mathcal{N}$ such that $x^n \overset{\mathcal{N}'}{\to} \tilde{x}$ for some $\tilde{x} \in \Phi^{-1}(y)$. For any $x \in \Phi^{-1}(y)$ and $i \in \mathcal{N}'$, we have

$$\begin{aligned}
J(x, \epsilon_i) - J(x^i, \epsilon_i) &\geq \left\langle x^i,\, x - x^i \right\rangle_{w^i} && (3.3.6) \\
&= \left\langle x^{i+1},\, x - x^i \right\rangle_{w^i} + \left\langle x^i - x^{i+1},\, x - x^i \right\rangle_{w^i} \\
&\geq \left\langle x^{i+1},\, x - x^i \right\rangle_{w^i} - \left\|x^i - x^{i+1}\right\|_{w^i} \left\|x - x^i\right\|_{w^i}, && (3.3.7)
\end{aligned}$$

where (3.3.6) follows from the convexity of $\sqrt{(\cdot)^2 + \epsilon_i^2}$ and (3.3.7) is the Cauchy-Schwartz inequality. Since $x^{i+1} = \operatorname{argmin}_{x \in \Phi^{-1}(y)} \|x\|_{w^i}^2$, we have $\left\langle x^{i+1},\, x - x^i \right\rangle_{w^i} = 0$. In addition,

since $\epsilon = \inf_n \epsilon_n$, we have $\|x - x^i\|_{w^i} \leq \epsilon^{-1} \|x - x^i\|$. By combining these two statements with (3.3.7), we obtain

$$J(x, \epsilon_i) - J(x^i, \epsilon_i) \geq -\epsilon^{-1} \left\|x^i - x^{i+1}\right\|_{w^i} \left\|x - x^i\right\|.$$

Since, by Lemma 3.3.3, $\|x^i - x^{i+1}\|_{w^i} \to 0$, we find that $J(x, \epsilon) \geq J(\tilde{x}, \epsilon)$. Consequently, $\tilde{x} = x^\epsilon$, that is, every subsequence of $\{x^n\}$ has a further subsequence convergent to x^ϵ which implies that the entire sequence converges to x^ϵ.

Now set $T := \{i | x_i^* \neq 0, 1 \leq i \leq N\}$ so that $|T| \leq K$, and observe that

$$\|x^\epsilon\|_1 \leq J(x^\epsilon, \epsilon) \leq J(x^*, \epsilon) \leq \|x^*\|_1 + N\epsilon. \tag{3.3.8}$$

In addition, we have

$$\begin{aligned}
\|x^\epsilon_{T^c}\|_1 &= \|x^\epsilon\|_1 - \|x^\epsilon_T\|_1 \\
&\leq \|x^*\|_1 + N\epsilon - (\|x^*_T\|_1 - \|x^*_T - x^\epsilon_T\|_1) && \text{(by (3.3.8) and } \Delta \text{ inequality)} \\
&\leq N\epsilon + \|x^*_T - x^\epsilon_T\|_1 && \text{(since } \|x^*\|_1 = \|x^*_T\|_1) \\
&\leq \gamma \|x^\epsilon_{T^c}\| + N\epsilon\,. && \text{(NSP)}
\end{aligned} \tag{3.3.9}$$

Next observe that

$$N\epsilon = \lim_{n\to\infty} N\epsilon_n \leq \lim_{n\to\infty} \eta(1-\gamma)\sigma_K(x^n) = \eta(1-\gamma)\sigma_K(x^\epsilon) \leq \eta(1-\gamma) \|x^\epsilon_{T^c}\|_1\,.$$

Plugging this into (3.3.9) gives

$$\|x^\epsilon_{T^c}\|_1 \leq \gamma \|x^\epsilon_{T^c}\| + \eta(1-\gamma) \|x^\epsilon_{T^c}\|_1\,. \tag{3.3.10}$$

If $\|x^\epsilon_{T^c}\|_1 = 0$, then $x^\epsilon = x^*$ and $\sigma_K(x^\epsilon) = 0$. But then $\lim_n \sigma_K(x^n) = \sigma_K(x^\epsilon) = 0$ which implies that $\epsilon_n \to 0$, a contradiction. Therefore, $\|x^\epsilon_{T^c}\|_1 > 0$. Dividing (3.3.10) by $\|x^\epsilon_{T^c}\|_1$ gives

$$1 \leq \gamma + \eta(1-\gamma) < \gamma + (1-\gamma) = 1 \qquad \text{(since } \eta \in (0,1)\text{)}$$

a contradiction. Therefore, ϵ must equal zero which returns us to Part (1) and completes the proof. □

We now establish the local linear convergence for Algorithm 1. Recall that a sequence $\{z^k\} \subset \mathbb{R}^N$ converges *locally linearly* to $z^* \in \mathbb{R}^N$ if there are constants $\kappa \geq 0$ and $\lambda \in (0,1)$ and an iteration $k_0 \in \mathbb{N}$ such that

$$\left\|z^k - z^*\right\| \leq \kappa \lambda^{k-k_0} \left\|z^{k_0} - z^*\right\| \quad \forall\, k \geq k_0.$$

In [32], the authors refer to linear convergence as *exponential convergence*.

Theorem 3.3.6 (The Local Linear Convergence of Algorithm 1)**.** *Let Assumption 3.3.1 hold, and suppose that $\Phi^{-1}(y)$ contains a K-sparse vector x^*. Set $T := \{i | x_i^* \neq 0, 1 \leq i \leq N\}$ and choose $\rho \in (0,\, 1-\gamma(1+\eta(1-\gamma)))$, where γ is given in A2 and $\eta \in (0,1)$ is initialized in Algorithm 1. Then there is a smallest $n_0 \in \mathbb{N}$ such that*

$$\left\|(x^{n_0} - x^*)_{T^c}\right\|_1 \leq \rho \min_{i \in T} |x_i^*| \ . \tag{3.3.11}$$

Moreover, for all $n \geq n_0$,

$$\left\|(x^{n+1} - x^*)_{T^c}\right\|_1 \leq \mu \left\|(x^n - x^*)_{T^c}\right\|_1 \qquad \text{and} \tag{3.3.12}$$

$$\left\|x^n - x^*\right\|_1 \leq (1+\gamma)\mu^{n-n_0} \left\|x^{n_0} - x^*\right\|_1, \tag{3.3.13}$$

where $\mu := \frac{\gamma(1+\eta(1-\gamma))}{1-\rho} < 1$.

Proof. By Theorem 3.3.5, $x^n \to x^*$ so that for every $\rho \in (0,\, 1-\gamma(1+\eta(1-\gamma)))$ there is a smallest $n_0 \in \mathbb{N}$ such that (3.3.11) holds. Consequently, n_0 exists.

We follow the proof in [32, Theorem 6.1]. We prove (3.3.12) by induction. Let $\hat{n} \geq n_0$ be such that (3.3.11) holds with n_0 replaced by $\hat{n}$. Since $x^{\hat{n}+1} = \operatorname{argmin}_{x \in \Phi^{-1}(y)} \|x\|_{w^{\hat{n}}}^2$, the optimality conditions for this problem tell us that

$$\left\langle x^{\hat{n}+1},\, x^{\hat{n}+1} - x^* \right\rangle_{w^{\hat{n}}} = 0.$$

Consequently,

$$\left\|x^{\hat{n}+1} - x^*\right\|_{w_{\hat{n}}}^2 = -\left\langle x^*,\, x^{\hat{n}+1} - x^* \right\rangle_{w_{\hat{n}}} = -\left\langle (x^*)_T,\, x^{\hat{n}+1} - x^* \right\rangle_{w_{\hat{n}}} \leq \sum_{i \in T} \frac{|x_i^*(x_i^{\hat{n}+1} - x_i^*)|}{\sqrt{(x_i^{\hat{n}})^2 + \epsilon_{\hat{n}}^2}}.$$

Note, for $i \in T$, NSP tells us that

$$|x_i^{\hat{n}} - x_i^*| \leq \left\|(x^{\hat{n}} - x^*)_T\right\|_1 \leq \gamma \left\|(x^{\hat{n}} - x^*)_{T^c}\right\| \leq \rho \min_{i \in T} |x_i^*|,$$

we have

$$\frac{|x_i^*|}{\sqrt{(x_i^{\hat{n}})^2 + \epsilon_{\hat{n}}^2}} \leq \frac{|x_i^*|}{|x_i^{\hat{n}}|} \leq \frac{|x_i^*|}{|x_i^*| - |x_i^{\hat{n}} - x_i^*|} \leq \frac{1}{1-\rho}.$$

Hence

$$\left\|x^{\hat{n}+1} - x^*\right\|_{w_{\hat{n}}}^2 \leq \frac{1}{1-\rho} \left\|(x^{\hat{n}+1} - x^*)_T\right\|_1 \leq \frac{\gamma}{1-\rho} \left\|(x^{\hat{n}+1} - x^*)_{T^c}\right\|_1.$$

Consequently, by Cauchy-Schwartz Inequality,

$$\begin{aligned}
\left\|(x^{\hat{n}+1} - x^*)_{T^c}\right\|_1^2 &= \left(\sum_{i \in T^c} \frac{|x_i^{\hat{n}+1} - x_i^*|}{((x_i^{\hat{n}})^2 + \epsilon_{\hat{n}}^2)^{1/4}} ((x_i^{\hat{n}})^2 + \epsilon_{\hat{n}}^2)^{1/4} \right)^2 \\
&= \left\|(x^{\hat{n}+1} - x^*)_{T^c}\right\|_{w_{\hat{n}}}^2 \left(\sum_{i \in T^c} \sqrt{(x_i^{\hat{n}})^2 + \epsilon_{\hat{n}}^2} \right) \\
&\leq \left\|x^{\hat{n}+1} - x^*\right\|_{w_{\hat{n}}}^2 \left(\sum_{i \in T^c} |x_i^{\hat{n}}| + \epsilon_{\hat{n}} \right) \\
&\leq \frac{\gamma}{1-\rho} \left\|(x^{\hat{n}+1} - x^*)_{T^c}\right\|_1 [\left\|(x^{\hat{n}} - x^*)_{T^c}\right\|_1 + N\epsilon_{\hat{n}}].
\end{aligned}$$

Therefore

$$\begin{aligned}
\left\|(x^{\hat{n}+1} - x^*)_{T^c}\right\|_1 &\leq \frac{\gamma}{1-\rho} [\left\|(x^{\hat{n}} - x^*)_{T^c}\right\|_1 + N\epsilon_{\hat{n}}] \\
&\leq \frac{\gamma}{1-\rho} [\left\|(x^{\hat{n}} - x^*)_{T^c}\right\|_1 + \eta(1-\gamma)\sigma_K(x^{\hat{n}})]. \quad \text{(Step 4 in Algorithm 1)}
\end{aligned}$$

Observe $\sigma_K(x^{\hat{n}}) \leq \left\|(x^{\hat{n}})_{T^c}\right\|_1 = \left\|(x^{\hat{n}} - x^*)_{T^c}\right\|_1$. Hence

$$\left\|(x^{\hat{n}+1} - x^*)_{T^c}\right\|_1 \leq \frac{\gamma(1 + \eta(1-\gamma))}{1-\rho} \left\|(x^{\hat{n}} - x^*)_{T^c}\right\|_1 = \mu \left\|(x^{\hat{n}} - x^*)_{T^c}\right\|_1. \tag{3.3.14}$$

Since n_0 satisfies (3.3.11), this shows that (3.3.12) is satisfied for $\hat{n} = n_0$.

Now assume (3.3.12) holds for $\{n_0, n_0 + 1, ..., n-1\}$. Then (3.3.12) tell us that

$$\left\|(x^n - x^*)_{T^c}\right\|_1 \leq \mu \left\|(x^{n-1} - x^*)_{T^c}\right\|_1 \leq ... \leq \mu^{n-n_0} \left\|(x^{n_0} - x^*)_{T^c}\right\|_1 \leq \rho \min_{i \in T} |x_i^*|, \tag{3.3.15}$$

where the last inequality is by (3.3.11) and $\mu < 1$. In particular, we have (3.3.11) with n_0 replaced by n, and so, by (3.3.14), (3.3.12) is satisfied at n which completes the induction.

Finally, the NSP for Φ tells us that

$$\begin{aligned}\|x^n - x^*\|_1 &\leq (1+\gamma)\,\|(x^n - x^*)_{T^c}\|_1 \\ &\leq (1+\gamma)\mu^{n-n_0}\,\|(x^{n_0} - x^*)_{T^c}\|_1 \leq (1+\gamma)\mu^{n-n_0}\,\|x^{n_0} - x^*\|_1 .\end{aligned}$$

□

3.4 Failure of DDFG-IRLS

We construct an example where the DDFG-IRLS algorithm provably fails for $K - 2\gamma/(1-\gamma) \leq \gamma \leq K$. However, we emphasize that, in general, the failure of this inequality does not imply the failure of the DDFG-IRLS algorithm.

The example is formulated in the context of the ℓ_1 regression problem ℓ_1R discussed in the introduction. It is well-known that BP is equivalent to this ℓ_1 regression problem under the corresponces $\mathrm{rge}\,(A) = \mathrm{Null}(\Phi)$ and $\Phi b = -y$ [22]. In addition, under these correspondences, the NSP for Φ of order K for $\gamma \in (0,1)$ is equivalent to the following condition on the matrix A:

$$\|(Az)_T\|_1 \leq \gamma\,\|(Az)_{T^c}\|_1 \quad \text{for all } z \text{ and all } |T| \leq K. \tag{3.4.1}$$

In terms of the DDFG-IRLS algorithm, when the matrix A has full column rank, then there is a 1-1 correspondence between the iterates of this algorithm and a corresponding IRLS algorithm for solving the ℓ_1R. If we denote the ith row of A by a_i, for given ϵ_0 and x^0, this correspondence is given by

$$x^n = Az^n - b \quad \forall\, n = 0, 1, \ldots,$$

where, for $n = 0, 1, \ldots,$

$$\text{DDFG-IRLS} \begin{cases} x^{n+1} := \min\limits_{x \in \Phi^{-1}(y)} \sum\limits_{i=1}^{N} \dfrac{x_i^2}{\sqrt{(x_i^n)^2 + \epsilon_n^2}} \\ \epsilon_{n+1} := \min\left\{\epsilon_n, \dfrac{r_{K+1}(x^{n+1})}{N}\right\} \end{cases} \tag{3.4.2}$$

$$\ell_1\text{R-IRLS} \begin{cases} z^{n+1} := \min\limits_{z} \sum\limits_{i=1}^{N} \dfrac{(a_i^T z - b_i)^2}{\sqrt{(a_i^T z^n - b_i)^2 + \epsilon_n^2}} \\ \epsilon_{n+1} := \min\left\{\epsilon_n, \dfrac{r_{K+1}(Az^{n+1} - b)}{N}\right\} \end{cases}.$$

Therefore, by Lemma 3.3.2, whenever Φ satisfies the NSP of order K for γ, or equivalently, A satisfies (3.4.1), if there exists z^* for which $Az^* - b$ is K-sparse, then $x^* := Az^* - b$ is the unique solution to BP. If, in addition, A has full column rank, then z^* is the unique solution to ℓ_1R.

We now construct our example. Given $k \geq 1$, set $\tilde{A} := (I_k, ..., I_k)^T \in \mathbb{R}^{(2k^2+k)\times k}$ with $2k+1$ blocks of the identity $k \times k$ matrix I_k. For any $z \in \mathbb{R}^k$ and any $T \subseteq [2k^2 + k]$ with $|T| = k$, let $i_0 \in \{i \,|\, |z_i| \geq |z_j| \;\forall\, 1 \leq j \leq k\}$. Then

$$\left\|(\tilde{A}z)_T\right\|_1 \leq k|z_{i_0}| = \frac{k}{k+1}(k+1)|z_{i_0}| \leq \frac{k}{k+1}\left\|(\tilde{A}z)_{T^c}\right\|_1.$$

Thus, for $K = k$, $\tilde{A}$ satisfies (3.4.1) with $\gamma = \frac{k}{k+1}$, and this value for γ is sharp. We now modify $\tilde{A}$ to obtain an A_γ whose γ is any element of $(\frac{k}{k+1},\ 1)$. To this end, let $\gamma \in (\frac{k}{k+1},\ 1)$ and define $A_\gamma \in \mathbb{R}^{(2k^2+k)\times k}$ so that $A_\gamma(ik+1, 1) := \frac{k+1}{k}\gamma$ for all $0 \leq i \leq k-1$, while all other components of A_γ coincide with those of $\tilde{A}$. That is, we only replace the $(1,1)$ entry in each of the first k identity matrices I_k of $\tilde{A}$ by $\frac{k+1}{k}\gamma$. By applying the same argument to A_γ as above for $\tilde{A}$, we find that A_γ satisfies (3.4.1) with $K = k$ and $\gamma = \frac{k}{k+1}$, and this γ is also sharp.

Next choose $z^* \in \mathbb{R}^k_{++}$. Given $\delta \in \mathbb{R}$, set $b := A_\gamma z^* + \delta\tilde{e}$, where $\tilde{e} := \sum_{j=0}^{k-1} e_{(jk+1)}$ with each $e_{(jk+1)}$ the $(jk+1)$th standard unit coordinate vector. Observe that $x^* := A_\gamma z^* - b$ is k-sparse and A_γ has full column rank. Hence, by our previous discussion, Lemma 3.3.2 implies that z^* is the unique solution to ℓ_1R for this choice of A and b.

Our goal is to show that there is an initialization for the ℓ_1R-IRLS algorithm in (3.4.2) such that the generated sequence $\{z^n\}$ satisfies $z^n \nrightarrow z^*$, and hence, the corresponding DDFG-IRLS iterates $x^n := A_\gamma z^n - b$ do not converge to the unique solution $x^* := A_\gamma z^* - b$ to BP.

Theorem 3.4.1. *Let $z^* \in \mathbb{R}^k$, $\delta \in (0, k(2k+1)]$, and $\gamma \in [\nu, 1)$, where*

$$\nu := \sqrt{\frac{1 + \frac{1}{4k^2(2k+1)^2}}{1 + \frac{1}{k^2(2k+1)^2}}} = \sqrt{\frac{4k^2(2k+1)^2 + 1}{4k^2(2k+1)^2 + 4}}.$$

For these values of z^*, γ and δ, let A_γ and b be as given above and consider the problem $\ell_1 R$ having unique solution z^*. Define

$$\alpha := \gamma \frac{k+1}{k} \quad \text{and} \quad \xi := \gamma \sqrt{1 + \frac{1}{k^2(2k+1)^2}}.$$

Then

$$\alpha > 1\,, \quad \xi \geq \sqrt{1 + (4k^2(2k+1)^2)^{-1}} > 1 \quad \text{and} \quad \gamma/(k(2k+1)\sqrt{\xi^2 - 1}) > 1. \tag{3.4.3}$$

Initialize $\epsilon_0 := 1$ and $z^0 \in \mathbb{R}^k_{++}$ componentwise by

$$\begin{aligned} &z_1^0 \in \left(z_1^* + \frac{\delta}{\alpha + \gamma/(k(2k+1)\sqrt{\xi^2 - 1})},\ z_1^* + \frac{\delta}{\alpha + 1} \right) \qquad \text{and} \\ &z_i^0 := z_i^*,\ i = 2, \ldots, k\,. \end{aligned} \tag{3.4.4}$$

If $\{z^n\}$ is the sequence generated by the $\ell_1 R$-IRLS algorithm in (3.4.2) with this initialization, then $z^n \nrightarrow z^*$.

Proof. We first prove the inequalities in (3.4.3). The first inequality follows since

$$\begin{aligned} \alpha > 1 &\Longleftarrow \nu^2 < \left(\frac{k}{k+1} \right)^2 \\ &\Longleftrightarrow (k+1)^2 \left(1 + \frac{1}{4k^2(2k+1)^2} \right) > k^2 \left(1 + \frac{1}{k^2(2k+1)^2} \right) \\ &\Longleftrightarrow 2k + 1 + \frac{(k+1)^2}{4k^2(2k+1)^2} > \frac{1}{(2k+1)^2}. \end{aligned}$$

The second inequality in (3.4.3) follows directly from the fact that $\gamma \geq \nu$. The third inequality in (3.4.3) follows since

$$\gamma / \left(k(2k+1)\sqrt{\xi^2 - 1} \right) > 1 \iff \xi^2 < 1 + \frac{\gamma^2}{k^2(2k+1)^2}$$
$$\iff \gamma^2 < 1.$$

Note that the third inequality in (3.4.3) implies that

$$\delta[\alpha + \gamma/(k(2k+1)\sqrt{\xi^2 - 1})]^{-1} < \delta(\alpha + 1)^{-1}$$

so that x_1^0 is well defined.

We establish the result by showing that $z_1^n \nrightarrow z_1^*$. Observe that

$$b_{k^2+1} = (A_\gamma z^*)_{k^2+1} = z_1^* \quad \text{and} \quad b_1 = \alpha z_1^* + \delta = \alpha b_{k^2+1} + \delta. \tag{3.4.5}$$

By the ℓ_1R-IRLS algorithm, z^{n+1} solves the least-squares problem

$$\min_z \ \frac{k(\alpha z_1 - b_1)^2}{\sqrt{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}} + \frac{(k+1)(z_1 - b_{k^2+1})^2}{\sqrt{(z_i^n - b_{k^2+1})^2 + \epsilon_n^2}} + (2k+1)\sum_{i=2}^{k} \frac{(z_i - b_i)^2}{\sqrt{(z_i^n - b_i)^2 + \epsilon_n^2}}.$$

Due to the separability of the objective in the variables $z_i,\ i = 2, \ldots, k$, we have $z_i^n = b_i = z_i^*,\ i = 2, \ldots, k$, for $n \geq 1$. The optimality conditions for each subproblem tells us that

$$z_1^{n+1} = \frac{\frac{\alpha b_1 k}{\sqrt{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}} + \frac{(k+1) b_{k^2+1}}{\sqrt{(z_1^n - b_{k^2+1})^2 + \epsilon_n^2}}}{\frac{k\alpha^2}{\sqrt{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}} + \frac{(k+1)}{\sqrt{(z_1^n - b_{k^2+1})^2 + \epsilon_n^2}}}. \tag{3.4.6}$$

By (3.4.6), we have

$$\begin{aligned} z_1^{n+1} - b_{k^2+1} &= \frac{\frac{\alpha k (b_1 - \alpha b_{k^2+1})}{\sqrt{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}}}{\frac{k\alpha^2}{\sqrt{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}} + \frac{(k+1)}{\sqrt{(z_1^n - b_{k^2+1})^2 + \epsilon_n^2}}} \\ &= \frac{\frac{\alpha k \delta}{\sqrt{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}}}{\frac{k\alpha^2}{\sqrt{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}} + \frac{(k+1)}{\sqrt{(z_1^n - b_{k^2+1})^2 + \epsilon_n^2}}} \geq 0 \end{aligned} \tag{3.4.7}$$

and

$$\begin{aligned} b_1 - \alpha z_1^{n+1} &= \frac{\frac{(k+1)(b_1 - \alpha b_{k^2+1})}{\sqrt{(z_1^n - b_{k^2+1})^2 + \epsilon_n^2}}}{\frac{k\alpha^2}{\sqrt{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}} + \frac{(k+1)}{\sqrt{(z_1^n - b_{k^2+1})^2 + \epsilon_n^2}}} \\ &= \frac{\frac{(k+1)\delta}{\sqrt{(z_1^n - b_{k^2+1})^2 + \epsilon_n^2}}}{\frac{k\alpha^2}{\sqrt{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}} + \frac{(k+1)}{\sqrt{(z_1^n - b_{k^2+1})^2 + \epsilon_n^2}}} \geq 0 \ . \end{aligned} \tag{3.4.8}$$

Hence,

$$z_1^{n+1} - b_{k^2+1} \geq 0,\ b_1 - \alpha z_1^{n+1} \geq 0, \text{ and } s_{n+1} = \gamma \sqrt{\frac{(z_1^n - b_{k^2+1})^2 + \epsilon_n^2}{(b_1 - \alpha z_1^n)^2 + \epsilon_n^2}}, \quad \forall\, n \geq 0, \tag{3.4.9}$$

where $s_{n+1} := (z_1^{n+1} - b_{k^2+1})/(b_1 - \alpha z_1^{n+1})$.

If we let $\varepsilon_n := z_1^n - b_{k^2+1}$, then $s_n = \varepsilon_n/(\delta - \alpha\varepsilon_n)$ by (3.4.5). For $n = 0$, (3.4.4) tells us that

$$s_0 = \frac{\varepsilon_0}{\delta - \alpha\varepsilon_0} = \frac{1}{(\delta/\varepsilon_0) - \alpha} \in \left(\frac{k(2k+1)\sqrt{\xi^2 - 1}}{\gamma}, 1 \right). \tag{3.4.10}$$

We now show by induction that

$$s_n > k(2k+1)\sqrt{\xi^2 - 1} \quad \text{and} \quad \epsilon_n = \varepsilon_n/(k(2k+1)) \qquad \forall\ n \geq 1\ . \tag{3.4.11}$$

First consider $n = 1$. Since $\epsilon_0 = 1$, the definition of ε_0 and s_0 in conjunction with (3.4.5) and (3.4.9) tell us that $s_1 = \gamma\sqrt{\frac{\varepsilon_0^2+1}{(\delta - \alpha\varepsilon_0)^2+1}}$ and so, by (3.4.10)

$$s_1 = \gamma \sqrt{\frac{\varepsilon_0^2 + 1}{(\delta - \alpha\varepsilon_0)^2 + 1}} \leq \gamma < 1. \tag{3.4.12}$$

Observe that

$$(A_\gamma z^n - b)_i = \begin{cases} \alpha z_1^n - b_1, & \text{if } i \in \{jk+1 \mid j \in \{0, \ldots, k-1\}\}\,, \\ z_1^n - b_{k^2+1}, & \text{if } i \in \{jk+1 \mid j \in \{k, \ldots, 2k\}\}\,, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, since $(z_1^1 - b_{k^2+1})/(b_1 - \alpha z_1^1) = s_1 < 1$, the $(k+1)$th largest magnitude of the entries of $A_\gamma z^1 - b$ is $|z_1^1 - b_{k^2+1}|$ with $|z_1^1 - b_{k^2+1}| = z_1^1 - b_{k^2+1}$ by (3.4.9). Thus $\epsilon_1 = \min\left\{\epsilon_0, \frac{z_1^1 - b_{k^2+1}}{k(2k+1)}\right\}$.

The given definitions and the inequality $s_1 < 1$, yield

$$z_1^1 - b_{k^2+1} = \varepsilon_1 = \frac{\delta s_1}{\alpha s_1 + 1} = \frac{\delta}{\alpha + (1/s_1)} \leq \frac{\delta}{\alpha + 1} \leq k(2k+1).$$

Therefore, $\epsilon_1 = \frac{\varepsilon_1}{k(2k+1)}$, since $\epsilon_0 = 1$, which proves the second part of (3.4.11) for $n = 1$. To obtain the first part of (3.4.11) for $n = 1$, observe that

$$\begin{aligned}
\frac{s_1^2}{k^2(2k+1)^2} + 1 &= \frac{\gamma^2}{k^2(2k+1)^2} \frac{\varepsilon_0^2 + 1}{(\delta - \alpha\varepsilon_0)^2 + 1} + 1 && \text{(by (3.4.12))} \\
&\geq \frac{\gamma^2}{k^2(2k+1)^2} \frac{\varepsilon_0^2}{(\delta - \alpha\varepsilon_0)^2} + 1 && \text{(since } \varepsilon_0 \leq \delta - \alpha\varepsilon_0 \text{ by (3.4.10))} \\
&> \xi^2 \,. && \text{(by lower bound in (3.4.10))}
\end{aligned} \tag{3.4.13}$$

Thus, $s_1 > k(2k+1)\sqrt{\xi^2 - 1}$.

Assume $s_n > k(2k+1)\sqrt{\xi^2 - 1}$ and $\epsilon_n = \frac{\varepsilon_n}{k(2k+1)}$. Plugging $\epsilon_n = \frac{\varepsilon_n}{k(2k+1)}$ into (3.4.9) gives

$$\begin{aligned}
s_{n+1} &= \gamma \sqrt{\frac{\varepsilon_n^2 + \frac{\varepsilon_n^2}{k^2(2k+1)^2}}{(b_1 - \alpha z_1^n)^2 + \frac{\varepsilon_n^2}{k^2(2k+1)^2}}} \\
&= \gamma \sqrt{1 + \frac{1}{k^2(2k+1)^2}} \sqrt{\frac{\varepsilon_n^2}{(b_1 - \alpha z_1^n)^2 + \frac{\varepsilon_n^2}{k^2(2k+1)^2}}} \\
&= \xi \frac{s_n}{\sqrt{1 + \frac{s_n^2}{k^2(2k+1)^2}}}.
\end{aligned} \tag{3.4.14}$$

Since the function $f(x) := \frac{x}{\sqrt{1 + x^2(k^2(2k+1)^2)^{-1}}}$ is increasing on $(0, \infty)$, we know

$$s_{n+1} = \xi f(s_n) \geq \xi f\left(k(2k+1)\sqrt{\xi^2 - 1}\right) = k(2k+1)\sqrt{\xi^2 - 1},$$

which established the first part of (3.4.11) for $n + 1$. To establish the second part, observe that (3.4.14) and the induction hypothesis gives

$$s_{n+1} = \xi \frac{s_n}{\sqrt{1 + s_n^2(k^2(2k+1)^2)^{-1}}} \leq \xi \frac{s_n}{\sqrt{1 + \xi^2 - 1}} = s_n,$$

Thus far, we have shown that $s_{n+1} \geq k(2k+1)\sqrt{\xi^2 - 1}$ and $s_{n+1} \leq s_n$. By combining these inequalities with the fact that $s_n = \frac{\varepsilon_n}{\delta_n - \alpha\varepsilon_n}$ for each $n \geq 1$, we have $\varepsilon_{n+1} \leq \varepsilon_n$ for all $n \geq$

1. Therefore, by the induction hypothesis, $\epsilon_{n+1} = \min\{\epsilon_n, \frac{\varepsilon_{n+1}}{k(2k+1)}\} = \min\{\frac{\varepsilon_n}{k(2k+1)}, \frac{\varepsilon_{n+1}}{k(2k+1)}\} = \frac{\varepsilon_{n+1}}{k(2k+1)}$. This concludes the proof of (3.4.11).

Observe that our induction proof also shows that $\{s_n\}$ is a non-increasing sequence bounded below by $k(2k+1)\sqrt{\xi^2-1}$. Therefore, there is an $s^* \geq k(2k+1)\sqrt{\xi^2-1} > 0$ such that $s_n \downarrow s^*$. In particular, by taking the limit in (3.4.14), we have $s^* = \xi s^*/\sqrt{1+\frac{(s^*)^2}{k^2(2k+1)^2}}$, or equivalently, $s^* = k(2k+1)\sqrt{\xi^2-1}$. The induction showed that $s_n = \varepsilon_n(\delta - \alpha\varepsilon_n)$ and so $\varepsilon_n = (\delta s_n)/(1+\alpha s_n)$ which tells us that

$$\begin{aligned} z_1^n - z_1^* = z_1^n - b_{k^2+1} = \varepsilon_n = (\delta s_n)/(1+\alpha s_n) \\ \to (\delta s^*)/(1+\alpha s^*) = \frac{k(2k+1)\sqrt{\xi^2-1}}{1+\alpha k(2k+1)\sqrt{\xi^2-1}} > 0. \end{aligned}$$

Consequently, $z_1^n \not\to z_1^*$, and we have arrive at the desired result. □

3.5 Numerical Examples

3.5.1 Failure of the DDFG-IRLS Algorithm

We present three numerical experiments illustrating the failure of the DDFG-IRLS algorithm for small perturbations of the example given in Theorem 3.4.1. Experiment 1 (see Figure 3.1) simply illustrates the content of Theorem 3.4.1 for $k = 5$, $\gamma = \sqrt{(4k^2(2k+1)^2+1)/(4k^2(2k+1)^2+4)} = 0.999876$, $\delta = k(2k+1) = 55$. The true solution of problem ℓ_1R, z^*, is sampled from $N(0, I_k)$. In both algorithms, x^0 is initialized as $x^0 := A_\gamma z^0 - b$ where z^0 satisfies (3.4.4), i.e., $z_1^0 = z_1^* + (\delta/(\alpha + \gamma/(k(2k+1)\sqrt{\xi^2-1})) + \delta/(\alpha+1))/2$. For Algorithm 1, $\eta = 0.9$.

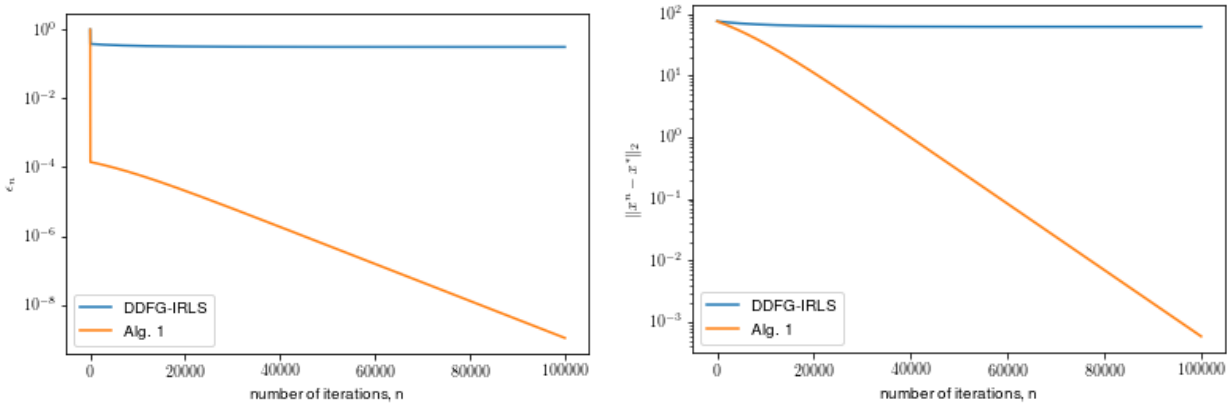

Figure 3.1: Experiment 1: The performance of DDFG-IRLS versus Algorithm 1 for the set-up in Theorem 3.4.1. The left figure is ϵ_n versus the number of iterations n. The right figure is $\|x^n - x^*\|_2$ versus the number of iterations n.

In experiment 2 (see Figure 3.2), we examine the sensitivity of the success/failure of the DDFG-IRLS algorithm to the selection of the parameter γ near the critical value $\gamma_0 := \sqrt{(4k^2(2k+1)^2+1)/(4k^2(2k+1)^2+4)} \approx 1 - 10^{-3.9}$. Again, we let $k = 5$. To illustrate the effect of the selection of γ, we run the DDFG-IRLS algorithm for $\gamma \in \{1-10^{-1}, 1-10^{-2}, 1-10^{-3}, 1-10^{-3.3}, 1-10^{-3.6}, 1-10^{-\gamma_0}, 1-10^{-4}, 1-10^{-5}\}$. Here, 20 instances of the random variable $N(0, 100 \cdot I_5)$ are chosen for the starting point z_0. All other parameters are the same as those of experiment 1. The iterations are terminated when either $\|x^n - x^*\| \leq 10^{-3}$ or the number of iterations exceeds 10^5. In the range $1 - 10^{-3.6} \leq \gamma < \gamma_0$, all the experiments fail to achieve the termination criteria $\|x^n - x^*\|_2 \leq 10^{-3}$. This illustrates the extremely slow rate of convergence of the DDFG-IRLS algorithm when the critical value γ_0 is approached from below.

In experiment 3 (see Figure 3.2), we examine the robustness of the success/failure of the DDFG-IRLS algorithm for small perturbations of the example given in Theorem 3.4.1 obtained by perturbing the matrix A_γ. Again, we let $k = 5$ and $\delta = k(2k+1) = 55$ and

use DDFG-IRLS to solve perturbed versions of our basic example with $A_{\gamma,\sigma} = A_\gamma + \sigma\mathcal{R}$, where $\mathcal{R} \in \mathbb{R}^{k(2k+1)\times k}$ is a random matrix with i.i.d. $N(0,1)$ entries and $b_\sigma := A_{\gamma,\sigma}z^* + \delta\tilde{e}$, where $\tilde{e} := \sum_{j=0}^{k-1} e_{(jk+1)}$ with each $e_{(jk+1)}$ the $(jk+1)$th standard unit coordinate vector. As in experiment 2, the entries of vector z^* are realizations of i.i.d. $N(0,1)$ random variables. For each $\sigma \in [10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}]$, construct 50 problems with the entries of $\mathcal{R}$ i.i.d. $N(0,1)$. The DDFG-IRLS algorithm is run on all 50 problems with each run of the algorithm initialized at a z^0 with components selected i.i.d. $N(0,100)$. The algorithm is terminated when either $\left\|x^k - x^*\right\|_2 < 10^{-3}$ or the number of iterations exceed 10^5. The results are presented on the right hand side of Figure 3.2. Each point with coordinates (x,y) represents the experiment with $\sigma = 10^{-y}$ terminated after x iterations. When $\sigma = 10^{-4}$, the DDFG-IRLS algorithm fails to recover the true x^* within 10^5 steps for all the 50 problems. In other words, the failure of DDFG-IRLS is robust to a small random normal perturbation of matrix A_γ and when it does succeed for slightly large perturbations of A_γ the convergence is still quite slow.

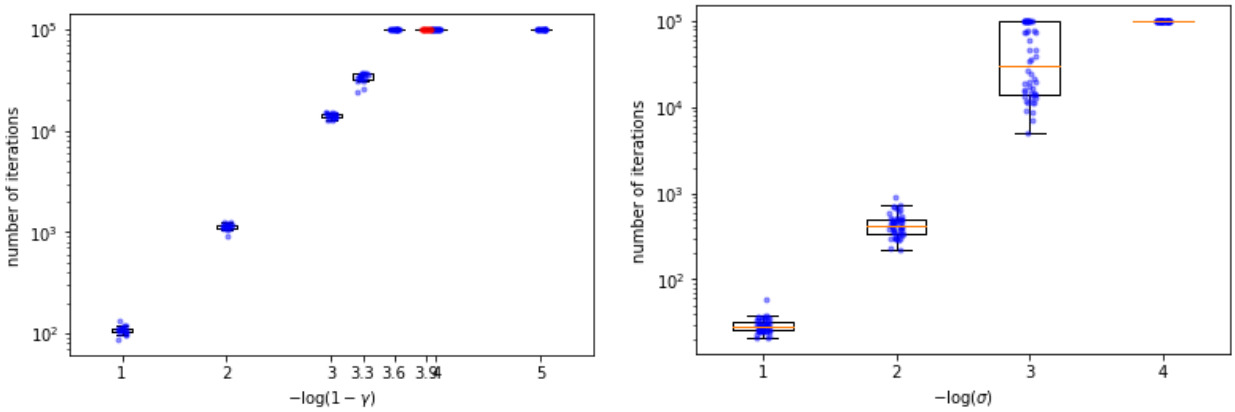

Figure 3.2: Experiment 2 is presented in the left figure. The red points represents the convergence result for $\gamma = \sqrt{(4k^2(2k+1)^2+1)/(4k^2(2k+1)^2+4)} = 1-10^{-3.9}$. Experiment 3 is presented in the right figure. In each experiment, every point is a single trial with the corresponding parameter (γ and σ respectively).

3.5.2 Comparison of DDFG-IRLS and Algorithm 1

In practice the DDFG-IRLS algorithm and Algorithm 1 have nearly identical performance on randomly generated problems. We illustrate this with two additional numerical experiments.

In experiment 4, the entries of $\Phi \in \mathbb{R}^{300\times 500}$ are chosen to be i.i.d. $N(0,1)$ with the solution $x_* \in \mathbb{R}^{500}$ chosen so that the first 100 entries are independent samples from $N(0,1)$ and the remaining components are taken to be 0. Set $y := \Phi x_*$. In practice, the NSP parameters K and γ are not known even though they appear explicitly in the updating policy for the smoothing parameter ϵ_k. All that is known is that if the NSP holds, then $K < N/2$ and $\gamma \in (0,1)$. In this regard, it may be that the DDFG-IRLS algorithm has an edge over Algorithm 1 since the performance of Algorithm 1 may be sensitive to the choice of γ. Consequently, in this experiment, we examine the robustness of the performance of both algorithms to an ad hoc choice of the NSP parameters K and γ. For each $K \in \{99, 100, 150, 200, 250, 300\}$ and $\gamma \in \{0.1, 0.5, 0.9\}$, we run Algorithm 1 one hundred times

with a random initialization $x_0 \sim N(0, 100 \cdot I_5)$ on each run. For each of these values of K, we also run the DDFG-IRLS algorithm for 100 times with same random initializations $x_0 \sim N(0, 100 \cdot I_5)$. The results are presented in Figure 3.3. The plot tells us that the success of both algorithms is robust with respect to the choice of K. When K is strictly smaller than the true number of the nonzero entries in the solution, both algorithms fail regardless of the choice of γ. On the other hand, if we take $K = 250 = N/2$ or $K = 300$, both algorithms succeed. In addition, the two algorithms have nearly identical performance regardless of the choice of K when γ is chosen to be 0.9. Overall, a degradation in the performance of Algorithm 1 for the smaller values of γ only occurs when K is poorly chosen. In practice, we recommend choosing K be a half of the columns of the measurement matrix Φ and set $\gamma \approx 0.9$. In this case, our experiment indicates that the performance of the two algorithms is essentially identical.

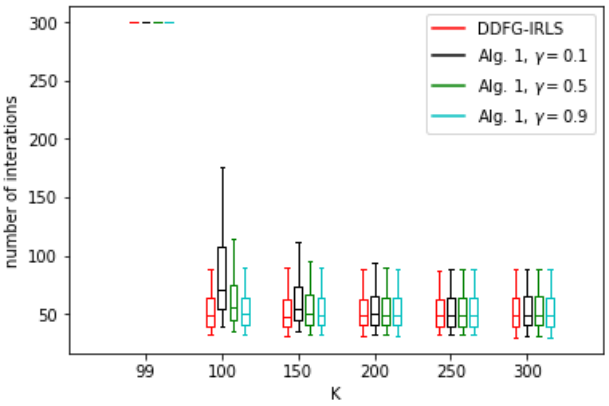

Figure 3.3: Experiment 4: Each box represent 100 times runs of the algorithm.

In the final numerical experiment 5, we briefly examine the efficiency of the DDFG-IRLS algorithm and Algorithm 1 in solving problems with randomly generated data. In this experiment we use the fixed parameter setting $(K, \gamma, \eta) = (N/2, 0.9, 0.9)$ with $(N, m) = (500, 300)$. In all of these experiments, the entries of Φ are independent samples from $N(0, 1)$. In all experiments, the first k entries of x^* are i.i.d. sampled from $N(0, 100)$ with remaining

entries set to zero. In the top row of Figure 3.4, $k = 100$ while in the bottom left $k = 120$ and in the bottom right $k = 50$. The experiment is repeated 50 times for each algorithm. On the top-left side of Figure 3.4 we graph the percentage of problems solved versus the number of iterations up to 120 iterations. On the right side of Figure 3.4 we graph the average error $(1/50)\sum_{i=1}^{50} \text{error}_i^k$ where error_i^k is the value of $\left\|x^k - x^*\right\|$ in the i$^{\text{th}}$ trial. On the bottom left and right of Figure 3.4, we again graph the percentage of problems solved versus the number of iterations up to 500 and 25 iterations, respectively.

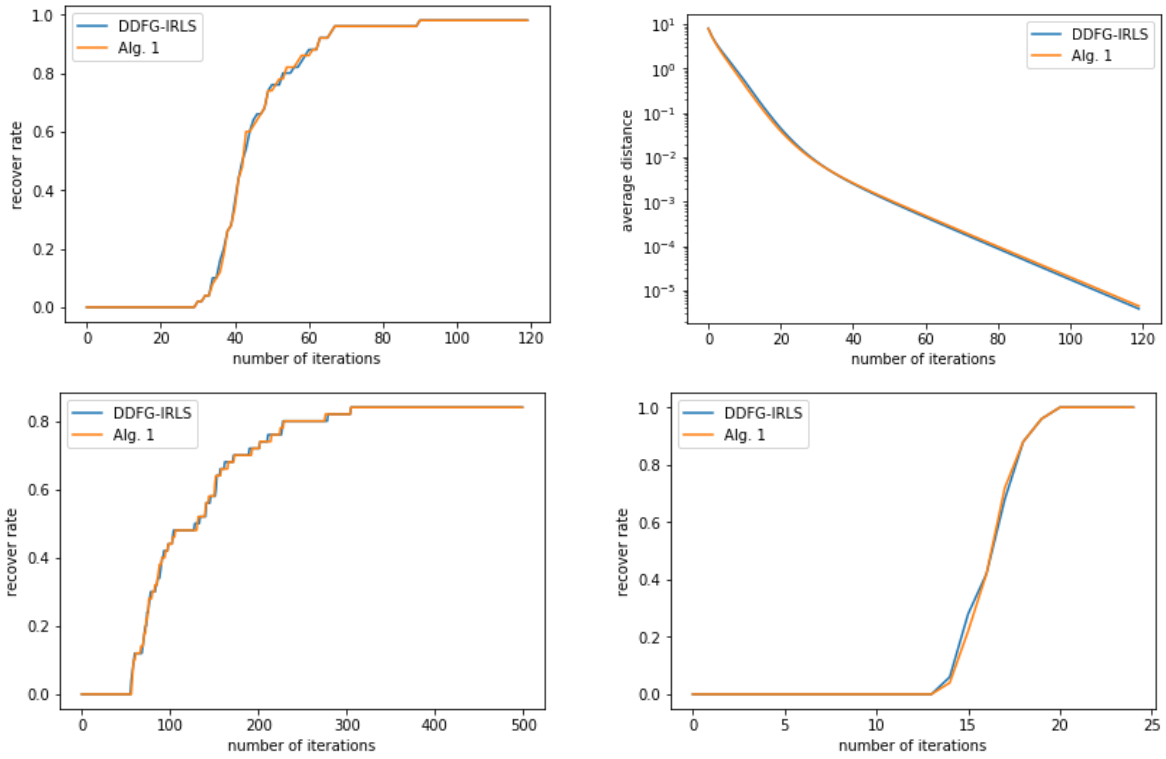

Figure 3.4:

First observe that the DDFG-IRLS and Algorithm 1 perform essentially the same in these random experiments. Next note that the number of iterations required depends on the sparsity of the solution with the iteration count decreasing with the sparsity k. This indicates that these algorithms are most useful when it is known that the sparsity is not too

great. Finally, the graph in the upper right of Figure 3.4 demonstrates the linear rate of convergence of these methods.

3.6 Discussion

In this contribution we provide a concrete example where the DDFG-IRLS fails when $k = K$, and provide a remedy by changing the updating strategy for the smoothing parameter ϵ_n. This remedy increases the range of values for both k and γ for which the algorithm provably converges with a local linear rate to the largest possible intervals $[1, K]$ and $(0, 1)$ for k and γ, respectively. We have also shown through our numerical experiments that on randomly generated problems both algorithms are robust to the choice of K and γ and that their performance is essentially identical. Therefore, if one is concerned about the possible failure DDFG-IRLS, then Algorithm 1 should be considered with recommended parameter choices $(K, \eta, \gamma) = (N/2, 0.9.0.9)$, or equivalently, $0.05 \leq \eta(1 - \gamma) \leq 0.09$ since knowledge of the product $\eta(1 - \gamma)$ is all that is required for implementation.

Chapter 4

ITERATIVELY RE-WEIGHTED LEAST SQUARES ALGORITHM FOR ROBUST PHASE RETRIEVAL

4.1 Introduction

Consider the measurement matrix $A := (a_1, a_2, ..., a_m)^T \in \mathbb{R}^{m\times n}$, with each measurement vector $a_i \in \mathbb{R}^n$, and the observation vector $b \in \mathbb{R}^m$. The *phase retrieval* problem is to recover the vector $x^* \in \mathbb{R}^n$, given the phase of each measurement, that is, $|\langle a_i, x^*\rangle| = b_i$ for $i = 1, 2, ..., m$.

We consider two cases of this problem. The first is the noiseless case, namely for each $i = 1, ..., m$, $|\langle a_i, x^*\rangle| = b_i$, while in the second case we allow sparse corrupted measurements. We know in advance that a constant fraction of the measurements are corrupted, but do not know which these are. Denote by T the indices of the corrupted measurements with $|T| \leq sm$ for some constant $s \in (0, 1)$. We hope to recover x^* given $|\langle a_i, x^*\rangle| = b_i + u_i$, where for the noise vector $u \in \mathbb{R}^m$, $u_i = 0$ for $i \in T$ and $u_i \neq 0$ otherwise. Our recovery should not depend on the location of the corrupted components of u in $[m]$.

Recently a two step procedure has been applied to solve the phase retrieval problem. First a spectral method is used to get a constant error initial estimate of x^*. Second, iterative algorithms are used to refine the estimate. Here we focus on the second step. In particular we apply the Iteratively Re-weighted Least Squares Algorithm (IRLS) to the problem

$$\min_x J(x) := \||Ax| - b\|_1 = \sum_{i=1}^{m} \text{dist}(\langle a_i, x\rangle \,|\{\pm b_i\}). \tag{4.1.1}$$

Throughout the paper we always assume that the measurements a_i are taken independently and $a_i \sim N(0, I_{n\times n})$ for each $i = 1, 2, ..., m$, as formalized in Assumption 4.1.1.

Assumption 4.1.1. *The measurements a_i are independent and $a_i \sim N(0, I_{n\times n})$. Equivalently the entries of $A \in \mathbb{R}^{m\times n}$ are i.i.d. $N(0,1)$.*

4.2 Notation

Let $\|\cdot\|_F$ denote the Frobenius norm of a matrix and let $\|\cdot\|$ and $\|\cdot\|_1$ be the Euclidean norm (ℓ_2 norm) and the ℓ_1 norm of a vector, respectively. Let $\|x\|_0 := \#\{i \mid x_i \neq 0\}$. For $A \in \mathbb{R}^{m\times n}$, a_i represents the ith row vector of A. For $b \in \mathbb{R}^m_+$, $P_i : \mathbb{R} \to \mathbb{R}$ represents the projection to $\{\pm b_i\}$. Let $P : \mathbb{R}^m \to \mathbb{R}^m$ denote the map $(P(z))_i := P_i(z_i)$ for $z \in \mathbb{R}^m$. Let w_i^z denote $((\langle a_i,\, z\rangle - P_i \langle a_i,\, z\rangle)^2 + \gamma^2)^{-1/2}$. Denote the diagonal matrix $\operatorname{diag}\{w_1^z, ..., w_m^z\}$ by W^z. For $a, b \in \mathbb{R}^m$, let $\langle a,\, b\rangle_{W^z}$ denote the weighted inner product $\sum_{i=1}^m w_i^z a_i b_i$. Denote the norm induced by the inner product $\langle \cdot,\, \cdot\rangle_{W^z}$ by $\|\cdot\|_{W^z}$. We use W^n and w_i^n as short for W^{x^n} and $w_i^{x^n}$ where x^n is the n^{th} iterates of the algorithms.

4.3 Algorithms

We first discuss algorithms for the exact case, i.e. we assume that there exists a vector $x^* \in \mathbb{R}^n$ such that $|Ax^*| = b$. For $x, y \in \mathbb{R}^n$ and $0 < p \leq 2$, let

$$G_p(x, y, \gamma) := \sum_{i=1}^m \frac{(\langle a_i,\, x\rangle - P_i \langle a_i,\, y\rangle)^2}{((\langle a_i,\, y\rangle - P_i \langle a_i,\, y\rangle)^2 + \gamma^2)^{1-p/2}}.$$

The Iteratively Re-weighted Least Squares Algorithms (IRLS) for general $p \in (0, 2]$ uses the update rule $x_{k+1} := \min_x G_p(x, x^k, \gamma_k)$ for an iterate-specific γ_k at step k. The method is

summarized in Algorithm 2.

Algorithm 2: IRLS algorithm for $0 < p \leq 2$.

Input : $x^0 \in \mathbb{R}^n$ and $p \in (0, 2]$.

while *not converge* **do**
  $x^{k+1} \leftarrow G_p(x, x^k, \gamma_k)$
  Update γ_{k+1}
  $k \leftarrow k + 1$
end

Output: x^k

Theoretically, it is unclear what the initialization and the update strategy for γ_k are for Algorithm 2 for general $p \in (0, 2]$. Furthermore, there are no theoretical guarantees for the convergence of Algorithm 2. However, when $p = 1$ and $p = 2$, we have analytical results for this algorithm. When $p = 2$, since the weights of the least square sub-problem equal 1 regardless of γ, Algorithm 2 reduces to the Gerchberg-Saxton Algorithm, summarized in Algorithm 3.

Algorithm 3: IRLS algorithm for $p = 2$ (Gerchberg-Saxton Algorithm)

Input : x^0 such that $\|x^0 - x^*\| \leq \frac{1}{50} \|x^*\|$.

while *not converge* **do**
  $x^{k+1} \leftarrow \operatorname{argmin}_x \left\|Ax - P(Ax^k)\right\|^2$
  $k \leftarrow k + 1$
end

Output: x^k

In the complex case, it is shown in [75] that under a weaker initialization ($\|x^0 - x^*\| \leq \frac{1}{10} \|x^*\|$), with high probability, the iterates of Algorithm 3 converge to x^* locally linearly. In Theorem 4.5.1, we provide a new proof of the locally linear convergence rate of Algorithm 3 in the real case. We also show that the problem we solve in this case is

$$\min_x \| |Ax| - b \|^2 .$$

When $p = 1$, For $x, y \in \mathbb{R}^n$, let

$$G(x, y, \gamma) := \sum_{i=1}^{m} \frac{(\langle a_i,\, x\rangle - P_i \langle a_i,\, y\rangle)^2}{\sqrt{(\langle a_i,\, y\rangle - P_i \langle a_i,\, y\rangle)^2 + \gamma^2}} = \|Ax - P(Ay)\|_{W^y}^2 .$$

We discuss two methods about the updating of γ_k when $p = 1$. The first method is to keep $\gamma_k = \gamma$ fixed throughout all the iterates.

Algorithm 4: IRLS algorithm for $p = 1$ with fixed γ

Input : x^0 such that $\|x^0 - x^*\| \leq \frac{1}{56} \|x^*\|$.
$\gamma = \left(\frac{\|b\|^2}{m}\right)^{\frac{1}{2}}$.

1 **while** *not converge* **do**
2 $\quad x^{k+1} \leftarrow \operatorname{argmin}_x G(x, x^k, \gamma)$
3 $\quad k \leftarrow k + 1$
4 **end**

Output: x^k

Since $|Ax^*| = b$, we see that $x^* \in \operatorname{argmin}_x J(x, \gamma)$. We show in Theorem 4.6.3 that Algorithm 9 solves the problem

$$\min_x J(x, \gamma) := \sum_{i=1}^{m} \sqrt{\langle a_i,\, x\rangle^2 + \gamma^2}$$

with high probability, and x^k converges to x^* locally linearly, i.e. $\left\|x^k - x^*\right\| \leq c\phi^k \|x^*\|$ for some $\phi \in (0, 1)$ and $c > 0$.

The second method is to update γ_k based to the iterates $\{x^k\}_{k\geq 0}$, detailed in Algorithm 5.

Algorithm 5: IRLS algorithm with reducing γ_k for noiseless case

Input : x^0 such that $\|x^0 - x^*\| \leq \frac{1}{28}\|x^*\|$.

$\gamma_0 = \frac{1}{m}\||Ax^0| - b\|_1$.

while *not converge* **do**

  $x^{k+1} \leftarrow \operatorname{argmin}_x G(x, x^k, \gamma_k)$

  $\gamma_{k+1} \leftarrow \frac{1}{m}\left\||Ax^{k+1}| - b\right\|_1$

  $k \leftarrow k + 1$

end

Output: x^k

We show in Theorem 4.7.1 that Algorithm 5 solves the problem

$$\min_x \||Ax| - b\|_1 ,$$

with high probability, and x^k converges to x^* locally linearly.

The second case we consider is the application of IRLS to phase retrieval with sparse noise. In this case, a fraction of the measurements is corrupted. In particular, there exist constants $s > 0$, such that for any noise vector $u \in \mathbb{R}^m$ satisfying $\|u\|_0 := \#\{i \mid u_i \neq 0\} \leq sm$ with $|Ax^*| = b + u$, the iterates of IRLS algorithm should also converge to x^* locally linearly, with high probability. For $z \in \mathbb{R}^n$, let T be the indices of the sm largest elements among

$\{|| \langle a_1,\, z\rangle | - b_1|, || \langle a_2,\, z\rangle | - b_2|, ..., || \langle a_m,\, z\rangle | - b_m|\}$. Set $\sigma_{sm}(z) := \|(|Az| - b)_{T^C}\|_1$.

Algorithm 6: IRLS algorithm with reducing γ_k for sparse noise

Input : Parameters: θ_0, θ, $s \in (0, 1)$.

x^0 such that $\|x^0 - x^*\| \leq \theta_0 \|x^*\|$.

$\gamma_0 = \frac{\theta}{m}\sigma_{sm}(x^0)$.

while *not converge* **do**

  $x^{k+1} \leftarrow \text{argmin}_x\, G(x, x^k, \gamma_k)$

  $\gamma_{k+1} \leftarrow \frac{\theta}{m}\sigma_{sm}(x^{k+1})$

  $k \leftarrow k + 1$

end

Output: x^k

Under Assumption 4.1.1, our argument in this paper can further prove the locally linear convergence of the general IRLS algorithm for $p \in [1, 2]$ detailed in Algorithm 7. However, the convergence analysis and the convergence rates of the general IRLS algorithm for $p \in (0, 1)$ and $p \in (2, \infty)$ remain open problems.

Algorithm 7: IRLS algorithm for $p > 0$:

Input : x^0 and $p > 0$

while *not converge* **do**

  $x^{k+1} \leftarrow \text{argmin}_x \sum_{i=1}^m \frac{(\langle a_i, x\rangle - P_i\langle a_i, x^k\rangle)}{(\langle a_i, x^k\rangle - P_i\langle a_i, x^k\rangle)^2 + \gamma_k^2)^{1-p/2}}$

  $\gamma_{k+1} \leftarrow \frac{1}{m}\left\||Ax^{k+1}| - b\right\|_1$

  $k \leftarrow k + 1$

end

Output: x^k

It is shown in [76] and [28] that truncation can help boost the algorithm for amplitude flow and reduce the complexity of the number of measurements for Wirtinger flow. Hence we expect the iterates of Algorithm 8 to still converge locally linearly for the truncated IRLS. We use a similar truncation approach as in the amplitude flow setting [76]. However, the

convergence analysis of the truncated IRLS remains an open problem.

Algorithm 8: Truncated IRLS algorithm for $p > 0$:

Input : x^0, $\tau > 0$ and $p > 0$

while *not converge* **do**

$x^{k+1} \leftarrow \operatorname{argmin}_x \sum_{i=1}^m \frac{(\langle a_i, x\rangle - P_i\langle a_i, x^k\rangle)}{(\langle a_i, x^k\rangle - P_i\langle a_i, x^k\rangle)^2+\gamma_k^2)^{1-p/2}} \cdot 1_{\{i||\langle a_i, x^k\rangle| \geq \frac{1}{1+\tau} b_i\}}$

$\gamma_{k+1} \leftarrow \frac{1}{m} \left\| |Ax^{k+1}| - b \right\|_1$

$k \leftarrow k + 1$

end

Output: x^k

We use the spectral algorithm in [76] to initialize x_0. Before the formal discussion of the convergence results, for $x, y \in \mathbb{R}^n$, define

$$\operatorname{dist}(x, y) := \min\{\|x + y\|, \|x - y\|\}. \tag{4.3.1}$$

We say $x \in B(y, r)$ for $r > 0$ if $\operatorname{dist}(x, y) < r$.

4.4 Preliminaries

In this section we consider robust phase retrieval without noise, in particular, there exists a vector x^* such that $|Ax^*| = b$. We need the following lemma from [21].

Lemma 4.4.1 (Lemma 3.1 [21]). *For any $1 < \epsilon < 1$, if $m > c_0 n\epsilon^{-2}$, then with probability at least $1 - 2\exp(-c_1\epsilon^2 m)$,*

$$(1 - \epsilon)\|h\|^2 \leq \frac{1}{m}\sum_{i=1}^m \langle a_i, h\rangle^2 \leq (1 + \epsilon)\|h\|^2 \tag{4.4.1}$$

holds for all non-zero vectors $h \in \mathbb{R}^n$. Here $c_0, c_1 > 0$ are some universal constants.

The following lemma is similar to Lemma 7 of [80], with a more careful selection of the radius parameter θ_0. The proof strategy is the same with [80].

Lemma 4.4.2. *For any $\epsilon > 0$ and $\theta_0 > 0$, if $m > c_0 n\epsilon^{-2}\log\epsilon^{-1}$, then with probability at least $1 - C\exp(-c_1\epsilon^2 m)$,*

$$\frac{1}{m}\sum_{i=1}^{m}\langle a_i,\, h\rangle^2 \cdot 1_{\{|\langle a_i, x\rangle|<|\langle a_i, h\rangle|\}} \le (1.9\theta_0 + \epsilon)\,\|h\|^2 \tag{4.4.2}$$

holds for all non-zero vectors $h \in \mathbb{R}^n$ satisfying $\|h\| \le \theta_0\,\|x\|$. Here $c_o, c_1, C > 0$ are some universal constants.

Proof. We follow the program and use notation of [80]. We must select a smaller $\theta := \frac{\|h\|}{\|x\|}$ and a smaller δ. Define γ_i by $\frac{|A_i h|^2}{\|h\|^2}\cdot 1_{\{(1-\delta)|A_i x|^2<|A_i h|^2\}}$. Let $f(\tau_1,\tau_2)$ be the density of two joint standard gaussian random variables with correlation $\rho = \frac{h^T x}{\|h\|\|x\|} \neq \pm 1$. We deduce

$$\begin{aligned}\mathbb{E}_{(\rho,\theta)}(\gamma_i) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\tau_2^2\cdot 1_{\{\sqrt{1-\delta}|\tau_1|<|\tau_2|\theta\}}f(\tau_1,\tau_2)d\tau_1 d\tau_2\\ &= \frac{1}{\sqrt{2\pi}}\int_0^{\infty}\tau_2^2\exp(-\frac{\tau_2^2}{2})(\mathrm{erf}(\frac{(\frac{\theta}{\sqrt{1-\delta}}-\rho)\tau_2}{\sqrt{1-\rho^2}}) + \mathrm{erf}(\frac{(\frac{\theta}{\sqrt{1-\delta}}+\rho)\tau_2}{\sqrt{1-\rho^2}}))d\tau_2\end{aligned}$$

where $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z \exp(-x^2)dx$. Denote $\mathrm{erf}\left(\frac{\left(\frac{\theta}{\sqrt{1-\delta}}-\rho\right)t}{\sqrt{1-\rho^2}}\right) + \mathrm{erf}\left(\frac{\left(\frac{\theta}{\sqrt{1-\delta}}+\rho\right)t}{\sqrt{1-\rho^2}}\right)$ by $f_t(\rho)$. We claim that

$$f_t(\rho) \le 2\mathrm{erf}\left(\frac{\theta\max\{t,1\}}{\sqrt{1-\delta}}\right). \tag{4.4.3}$$

We begin the proof of the claim (4.4.3). By symmetry we only need to prove the claim (4.4.3) for $0 \le \rho < 1$. By changing variables, let $r := \frac{\rho}{\sqrt{1-\rho^2}} \in [0,\infty)$, and denote $g_t(r) = f_t(\rho) = \mathrm{erf}((a\sqrt{1+r^2}+r)t) + \mathrm{erf}((a\sqrt{1+r^2}-r)t)$ where $0 \le a = \frac{\theta}{\sqrt{1-\delta}} < 1$(we will choose θ and δ later to ensure $a \in [0,1)$). When $t \ge \frac{\sqrt{2}}{2}$,

$g_t(r)$ is a decreasing function for $r \in [0,\infty] \Leftrightarrow$

$$\begin{aligned}&\frac{\sqrt{\pi}}{2}g_t'(r) = t\exp(-t^2(a\sqrt{1+r^2}+r)^2)(\frac{ar}{\sqrt{1+r^2}}+1) + t\exp(t^2(a\sqrt{1+r^2}-r)^2)(\frac{ar}{\sqrt{1+r^2}}-1) \le 0\\ &\Leftrightarrow \frac{ar}{\sqrt{1+r^2}} - 1 + \exp(-4t^2 ar\sqrt{1+r^2})(\frac{ar}{\sqrt{1+r^2}}+1) \le 0 \qquad (4.4.4)\end{aligned}$$

In order to prove (4.4.4), it suffices to prove

$$\frac{ar}{\sqrt{1+r^2}} - 1 + \frac{1}{1+2ar\sqrt{1+r^2}+2a^2r^2(1+r^2)}(\frac{ar}{\sqrt{1+r^2}}+1) \le 0 \tag{4.4.5}$$

since $t^2 \geq \frac{1}{2}$ and $\exp(x) \geq 1 + x + \frac{x^2}{2}$ for $x \geq 0$. (4.4.5) is equivalent to

$$ar\sqrt{1+r^2} - r^2 + a^2r^2(1+r^2) - ar(1+r^2)^{\frac{3}{2}} \leq 0 \tag{4.4.6}$$

We show (4.4.6) holds by considering two cases.

Case 1. If $a\sqrt{1+r^2} - r \leq 0$, the left side of (4.4.6) can be rewritten as $r(a\sqrt{1+r^2} - r) + ar(1+r^2)(ar - \sqrt{1+r^2})$. But this implies (4.4.6) since $a\sqrt{1+r^2} - r \leq 0$ and $ar - \sqrt{1+r^2} \leq ar - r \leq 0$ (since $a < 1$).

Case 2. If $a\sqrt{1+r^2} - r \geq 0$, since $a < 1$, the following inequality implies (4.4.6).

$$ar\sqrt{1+r^2} - r^2 + r^2(1+r^2) - ar(1+r^2)^{\frac{3}{2}} = -r^3(a\sqrt{1+r^2} - r) \leq 0$$

Therefore we have proved for $t \geq \frac{\sqrt{2}}{2}$, $g_t(r)$ is a decreasing function for $r \in [0, \infty]$, which implies $f_t(\rho) = g_t(r) \leq g_t(0) = 2\text{erf}(at)$. Since $\text{erf}(\cdot)$ is a increasing function, $f_t(\rho) \leq 2\text{erf}(a\max\{t,1\}) = 2\text{erf}(\frac{\theta\max\{t,1\}}{\sqrt{1-\delta}})$.

When $0 < t \leq \frac{\sqrt{2}}{2}$, since $\text{erf}(\cdot)$ is a increasing function on $(-\infty, \infty)$,

$$\begin{aligned} g_t(r) &= \text{erf}((a\sqrt{1+r^2} + r)t) + \text{erf}((a\sqrt{1+r^2} - r)t) \\ &= \text{erf}(a\sqrt{t^2 + (tr)^2} + tr) + \text{erf}(a\sqrt{t^2 + (tr)^2} - tr) \\ &\leq \text{erf}(a\sqrt{1 + (tr)^2} + tr) + \text{erf}(a\sqrt{1 + (tr)^2} - tr) = g_1(tr) \end{aligned} \tag{4.4.7}$$

By the proof of the first part we know $g_1(tr) \leq 2\text{erf}(a)$, which concludes the proof of claim (4.4.3).

By the claim (4.4.3), together with the fact that $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt \leq \frac{2x}{\sqrt{\pi}}$ for $x \geq 0$, we deduce

$$\begin{aligned} \mathbb{E}_{(\rho,\theta)}(\gamma_i) &\leq \sqrt{\frac{2}{\pi}}\left(\int_0^1 \tau^2 \exp(-\frac{\tau^2}{2})\text{erf}(a)d\tau + \int_1^\infty \tau^2 \exp(-\frac{\tau^2}{2})\text{erf}(a\tau)d\tau\right) \\ &\leq \frac{2\sqrt{2}a}{\pi}\left(\int_0^1 \tau^2 \exp(-\frac{\tau^2}{2})d\tau + \int_1^\infty \tau^3 \exp(-\frac{\tau^2}{2})d\tau\right) \\ &\leq \frac{2\sqrt{2}a}{\pi}\left(\int_0^1 \tau^2 \exp(-\frac{\tau^3}{2})d\tau + \int_1^\infty \tau^3 \exp(-\frac{\tau^2}{2})d\tau\right) \qquad \text{since } \tau^2 > \tau^3 \text{ for } \tau \in (0,1) \\ &= \frac{2\sqrt{2}a}{\pi}\left(\frac{2}{3}(1 - e^{-\frac{1}{2}}) + 3e^{-\frac{1}{2}}\right) \leq \frac{4.2 \times \sqrt{2}\theta}{\pi\sqrt{1-\delta}} \end{aligned} \tag{4.4.8}$$

Let $\delta = 0.005$, for any $\theta < \theta_0$, $\rho \in [-1, 1]$ (for $\rho = \pm 1$, $\gamma_i = 0$), we know $\mathbb{E}_{(\rho,\theta)}(\gamma_i) < 1.9\theta_0$. The rest of the proof goes through analogously to [80].

□

Let's recall Lemma 1 from [28].

Lemma 4.4.3 (Lemma 1 [28]). *For $\epsilon \in (0, 1)$. If $m > c_1 n\epsilon^{-2} \log(\frac{1}{\epsilon})$, then with probability at least $1 - c_2 \exp(-c_3\epsilon^2 m)$,*

$$0.9(1-\epsilon)\left\|xx^T - yy^T\right\|_F \leq \frac{1}{m}\sum_{i=1}^{m} |\langle a_i,\, x\rangle^2 - |\langle a_i,\, y\rangle|^2| \leq \sqrt{2}(1-\epsilon)\left\|xx^T - yy^T\right\|_F$$

holds for any $a, b \in \mathbb{R}^n$. Here $c_1, c_2, c_3 > 0$ are some universal constants.

The next lemma gives an estimation of $\frac{1}{m}\sum_{i=1}^{m} |A_i h|$, as opposed to $\frac{1}{m}\sum_{i=1}^{m} (A_i h)^2$ of lemma (4.4.1). A proof of it can be found in [2].

Lemma 4.4.4 (Lemma 4.3 [2]). *Under assumption G, there exist universal constants C_0, C_1, C_2 such that for $\epsilon > 0$ sufficiently small, if $m > C_0 n\epsilon^{-4} \log \epsilon^{-1}$, then with probability at least $1 - C_1 \exp(-C_2\epsilon^4 m)$,*

$$(1-\tilde{\epsilon})\sqrt{\frac{2}{\pi}}\,\|h\| \leq \frac{1}{m}\sum_{i=1}^{m} |\langle a_i,\, h\rangle| \leq (1+\tilde{\epsilon})\sqrt{\frac{2}{\pi}}\,\|h\| \quad \forall\, h \in \mathbb{R}^n. \tag{4.4.9}$$

Let's recall Lemma 4.7 from [2].

Lemma 4.4.5 (Lemma 4.7 [2]). *Under Assumption 4.1.1, there exist universal constants $\tilde{C}_0, \tilde{C}_1, \tilde{C}_2 > 0$ such that for $\epsilon > 0$ sufficiently small, if $m > \tilde{C}_0\epsilon^{-4} \log \epsilon^{-1}$, then with probability at least $1 - \tilde{C}_1 \exp(-\tilde{C}_2\epsilon^4 m)$,*

$$\mu_1\, dist(x, y) \leq \frac{1}{m}\,\big\|\,|Ax| - |Ay|\,\big\|_1 \leq \mu_2\, dist(x, y) \quad \forall\, x, y \in \mathbb{R}^n, \tag{4.4.10}$$

where $dist(x, y) := \min\{\|x + y\|, \|x - y\|\}$, $\mu_1 = \sqrt{\frac{2}{\pi}}(2 - \sqrt{2} - \epsilon)$ and $\mu_2 = \sqrt{\frac{2}{\pi}}(1 + \epsilon)$.

4.5 A Warm-up: Gerchberg-Saxton Algorithm

In this section we prove the locally linear convergence of the Gerchberg-Saxton Algorithm.

Theorem 4.5.1. *If $x^0 \in B(x^*, \frac{\theta_0}{2.3}\|x_*\|)$, we do the following updates, for $k = 1, 2, 3, ...$:*

$$x^{k+1} := \operatorname*{argmin}_x \left\| Ax - P(Ax^k) \right\|^2$$

Then there exist universal constants C_0, C_1 and C_2, such that if $m \geq C_0 n$, then with probability at least $1 - C_1 \exp(-C_2 m)$, the following statements hold: $\gamma_k \searrow 0$ as $k \to \infty$ and for all $k = 0, 1, 2, ...,$

$$\left\| |Ax^{k+1}| - |Ax^*| \right\|^2 \leq \phi \left\| |Ax^k| - |Ax^*| \right\|^2 \tag{4.5.1}$$

for $\phi := \frac{\kappa(\theta_0)}{2} = \frac{\pi(1.9\theta_0 + 0.001)}{(2-\sqrt{2}-0.001)^2} < 1$. Moreover, we have

$$dist(x^0, x^*) \leq \theta_0 \cdot \phi^k \|x^*\| \tag{4.5.2}$$

An example of such parameter pair is $\theta_0 = \frac{1}{20}$. In this case $x_0 \in B(x^*, \frac{1}{50}\|x^*\|)$ is enough to guarantee the convergence rate results (4.5.1) and (4.5.2).

Proof. Let $\epsilon = 0.001$ in Lemma 4.4.5, Lemma 4.4.2 and Lemma 4.4.1. Let C_0, C_1 and C_2 be the common universal constants such that if $m \geq C_0 n$, then with probability at least $1 - C_1 \exp(-C_2 m)$, (4.4.1), (4.4.10) and (4.4.2) holds. The proof below is conditioned on this event.

We prove the theorem by induction. We want to show for $k = 0, 1, ...$

$$\left\| |Ax^k| - |Ax^*| \right\|^2 \leq \phi \left\| |Ax^{k-1}| - |Ax^*| \right\|^2 \tag{4.5.3}$$

and

$$\rho_1(x^k, x^*) \leq \theta_0 \|x^*\|. \tag{4.5.4}$$

When $k = 0$, (4.5.4) is trivial by the initialization of x^0. Assume induction assumptions (4.5.3) and (4.5.4) hold for all $j = 0, ..., k$ where $k \geq 0$. We consider the $j = k + 1$ case. Since

$$x^{k+1} = \operatorname{argmin}_x \left\| Ax - P(Ax^k) \right\|^2,$$

by stationarity we know $A^T(Ax^{k+1} - P(Ax^k)) = 0$. Without loss of generality we assume $\text{dist}(x^k, x^*) = \left\|x^k - x^*\right\|$. If not, we substitute x^* with $-x^*$ in the following argument. Hence

$$\left\langle Ax^{k+1} - P(Ax^k),\, Ax^{k+1} - Ax^* \right\rangle = \left\langle A^T(Ax^{k+1} - P(Ax^k)),\, x^{k+1} - x^* \right\rangle = 0. \tag{4.5.5}$$

By completing squares of (4.5.5) we know

$$\left\|Ax^{k+1} - P(Ax^k)\right\|^2 + \left\|Ax^{k+1} - Ax^*\right\|^2 - \left\|Ax^* - P(Ax^k)\right\|^2 = 0. \tag{4.5.6}$$

Let

$$\begin{aligned} S_k := \{i|\left\langle a_i,\, x^k \right\rangle \left\langle a_i,\, x^* \right\rangle \le 0\} &= \{i|\left\langle a_i,\, x^* \right\rangle^2 \le -\left\langle a_i,\, x^k - x^* \right\rangle \left\langle a_i,\, x^* \right\rangle\} \\ &\subseteq \{i||\left\langle a_i,\, x^k - x^* \right\rangle| \ge |\left\langle a_i,\, x^* \right\rangle|\} \end{aligned} \tag{4.5.7}$$

By the Cauchy-Schwartz inequality and (4.4.10), we have

$$\left\||Ax| - |Ay|\right\|^2 \ge \frac{1}{m}\left\||Ax| - |Ay|\right\|_1^2 \ge \frac{2m}{\pi}(2 - \sqrt{2} - 0.001)^2 \text{dist}^2(x, y) \tag{4.5.8}$$

Hence

$$\begin{aligned} \left\|Ax^* - P(Ax^k)\right\|^2 &= 4\sum_{i=1}^m \left\langle a_i,\, x^* \right\rangle^2 \cdot 1_{\{i \in S_k\}} \\ &\le \sum_{i=1}^m \left\langle a_i,\, x^* \right\rangle^2 \cdot 1_{\{i||\langle a_i, x^*\rangle| \le |\left\langle a_i, x^k - x^* \right\rangle|\}} \\ &\le 4\sum_{i=1}^m \left\langle a_i,\, h^k \right\rangle^2 \cdot 1_{\{i||\langle a_i, x^*\rangle| \le |\left\langle a_i, x^k - x^* \right\rangle|\}} \\ &\le 4m(1.9\theta_0 + 0.001)\left\|x^k - x^*\right\|^2 \qquad \text{(By (4.4.2))} \\ &\le \kappa(\theta_0)\left\||Ax^k| - |Ax^*|\right\|^2 \qquad \text{(By (4.5.8))} \end{aligned} \tag{4.5.9}$$

where $\kappa(\theta_0) := \frac{2\pi(1.9\theta_0 + 0.001)}{(2-\sqrt{2}-0.001)^2}$. Consequently,

$$\begin{aligned} \left\||Ax^{k+1}| - |Ax^*|\right\|^2 &\le \frac{1}{2}\left(\left\|Ax^{k+1} - P(Ax^k)\right\|^2 + \left\|Ax^{k+1} - Ax^*\right\|^2\right) \\ &= \frac{1}{2}\left\|Ax^* - P(Ax^k)\right\|^2 \qquad \text{(By (4.5.6))} \\ &\le \phi\left\||Ax^k| - |Ax^*|\right\|^2 \qquad \text{(By (4.5.9))} \end{aligned} \tag{4.5.10}$$

where $\phi := \frac{\kappa(\theta_0)}{2} = \frac{\pi(1.9\theta_0+0.001)}{(2-\sqrt{2}-0.001)^2} < 1$. Notice

$$\begin{aligned}
\rho_1(x^{k+1}, x^*)^2 &\le \frac{\pi \left\| |Ax^{k+1}| - |Ax^*| \right\|^2}{2m(2-\sqrt{2}-0.001)^2} && \text{(By (4.5.8))} \\
&\le \frac{\pi\phi^{k+1} \left\| |Ax^0| - |Ax^*| \right\|^2}{2m(2-\sqrt{2}-0.001)^2} && \text{(By induction hypothesis and (4.5.10))} \\
&\le \frac{\pi\phi^{k+1} \left\| A(x^0 - x^*) \right\|^2}{2m(2-\sqrt{2}-0.001)^2} && \text{(Triangle inequality)} \\
&\le \frac{\pi\phi^{k+1}(1+0.001)}{2(2-\sqrt{2}-0.001)^2} \left\| x^0 - x^* \right\|^2 && \text{(By (4.4.1))}
\end{aligned} \tag{4.5.11}$$

Since $\phi < 1$, (4.5.11) tells us that

$$\text{dist}_1(x^{k+1}, x^*) \le \left(\frac{\pi(1+0.001)}{2(2-\sqrt{2}-0.001)^2} \right)^{1/2} \left\| x^0 - x^* \right\| < 2.3 \cdot \left\| x^0 - x^* \right\| \le \theta_0 \left\| x^* \right\|.$$

The convergence rate inequality (4.5.2) follows from (4.5.11) in each step. □

4.6 IRLS with Fixed γ for Noiseless Phase Retrieval

The following lemma is key in showing the geometric convergence of $J(x_n, \gamma) - m\gamma$.

Lemma 4.6.1. *Let $\gamma > 0$, $0 < \epsilon < 1$ and $\alpha = \frac{2(1-\epsilon)^2}{\pi}$. Assume $\|z - x^*\| \le \theta_0 \|x_*\|$ and there exists constant $\tilde{C}(\epsilon, \gamma, \theta_0)$ depending on $\epsilon, \gamma, \theta_0$, such that*

$$g(\|z - x^*\|, \epsilon, \gamma, \theta_0) := \frac{(1.9\theta_0 + \epsilon)(\sqrt{\alpha \|z - x^*\|^2 + \gamma^2} + \gamma)}{\alpha\gamma} < \tilde{C}(\epsilon, \gamma, \theta_0) < \frac{1}{4}.$$

For example, if $\theta_0 = \frac{1}{28}$, $\|z - x^*\| \le \gamma$, $\epsilon = 0.001$. Then under the condition of Lemma 4.4.4, if $m > C_0 n \epsilon^{-4} log \epsilon^{-1}$, with probability at least $1 - C_1 \exp(-C_2 \epsilon^4 m)$, we have

$$\langle Az - Ax^*, (I - P)Az \rangle_{W^z} \ge c \|Az - Ax^*\|_{W^z}^2,$$

where $c = 1 - 2\tilde{C}(\epsilon, \gamma, \theta_0) \in \left(\frac{1}{2}, 1\right)$, and, C_0, C_1 and C_2 are some universal constants.

Proof. Let $h = z - x^*$ and $S := \{1 \le i \le m : \langle a_i, x^* \rangle \langle a_i, z \rangle < 0\}$. Recall that w_i^z denotes $\frac{1}{\sqrt{(\langle a_i, z \rangle - P_i \langle a_i, z \rangle)^2 + \gamma^2}}$ and $\langle \cdot, \cdot \rangle_{W^z}$ is the weighted inner product with $W^z = \text{diag}\{w_1^z, w_2^z, ..., w_m^z\}$.

Observe for $i \in S$, $\langle a_i,\, x^* \rangle = -P_i \langle a_i,\, z \rangle$ and for $i \in S^c$, $\langle a_i,\, x^* \rangle = P_i \langle a_i,\, z \rangle$. Then we know

$$\begin{aligned}
\langle Ah,\, (I-P)Az \rangle_{W^z} &= \|Ah\|_{W^z}^2 + \langle Ah,\, Ax^* - PAz \rangle_{W^z} \\
&= \|Ah\|_{W^z}^2 + 2\sum_{i\in S} w_i^z \langle a_i,\, h \rangle \langle a_i,\, x^* \rangle \\
&\geq \|Ah\|_{W^z}^2 - 2\sum_{i\in S} w_i^z |\langle a_i,\, h \rangle \langle a_i,\, x^* \rangle| \qquad (4.6.1)
\end{aligned}$$

Note

$$\begin{aligned}
2\sum_{i\in S} w_i^z |\langle a_i,\, h \rangle \langle a_i,\, x^* \rangle| &= 2\sum_{i=1}^m w_i^z |\langle a_i,\, h \rangle \langle a_i,\, x^* \rangle| \cdot 1_{\{\langle a_i, x^*\rangle^2 < -\langle a_i, x^*\rangle\langle a_i, h\rangle\}} \\
&\leq 2\sum_{i=1}^m w_i^z |\langle a_i,\, h \rangle \langle a_i,\, x^* \rangle| \cdot 1_{\{|\langle a_i, x^*\rangle| < |\langle a_i, h\rangle|\}} \\
&\leq 2\sum_{i=1}^m w_i^z \langle a_i,\, h \rangle^2 \cdot 1_{\{|\langle a_i, x^*\rangle| < |\langle a_i, h\rangle|\}} \\
&\leq \frac{2}{\gamma}\sum_{i=1}^m \langle a_i,\, h \rangle^2 \cdot 1_{\{|\langle a_i, x^*\rangle| < |\langle a_i, h\rangle|\}}, \qquad (4.6.2)
\end{aligned}$$

where the last inequality is by $w_i^z \leq \frac{1}{\gamma}$ for each $i = 1, 2, ..., m$. Let $\alpha = \frac{2(1-\epsilon)^2}{\pi}$. Together

with lemma (4.4.1) and (4.4.2), with large probability, (4.6.2) yields

$$\gamma \sum_{i\in S} w_i^z |\langle a_i,\, h\rangle \langle a_i,\, x^*\rangle| \le m(1.9\theta_0+\epsilon)\|h\|^2 = m(1.9\theta_0+\epsilon)\frac{(\sqrt{\alpha\|h\|^2+\gamma^2}+\gamma)\|h\|^2}{\sqrt{\alpha\|h\|^2+\gamma^2}+\gamma} \tag{4.6.3}$$

$$= \frac{m(1.9\theta_0+\epsilon)(\sqrt{\alpha\|h\|^2+\gamma^2}+\gamma)}{\alpha}(\sqrt{\alpha\|h\|^2+\gamma^2}-\gamma)$$

$$\le \frac{m(1.9\theta_0+\epsilon)(\sqrt{\alpha\|h\|^2+\gamma^2}+\gamma)}{\alpha}(\sqrt{(\frac{1}{m}\sum_{i=1}^m |\langle a_i,\, h\rangle|)^2+\gamma^2}-\gamma) \tag{4.6.4}$$

$$\le \frac{(1.9\theta_0+\epsilon)(\sqrt{\alpha\|h\|^2+\gamma^2}+\gamma)}{\alpha}\sum_{i=1}^m(\sqrt{\langle a_i,\, h\rangle^2+\gamma^2}-\gamma) \tag{4.6.5}$$

$$\le \frac{(1.9\theta_0+\epsilon)(\sqrt{\alpha\|h\|^2+\gamma^2}+\gamma)}{\alpha}\|Ah\|_{W_z}^2 \quad , \tag{4.6.6}$$

where (4.6.4) is by Lemma 4.4.4, (4.6.5) is by convexity of the function $\sqrt{(\cdot)^2+\gamma^2}$, and the last inequality (4.6.6) is due to

$$\sqrt{\langle a_i,\, h\rangle^2+\gamma^2}-\gamma = \frac{\langle a_i,\, h\rangle^2}{\sqrt{\langle a_i,\, h\rangle^2+\gamma^2}+\gamma} \le \frac{\langle a_i,\, h\rangle^2}{\sqrt{(\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle)^2+\gamma^2}} = w_i^z \langle a_i,\, h\rangle^2 . \tag{4.6.7}$$

From the way we choose $\epsilon, \gamma, \theta_0$ ($\theta_0 = \frac{1}{28}$, $\|z-x^*\| \le \gamma$, $\epsilon = 0.001$),

$$g(\|h\|, \epsilon, \gamma, \theta_0) = \frac{(1.9\theta_0+\epsilon)(\sqrt{\alpha\|h\|^2+\gamma^2}+\gamma)}{\alpha\gamma} < \tilde{C}(\epsilon,\gamma,\theta_0) = 0.247 < \frac{1}{4}.$$

By (4.6.1) and (4.6.6), we know

$$\langle Ah,\, (I-P)Az\rangle_{W^z} \ge (1-2\tilde{C}(\epsilon,\gamma,\theta_0))\|Ah\|_{W^z}^2$$

Since $1-2C(\epsilon,\gamma,\theta_0) > \frac{1}{2}$, letting $c = 1-2\tilde{C}(\epsilon,\gamma,\theta_0)$, we have the desired result. □

Lemma 4.6.2. *Under the assumptions of Lemma 4.6.1, let the constants be as in Lemma 4.6.1 and* $z' := \operatorname{argmin}_x G(x, z, \gamma)$. *Then there exists universal constants* C_0, C_1, C_2, *and constant* $\phi \in \left(\frac{1}{2}, 1\right)$ *depending on* β, γ *and* ϵ, *if* $m > C_0 n \epsilon^{-4} \log \epsilon^{-1}$, *then with probability at least* $1 - C_1 \exp(-C_2 \epsilon^4 m)$, *when* $\|z - x^*\| \le \theta_0 \|x_*\|$,

$$J(z', \gamma) - m\gamma \le \phi(J(z, \gamma) - m\gamma) \tag{4.6.8}$$

holds, where $\phi := 2\tilde{C}(\epsilon, \gamma, \theta_0) + \frac{1}{2}$.

Proof. First observe that

$$\begin{aligned}
&2\langle Ax^* - Az,\, (I-P)Az\rangle_{W^z} \\
&= \|Ax^* - PAz\|_{W^z}^2 - \|Ax^* - Az\|_{W^z}^2 - \|(I-P)Az\|_{W^z}^2 \\
&\ge \|Az' - PAz\|_{W^z}^2 - \|Az - PAz\|_{W^z}^2 - \|Ax^* - Az\|_{W^z}^2 && (4.6.9) \\
&\ge 2\sum_{i=1}^{m}(\sqrt{(\langle a_i,\, z'\rangle - P_i\langle a_i,\, z\rangle)^2 + \gamma^2} - \sqrt{(\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle)^2 + \gamma^2}) - \|Az - Ax^*\|_{W^z}^2 && (4.6.10) \\
&\ge 2(J(z', \gamma) - J(z, \gamma)) - \|Az - Ax^*\|_{W^z}^2, && (4.6.11)
\end{aligned}$$

where (4.6.9) is by the definition of z' and (4.6.10) is by the concavity of the square root function. By Lemma 4.6.1 and (4.6.11) we know $J(z, \gamma)$ has sufficient decrease in each iteration. In particular,

$$\begin{aligned}
2(J(z', \gamma) - J(z, \gamma)) &\le -(2c-1)\|Az - Ax^*\|_{W^z}^2 \\
&\le -(2c-1)(\sum_{i=1}^{m}\sqrt{(\langle a_i,\, z\rangle - \langle a_i,\, x^*\rangle)^2 + \gamma^2} - m\gamma) && (4.6.12) \\
&\le -(2c-1)(\sum_{i=1}^{m}\sqrt{(\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle)^2 + \gamma^2} - m\gamma) && (4.6.13) \\
&= -(2c-1)(J(z, \gamma) - m\gamma),
\end{aligned}$$

where (4.6.12) is by (4.6.7) and (4.6.13) is by the definition of the projection operator P_i. Hence

$$J(z', \gamma) - m\gamma \le \phi(J(z, \gamma) - m\gamma) \tag{4.6.14}$$

where $\phi := \frac{3}{2} - c$. Observe $\frac{1}{2} < \phi < 1$ since $c > \frac{1}{2}$ by Lemma 4.6.1. □

Therefore we prove the contraction inequality for the objective $J(z,\gamma) - m\gamma$. In order to show the local convergence rate,z' should also satisfy $\|z' - x^*\| \le \theta_0 \|x^*\|$ and $g(\|z' - h\|, \epsilon, \gamma, \theta_0)$.

Theorem 4.6.3. *Under Assumption 4.1.1, suppose* $x_0 \in B(x^*, \frac{1}{56}\|x^*\|)$. *Let* $\gamma = \left(\frac{1}{m}\|b\|^2\right)^{1/2}$ *and let* x^k *be the iterates of IRLS algorithm 9. Then there exist universal constants* C_0, C_1 *and* C_2, *if* $m \ge C_1 n$, *with probability at least* $1 - C_1\exp(-C_2 m)$, *for each* $k \ge 0$, *the following contraction inequality holds,*

$$J(x^{k+1},\gamma) - m\gamma \le 0.984\left(J(x^k,\gamma) - m\gamma\right). \tag{4.6.15}$$

Moreover, we have

$$dist(x^k, x^*) \le \frac{0.992^k}{25}\|x^*\|. \tag{4.6.16}$$

Proof. Let h^k be $x^k - x^*$ and $\theta_0 = \frac{1}{25}$. By Lemma 4.4.1, letting $\epsilon = 0.001$, there exists constants C_0, C_1 and C_2, such that if $m \ge C_0 m$, with probability at least $1 - C_1\exp(-C_2 m)$, we have

$$\gamma = \left(\frac{1}{m}\|b\|^2\right)^{1/2} = \left(\frac{1}{m}\sum_{i=1}^{m}\langle a_i,\, x^*\rangle^2\right)^{1/2} \ge \sqrt{0.999}\,\|x^*\|. \tag{4.6.17}$$

The event we consider in this lemma is the intersection of the event where (4.6.17) holds, Lemma 4.4.5 and Lemma 4.6.2 holds with $\epsilon = 0.001$. Next we want to show that $x^k \in B(x^*, \frac{0.992^k}{25}\|x^*\|)$ for $k = 1, 2, ...$ by induction.

For the base case $k = 1$, we assume $\text{dist}(x^0, x^*) = \|x^0 - x^*\|$. If not, we replace x^* with $-x^*$. Since $x^0 \in B(x^*, \frac{1}{56}\|x^*\|)$, by (4.6.17), we know

$$\left\|h^0\right\| = \left\|x^0 - x^*\right\| \le \frac{1}{25}\|x^*\| \le \frac{\gamma}{25\cdot\sqrt{0.999}}. \tag{4.6.18}$$

For all $z \in \mathbb{R}^n$,

$$\begin{aligned} J(z,\gamma) - m\gamma &= \sum_{i=1}^{m} \frac{(\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle)^2}{\sqrt{(\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle)^2 + \gamma^2} + \gamma} \\ &\le \frac{1}{2\gamma}\sum_{i=1}^{m}(\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle)^2 \\ &\le \frac{1}{2\gamma}\min\{\sum_{i=1}^{m}\langle a_i,\, z - x^*\rangle^2, \sum_{i=1}^{m}\langle a_i,\, z + x^*\rangle^2\} \\ &\le \frac{1.001m}{2\gamma}\mathrm{dist}(z,x^*)^2 \qquad \text{(by Lemma 4.4.1)} \end{aligned} \tag{4.6.19}$$

When $\epsilon = 0.001$ and $\theta_0 = \frac{1}{25}$, we have

$$\begin{aligned} g(\left\|x^0 - x^*\right\|, \epsilon, \gamma, \theta_0) &= \frac{(1.9\theta_0 + 0.001)(\sqrt{\alpha\left\|x^0 - x^*\right\|^2 + \gamma^2} + \gamma)}{\alpha\gamma} \\ &< (1.9\theta_0 + 0.001)\left(\sqrt{\frac{\theta_0^2}{0.999\alpha} + \frac{1}{\alpha^2}} + \frac{1}{\alpha}\right) \qquad \text{(By (4.6.18))} \\ &:= \tilde{C}(\epsilon,\gamma,\theta_0) = 0.242 < \frac{1}{4}, \end{aligned} \tag{4.6.20}$$

then by Lemma 4.6.2, and (4.6.19) with $z = x^0$, since $\left\|x^0 - x^*\right\| \le \frac{1}{56}\left\|x^*\right\|$,

$$J(x^1,\gamma) - m\gamma \le J(x^0,\gamma) - m\gamma \le \frac{0.102\cdot m}{\gamma}\theta_0^2\left\|x^*\right\|^2, \tag{4.6.21}$$

On the other hand, for all $z \in \mathbb{R}^n$,

$$\begin{aligned} J(z,\gamma) - m\gamma &= \sum_{i=1}^{m} \frac{(\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle)^2}{\sqrt{(\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle)^2 + \gamma^2} + \gamma} \\ &\ge \frac{(\sum_{i=1}^{m}|\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle|)^2}{\sum_{i=1}^{m}\left(\sqrt{(\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle)^2 + \gamma^2} + \gamma\right)} \qquad \text{(By Cauchy-Schwartz)} \\ &\ge \frac{(\sum_{i=1}^{m}|\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle|)^2}{\sum_{i=1}^{m}|\langle a_i,\, z\rangle - P_i\langle a_i,\, z\rangle| + 2m\gamma} \qquad \text{(By } \sqrt{a^2+b^2} \le a+b \text{ for } a,b \ge 0) \\ &\ge \frac{\zeta^2 m\mathrm{dist}(z,x^*)^2}{\zeta\mathrm{dist}(z,x^*) + 2\gamma}, \end{aligned} \tag{4.6.22}$$

where $\zeta = \sqrt{\frac{2}{\pi}}(2 - \sqrt{2} - 0.001)$. Combining (4.6.21), and (4.6.22) with $z = x^1$ we have

$$\frac{\zeta^2 m\mathrm{dist}(x^1,x^*)^2}{\zeta\mathrm{dist}(x^1,x^*) + 2\gamma} \le \frac{0.102\cdot m}{\gamma}\theta_0^2\left\|x^*\right\|^2,$$

which implies

$$\begin{aligned}\operatorname{dist}(x^1, x^*) &\le \frac{1}{\zeta} \cdot \frac{\frac{0.102\theta_0^2\|x^*\|^2}{\gamma} + \sqrt{\left(\frac{0.102\theta_0^2\|x^*\|^2}{\gamma}\right)^2 + 0.816\theta_0^2\|x^*\|^2}}{2} \\ &\le \frac{1}{\zeta} \cdot \frac{\frac{0.102\theta_0}{\sqrt{0.999}} + \sqrt{\frac{0.102^2\theta_0^2}{0.999} + 0.816}}{2}\theta_0\|x^*\| \qquad \text{(By (4.6.17))} \\ &\le 0.98\theta_0\|x^*\|\end{aligned} \tag{4.6.23}$$

We finish the proof for the base case $x^1 \in B(x^*, \frac{0.992}{25}\|x^*\|)$.

Next we assume $x^j \in B(x^*, \frac{0.992^j}{25}\|x^*\|)$ holds for $j = 1, ..., k$. The we consider the case for x^{k+1}. Same as (4.6.20), since by induction assumption, $\operatorname{dist}(x^j, x^*) \le \frac{1}{25}\|x^*\| \le \frac{\gamma}{25\cdot\sqrt{0.999}}$, for all $j = 1, ..., k$, we have

$$g(\operatorname{dist}(x^j, x^*), \epsilon, \gamma, \theta_0) < \tilde{C}(\epsilon, \gamma, \theta_0) = 0.242.$$

By Lemma 4.6.2 and (4.6.21),

$$J(x^{k+1}, \gamma) - m\gamma \le 0.984^k \left(J(x^1, \gamma) - m\gamma\right) \le 0.984^k \cdot \frac{0.102m}{\gamma}\theta_0^2\|x^*\|^2$$

Combining this with (4.6.22) with $z = x^{k+1}$,

$$\frac{\zeta^2 m \operatorname{dist}(x^{k+1}, x^*)^2}{\zeta\operatorname{dist}(x^{k+1}, x^*) + 2\gamma} \le \frac{0.102 \cdot m}{\gamma} 0.984^k\theta_0^2\|x^*\|^2$$

Consequently,

$$\operatorname{dist}(x^{k+1}, x^*) \le 0.98 \cdot 0.992^k\theta_0\|x^*\| \le \frac{0.992^k}{25}\|x^*\|.$$

This completes the inductive step. By Lemma 4.6.2, (4.6.15) holds for each $k \ge 0$.

□

4.7 IRLS with reduced γ for Noiseless Phase Retrieval

In this section we derive the convergence rate of Algorithm 5, using an iterate-specific γ_k. Intuitively, each time after we find x_{k+1}, γ_{k+1} should be updated so that x_{k+1} still lies in the convergence region of $J(x, \gamma_{k+1}) - m\gamma_{k+1}$.

Theorem 4.7.1. *Let $\theta_0, \theta > 0$ be conconstants such that*

$$\phi := \sqrt{\frac{\pi(1.9\theta_0 + 0.001)}{(2-\sqrt{2}-0.001)^2}\left(1+\frac{1}{\theta}\right)} \in (0,1).$$

If $x^0 \in B(x^*, \frac{\theta_0}{1.72}\|x_*\|)$, let $\gamma_0 := \frac{\theta}{m}\||Ax^0| - b\|_1$. We do the following updates, for $k \geq 1$:

1. $x^{k+1} := argmin_x G(x, x^k, \gamma_k)$.

2. $\gamma_{k+1} = \frac{\theta}{m}\left\||Ax^{k+1}| - b\right\|_1$

Then there exist universal constants C_0, C_1 and C_2, such that if $m \geq C_0 n$, then with probability at least $1 - C_1\exp(-C_2 m)$, the following statements hold: $\gamma_k \searrow 0$ as $k \to \infty$ and for all $k = 0, 1, 2, ...,$

$$\left\||Ax^{k+1}| - |Ax^*|\right\|_1 \leq \phi\left\||Ax^k| - |Ax^*|\right\|_1. \tag{4.7.1}$$

Moreover, similar to Theorem (4.6.3), x_k converges to x_* geometrically, namely

$$dist(x^k, x^*) \leq \theta_0 \cdot \phi^k \|x^*\| \tag{4.7.2}$$

An example of such a parameter pair is $\theta_0 = \frac{1}{36}$ and $\theta \in (0.98, \infty)$. In this case $x_0 \in B(x^*, \frac{1}{62}\|x^*\|)$ is enough to guarantee the convergence rate results (4.7.1) and (4.7.2).

Proof. Let $\epsilon = 0.001$ in Lemma 4.4.5, Lemma 4.4.2 and Lemma 4.4.1. Let C_0, C_1 and C_2 be the common universal constants such that if $m \geq C_0 n$, then with probability at least $1 - C_1\exp(-C_2 m)$, (4.4.1), (4.4.10) and (4.4.2) hold. The proof below is conditioned on this event.

We prove the theorem by induction. We want to show for $k = 0, 1, ...$

$$\left\||Ax^k| - |Ax^*|\right\|_1 \leq \phi\left\||Ax^{k-1}| - |Ax^*|\right\|_1 \tag{4.7.3}$$

and

$$\left\|x^k - x^*\right\| \leq \theta_0\|x^*\|. \tag{4.7.4}$$

When $k = 0$, (4.7.9) is trivial by the initialization of x^0.

Assume that (4.7.3) and (4.7.4) hold for all $j = 0, ..., k$ where $k \geq 0$, and consider the $j = k+1$ case. Since

$$x^{k+1} = \underset{x}{\operatorname{argmin}}\, G(x, x^k, \gamma_k) = \left\|\langle a_i, x\rangle - P_i\left\langle a_i, x^k\right\rangle\right\|^2_{W^k},$$

x^{k+1} is a stationary point of $G(x, x^k, \gamma_k)$. Thus $A^T W^k(Ax^{k+1} - P(Ax^k)) = 0$. Without loss of generality we assume $\text{dist}(x^k, x^*) = \left\|x^k - x^*\right\|$. If not, we substitute x^* with $-x^*$ in the following argument. Hence,

$$\left\langle Ax^{k+1} - P(Ax^k),\, Ax^{k+1} - Ax^* \right\rangle_{W^k} = \left\langle A^T W^k(Ax^{k+1} - P(Ax^k)),\, x^{k+1} - x^* \right\rangle_{W^k} = 0. \tag{4.7.5}$$

By completing squares of (4.7.5), we know

$$\left\|Ax^{k+1} - P(Ax^k)\right\|_{W^k}^2 + \left\|Ax^{k+1} - Ax^*\right\|_{W^k}^2 - \left\|Ax^* - P(Ax^k)\right\|_{W^k}^2 = 0. \tag{4.7.6}$$

Set $\beta_k = \left\|x^k - x^*\right\| = \text{dist}(x^k, x^*)$. Let $\kappa(\theta_0)$ be $\frac{4(1.9\theta_0+0.001)}{2-\sqrt{2}-0.001} \cdot \sqrt{\frac{\pi}{2}}$. Let

$$\begin{aligned} S_k := \{i |\left\langle a_i,\, x^k \right\rangle \left\langle a_i,\, x^* \right\rangle \le 0\} &= \{i |\left\langle a_i,\, x^* \right\rangle^2 \le -\left\langle a_i,\, h^k \right\rangle \left\langle a_i,\, x^* \right\rangle\} \\ &\subseteq \{i ||\left\langle a_i,\, h^k \right\rangle| \ge |\left\langle a_i,\, x^* \right\rangle|\}. \end{aligned} \tag{4.7.7}$$

Hence

$$\begin{aligned} \left\|Ax^* - P(Ax^k)\right\|_{W^k}^2 &= 4\sum_{i=1}^m w_i^k \left\langle a_i,\, x^* \right\rangle^2 \cdot 1_{\{i \in S_k\}} \\ &\le 4\sum_{i=1}^m w_i^k \left\langle a_i,\, x^* \right\rangle^2 \cdot 1_{\{i ||\left\langle a_i, h^k \right\rangle| \ge |\left\langle a_i, x^* \right\rangle|\}} \\ &\le \frac{4}{\gamma_k} \sum_{i=1}^m \left\langle a_i,\, h^k \right\rangle^2 \cdot 1_{\{i ||\left\langle a_i, h^k \right\rangle| \ge |\left\langle a_i, x^* \right\rangle|\}} \\ &\le \frac{4m(1.9\theta_0 + 0.001)}{\gamma_k} \left\|x^k - x^*\right\|^2 \qquad \text{(By Lemma 4.4.2)} \\ &= \frac{4m(1.9\theta_0 + 0.001)}{\gamma_k} \text{dist}(x^k, x^*)^2 \\ &\le \frac{\kappa(\theta_0)\beta_k}{\gamma_k} \left\||Ax^k| - |Ax^*|\right\|_1 \qquad \text{(By Lemma 4.4.5)} \end{aligned} \tag{4.7.8}$$

By (4.7.6) and (4.7.8), we have

$$\left\|Ax^{k+1} - P(Ax^k)\right\|_{W^k}^2 + \left\|Ax^{k+1} - Ax^*\right\|_{W^k}^2 \le \frac{\kappa(\theta_0)\beta_k}{\gamma_k} \left\||Ax^k| - |Ax^*|\right\|_1 \tag{4.7.9}$$

Since for $i = 1, 2, ..., m$,

$$|| \left\langle a_i,\, x^{k+1} \right\rangle | - | \left\langle a_i,\, x^* \right\rangle || = | \left\langle a_i,\, x^{k+1} \right\rangle - P_i \left\langle a_i,\, x^{k+1} \right\rangle | \le | \left\langle a_i,\, x_{k+1} \right\rangle - P_i \left\langle a_i,\, x^k \right\rangle |,$$

combining with (4.7.9) we know

$$\left\| |Ax^{k+1}| - |Ax^*| \right\|_{W^k}^2 \leq \frac{\kappa(\theta_0)\beta_k}{2\gamma_k} \left\| |Ax^k| - |Ax^*| \right\|_1 \tag{4.7.10}$$

For all $k = 0, 1, 2, ...,$

$$\beta^k = \operatorname{dist}(x^k, x^*) \leq \frac{\left\| |Ax^k| - |Ax^*| \right\|_1}{Cm} = \frac{\gamma_k}{C\theta} \tag{4.7.11}$$

for $C = \sqrt{\frac{2}{\pi}}(2 - \sqrt{2} - 0.001)$. By the Cauchy-Schwartz inequality, we have

$$\begin{aligned} &(\left\| |Ax^{k+1}| - |Ax^*| \right\|_1)^2 \\ &\leq (\left\| |Ax^{k+1}| - |Ax^*| \right\|_{W^k}^2)(\sum_{i=1}^{m} \sqrt{(\langle a_i,\, x^k\rangle - P_i\,\langle a_i,\, x^k\rangle)^2 + \gamma_k^2}) \end{aligned} \tag{4.7.12}$$

On the other hand, since $m\gamma_k := \theta \left\| |Ax^k| - |Ax^*| \right\|_1 = \theta \left\| Ax^k - P(Ax^k) \right\|_1$ for $k \geq 0$, we have

$$\begin{aligned} \sum_{i=1}^{m} \sqrt{(\langle a_i,\, x^k\rangle - P_i\,\langle a_i,\, x^k\rangle)^2 + \gamma_k^2} &\leq \left\| Ax^k - P(Ax^k) \right\|_1 + m\gamma_k \\ &= (1+\theta) \left\| |Ax^k| - |Ax^*| \right\|_1 \end{aligned} \tag{4.7.13}$$

By (4.7.10), (4.7.11), (4.7.12) and (4.7.13), we have

$$(\left\| |Ax^{k+1}| - |Ax^*| \right\|_1)^2 \leq \frac{(1+\theta)\kappa(\theta_0)}{2C\theta} (\left\| |Ax^k| - |Ax^*| \right\|_1)^2 \tag{4.7.14}$$

Therefore we have

$$\left\| |Ax^{k+1}| - |Ax^*| \right\|_1 \leq \phi(\left\| |Ax^k| - |Ax^*| \right\|_1) \tag{4.7.15}$$

where $\phi := \sqrt{\frac{\kappa(\theta_0)}{2C} \cdot \frac{1+\theta}{\theta}} < 1$.

$$\begin{aligned} \operatorname{dist}(x^{k+1}, x^*) &\leq \sqrt{\frac{\pi}{2}} \frac{\left\| |Ax^{k+1}| - |Ax^*| \right\|_1}{1.999 - \sqrt{2}} \\ &\leq \sqrt{\frac{\pi}{2}} \frac{\phi^{k+1} \left\| |Ax^0| - |Ax^*| \right\|_1}{1.990 - \sqrt{2}} \\ &\leq \frac{1.001 \cdot \phi^{k+1}}{1.99 - \sqrt{2}} \operatorname{dist}(x^0, x^*) \leq 1.72 \cdot \phi^{k+1} \left\| x^0 - x^* \right\| \end{aligned} \tag{4.7.16}$$

Therefore $\left\| x^{k+1} - x^* \right\| < 1.72 \cdot \|x^0 - x^*\| \leq \theta_0 \|x^*\|$. We finish the proof of (4.7.3) and (4.7.4) for x^{k+1}. The convergence rate inequality (4.7.2) follows from (4.7.16) in each step. □

4.8 IRLS for Phase Retrieval with Sparse Noise

In this section we consider real robust phase retrieval with sparse noise. In particular, we consider the problem

$$\min_x \| |Ax| - b \|_1 \tag{4.8.1}$$

for $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, where the measurement matrix A satisfies Assumption 4.1.1. We aim to recover a vector $x_* \in \mathbb{R}^n$ where $|Ax_*| = b + u$ where the noise vector $u \in \mathbb{R}^m$ is a sparse vector. The following lemma and theorem from [2] tell us that under Assumption 4.1.1, if the noise vector is sufficiently sparse, in the sense that $\|u\|_0 \leq cm$ for some universal constant c, then $x_* \in \operatorname{argmin}_x \| |Ax| - b \|_1$, with high probability.

Lemma 4.8.1 (1-ARP, Lemma 4.8 [2]). *Under Assumption 4.1.1, there exist universal constants $C_0, C_1, C_2 > 0, s \in (0, 0.04), \psi_0 \in (0, 0.72)$ such that if $m > c_0 n$, then*

$$\|(|Ax| - |Ay|)_T\|_1 \leq \psi_0 \|(|Ax| - |Ay|)_{T^c}\|_1 \quad \forall\, x, y \in \mathbb{R}^n \text{ and } T \subseteq [m] \text{ with } |T| \leq sm$$

holds with probability at least $1 - C_1 \exp(-C_2 m)$.

Proof. By taking $\epsilon = 0.001$ and $s = 0.001$ in the proof of Lemma 4.8 of [2]. □

This is called the 1-Absolute Range Property of measurement matrix A by the authors of [2].

Theorem 4.8.2 (Theorem 3.2 [2]). *Let $L \in (0, m)$. Suppose $x_* \in \mathbb{R}^n$ is such that $(|Ax_*| - b)$ is L sparse. Assume the measurement matrix A satisfies 1-ARP. Then x_* is a global minimizer of the robust phase retrieval problem (4.8.1). If $\tilde{x}$ is another global minimizer, then $|A\tilde{x}| = |Ax_*|$. If we further assume that the entries of A satisfy Assumption 4.1.1 and $m \geq 2n - 1$, then, with probability 1, x_* is the unique solution of (4.8.1) up to multiplication by -1.*

Consequently, recovering x_* is equivalent to finding the global minimizer problem (4.8.1). The following lemma is plays an important role in proving the convergence of IRLS for robust phase retrieval with sparse noise.

Lemma 4.8.3. *Under Assumption 4.1.1, let $p \in \{1,2\}$. For any $\psi \in (0,1)$, there exist constants $C_0, C_1, C_2 > 0, s \in (0,1)$ such that if $m > C_0 n$, then*

$$\|(Ax)_T\|_p \leq \psi \|(Ax)_{T^c}\|_p, \ \forall x \in \mathbb{R}^n, p \in \{1,2\}, \text{and } T \subseteq [m] \text{ with } |T| \leq sm$$

holds with probability at least $1 - C_1 \exp(-C_2 m)$.

Proof. We first derive conditions on $\epsilon, \ s \in (0,1)$ so that $\psi \in (0,1)$ exists. To this end let $\epsilon, s \in (0,1)$ be given. Let $T \subset [m]$ be any subset of sm indices. By applying Lemma 4.4.1 and Lemma 4.4.4 to the rows of A with row indices in T^c, we know there exist universal constants $c_0, c_1, C > 0$ such that if $(1-s)m \geq c_0 \epsilon^{-4} \log(\epsilon^{-1})$,

$$(1-\epsilon)\tilde{c}_p \|h\|^p \leq \frac{1}{(1-s)m} \sum_{i \in T^c} |A_i h|^p \leq (1+\epsilon) \cdot \tilde{c}_p \|h\|^p, \ \forall \ h \in \mathbb{R}^n \tag{4.8.2}$$

holds with probability at least $1 - C\exp(-c_1 \epsilon^4 (1-s)m)$, where $\tilde{c}_p = 1$ for $p = 2$ and $\tilde{c}_p = \sqrt{\frac{2}{\pi}}$ for $p = 1$. We also know with probability at least $1 - C\exp(-c_1 \epsilon^4 m)$,

$$(1-\epsilon) \cdot \tilde{c}_p \|h\|^p \leq \frac{1}{m} \sum_{i=1}^{m} |A_i h|^p \leq (1+\epsilon) \cdot \tilde{c}_p \|h\|^p, \ \forall \ h \in \mathbb{R}^n \tag{4.8.3}$$

Since the number of such $T's$ is given by

$$\binom{m}{(1-s)m} = \binom{m}{sm} \leq \left(e\frac{m}{sm}\right)^{sm} = \left(\frac{e}{s}\right)^{sm},$$

the event

$$B := \{(4.8.2) \text{ holds for every } T \subseteq [m] \text{ with } |T| = sm\} \cap \{(4.8.3) \text{ holds}\},$$

satisfies

$$\begin{aligned}
\mathbb{P}(B) &\geq 1 - C(e/s)^{sm} \exp(-c_1 \epsilon^4 (1-s)m) - C\exp(-c_1 \epsilon^4 m) \\
&= 1 - C\exp((1 + c_1\epsilon^4)sm + sm\log(\frac{1}{s}) - c_1\epsilon^4 m) - C\exp(-c_1\epsilon^4 m).
\end{aligned}$$

Choose $\hat{s} > 0$ so that $(1 + c_1\epsilon^4)\hat{s} + \hat{s}\log(\frac{1}{\hat{s}}) < \frac{c_1}{2}\epsilon^4$. Then, for all $s \in (0, \hat{s})$, $\mathbb{P}(B) \geq 1 - 2C\exp(-c_1\epsilon^4 m/2)$. Thus, if event B occurs, we have

$$\begin{aligned}\sum_{i\in T}|A_i h|^p &= \sum_{i=1}^{m}|A_i h|^p - \sum_{i\in T^c}|A_i h|^p \\ &\leq \tilde{c}_p(1+\epsilon)m\left\|h\right\|^p - \tilde{c}_p(1-\epsilon)(1-s)m\left\|h\right\|_p \\ &\leq \frac{2\epsilon + s - \epsilon s}{(1-\epsilon)(1-s)}\sum_{i\in T^c}|A_i h|^p,\end{aligned} \tag{4.8.4}$$

where the first inequality follows from (4.8.3) applied to the first term and (4.8.2) applied to the second, and the second inequality follows by (4.8.2). Consequently, as long as $s \in (0, \hat{s})$ is chosen so that $\frac{2\epsilon+s-\epsilon s}{(1-\epsilon)(1-s)} < \psi$, the conclusion follows. This can be accomplished by choosing ϵ so that $\epsilon < \frac{\psi}{5}$ and then choosing $s = \min\{\hat{s}, \frac{\psi}{5}\}$, since in this case $\frac{2\epsilon+s-\epsilon s}{(1-\epsilon)(1-s)} < \frac{2\epsilon+s}{(1-\epsilon)(1-s)} < \frac{3\psi/5}{16/25} = \frac{15}{16}\psi < \psi$. □

Before the proof of the convergence rate theorem of IRLS for robust phase retrieval with sparse noise, let's recall a lemma from [2]. Let $\sigma_L(y)$ be the sum of the $m - L$ smallest elements from the set $\{||A_1 y| - b_1|, ||A_2 y| - b_2|, ..., ||A_m y| - b_m|\}$.

Lemma 4.8.4 (Lemma 3.1). *Let $A \in \mathbb{R}^{m\times n}$, and $L \in (0, m)$. If the matrix A satisfies 1-ARP of order L for $\psi_0 \in (0, 1)$, then*

$$\left\|\,|Ax| - |Ay|\,\right\|_1 \leq \frac{1+\psi_0}{1-\psi_0}\left(\left\|\,|Ax| - b\right\|_1 - \left\|\,|Ay| - b\right\|_1 + 2\sigma_L(y)\right), \tag{4.8.5}$$

for all $x, y \in \mathbb{R}^n$.

Theorem 4.8.5. *Under Assumption 4.1.1, there exist constants $\theta_0, c, C_0, C_1, C_2 > 0$, $s \in (0, 1)$ and $\phi \in (0, 1)$, such that the following statements hold. If $x_0 \in B(x_*, \frac{\theta_0}{2.18}\left\|x_*\right\|)$ and $m \geq C_0 n$, let $\gamma_0 := \frac{1}{m}\sigma_{sm}(x_{k+1})$. We do the following updates, for $k \geq 1$:*

1. $x_{k+1} := argmin_x G(x, x_k, \gamma_k)$.

2. $\gamma_{k+1} = \frac{1}{m}\sigma_{sm}(x_{k+1})$

Then with probability at least $1 - C_1\exp(-C_2 m)$, we have $\gamma_k \searrow 0$ as $k \to \infty$ and for all

$k = 0, 1, 2, ...,$

$$\begin{aligned}&\left\||Ax^{k+1}| - |Ax^*|\right\|_1 + \left\|Ax^{k+1} - AQ^*x^{k+1}\right\|_1 \\ &\qquad \le \phi(\left\||Ax^k| - |Ax^*|\right\|_1 + \left\|Ax^k - AQ^*x^k\right\|_1),\end{aligned}$$

where $Q^* : \mathbb{R}^n \to \mathbb{R}^n$ *represents the projection to* $\{\pm x^*\}$. *Moreover, similar to Theorem 4.7.1, we have*

$$dist(x^k, x^*) \le c \cdot \phi^k \left\|x_*\right\|. \tag{4.8.6}$$

An example of the constants is $\theta_0 = 1/400$, s *is the constant fraction in Lemma 4.8.1 holds for* $\psi_0 = 0.72$ *and Lemma 4.8.3 holds for* $\psi = 0.04$.

Proof. Let $P_i : \mathbb{R} \to \mathbb{R}$ be the projection to $\{\pm b_i\}$ and $P_i^* : \mathbb{R} \to \mathbb{R}$ be the projection to $\{\pm|A_i x_*|\}$. Similarly, let

$$P : \mathbb{R}^m \to \mathbb{R}^m \text{ be the map } P(x) := (P_1(x_1), P_2(x_2), ..., P_m(x_m))$$

and

$$P^* : \mathbb{R}^m \to \mathbb{R}^m \text{ be the map } P^*(x) := (P_1^*(x_1), P_2^*(x_2), ..., P_m^*(x_m)).$$

Let $\epsilon = 0.001$ in Lemma 4.4.5, Lemma 4.4.2 and Lemma 4.4.1. Let C_0, C_1 and C_2 be the common universal constants such that if $m \ge C_0 n$, then with probability at least $1 - C_1 \exp(-C_2 m)$, (4.4.1), (4.4.10), (4.4.2) hold, Lemma 4.8.1 holds for $\psi_0 = 0.72$ and Lemma 4.8.3 holds for $\psi = 0.04$. The proof below is conditioned on this event.

We prove the theorem by induction. We want to show for $k = 0, 1, ...$

$$\begin{aligned}&\left\||Ax^k| - |Ax^*|\right\|_1 + \left\|Ax^k - AQ^*x^k\right\|_1 \\ &\qquad \le \phi(\left\||Ax^{k-1}| - |Ax^*|\right\|_1 + \left\|Ax^{k-1} - AQ^*x^{k-1}\right\|_1),\end{aligned} \tag{4.8.7}$$

and

$$\left\|x^k - x^*\right\| \le \theta_0 \left\|x^*\right\|. \tag{4.8.8}$$

When $k = 0$, (4.8.8) is trivial by the initialization of x^0. Assume induction assumptions (4.8.7) and (4.8.8) hold for all $j = 0, ..., k$ where $k \ge 0$, we consider the $j = k + 1$ case. Let

h^k denote $x^k - x^*$. Without loss of generality we assume $\text{dist}(x^k, x^*) = \left\|x^k - x^*\right\|$. If not we substitute x^* with $-x^*$ in the following proof. Since x^{k+1} is a stationary point of $G(x, x^k, \gamma_k)$, we have

$$\sum_{i=1}^{m} w_i^k(\left\langle a_i,\, x^{k+1}\right\rangle - P_i\left\langle a_i,\, x^k\right\rangle)\left\langle a_i,\, h^{k+1}\right\rangle = 0$$

By rearranging terms, we kown

$$\begin{aligned} A &:= \sum_{i=1}^{m} w_i^n(\left\langle a_i,\, x^{k+1}\right\rangle - P_i^*\left\langle a_i,\, x^k\right\rangle)\left\langle a_i,\, h^{k+1}\right\rangle \\ &= -\sum_{i=1}^{m} w_i^n(P_i^*\left\langle a_i,\, x^k\right\rangle - P_i\left\langle a_i,\, x^k\right\rangle)\left\langle a_i,\, h^{k+1}\right\rangle \end{aligned} \tag{4.8.9}$$

Let $T := \{i|| \left\langle a_i,\, x^*\right\rangle | \neq b_i, 1 \leq i \leq m\}$. For the right hand side of (4.8.9), we deduce that

$$\begin{aligned} A &\leq \sum_{i\in T} w_i^k|P_i^*\left\langle a_i,\, x^k\right\rangle - P_i\left\langle a_i,\, x^k\right\rangle||\left\langle a_i,\, h^{k+1}\right\rangle| \\ &\leq \sum_{i\in T} w_i^k|\left\langle a_i,\, x^k\right\rangle - P_i\left\langle a_i,\, x^k\right\rangle||\left\langle a_i,\, h^{k+1}\right\rangle| + \sum_{i\in T} w_i^k|\left\langle a_i,\, x^k\right\rangle - P_i^*\left\langle a_i,\, x^k\right\rangle||\left\langle a_i,\, h^{k+1}\right\rangle| \\ &\leq \left\|(Ah^{k+1})_T\right\|_1 + \sum_{i\in T} w_i^k|\left\langle a_i,\, x^k\right\rangle - P_i^*\left\langle a_i,\, x^k\right\rangle||\left\langle a_i,\, h^{k+1}\right\rangle| \\ &\leq \psi\left\|(Ah^{k+1})_{T^c}\right\|_1 + \sum_{i\in T} w_i^k|\left\langle a_i,\, h^k\right\rangle||\left\langle a_i,\, h^{k+1}\right\rangle| \\ &\leq \psi\left\|(Ah^{k+1})_{T^c}\right\|_1 + \left\|(Ah^k)_T\right\|_{W^k}\left\|(Ah^{k+1})_T\right\|_{W^k}. \end{aligned} \tag{4.8.10}$$

For the above computation, the second line follows by the triangle inequality,

$$|P_i^*\left\langle a_i,\, x^k\right\rangle - P_i\left\langle a_i,\, x^k\right\rangle| \leq |\left\langle a_i,\, x^k\right\rangle - P_i\left\langle a_i,\, x^k\right\rangle| + |\left\langle a_i,\, x^k\right\rangle - P_i^*\left\langle a_i,\, x^k\right\rangle|$$

for each i. The third line follows by

$$w_i^k|\left\langle a_i,\, x^k\right\rangle - P_i\left\langle a_i,\, x^k\right\rangle| = |\left\langle a_i,\, x^k\right\rangle - P_i\left\langle a_i,\, x^k\right\rangle|/\sqrt{(\left\langle a_i,\, x^k\right\rangle - P_i\left\langle a_i,\, x^k\right\rangle)^2 + \gamma_k^2} \leq 1$$

The fourth line is from Lemma 4.8.3 and $|\left\langle a_i,\, x^k\right\rangle - P_i^*\left\langle a_i,\, x^k\right\rangle| \leq |\left\langle a_i,\, x^k\right\rangle - \left\langle a_i,\, x^*\right\rangle|$. The last line is by the Cauchy-Schwartz inequality

$$\sum_{i\in T} w_i^k|\left\langle a_i,\, h^k\right\rangle||\left\langle a_i,\, h^{k+1}\right\rangle| \leq \left\|(Ah^k)_T\right\|_{W^k}\left\|(Ah^{k+1})_T\right\|_{W^k}.$$

Let T_k be the indices of the sm largest elements among $\{|| \langle a_1,\, x^k\rangle | - b_1|, || \langle a_2,\, x^k\rangle | - b_2|, ..., || \langle a_m,\, x^k\rangle | - b_m|\}$ and $\sigma_{sm}(x^k) := \left\|(|Ax^k| - b)_{T_k^C}\right\|_1$. By Lemma 4.8.4, we have

$$\begin{aligned}\left\||Ax^k| - |Ax^*|\right\|_1 &\le \||Ax^*| - b\|_1 - \left\||Ax^k| - b\right\|_1 + \frac{2(1+\psi_0)}{1-\psi_0}\sigma_{sm}(x^k) \\ &\le 12.3\sigma_{sm}(x^k)\end{aligned} \tag{4.8.11}$$

where the last inequality follows from $x^* \in \operatorname{argmin}_x \||Ax| - b\|_1$ and $\psi_0 \in (0, 0.72)$. Therefore

$$\begin{aligned}\gamma_k = \frac{1}{m}\sigma_{sm}(x^k) &\ge \frac{1}{12.3m}\left\||Ax^k| - |Ax^*|\right\|_1 && \text{(By (4.8.11))} \\ &\ge \frac{C}{12.3}\left\|x^k - x^*\right\| && \text{(By Lemma 4.4.5)}\end{aligned} \tag{4.8.12}$$

where $C := \sqrt{\frac{2}{\pi}}(2 - \sqrt{2} - 0.001)$. By Lemma 4.8.3, for $p \in \{1, 2\}$, $(1+\psi)\left\|(Ah^k)_T\right\|_p \le \psi\left\|(Ah^k)_{T^c}\right\|_p + \psi\left\|(Ah^k)_T\right\|_p = \psi\left\|Ah^k\right\|_p$. Thus

$$\left\|(Ah^k)_T\right\|_p \le \frac{\psi}{1+\psi}\left\|Ah^k\right\|_p. \tag{4.8.13}$$

Consequently, we know

$$\begin{aligned}\left\|(Ah^k)_T\right\|_{W^k}^2 &= \sum_{i\in T}\frac{|Ah^k|^2}{\sqrt{(\langle a_i,\, x^k\rangle - P_i\langle a_i,\, x^k\rangle)^2 + \gamma_k^2}} \\ &\le \sum_{i\in T}\frac{|Ah^k|^2}{\gamma_k} = \frac{\left\|(Ah^k)_T\right\|^2}{\gamma_k} \\ &\le \frac{12.3}{C}\cdot\frac{\psi^2}{(1+\psi)^2}\frac{\left\|Ah^k\right\|^2}{\|h^k\|} && \text{(By (4.8.12) and (4.8.13))} \\ &\le \frac{12.32}{C}\cdot\frac{m\psi^2}{(1+\psi)^2}\left\|h^k\right\| && \text{(By Lemma 4.4.1)} \\ &\le \sqrt{\frac{\pi}{2}}\frac{12.34}{C}\cdot\frac{\psi^2}{(1+\psi)^2}\left\|Ah^k\right\|_1 && \text{(By Lemma 4.4.4)} \\ &\le \frac{C_1\psi^2}{(1+\psi)}\left\|(Ah^k)_{T^c}\right\|_1, && \text{(By Lemma 4.8.3)}\end{aligned} \tag{4.8.14}$$

where $C_1 := \sqrt{\frac{\pi}{2}}\frac{12.34}{C} = \frac{12.34\pi}{2(2-\sqrt{2}-0.001)} < 33.15$. Hence by (4.8.10) and (4.8.14) , we have

$$\begin{aligned}A \le& \psi\left\|(Ah^{k+1})_{T^c}\right\|_1 + \left(\frac{C_1\psi^2}{1+\psi}\right)^{\frac{1}{2}}\left\|(Ah^k)_{T^c}\right\|_1^{\frac{1}{2}}\left\|(Ah^{k+1})_T\right\|_{W^k} \\ &\le \psi\left\|(Ah^{k+1})_{T^c}\right\|_1 + \frac{C_1\psi^2}{2(1+\psi)}\left\|(Ah^k)_{T^c}\right\|_1 + \frac{1}{2}\left\|(Ah^{k+1})_T\right\|_{W^k}^2,\end{aligned} \tag{4.8.15}$$

where the last inequality is by the average inequality $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ for $a, b \geq 0$.

Let β_k be $\left\|x^k - x^*\right\|$ and l be $\frac{4}{1-0.001} \cdot \sqrt{\frac{\pi}{2}} < 5.02$. Let

$$\begin{aligned} S_k := \{i | \left\langle a_i,\, x^k \right\rangle \left\langle a_i,\, x^* \right\rangle < 0\} &= \{i | \left\langle a_i,\, x^* \right\rangle^2 < -\left\langle a_i,\, h^k \right\rangle \left\langle a_i,\, x^* \right\rangle\} \\ &\subseteq \{i | | \left\langle a_i,\, h^k \right\rangle | \geq | \left\langle a_i,\, x^* \right\rangle |\} \end{aligned} \tag{4.8.16}$$

Hence

$$\begin{aligned} &\left\|Ax^* - P(Ax^k)\right\|_{W^k}^2 = 4\sum_{i=1}^m w_i^k \left\langle a_i,\, x^* \right\rangle^2 \cdot 1_{\{i \in S_k\}} \\ &\leq 4\sum_{i=1}^m w_i^k \left\langle a_i,\, x^* \right\rangle^2 \cdot 1_{\{i | | \left\langle a_i, h^k \right\rangle | \geq | \left\langle a_i, x^* \right\rangle |\}} \\ &\leq \frac{4}{\gamma_k}\sum_{i=1}^m \left\langle a_i,\, h^k \right\rangle^2 \cdot 1_{\{i | | \left\langle a_i, h^k \right\rangle | \geq | \left\langle a_i, x^* \right\rangle |\}} \\ &\leq \frac{4m(1.9\theta_0 + 0.001)}{\gamma_k} \left\|x^k - x^*\right\|^2 \qquad \text{(By Lemma 4.4.2)} \\ &\leq \frac{l(1.9\theta_0 + 0.001)\beta_k}{\gamma_k} \left\|Ax^k - Ax^*\right\|_1 \qquad \text{(By Lemma 4.4.4)} \\ &\leq \frac{12.3(1+\psi)(1.9\theta_0 + 0.001)l}{C} \left\|(Ax^k - Ax^*)_{T^c}\right\|_1 \qquad \text{(By Lemma 4.8.3)} \\ &\leq 132.34(1.9\theta_0 + 0.001)(1+\psi) \left\|(Ax^k - Ax^*)_{T^c}\right\|_1 \end{aligned} \tag{4.8.17}$$

On the other hand, by completing the left hand side of (4.8.9), we have

$$2A = \left\|Ax^{k+1} - P^*(Ax^k)\right\|_{W^k}^2 + \left\|Ax^{k+1} - Ax^*\right\|_{W^k}^2 - \left\|Ax^* - P^*(Ax^k)\right\|_{W^k}^2. \tag{4.8.18}$$

By (4.8.15), (4.8.17) and (4.8.18), since $C_1 < 33.15$, we obtain

$$\begin{aligned} &\left\|Ax^{k+1} - P^*(Ax^k)\right\|_{W^k}^2 + \left\|(Ax^{k+1} - Ax^*)_{T^c}\right\|_{W^k}^2 \leq \\ &\left(132.34(1.9\theta_0 + 0.001)(1+\psi) + \frac{33.15\psi^2}{1+\psi}\right)\left\|(Ax^k - Ax^*)_{T^c}\right\|_1 + 2\psi\left\|(Ax^{k+1} - Ax^*)_{T^c}\right\|_1. \end{aligned} \tag{4.8.19}$$

Let $\alpha(\psi, \theta) := 132.34(1.9\theta_0 + 0.001)(1+\psi) + \frac{33.15\psi^2}{1+\psi}$. By (4.8.19) and Cauchy-Schwartz

inequality we know

$$\begin{aligned}
&\left(\left\|(|Ax^{k+1}|-|Ax^*|)_{T^c}\right\|_1+\left\|(Ah^{k+1})_{T^c}\right\|_1\right)^2\\
&\le(\left\|Ax^{k+1}-P^*(Ax^k)\right\|_{W^k}^2+\left\|(Ah^{k+1})_{T^c}\right\|_{W^k})(2\sum_{i\in T^c}\sqrt{(\langle a_i,\,x^k\rangle-P_i\,\langle a_i,\,x^k\rangle)^2+\gamma_k^2})\\
&\le 2(\alpha\left\|(Ah^k)_{T^c}\right\|_1+2\psi\left\|(Ah^{k+1})_{T^c}\right\|_1)(\left\|(|Ax^k|-|Ax^*|)_{T^c}\right\|_1+m\gamma_k)\\
&\le 4(\alpha\left\|(Ah^k)_{T^c}\right\|_1+2\psi\left\|(Ah^{k+1})_{T^c}\right\|_1)\left\|(|Ax^k|-|Ax^*|)_{T^c}\right\|_1,
\end{aligned}\tag{4.8.20}$$

where the last inequality is by

$$m\gamma_n=\sigma_{sm}(x^k)\le\left\|(|Ax^k|-b)_{T^c}\right\|_1=\left\|(|Ax^k|-|Ax^*|)_{T^c}\right\|_1.$$

By the average inequality $4ab\le(a+b)^2$ for $a=\alpha^{-1/2}(\alpha\left\|(Ah^k)_{T^c}\right\|_1+2\psi\left\|(Ah^{k+1})_{T^c}\right\|_1)$ and $b=\alpha^{1/2}\left\|(|Ax^k|-|Ax^*|)_{T^c}\right\|$, we know

$$\begin{aligned}
&\left\|(|Ax^{k+1}|-|Ax^*|)_{T^c}\right\|_1+\left\|(Ah^{k+1})_{T^c}\right\|\\
&\qquad\le\sqrt{\alpha}\left\|(Ah^k)_{T^c}\right\|_1+\frac{2\psi}{\sqrt{\alpha}}\left\|(Ah^{k+1})_{T^c}\right\|+\sqrt{\alpha}\left\|(|Ax^k|-|Ax^*|)_{T^c}\right\|_1
\end{aligned}\tag{4.8.21}$$

If

$$1-2\psi/\alpha^{1/2}>\alpha^{1/2},\tag{4.8.22}$$

we have

$$\left\|(|Ax^{k+1}|-|Ax^*|)_{T^c}\right\|_1+\left\|(Ah^{k+1})_{T^c}\right\|_1\le\phi\left(\left\|(|Ax^k|-|Ax^*|)_{T^c}\right\|_1+\left\|(Ah^k)_{T^c}\right\|_1\right).\tag{4.8.23}$$

for $\phi:=\alpha^{1/2}/(1-2\psi/\alpha^{1/2})<1$. We let $\theta_0=\frac{1}{400}$ and $\psi=0.04$ to make (4.8.22) hold. Consequently, (4.8.23) implies

$$\begin{aligned}
&\left\|(|Ax^{k+1}|-|Ax^*|)_{T^c}\right\|_1+\left\|(Ax^{k+1}-AQ^*x^{k+1})_{T^c}\right\|_1\\
&\qquad\le\phi\left(\left\|(|Ax^k|-|Ax^*|)_{T^c}\right\|_1+\left\|(Ax^k-AQ^*x^k)_{T^c}\right\|_1\right)
\end{aligned}\tag{4.8.24}$$

and we have

$$\begin{aligned}
\operatorname{dist}(x^{k+1}, x^*) &\leq \sqrt{\frac{\pi}{2}} \frac{\left\| |Ax^{k+1}| - |Ax^*| \right\|_1 + \left\| Ax^{k+1} - AQ^* x^{k+1} \right\|_1}{2.998 - \sqrt{2}} \\
&\leq (1 + 0.72) \sqrt{\frac{\pi}{2}} \frac{\left(\left\| |Ax^{k+1}| - |Ax^*| \right)_{T^c} \right\|_1 + \left\| (Ax^{k+1} - AQ^* x^{k+1})_{T^c} \right\|_1}{2.998 - \sqrt{2}} \\
&\leq 1.72 \sqrt{\frac{\pi}{2}} \frac{\phi^{k+1} \left(\| (|Ax^0| - |Ax^*|)_{T^c} \|_1 + \| (Ax^0 - AQ^* x^0)_{T^c} \|_1 \right)}{2.998 - \sqrt{2}} \\
&\leq 1.72 \sqrt{\frac{\pi}{2}} \frac{\phi^{k+1} \left(\| |Ax^0| - |Ax^*| \|_1 + \| Ax^0 - AQ^* x^0 \|_1 \right)}{2.998 - \sqrt{2}} \\
&\leq \frac{1.72 \cdot 2.002 \cdot \phi^{k+1}}{2.998 - \sqrt{2}} \operatorname{dist}(x^0, x^*) \leq 2.18 \cdot \phi^{k+1} \left\| x^0 - x^* \right\|,
\end{aligned} \tag{4.8.25}$$

where the first inequality is from Lemma 4.4.4 and Lemma 4.4.5, the second inequality follows from Lemma 4.8.1 for $\psi_0 = 0.72$, Lemma 4.8.3 for $\psi = 0.04$ and $0.04 = \psi < \psi_0 = 0.72$, the third inequality is due to the induction assumption and the fifth inequality follows from Lemma 4.4.4 and Lemma 4.4.5.

Therefore $\left\| x^{k+1} - x^* \right\| < 2.18 \cdot \| x^0 - x^* \| \leq \theta_0 \| x^* \|$. We finish the proof of (4.8.7) and (4.8.8) for x^{k+1}. The convergence rate inequality (4.8.6) follows from (4.8.25) in each step. □

4.9 Numerical Experiments

In all examples, we use the conjugate gradient method (with 30 iterations) to solve the least square sub-problem. The initialization algorithm we use is from [76].

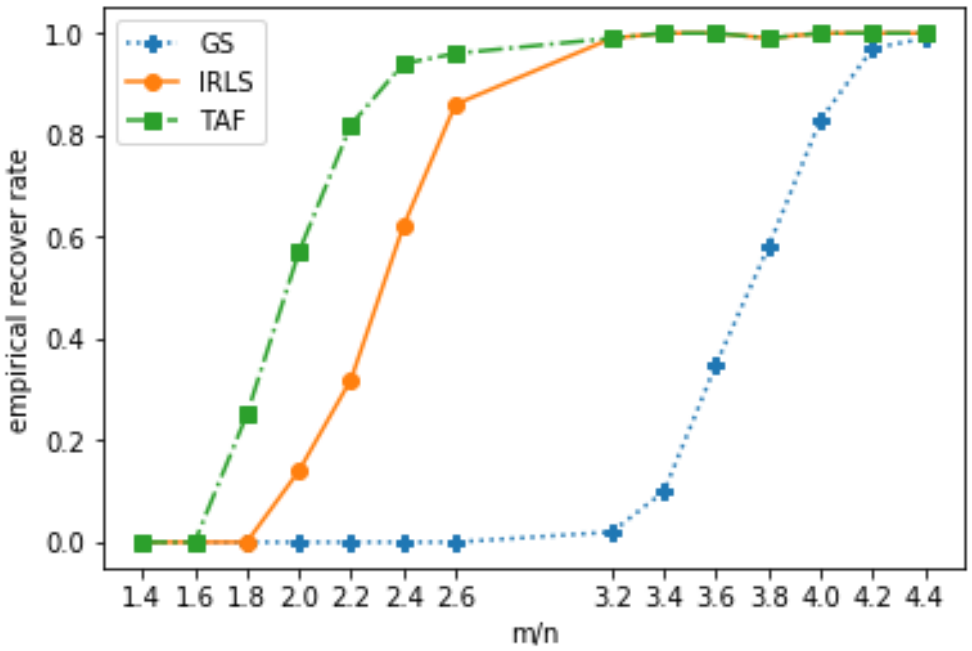

Figure 4.1: Empirical recover rate for IRLS with $p = 1$, Gerchberg-Saxton, TAF with $n = 200$ for different m/n. Noiseless Gaussian model with $a_i \sim N(0, I_n)$ and $x^* \sim N(0, I_n)$.

In the first experiment we compare the empirical successful recovery rate of three algorithms under Assumption 4.1.1 with n = 200 and m/n varying by 0.2 from 1.4 to 2.6 and from 3.2 to 4.4, where a trial is defined as a success if the estimate has a relative error less than 10^{-5}, namely $\text{dist}(x^k, x^*) < 10^{-5} \|x^*\|$. For each ratio, we perform the IRLS with $p = 1$, IRLS with $p = 2$ (Gerchberg-Saxton Algorithm) and TAF 100 times each, using random data. The plot 4.1 shows the empirical recovery rate of the three algorithms out of 100 experiments for each m/n. We see that TAF obtains better rates of recovery than the IRLS with $p = 1$ or $p = 2$.

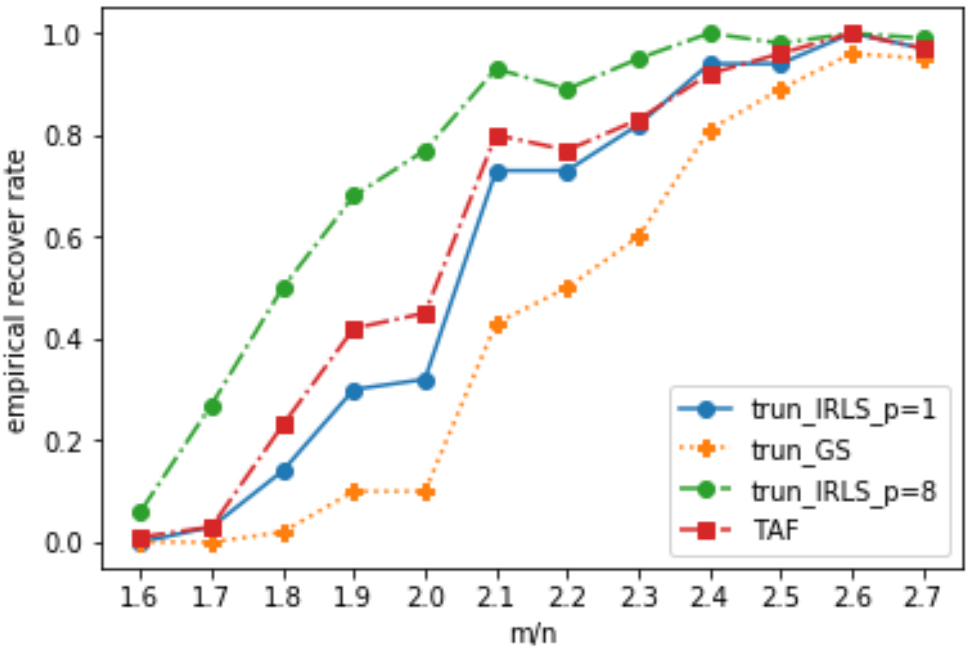

Figure 4.2: Empirical recover rate for truncated IRLS with $p = 1$, truncated IRLS with $p = 8$, truncated Gerchberg-Saxton, TAF with $n = 200$ for different m/n. Noiseless Gaussian model with $a_i \sim N(0, I_n)$ and $x^* \sim N(0, I_n)$.

In the second experiment we compare the empirical recover rate of truncated IRLS for $p = 1$, IRLS for $p = 2$ (Gerchberg-Saxton) with truncation, truncated IRLS for $p = 8$, and TAF. The truncation parameter is take to be $\tau = 0.7$. The range of the ratios m/n is from 1.6 to 2.7 varying by 0.1. We see that there is a boost of the recovery rate after we apply the truncation to IRLS. IRLS with $p = 8$ outperforms TAF all the way from $m/n = 1.6$ to $m/n = 2.7$. Even when $m = 2n$, which is almost the critical case[1], IRLS with $p = 8$ reaches an empirical recovery rate of approximately 0.8.

[1] $m = 2n - 1$ is the condition to guarantee unique recovery up to global sign change in the Gaussian case

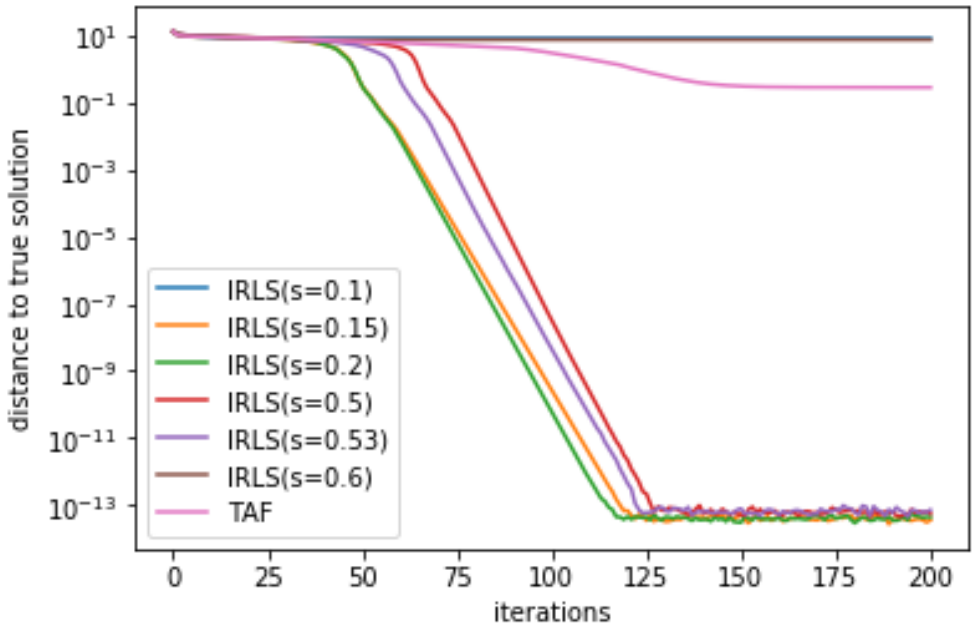

Figure 4.3: The convergence curve in the case with sparse noise. $A \in \mathbb{R}^{400\times 200}$ and the first 40 entries of the noise vector is non-zero.

In the third experiment we consider the convergence of IRLS with $p = 1$ for the case with sparse noise. In this experiment $A \in \mathbb{R}^{400\times 200}$, with each entry independently sampled from $N(0, 1)$. The first 40 entries of the noise vector $u \in \mathbb{R}^{200}$ and all the entries of $x^* \in \mathbb{R}^{200}$ are independently sampled from $N(0, 1)$. The plot 4.3 shows the convergence of IRLS with different s. Since $40/200 = 0.2$, it is expected that $s \geq 0.2$. However, we see even when $s = 0.15$, IRLS still converges. IRlS fail to converge to x^* when $s = 0.1$. When s is slightly greater than 0.5, i.e.$s = 0.53$, IRLS is still able to recover x^*, while fails when $s = 0.6$. We recommend to use $s = 0.5$ in practice.

Chapter 5

ITERATIVELY RE-WEIGHTED LEAST SQUARES ALGORITHM FOR DISTANCE FUNCTIONS TO NON-CONVEX SETS

5.1 Introduction

The iteratively re-weighted least square method (IRLS) was first introduced to solve l_p regression problems for $1 \leq p \leq \infty$ by solving a sequence of l_2 problems [**?**]. Recently, this methodology has been extended to l_p regression problems for $0 \leq p < 1$. Constrained l_1 and l_p regression problems appearing in compressed sensing can also be solved by IRLS [25, 32]. In [32], it is shown that IRLS methods can recover sparse solutions and if the underlying linear mapping admits the isometry property, then IRLS for l_1 regression converges at a local linear rate while that for l_p regression is locally super-linearly convergent. However, the isometry property does not hold in general. In [59, 44], an IRLS algorithm is shown to solve the low-rank matrix recovery problem when the mapping matrix satisfies restricted isometry and null space properties. In [55], IRLS is used to solve low rank and sparse matrix recovery problem but no convergence rate result is provided. In [11] it is shown that IRLS can be applied to the convex problems formalized as the sum of distance functions to convex sets with an iteration complexity of $O(\frac{1}{\epsilon^2})$. The problem formalization studied here originates from that considered in [11].

This chapter concerns the development of an IRLS algorithms (IRLS) designed to solve problems of the form.

$$J(x) := \min_x \|A_0 x - b\|^2 + \sum_{i=1}^{\ell} \text{dist}(A_i x | C_i) + \sum_{i=\ell+1}^{\ell+h} \text{dist}(A_i x | C_i)\text{dist}(A_{i+h} x | C_{i+h}), \quad (5.1.1)$$

where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^{n_0}$, A_i is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^{n_i}$ and C_i's are non-empty

closed sets that can be non-convex. Here the $\|\cdot\|$ is 2-norm and the distance function is induced by this norm. A few examples of problems of this type are listed below.

Example 5.1.1. l_1-regression. *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Let A_i be the i^{th} row vector of A and b_i be the i^{th} entry of b, then*

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_1 = \sum_{i=1}^{m} dist(A_i x|C_i),$$

where $C_i := \{b_i\}$.

Example 5.1.2. Real robust phase retrieval 1. *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m_{++}$. Let A_i be the i^{th} row vector of A and b_i be the i^{th} entry of b, then*

$$\min_{x\in\mathbb{R}^n} \||Ax| - b\|_1 = \sum_{i=1}^{m} dist(A_i x|C_i)$$

where $C_i := \{b_i, -b_i\}$.

Example 5.1.3. Smoothed real robust phase retrieval. *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m_{++}$. Let A_i be the i^{th} row vector of A and b_i be the i^{th} entry of b, then*

$$\begin{aligned}\min_{x\in\mathbb{R}^n} \big\||Ax|^2 - b\big\|_1 &= \sum_{i=1}^{m}(|A_i x||A_i x| - \sqrt{b_i}| + \sqrt{b_i}||A_i x| - \sqrt{b_i}|) \\ &= \sum_{i=1}^{m} dist((\sqrt{b_i})A_i x|C_i^1) + dist(A_i x|C^2)\, dist(A_i x|C_i^3),\end{aligned} \tag{5.1.2}$$

where $C_i^1 := \{\pm b_i\}$, $C^2 = \{0\}$ and $C_i^3 := \{\pm\sqrt{b_i}\}$.

Example 5.1.4. Nesterov's Chebyshev-Rosenbrock function 1. *Nesterov considered the following non-smooth function:*

$$\begin{aligned}\min_{x\in\mathbb{R}^n} f(x) &= \frac{1}{4}|x_1 - 1| + \sum_{i=1}^{n-1} |x_{i+1} - 2|x_i| + 1| \\ &= \frac{1}{4} dist(x_1|C^0) + \sum_{i=1}^{n-1}[dist((\sqrt{2}x_{i+1}, 2\sqrt{2}x_i)|C^1) + dist((2x_{i+1}, 4x_i)|C^2)],\end{aligned} \tag{5.1.3}$$

where $C^0 := \{1\}$, $C^1 = \{(x,y)|\sqrt{2}x + \sqrt{2}y + 1 = 0 \text{ or } \sqrt{2}x - \sqrt{2}y + 1 = 0\}$ *and* $C = \{(x,y)|x \geq -\frac{1}{2} \text{ or } y = 0\}$. *The second equality of* (5.1.3) *is by*

$$|a - 2|b| + 1| = \begin{cases} ||a+1| - 2|b|| & a \geq -1 \\ ||a+1| - 2|b|| + 2\min\{|a+1|, 2|b|\} & a < -1 \end{cases}$$
$$= \sqrt{2}dist((a,2b)|\tilde{C}_1) + 2dist((a,2b)|\tilde{C}_2)$$

where $\tilde{C}_1 = \{(x,y)|x + y + 1 = 0 \text{ or } x - y + 1 = 0\}$ *and* $\tilde{C}_2 = \{(x,y)|x \geq -1 \text{ or } y = 0\}$.

Example 5.1.5. Nesterov's Chebyshev-Rosenbrock function 2. *A second non-smooth variation of the Rosenbrock function proposed by Nesterov is*

$$\tilde{f}(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|$$
$$= \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{m} dist((2\sqrt{2}x_{i+1}, 2\sqrt{2}x_i, 2\sqrt{2}x_{i+1})|C^3)dist((x_{i+1}, x_i)|\{(-\frac{7}{8}, 0)\})$$
$$+ \sum_{i=1}^{m} dist((2\sqrt{2}x_{i+1}, 2\sqrt{2}x_i, 2\sqrt{2}x_{i+1})|C^3)dist(x_{i+1}|\{-\frac{9}{8}\})$$

We know

$$|a - 2b^2 + 1| = 2|b^2 + (a + \frac{7}{8})^2 - (a + \frac{9}{8})^2|$$
$$= 2|\sqrt{b^2 + (a + \frac{7}{8})^2} + |a + \frac{9}{8}||\sqrt{b^2 + (a + \frac{7}{8})^2} - |a + \frac{9}{8}||$$
$$= 2\sqrt{2}[dist((a,b)|\{(-\frac{7}{8}, 0)\}) + dist(a|\{-\frac{9}{8}\})]dist((a,b,a)|C_3),$$

where the cone $C_3 = \{(x,y,z)|(x + \frac{7}{8})^2 + y^2 = (z + \frac{9}{8})^2\}$. *The minimizers of both problems are* $\bar{x} = (1, 1, ..., 1)^T$

5.2 Notation

Let $\mathbb{E}$ be a Euclidean space with Euclidean norm $\|\cdot\|$. For $C \subset \mathbb{E}$, the distance function for C is defined to be

$$\text{dist}\,(a \mid C) := \inf_{b \in C} \|a - b\|\,.$$

In this chapter we assume that $C \subset \mathbb{E}$ is a closed subset of the Euclidean space $\mathbb{E}$. In this case the infimum defining the distance to C is a minimum. If C is also convex, the distance function is convex. The projection mapping for is given by

$$P_C(a) := \operatorname*{argmin}_{b \in C} \|a - b\| .$$

In general, $P_C : \mathbb{E} \Rightarrow C$ is multivalued with P_C being everywhere single valued if and only if C is closed and convex. We define $p_C : \mathbb{E} \to C$ to be a selection from P_C, that is, $p_C(y) \in P_C(y)$ for all $y \in \mathbb{E}$. Hence $\operatorname{dist}(y|C) = \|y - p_C(y)\|$ on $\mathbb{E}$. The distance function is Lipschitz continuous Lipschitz constant 1, i.e.,

$$|\operatorname{dist}(a \,|\, C) - \operatorname{dist}(b \,|\, C)| \leq \|a - b\| \quad \forall\ a, b \in \mathbb{E}.$$

The set of *proximal* normals to C at a points $\bar{y} \in C$ is given by

$$\begin{aligned} N_C^p(\bar{y}) &:= \{\lambda(y - \bar{y}) \mid \lambda \geq 0,\ \bar{y} \in P_C(y)\} \\ &= \{v \mid \exists\, \bar{\tau} > 0 \ s.t.\ \bar{y} \in P_C(\bar{y} + \tau v)\ \forall\, \tau \in [0, \bar{\tau}]\}. \end{aligned}$$

The next lemma speaks to the structure of the directional derivative of the distance function to an arbitrary non-empty closed set. For this purpose, we define the multivalued mapping $G : \mathbb{E} \Rightarrow \mathbb{E}$ by

$$G(x) := \begin{cases} \operatorname{clconv}\left(\bigcup_{y \in P_C(x)} \mathbb{S} \cap N_C^p(y)\right), & x \notin C, \\ \mathbb{B} \cap N_C(x) & , x \in C, \end{cases}$$

where $\mathbb{S} := \{v \mid \|v\| = 1\}$ is the unit sphere in $\mathbb{E}$ and $N_C(x)$ is the normal cone to C at x.

Lemma 5.2.1. *[46, Lemma 2.19] Let C be a closed set. if $x \notin C$, then*

$$\operatorname{dist}(x + h \,|\, C) = \operatorname{dist}(x \,|\, C) + L(h; x) + o(\|h\|) \ \text{as}\ h \to 0 \tag{5.2.1}$$

where

$$\begin{aligned} L(h; x) &:= \min\left\{\left\langle h,\ \frac{x - z}{\|x - z\|}\right\rangle \mid z \in P_C(x)\right\}. \\ &= -\sigma_{G(x)}(-h), \end{aligned}$$

In particular, $\operatorname{dist}(\cdot \,|\, C)$ *is differentiable at x if and only if there is a unique nearest point from x in C.*

For a function f we introduce the Fretchet subdifferential.

$$\hat{\partial} f(\bar{x}) = \{v | f(x) \geq f(\bar{x}) + \langle v,\, x - \bar{x}\rangle + o(\|x - \bar{x}\|) \text{ as } x \to \bar{x}.\}$$

The Fretchet subdifferential is extended to limiting subdifferential which is a substitute of subdifferential in convex case. The limiting subdifferential is defined to be

$$\partial f(\bar{x}) := \{v | \exists \text{ a sequence } \{x_i, v_i\}_{i=1}^{\infty} \text{ such that } x_i \to x, v_i \to v, f(x_i) \to f(x) \text{ and } v_i \in \hat{\partial} f(x_i)\}$$

Obviously $\hat{\partial} f(x) \subseteq \partial f(x)$. The subdifferential of the distance function is given in the following theorem.

Theorem 5.2.2. *[?, Example 8.53] Let $C \subset \mathbb{E}$ be non-empty and closed, and set $f(y) :=$ dist $(y \,|\, C)$. Then*

$$\partial f(x) = \begin{cases} N_C(x) \cap \mathbb{B} & ,\ x \in C, \\ \frac{x - P_C(x)}{\text{dist}(x \,|\, C)} & , x \notin C. \end{cases}$$

5.3 Problem

Let $x \in \mathbb{R}^n$, $b \in \mathbb{R}^{n_0}$, A_i is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^{n_i}$ and C_i's are closed sets.

$$J(x) := \min_x \|A_0 x - b\|_2^2 + \sum_{i=1}^{\ell} \text{dist}(A_i x | C_i) + \sum_{i=\ell+1}^{\ell+h} \text{dist}(A_i x | C_i)\text{dist}(A_{i+h} x | C_{i+h}) \qquad (5.3.1)$$

where C_i is a closed set in $\mathbb{R}^{n_i}$ and $b \in \mathbb{R}^{n_0}$. Note the above problem is not a convex problem even when all the C_i's are convex and H is positive semidefinite. For example $|(x-1)(x-2)|$ for $x \in \mathbb{R}$ is not convex.

Instead of minimizing the non-smooth problem (5.3.1), we solve the following approximation of (5.3.1).

$$J(x, \epsilon) := \|A_0 x - b\|^2 + \sum_{i=1}^{\ell} \sqrt{\text{dist}^2(A_i x | C_i) + \epsilon^2} \qquad (5.3.2)$$
$$+ \sum_{i=\ell+1}^{\ell+h} \sqrt{\text{dist}^2(A_i x | C_i) + \epsilon^2}\sqrt{\text{dist}^2(A_{i+h} x | C_{i+h}) + \epsilon^2}$$

Set $p_i := p_{C_i} \in P_{C_i}$ and $P_i := P_{C_i}$. For a fixed ϵ, define the weight vector to be:

$$w_i(x,\epsilon) := \begin{cases} \frac{1}{\sqrt{\text{dist}^2(A_i x|C_i)+\epsilon^2}} & 1 \le i \le \ell \\ \frac{\sqrt{\text{dist}^2(A_{i+h} x|C_{i+h})+\epsilon^2}}{\sqrt{\text{dist}^2(A_i x|C_i)+\epsilon^2}} & \ell+1 \le i \le \ell+h \\ \frac{\sqrt{\text{dist}^2(A_{i-h} x|C_{i-h})+\epsilon^2}}{\sqrt{\text{dist}^2(A_i|C_i)+\epsilon^2}} & \ell+h+1 \le i \le \ell+2h. \end{cases}$$

Notice $w_i(x,\epsilon)w_{i+h}(x,\epsilon) = 1$ for $\ell+1 \le i \le \ell+h$. When C_i is convex for each $1 \le i \le \ell+2h$, $J(x,\epsilon)$ is differentiable and $\nabla J(x,\epsilon) = \sum_{i=1}^{\ell+2h} w_i(x,\epsilon)A_i^T(A_i x - p_i(A_i x))$. For the general case, we have the following characterization of the limiting subdifferential of $J(x,\epsilon)$.

Lemma 5.3.1. *Consider the function $J(x,\epsilon)$ defined in (5.3.2) for some $\epsilon > 0$. If $A_i x$ has a unique nearest point in C_i for each $1 \le i \le \ell+2h$, then $J(x,\epsilon)$ is differentiable at x and*

$$\nabla J(x,\epsilon) = 2A_0^T(A_0 x - b) + \sum_{i=1}^{\ell+2h} w_i(x,\epsilon)A_i^T(A_i x - p_i(A_i x)). \tag{5.3.3}$$

Remark 5.3.2. *In the case where all of the sets C_i are convex the projections are always unique and so the function $J(\cdot,\epsilon)$ is everywhere differentiable.*

Proof. For (5.3.3), we only need to show that for a closed set $C \subset \mathbb{R}^n$, if $a \in \mathbb{R}^n$ has a unique nearest point in C, then $f(x) := \sqrt{\text{dist}^2(x|C)+\epsilon^2}$ is differentiable at a and $\nabla f(a) = \frac{a-P_C(a)}{\sqrt{\text{dist}^2(a|C)+\epsilon^2}}$. If $a \notin C$, Lemma 5.2.1 and the differentibility of the mapping $y \mapsto \sqrt{(y)^2+\epsilon^2}$ yield the result. If $a \in C$,

$$|f(b)-f(a)| = \sqrt{\text{dist}^2(b|C)+\epsilon^2} - \epsilon = \frac{\text{dist}^2(b)}{\epsilon + \sqrt{\text{dist}^2(b|C)+\epsilon^2}} \le \frac{1}{2\epsilon}\|a-b\|^2 = o(\|a-b\|).$$

Hence $f(x)$ is differentiable and $\nabla f(a) = \frac{a-P_C(a)}{\sqrt{\text{dist}^2(a|C)+\epsilon^2}}$. □

We also require knowledge of the subdifferential $\partial_x J(x,\epsilon)$ when the projections on to each of the sets C_i is not unique. For this we make use of a simplified version of the subdifferential product rule.

Lemma 5.3.3. *For $i = 1, 2$, let $\phi_i : \mathbb{E}_i \to \mathbb{R}_+$ be locally Lipschitz on the Euclidean spaces $\mathbb{E}_i$, and define $\phi : \mathbb{E}_1 \times \mathbb{E}_2 \to \mathbb{R}_+$ by $\phi(y_1, y_2) = \phi_1(y_1)\phi_2(y_2)$. Then, for all $(y_1, y_2) \in \mathbb{E}_1 \times \mathbb{E}_2$,*

$$\partial\phi(y_1, y_2) = \phi_2(y_2)\partial\phi_1(y_1) \times \phi_1(y_1)\partial\phi_2(y_2).$$

Proof. By [60] page 168,

$$\partial\phi(y_1, y_2)|_{(y_1,y_2)=(\bar{y}_1,\bar{y}_2)} = \partial(\phi_2(\bar{y}_2)\phi_1(\cdot) + \phi_1(\bar{y}_1)\phi_2(\cdot))(\bar{y}_1, \bar{y}_2).$$

By [68, Proposition 10.5] and the non-negativity and Lipschitz continuity of the functions ϕ_1,

$$\partial(\phi_2(\bar{y}_2)\phi_1(\cdot) + \phi_1(\bar{y}_1)\phi_2(\cdot))(\bar{y}_1, \bar{y}_2) = \phi_2(\bar{y}_2)\partial\phi_1(\bar{y}_1) \times \phi_1(\bar{y}_1)\partial\phi_2(\bar{y}_2).$$

□

Lemma 5.3.4. *Consider the function $J(x, \epsilon)$ defined in (5.3.2) for some $\epsilon > 0$. Then, for every $x \in \mathbb{R}^n$,*

$$\partial_x J(x, \epsilon) \subset 2A_0^T(A_0x - b) + \sum_{i=1}^{\ell+2h} w_i(x, \epsilon)A_i^T(A_ix - P_i(A_ix))$$

with equality holding if A_ix has a unique nearest point in C_i for each $1 \leq i \leq \ell + 2h$.

Proof. Define $h : \mathbb{R}^{n_0} \times \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_{\ell+2h}} \to \mathbb{R}$ by

$$\begin{aligned}\psi(y) := \|y_0 - b\|^2 &+ \sum_{i=1}^{\ell} \sqrt{\operatorname{dist}^2(y_i|C_i) + \epsilon^2} \\ &+ \sum_{i=\ell+1}^{\ell+h} \sqrt{\operatorname{dist}^2(y_i|C_i) + \epsilon^2}\sqrt{\operatorname{dist}^2(y_{i+h}|C_{i+h}) + \epsilon^2}.\end{aligned}$$

By [**?**, Proposition 10.5] and Lemma 5.3.3,

$$\partial\psi(y) = \{2(y_0 - b)\} \times w_1(x, \epsilon)(y_1 - P_1(y_1)) \times \cdots \times w_{\ell+2h}(x, \epsilon)(y_{\ell+2h} - P_{\ell+2h}(y_{\ell+2h})).$$

The result now follows by [68, Theorem 10.6]. □

Lemma 5.3.5. *Let x_* be the solution of (5.3.1), and let $\epsilon \in (0,1)$ and $\varepsilon > 0$. Suppose x_ϵ is a global solution to $\min_x J(x,\epsilon)$ and x_ϵ^ε is a ε-solution, i.e. $J(x_\epsilon^\varepsilon,\epsilon) - J(x_\epsilon,\epsilon) < \varepsilon$. When $h = 0$ in (5.3.1), we have*

$$J(x_\epsilon^\varepsilon) - J(x_*) \le \ell\epsilon + \varepsilon.$$

For the case when $h \neq 0$ in (5.3.1), assume that C_i, C_{i+h} are compact, $A_i = A_{i+h}$ for each $\ell+1 \le i \le \ell+h$, set $K_i := \max\{\|a-b\| \,|a \in C_i, b \in C_{i+h}\} < \infty$ and $K := \sum_{i=1}^{\ell+h} K_i$, and let $\tilde{x}$ be an arbitrary vector in $\mathbb{R}^n$. Then

$$J(x_\epsilon^\varepsilon) - J(x_*) \le \varepsilon + (3\ell + h + K + J(\tilde{x}))\epsilon.$$

Proof. When $h = 0$, the convexity of $J(x,\epsilon)$ in ϵ implies that

$$0 \le J(x_*,\epsilon) - J(x_*) \le \nabla_\epsilon J(x_*,\epsilon)(\epsilon - 0) = \sum_{i=1}^{\ell} \frac{\epsilon^2}{\sqrt{\mathrm{dist}^2(A_i x_*)|C_i) + \epsilon^2}} \le \ell\epsilon$$

The definitions of x_ϵ^ε, x_ϵ and x_* yield

$$J(x_\epsilon^\varepsilon) - J(x_*) \le (J(x_\epsilon^\varepsilon,\epsilon) - J(x_\epsilon,\epsilon)) + (J(x_\epsilon,\epsilon) - J(x_*,\epsilon)) + (J(x_*,\epsilon) - J(x_*)) \le \varepsilon + \ell\epsilon \tag{5.3.4}$$

When $h \neq 0$ and C_i is compact for each $\ell+1 \le i \le \ell+h$, by repeating the above procedure, since for $a, b \ge 0$,

$$\begin{aligned} \sqrt{a^2+\epsilon^2}\sqrt{b^2+\epsilon^2} - ab &= \frac{(a^2+b^2)\epsilon^2 + \epsilon^4}{\sqrt{(a^2+\epsilon^2)(b^2+\epsilon^2)} + ab} \le \sqrt{(a^2+b^2)\epsilon^2+\epsilon^4} \\ &= \epsilon\sqrt{a^2+b^2+\epsilon^2} \le (a+b+\epsilon)\epsilon \end{aligned} \tag{5.3.5}$$

we have

$$0 \le J(x_*,\epsilon) - J(x_*) \le (\ell + h\epsilon + \sum_{i=\ell+1}^{\ell+2h} \mathrm{dist}(A_i x_*|C_i))\epsilon. \tag{5.3.6}$$

For $\ell+1 \le i \le \ell+h$ and $z \in \mathbb{R}^{n_i}$, if either $\mathrm{dist}(z|C_i) \le 1$ or $\mathrm{dist}(z|C_{i+h}) \le 1$, by triangle inequality, we have

$$\mathrm{dist}(z|C_i) + \mathrm{dist}(z|C_{i+h}) \le 2 + \max\{\|a-b\| \,|a \in C_i, b \in C_{i+h}\} = 2 + K_i. \tag{5.3.7}$$

Since $\text{dist}(z|C_i)+\text{dist}(z|C_{i+h}) < 2\text{dist}(z|C_i)\text{dist}(z|C_{i+h})$ when $\text{dist}(z|C_i) > 1$ and $\text{dist}(z|C_{i+h}) > 1$, combining with (5.3.7), we know

$$\text{dist}(z|C_i) + \text{dist}(z|C_{i+h}) \leq 2 + K_i + 2\text{dist}(z|C_i)\text{dist}(z|C_{i+h}). \tag{5.3.8}$$

With $K := \sum_{i=\ell+1}^{\ell+h} K_i$ as defined above, we have

$$\begin{aligned}
\sum_{i=\ell+1}^{\ell+2h} \text{dist}(A_i x_*|C_i) &\leq 2h + K + 2\sum_{i=\ell+1}^{\ell+h} \text{dist}(A_i x_*|C_i)\text{dist}(A_i x_*|C_{i+h}) && \text{(By (5.3.8))} \\
&\leq 2h + K + 2J(x_*) \\
&\leq 2h + K + 2J(\tilde{x}). && (x_* \text{ minimize } J(x))
\end{aligned} \tag{5.3.9}$$

Similar to (5.3.4), combining with (5.3.6) and (5.3.9), we have

$$J(x_\epsilon^\varepsilon) - J(x_*) \leq \varepsilon + (3\ell + h + K + J(\tilde{x}))\epsilon.$$

□

Hence when $h = 0$, in order to find an ε-optimal solution of (5.3.1), we only need to find an $\frac{\varepsilon}{2}$-optimal solution of $J(x, \frac{\varepsilon}{2\ell})$. When $h \neq 0$, in order to find an ε-optimal solution of (5.3.1), we only need to find an $\frac{\varepsilon}{2}$-optimal solution of $J(x, \varepsilon/2(3\ell + h + K + J(\tilde{x})))$, where $\tilde{x}$ is some arbitrary vector we select beforehand. All the following algorithms are devoted to solve (5.3.2).

5.4 Algorithms

Define

$$G(x, x_k, \epsilon) := \alpha \|A_0 x - b\|^2 + \frac{1}{2}\sum_{i=1}^{\ell} w_i(x_k, \epsilon) \|A_i x - \alpha P_i(A_i x_k) - (1-\alpha)A_i x_k\|^2$$

The first algorithm is as follows:

Algorithm 9: Iteratively re-weighted least square algorithm with fixed α

Input : x_0

Initialize ϵ, $0 < \alpha < 2$.

while *not converge* **do**

  $x_{k+1} \leftarrow \operatorname{argmin}_x G(x, x_k, \epsilon)$

  $k \leftarrow k + 1$

end

Output: x_k

For the later discussion, without loss of generality, we always assume A is of full column rank, since we can preprocess the problem so that this condition is satisfied.

Lemma 5.4.1. *For fixed ϵ and $0 < \alpha < 2$. Suppose $\{x_k\}_{k\geq 0}$ be the sequence generated by Algorithm 1. Then for every $k \geq 1$,*

$$J(x_{k+1}, \epsilon) - J(x_k, \epsilon) \leq - \|A_0 x_{k+1} - A_0 x_k\|^2 - \frac{2-\alpha}{2\alpha} \sum_{i=1}^{\ell+2h} w_i(x_k, \epsilon) \|A_i x_{k+1} - A_i x_k\|^2$$

Proof. By definition

$$\begin{aligned} 2\alpha A_0^T(A_0 x_{k+1} - b) + \sum_{i=1}^{\ell+2h} w_i(x_k, \epsilon)[\alpha A_i^T(A_i x_{k+1} - P_i(A_i x_k)) \\ + (1-\alpha) A_i^T(A_i x_{k+1} - A_i x_k)] = 0 \end{aligned} \tag{5.4.1}$$

By applying inner product $\langle \cdot, x_{k+1} - x_k \rangle$ on both sides,

$$\begin{aligned} &2\alpha \langle A_0 x_{k+1} - b, A_0 x_{k+1} - A_0 x_k \rangle \\ &+ \sum_{i=1}^{\ell+2h} w_i(x_k, \epsilon)[\alpha \langle A_i x_{k+1} - P_i(A_i x_k), A_i x_{k+1} - A_i x_k \rangle + (1-\alpha) \|A_i x_{k+1} - A_i x_k\|^2] = 0 \end{aligned} \tag{5.4.2}$$

By concavity of the square root function $f(a) := \sqrt{a},\ a \geq 0$ for $1 \leq i \leq \ell$ and the joint concavity of $f(a, b) := \sqrt{a \cdot b},\ a, b \geq 0$ for $\ell + 1 \leq i \leq \ell + 2h$ (which follows from $\frac{a_1 b_2 + a_2 b_1}{2} \geq$

$\sqrt{a_1 b_1 a_2 b_2}$ for $a_1, a_2, b_1, b_2 \in \mathbb{R}_+$),

$$J(x_{k+1}, \epsilon) - J(x_k, \epsilon) \leq \|A_0 x_{k+1} - b\|^2 - \|A_0 x_k - b\|^2 + \frac{1}{2} \sum_{i=1}^{\ell+2h} w_i(x_k, \epsilon)(\|A_i x_{k+1} - P_i(A_i x_{k+1})\|^2 - \|A_i x_k - P_i(A_i x_k)\|^2). \tag{5.4.3}$$

Observe that

$$\sum_{i=1}^{\ell+2h} w_i(x_k, \epsilon) \langle A_i x_{k+1} - P_i(A_i x_k),\, A_i x_{k+1} - A_i x_k \rangle = \frac{1}{2} \sum_{i=1}^{\ell+2h} w_i(x_k, \epsilon) \|A_i x_{k+1} - P_i(A_i x_k)\|^2 - \frac{1}{2} \sum_{i=1}^{\ell+2h} w_i(x_k, \epsilon) \|A_i x_k - P_i(A_i x_k)\|^2 + \frac{1}{2} \sum_{i=1}^{\ell+2h} w_i(x_k, \epsilon) \|A_i x_{k+1} - A_i x_k\|^2 . \tag{5.4.4}$$

By (5.4.2), (5.4.3), (5.4.4) and $\|A_i x_{k+1} - P_i(A_i x_{k+1})\| \leq \|A_i x_{k+1} - P_i(A_i x_k)\|$,

$$J(x_{k+1}, \epsilon) - J(x_k, \epsilon) \leq - \|A_0 x_{k+1} - A_0 x_k\|^2 - \frac{2 - \alpha}{2\alpha} \sum_{i=1}^{\ell+2h} w_i(x_k, \epsilon) \|A_i x_{k+1} - A_i x_k\|^2 .$$

□

Convex setting

First we prove the complexity result of the convex case. In this subsection we assume $h = 0$ and all the sets C_i, $i = 1, \ldots, \ell$, are convex. For $Y \in \mathbb{R}^m$, define the horizon cone to be

$$Y^\infty := \left\{ z \;\middle|\; \exists t^k \downarrow 0, \{y^k\} \subset Y \text{ such that } t^k y^k \to z \right\}.$$

A standard result on the horizon cone [69, Theorem 8.1] is that if $Y \subset \mathbb{R}^m$ is nonempty, closed and convex, then

$$Y^\infty = \{ z \mid Y + z \subset Y \}.$$

Set $C := C_1 \times \ldots \times C_\ell$ which is a convex set in $\mathbb{E}_1 \times \ldots \times \mathbb{E}_\ell$. We make use of the following theorem from [11].

Theorem 5.4.2. *[11, Theorem 2.8] Let $\alpha > 0$ and $\epsilon > 0$ be such that the set*

$$L(\alpha, \epsilon) := \{x | J(x, \epsilon) \leq \alpha\}$$

is nonempty. Then $L(\alpha, \epsilon)$ is compact for all $(\alpha, \epsilon) \in \mathbb{R}^2$ if and only if

$$[\bar{x} \in ker(A_0^T A_0) \cap A^{-1} C^\infty] \Leftrightarrow \bar{x} = 0.$$

Hence if either C is compact or $A_0^T A_0$ is positive definite are satisfied, then the set $K := \{x | J(x, \epsilon) \leq J(x_0, \epsilon)\}$ is compact.

Assumption 1: *The integer $h = 0$, all of the sets C_i, $i = 1, \ldots, \ell$, are non-empty, closed and convex and either $C := C_1 \times ... \times C_\ell$ is compact or $A_0^T A_0$ is positive definite in which case the diameter of the set $K := \{x | J(x, \epsilon) \leq J(x_0, \epsilon)\}$ is finite, i.e.*

$$+\infty > M := \sup_{x,y \in K} \|x - y\| \ .$$

Theorem 5.4.3. *Under Assumption 1, let $\epsilon \in (0, 1)$ and x_ϵ be the solution to $\min_x J(x, \epsilon)$. Suppose $\{x_k\}_{k \geq 0}$ be the sequence generated by Algorithm 1. Then*

$$J(x_k, \epsilon) - J(x_\epsilon, \epsilon) \leq \frac{J(x_0, \epsilon)}{k\epsilon\alpha(2 - \alpha)J(x_0, \epsilon)/(\sigma_1^2 M^2(2\ell + 4)) + 1},$$

where σ_1 is great than the largest singular value of all A_i for $0 \leq i \leq \ell + 2h$.

Proof. Set $\delta_k := J(x_k, \epsilon) - J(x_\epsilon, \epsilon)$. By (5.4.1),

$$\begin{aligned} \nabla_x J(x_k, \epsilon) &= 2A_0^T(A_0 x_k - b) + \sum_{i=1}^{\ell} w_i(x_k, \epsilon) A_i^T(A_i x_k - P_i(A_i x_k)) \\ &= -2A_0^T(A_0 x_{k+1} - A_0 x_k) - \frac{1}{\alpha} \sum_{i=1}^{\ell} w_i(x_k, \epsilon) A_i^T(A_i x_{k+1} - A_i x_k) \end{aligned} \tag{5.4.5}$$

By (5.4.5), the convexity of $J(\cdot, \epsilon)$ and the fact that $\|x_k - x_\epsilon\| \leq M$,

$$\begin{aligned} \delta_k &\leq \langle \nabla_x J(x_k, \epsilon), \, x_k - x_\epsilon \rangle \leq M \, \|\nabla_x J(x_k, \epsilon)\| \\ &\leq \frac{\sigma_1 M}{\alpha} \sum_{i=1}^{\ell} w_i(x_k, \epsilon) \, \|A_i x_{k+1} - A_i x_k\| + 2\sigma_1 M \, \|A_0 x_{k+1} - A_0 x_k\| \, . \end{aligned} \tag{5.4.6}$$

Recall Cauchy-Schwartz inequality is

$$(\sum_{i=0}^{\ell} a_i b_i)^2 \leq (\sum_{i=0}^{\ell} a_i^2)(\sum_{i=0}^{\ell} b_i^2)$$

Letting $a_0 = 2$, $b_0 = \|A_0 x_{k+1} - A_0 x_k\|$, and for $1 \leq i \leq \ell$, $a_i := \sqrt{\frac{2w_i(x_k,\epsilon)}{\alpha(2-\alpha)}}$, $b_i := \sqrt{\frac{(2-\alpha)w_i(x_k,\epsilon)}{2\alpha}}\,\|A_i x_{k+1} - A_i x_k\|$, we have

$$\begin{aligned}
&\left(2\,\|A_0 x_{k+1} - A_0 x_k\| + \frac{1}{\alpha}\sum_{i=1}^{\ell} w_i(x_k,\epsilon)\,\|A_i x_{k+1} - A_i x_k\|\right)^2 \\
&\leq \left[\|A_0 x_{k+1} - A_0 x_k\|^2 + \frac{2-\alpha}{2\alpha}\sum_{i=1}^{\ell} w_i(x_k,\epsilon)\,\|A_i x_{k+1} - A_i x_k\|^2\right]\left[4 + \frac{2}{\alpha(2-\alpha)}\sum_{i=1}^{\ell} w_i(x_k,\epsilon)\right] \\
&\leq (J(x_k,\epsilon) - J(x_{k+1},\epsilon))\left[4 + \frac{2\ell}{\alpha(2-\alpha)\epsilon}\right],
\end{aligned} \tag{5.4.7}$$

where the last inequality follows from Lemma 5.4.1 and the observation that

$$\sum_{i=1}^{\ell} w_i(x_k,\epsilon) = \sum_{i=1}^{\ell} (\mathrm{dist}^2(A_i x_k | C_i) + \epsilon^2)^{-1/2} \leq \ell/\epsilon\,.$$

By combining (5.4.6) and (5.4.7) we know

$$\delta_k^2 \leq \sigma_1^2 M^2 \left(4 + \frac{2\ell}{\alpha(2-\alpha)\epsilon}\right)(J(x_k,\epsilon) - J(x_{k+1},\epsilon)) = C(\epsilon)(\delta_k - \delta_{k+1}) \tag{5.4.8}$$

where $C(\epsilon) := \sigma_1^2 M^2 \left(4 + \frac{2\ell}{\alpha(2-\alpha)\epsilon}\right)$ for simplicity.

By Lemma 5.4.1, $\delta_{k+1} \leq \delta_k$, and so, by dividing (5.4.8) by $\delta_k \delta_{k+1}$, we find that

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq \frac{\delta_k}{C(\epsilon)\delta_{k+1}} \geq \frac{1}{C(\epsilon)}.$$

Summing this inequality over k and yields the inequality

$$\frac{1}{\delta_k} \geq \frac{k}{C(\epsilon)} + \frac{1}{\delta_0}. \tag{5.4.9}$$

Since $\delta_0 \leq J(x_0,\epsilon)$, this gives the bound

$$\delta_k \leq \frac{C(\epsilon)J(x_0,\epsilon)}{kJ(x_0,\epsilon) + C(\epsilon)} = \frac{J(x_0,\epsilon)}{kJ(x_0,\epsilon)/C(\epsilon) + 1}.$$

Finally, since $\alpha(2-\alpha) \leq 1$ and $\epsilon \in (0,1)$, we have $C(\epsilon) = \sigma_1^2 M^2 \left(4 + \frac{2\ell}{\alpha(2-\alpha)\epsilon}\right) \leq \frac{\sigma_1^2 M^2 (2\ell+4)}{\alpha(2-\alpha)\epsilon}$. Therefore

$$J(x_k,\epsilon) - J(x_\epsilon,\epsilon) = \delta_k \leq \frac{J(x_0,\epsilon)}{kJ(x_0,\epsilon)/C(\epsilon)+1} \leq \frac{J(x_0,\epsilon)}{k\epsilon\alpha(2-\alpha)J(x_0,\epsilon)/(\sigma_1^2 M^2(2\ell+4))+1}.$$

□

Corollary 5.4.4. *Under the assumptions of Theorem 5.4.3, if $\epsilon = \frac{\varepsilon}{2\ell}$, Algorithm 1 requires $O(\frac{1}{\varepsilon^2})$ iterations k to attain ε-optimality for $J(x)$.*

Proof. This follows directly from Lemma (5.3.5) and Theorem (5.4.3). □

Non-convex setting

Next we consider the non-convex setting, and consider the measure of proximity to optimality given by

$$T(x) := \operatorname{dist}_2^2\left(0 \mid R(x)\right).$$

where

$$R(x) := 2A_0^T(A_0x - b) + \sum_{i=1}^{\ell+2h} w_i(x,\epsilon)A_i^T(A_ix - P_i(A_ix)).$$

Recall from Lemma 5.3.4 that $\partial_x J(x,\epsilon) \subset R(x)$ with equality if A_ix has a unique projection onto C_i for all $i = 1,\ldots,\ell+2h$. For this reason, the condition $0 \in R(x)$, or $T(x) = 0$, is a weak form of first-order stationarity for the problem $\min_x J(x,\epsilon)$.

Theorem 5.4.5. *Let $\epsilon \in (0,1]$, x_ϵ be the solution to $\min_x J(x,\epsilon)$, $\{x_k\}_{k\geq 0}$ be a sequence generated by Algorithm 1. Then*

$$\min_{1\leq j\leq k} T(x_j) \leq \frac{\sigma_1}{k}\left(4 + \frac{2}{\alpha(2-\alpha)}\left(\frac{\ell}{\epsilon} + \frac{2J(x_0,\epsilon)}{\epsilon^2}\right)\right) J(x_0,\epsilon).$$

Proof. Set $u_k := 2A_0^T(A_0x_k - b) + \sum_{i=1}^{\ell+2h} w_i(x_k,\epsilon)A_i^T(A_ix_k - p_i(A_ix_k))$. As in (5.4.5), we have

$$\begin{aligned} u_k &= 2A_0^T(A_0x_k - b) + \sum_{i=1}^{\ell+2h} w_i(x_k,\epsilon)A_i^T(A_ix_k - p_i(A_ix_k)) \\ &= -2A_0^T(A_0x_{k+1} - A_0x_k) - \frac{1}{\alpha}\sum_{i=1}^{\ell+2h} w_i(x_k,\epsilon)A_i^T(A_ix_{k+1} - A_ix_k). \end{aligned} \tag{5.4.10}$$

Next, observe that

$$\sum_{i=1}^{\ell+2h} w_i(x_k,\epsilon) \le \sum_{i=1}^{\ell} w_i(x_k,\epsilon) + \sum_{i=\ell+1}^{\ell+2h} w_i(x_k,\epsilon) \le \frac{\ell}{\epsilon} + \frac{2J(x_0,\epsilon)}{\epsilon^2},$$

since, for $1 \le i \le \ell$, $w_i(x_k,\epsilon) \le \frac{1}{\epsilon}$ and for $\ell+1 \le i \le \ell+h$, one can show that

$$w_i(x_k,\epsilon) + w_{i+h}(x_k,\epsilon) \le \frac{2\sqrt{\operatorname{dist}^2(A_i x_k|C_i)+\epsilon^2}\sqrt{\operatorname{dist}^2(A_{i+h}x_k|C_{i+h})+\epsilon^2}}{\epsilon^2},$$

so that

$$\begin{aligned}\sum_{i=\ell+1}^{\ell+2h} w_i(x_k,\epsilon) &\le \frac{2}{\epsilon^2}\sum_{i=\ell+1}^{\ell+h} \sqrt{\operatorname{dist}^2(A_i x_k|C_i)+\epsilon^2}\sqrt{\operatorname{dist}^2(A_{i+h}x_k|C_{i+h})+\epsilon^2} \\ &\le \frac{2}{\epsilon^2}J(x_k,\epsilon) \\ &\le \frac{2}{\epsilon^2}J(x_0,\epsilon),\end{aligned}$$

where the final inequality follows from Lemma 5.4.1 which shows that $\{J(x_k,\epsilon)\}$ is a decreasing sequence. Then, as in (5.4.7), we use (5.4.10) to find that

$$\begin{aligned}\|u_k\|^2 &\le \sigma_1^2\left(4 + \frac{2}{\alpha(2-\alpha)}\sum_{i=1}^{\ell+2h} w_i(x_k,\epsilon)\right)(J(x_k,\epsilon) - J(x_{k+1},\epsilon)) \\ &\le \sigma_1^2\left(4 + \frac{2}{\alpha(2-\alpha)}\left(\frac{\ell}{\epsilon} + \frac{2J(x_0,\epsilon)}{\epsilon^2}\right)\right)(J(x_k,\epsilon) - J(x_{k+1},\epsilon)) \\ &= C(\epsilon)(J(x_k,\epsilon) - J(x_{k+1},\epsilon))\end{aligned} \tag{5.4.11}$$

where $C(\epsilon) := \sigma_1^2\left(4 + \frac{2}{\alpha(2-\alpha)}\left(\frac{\ell}{\epsilon} + \frac{2J(x_0,\epsilon)}{\epsilon^2}\right)\right)$ for simplicity. By summing over k and telescoping we see that

$$\sum_{j=1}^{k} \|u_j\|^2 \le C(\epsilon)(J(x_0,\epsilon) - J(x_\epsilon,\epsilon)).$$

By Lemma 5.3.4, we arrive at the desired result. □

Corollary 5.4.6. *In Theorem (5.4.5), if $h = 0$, then*

$$\min_{1\le j\le \ell} T(x_j) \le \frac{\sigma_1}{k}\left(4 + \frac{2\ell}{\alpha(2-\alpha)\epsilon}\right)J(x_0,\epsilon)$$

Proof. When $h = 0$, $\sum_{i=1}^{\ell} w_i(x_k, \epsilon) \leq \frac{\ell}{\epsilon}$. The result follows similarly. □

Corollary 5.4.7. *In theorem (5.4.5), if for $\ell + 1 \leq i \leq \ell + h$, $A_i = A_{i+h}$ and C_i, C_{i+h} is compact. Hence $K_i := \max\{\|a - b\| \, | a \in C_i, b \in C_{i+h}\} < \infty$. Let $K := \sum_{i=1}^{m} K_i$. Then*

$$\min_{1 \leq j \leq \ell} T(x_j) \leq \frac{\sigma_1}{k} \left(4 + \frac{2\ell + 4h + K + 2J(x_0, \epsilon)}{\alpha(2 - \alpha)\epsilon}\right) J(x_0, \epsilon)$$

Proof. For $\ell + 1 \leq i \leq \ell + h$ and any $z \in \mathbb{R}^{n_i}$, we have

$$\begin{aligned} \sqrt{\text{dist}^2(z|C_i) + \epsilon^2} + \sqrt{\text{dist}^2(z|C_{i+h}) + \epsilon^2} &\leq 2\epsilon + \text{dist}(z|C_i) + \text{dist}(z|C_{i+h}) \\ &\leq 4 + K_i + 2\text{dist}(z|C_i)\text{dist}(z|C_{i+h}) \end{aligned} \tag{5.4.12}$$

Then for $k \geq 0$, letting $z = A_i x_k$ in (5.4.12), we have

$$\begin{aligned} w_i(x_k, \epsilon) + w_{i+h}(x_k, \epsilon) &\leq \frac{1}{\epsilon} \left(\sqrt{\text{dist}^2(A_i x_k|C_i) + \epsilon^2} + \sqrt{\text{dist}^2(A_i x_k|C_{i+h}) + \epsilon^2}\right) \\ &\leq \frac{1}{\epsilon} \left(4 + K_i + 2\text{dist}(A_i x_k|C_i)\text{dist}(A_i x_k|C_{i+h})\right) \\ &\leq \frac{1}{\epsilon} \left(4 + K_i + 2\sqrt{(\text{dist}^2(A_i x_k|C_i) + \epsilon^2)(\text{dist}^2(A_i x_k|C_{i+h}) + \epsilon^2)}\right) \end{aligned}$$

Therefore

$$\begin{aligned} \sum_{i=\ell+1}^{\ell+h} w_i(x_k, \epsilon) &\leq \frac{1}{\epsilon} \left(4h + K + 2 \sum_{i=\ell+1}^{\ell+h} \sqrt{(\text{dist}^2(A_i x_k|C_i) + \epsilon^2)(\text{dist}^2(A_i x_k|C_{i+h}) + \epsilon^2)}\right) \\ &\leq \frac{1}{\epsilon} \left(4h + K + 2J(x_k, \epsilon)\right) \\ &\leq \frac{1}{\epsilon} (4h + K + 2J(x_0, \epsilon)), \end{aligned} \tag{5.4.13}$$

where the last inequality is by decreasing of $J(x_k, \epsilon)$. Repeat the same procedure as in Theorem (5.4.5), we obtain the desired result. Follow the same procedure we can arrive at the desired result. □

5.5 *Interpretation of IRLS and an IRLS with line search*

Next we give an interpretation of IRLS. Define the following function

$$G(x,s,w,p) := \alpha \|A_0x-b\|^2 + \frac{\alpha}{2}\sum_{i=1}^{\ell}\frac{1}{s_i}(\|A_ix-p_i\|^2+\epsilon^2)+s_i$$
$$+\frac{\alpha}{2}\sum_{i=\ell+1}^{\ell+h}\frac{1}{s_i}(\|A_ix-p_i\|^2+\epsilon^2)+s_i(\|A_{i+h}x-p_{i+h}\|^2+\epsilon^2)+\frac{1-\alpha}{2}\sum_{i=1}^{\ell+2h}\|A_ix-w_i\|^2 \tag{5.5.1}$$

One can check this is a convex function. Consider the following optimization problem:

$$\min_{x,s,w,p} G(x,s,w,p) \text{ such that } p_i \in C_i \text{ and } s_i > 0 \text{ for all } 1 \le i \le \ell$$

Note if x^*, s^*, w^*, $p^* \in \operatorname{argmin}_{x,s,w,p} G(x,s,w,p)$, then $p_i^* = p_i(A_ix^*)$, $w_i^* = A_ix_*$, $s_i^* = \sqrt{\|A_ix^*-p_i(A_ix^*)\|^2+\epsilon^2}$ for $1 \le i \le \ell$, $s_i^* = \frac{\sqrt{\|A_ix^*-p_i(A_ix^*)\|^2+\epsilon^2}}{\sqrt{\|A_{i+h}x^*-P_{i+h}(A_{i+h}x^*)\|^2+\epsilon^2}}$ for $\ell+1 \le i \le \ell+h$. Furthermore, $G(x^*,s^*,w^*,p^*) = \alpha J(x^*,\epsilon)$. Thus x^* is also the minimizer of $J(x,\epsilon)$. For $0 < \alpha \le 1$, we have the following alternative direction block coordinate descent interpretation of IRLS.

Algorithm 10: A block coordinate descent interpretation of iteratively re-weighted least square algorithm with fixed $0 < \alpha \le 1$

Input : x_0, s_0, w_0, p_0

Initialize ϵ, $0 < \alpha \le 1$.

1 **while** *not converge* **do**

2 $\quad w_{k+1}, p_{k+1} \leftarrow \operatorname{argmin}_{p_i \in C_i, w} G(x_k, s_k, w, p)$

3 $\quad s_{k+1} \leftarrow \operatorname{argmin}_{s>0} G(x_k, s, w_{k+1}, p_{k+1})$

4 $\quad x_{k+1} \leftarrow \operatorname{argmin}_x G(x, s_{k+1}, w_{k+1}, p_{k+1})$

5 $\quad k \leftarrow k+1$

6 **end**

Output: x_k

Step 2 gives the least square sub-problem in each sub-step of Algorithm 1. Let

$$W(\alpha) = \operatorname{diag}\{w_1(x(\alpha),\epsilon)I_{n_1}, ..., w_\ell(x(\alpha),\epsilon)I_{n_\ell}\}$$

and

$$W = \text{diag}\{w_1(x_0, \epsilon)I_{n_1}, ..., w_{n_\ell}(x_0, \epsilon)I_{n_\ell}\}.$$

Let the operator

$$f(y) := \begin{bmatrix} (I - P_1)A_1 \\ \vdots \\ (I - P_\ell)A_\ell \end{bmatrix} y$$

For a fixed x_0, let

$$x(\alpha) = \underset{x}{\text{argmin}} \sum_{i=1}^{\ell} w_i(x_k, \epsilon) \left\| A_i x - \alpha p_i(A_i x_0) - (1 - \alpha)A_i x_0 \right\|^2$$

One can obtain $x(\alpha) = x_0 + \alpha v$ where $v = -(A^T W A)^{-1} A^T W f(x_0)$. By Lemma (5.4.1) we know v is a descent direction of $J(x, \epsilon)$ at x_0. Ideally, to make function value decrease most, we can let

$$\alpha_k = \underset{\alpha > 0}{\text{argmin}}\, J(x_{k-1} + \alpha v_{k-1}, \epsilon)$$

where $v_{k-1} = -(A^T W A)^{-1} A^T W f(x_{k-1})$. Note the above problem is a 1 dimensional convex optimization problem since it's the composition of a convex function and a linear function. This problem may still be costly to solve. In practice we may use a line search method, i.e. we record a geometric sequence of length L on this line and pick the α making the function value $J(x_{k-1} + \alpha v_{k-1}, \epsilon)$ decrease the most. The algorithm is as follows:

Algorithm 11: Iteratively re-weighted least square algorithm with line search

Input : x_0

Initialize $\epsilon > 0$, $L \in \mathbb{N}$, $\eta > 1$.

while *not converge* **do**
  $v = -(A^T W A)^{-1} A^T W f(x_k)$
  if $J(x_k + v, \epsilon) > J(x_k + \eta v, \epsilon)$ **then**
    $i \leftarrow 0$
    while $i < L$ *and* $J(x_k + \eta^i v, \epsilon) > J(x_k + \eta^{i+1} v, \epsilon)$ **do**
      $i \leftarrow i + 1$
    end
    $\alpha = \eta^i$
  else if $J(x_k + v, \epsilon) > J(x_k + v/\eta, \epsilon)$ **then**
    $i \leftarrow 0$
    while $i < L$ *and* $J(x_k + \eta^{-i} v, \epsilon) > J(x_k + \eta^{-(i+1)} v, \epsilon)$ **do**
      $i \leftarrow i + 1$
    end
    $\alpha = \eta^{-i}$
  else
    $\alpha = 1$
  end
  $x_{k+1} \leftarrow x_k + \alpha v$
  $k \leftarrow k + 1$
end

Output: x_k

For each sub-step, we select α in this way. Fix a constant $\eta > 1$ and $L \in \mathbb{N}$. First let $\alpha = 1$. If the minimizer is greater than 1, we choose the smallest among $\{J(x(\eta^j), \epsilon) | 0 \leq j \leq L - 1\}$. If the minimizer is smaller than a, we choose the smallest among $\{J(x(\eta^{-j}), \epsilon) | 0 \leq j \leq L - 1\}$.

The IRLS with line search reduces the number of least square sub-problems, though it will increase the time in each sub-step due to line search. However, in practice, line search helps accelerate convergence a lot.

5.6 Practical considerations

In this section we discuss about the practical issues when we implement the algorithms. We illustrate the numerical issues through a simple l_1-regression example, i.e., $J(x) = \|Ax - b\|_1$. A and b are randomly generated following independent normal distributions. For each entry a_{ij} of A, $a_{ij} \sim N(1, 5)$ and for each b_i of b, $b_i \sim N(0, 0.1)$. The sizes of A and b vary due to each different task.

5.6.1 numerical instability

The first question is which ϵ we should select. By lemma (5.3.5), in order to find an ε-optimal solution of (5.3.1), we need to find an $\frac{\varepsilon}{2}$-optimal solution of $J(x, \frac{\varepsilon}{2\ell})$. Thus we cannot choose a large ϵ.

Left side of figure (5.1) shows how $J(x_k) - J(x_{min})$ decreases with respect to different ϵ's using algorithm (12), where $x_{\min}$ is the true minimizer of $J(\cdot)$.

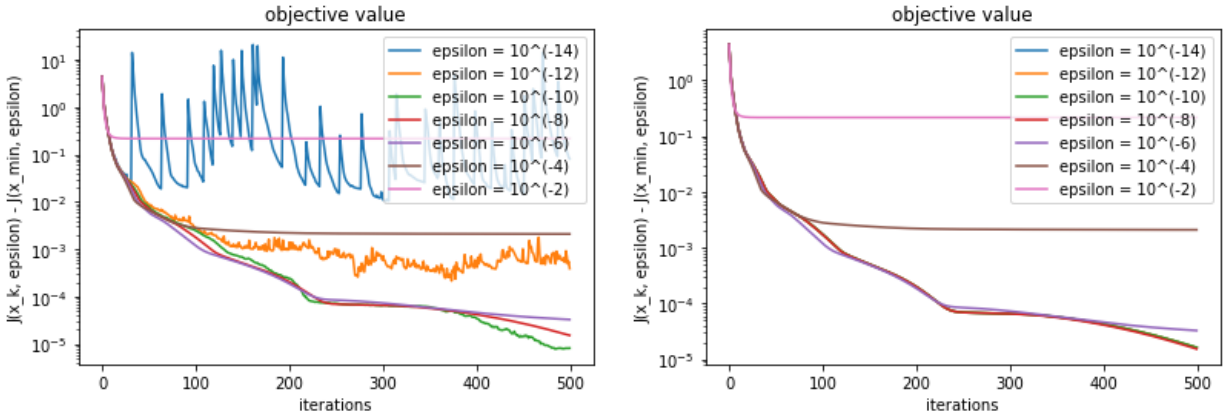

Figure 5.1: $A \in \mathbb{R}^{500\times 100}$ and $b \in \mathbb{R}^{500}$. Left side is $J(x_k) - J(x_{\min})$ vs number of iterations using algorithm (9) for different ϵ's. Right side is $J(x_k) - J(x_{\min})$ vs number of iterations using algorithm (12) for different ϵ's, where $M = 10^7$. In both experiments $\alpha = 1$.

IRLS converges fast for large ϵ, while for smaller ϵ it converges slowly. However, when ϵ is very small, $\epsilon = 10^{-12}$ or $\epsilon = 10^{-14}$ in this case, the function value doesn't decrease as expected but oscillates around a certain value. When ϵ is too small, as the algorithm proceed, it may generates x_k with small $|A_i x - b_i|$ for some i. In this case the reciprocal of a small number would devote to numerical instability. One way out is to perform a modified version of Algorithm (9) instead. We truncate the scaled weights if the scaled weights exceed

some large number.

Algorithm 12: Modified iteratively re-weighted least square algorithm with fixed α

Input : x_0

Initialize $\epsilon > 0$, $0 < \alpha < 2$, a large M.

while *not converge* **do**

$\tilde{w}_i(x_k, \epsilon) \leftarrow w_i(x_k, \epsilon) / \min\{w_i(x_k, \epsilon) | 1 \le i \le \ell\}$

$\tilde{w}_i(x_k, \epsilon) \leftarrow \min\{\tilde{w}_i(x_k, \epsilon), M\}$

$x_{k+1} \leftarrow$

$\operatorname{argmin}_x \alpha \|A_0 x - b\|^2 + \frac{1}{2} \sum_{i=1}^{\ell} \tilde{w}_i(x_k, \epsilon) \|A_i x - \alpha P_i(A_i x_k) - (1-\alpha) A_i x_k\|^2$

$k \leftarrow k + 1$

end

Output: x_k

On one hand algorithm (12) will stablize the convergence curve when ϵ is small (see left of figure (5.1)); but one the other hand, the truncation will prevent the curve from converging to the true optimal solution, especially when ϵ is small (see left side of figure (5.2)). The reason is that the weights after truncation might not give a descent direction for $J(\cdot, \epsilon)$ for a fixed α. However, IRLS with line search (algorithm (11)) will make $J(\cdot, \epsilon)$ increase less if the direction is not a descent direction and will make $J(\cdot, \epsilon)$ decease more if the direction is indeed a descent direction. Algorithm (11) has the ability to break the stability issue for the truncated IRLS with fixed α when ϵ is small. See the right side of figure (5.2).

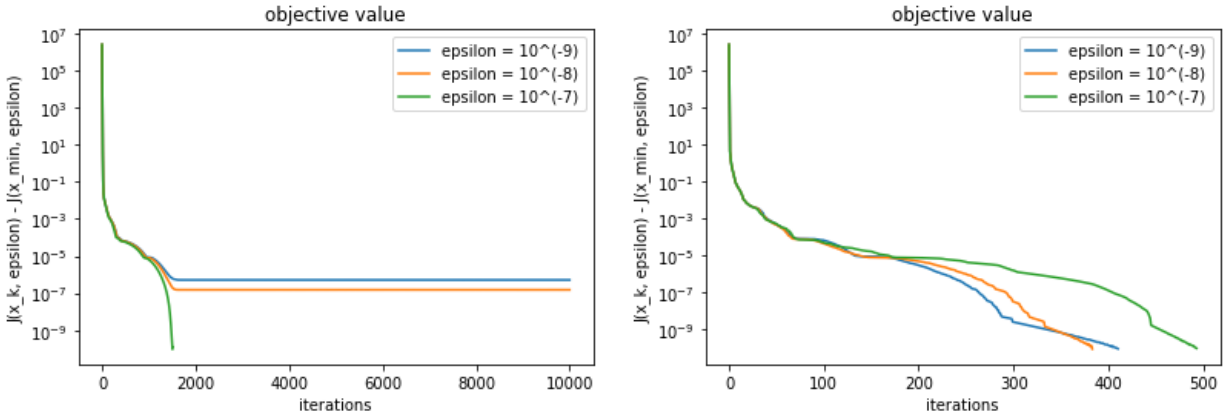

Figure 5.2: $A \in \mathbb{R}^{500\times 100}$ and $b \in \mathbb{R}^{500}$. Left side is $J(x_k, \epsilon) - J(x_{\min}, \epsilon)$ vs number of iterations using algorithm (12) for different ϵ's where $M = 10^7$ and $\alpha = 1$. Right side is $J(x_k, \epsilon) - J(x_{\min}, \epsilon)$ vs number of iterations using algorithm (11) with $L = 10$, $\eta = 1.1$.

5.6.2 different α's

The second question is which α we should choose in practice. Recall that α can be interpreted as the step size in the descent direction. Since our goal is to evaluate the optimization problem (5.3.1), the objective value we take is $J(x) - J(x_{\min})$ where $x_{\min}$ is true minimizer of $J(\cdot)$. Let $\epsilon = 10^{-9}$ and $M = 10^7$. We perform algorithm (12).

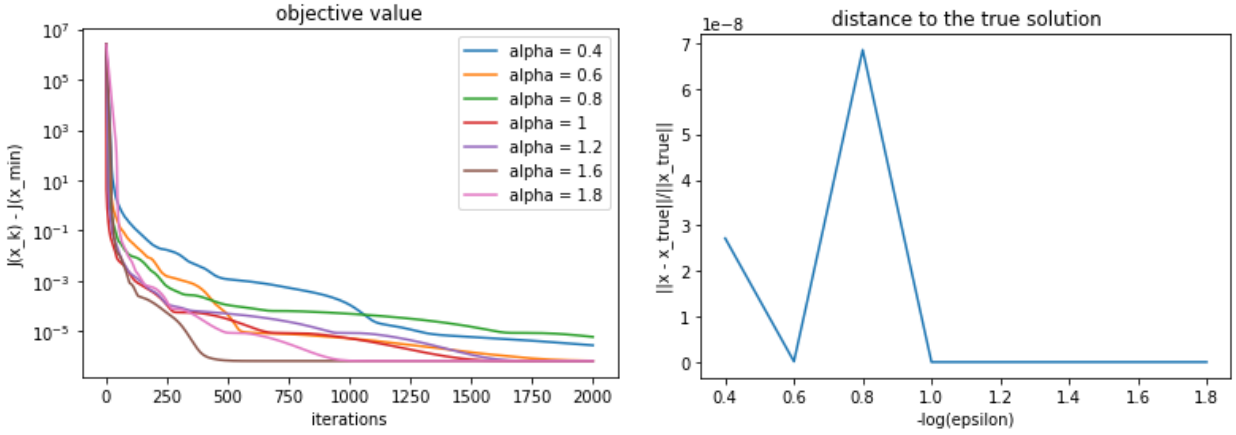

Figure 5.3: $A \in \mathbb{R}^{500\times100}$, $b \in \mathbb{R}^{500}$ and $\epsilon = 10^{-9}$. Left side is $J(x_k) - J(x_{\min})$ vs number of iterations using algorithm (12) for $\alpha \in [0.4, 0.6, 0.8, 1, 1.2, 1.6, 1.8]$ where $M = 10^7$. Right side is $\|x_k - x_{\min}\| / \|x_{\min}\|$ vs different α's using algorithm (11) with $L = 10$, $\eta = 1.1$. The number of the iteration steps is 2000.

[?] considered only the case when $\alpha = 1$. In figure (5.3), one can observe $\alpha = 1$ might not perform best among all $0 \leq \alpha \leq 2$. Empirically, when α is small, IRLS cannot obtain a good accuracy and converges slowly; when α is large, IRLS can obtain a better accuracy and converges faster. Also, a larger α provides with a closer solution to the true optimal point. In practice $\alpha \in [1.5, 1.8]$ gives the fastest convergence rate and highest accuracy.

5.6.3 convergence speed and accuracy

We compare the convergence speed of IRLS with line search (algorithm (11)), IRLS with fixed α and the l_1 solver of CVXPY package in python when the size of A and b varies. By default CVXPY uses a fast interior point method to solve l_1 regression since it can be transformed to a linear programming problem.

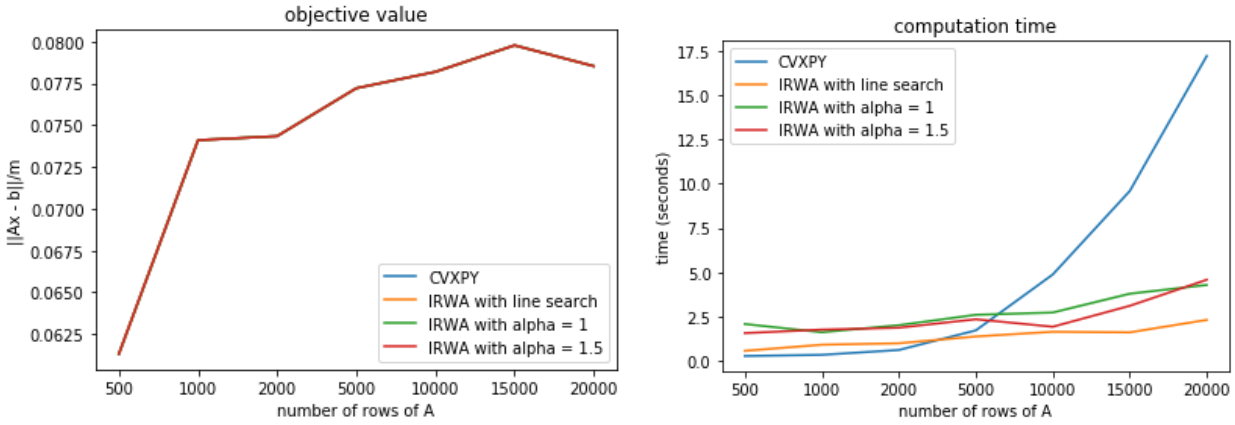

Figure 5.4: $A \in \mathbb{R}^{m\times 100}$, $b \in \mathbb{R}^m$ and $\epsilon = 10^{-9}$. $m \in [500, 1000, 2000, 5000, 10000, 15000, 20000]$. Left side is $\|Ax-b\|/m$ vs m using algorithm.$M = 10^7$. Right side is time vs different α's. The stopping condition is $\|x_k - x_{k-1}\| < \tau$ for $\tau = 10^{-5}$. When algorithm (12) is used, $M = 10^7$. When algorithm (11) is used, $L = 30$, $\eta = 1.1$.

CVXPY performs a little better for small scale problem. For large scale problems, the running times of IRLS's didn't increase too much while the running time of CVXPY increases drastically. For the accuracy, all the four algorithms gives approximately the same objective value. IRLS with line search always outperforms IRLS with $\alpha = 1$ or $\alpha = 1.5$ in terms of running time.

5.6.4 IRLS with line search

In non-convex optimization, IRLS with line search can help escape from local minima/saddle point in practice, compared to IRLS with fixed step size. See more in the numerical experiments section.

5.7 Numerical Experiments

5.7.1 Robust Phase Retrieval 1

We consider the robust phase retrieval problem:

$$\min_{x\in\mathbb{R}^n} J(x) := \||Ax| - b\|_1 = \sum_{i=1}^{\ell} \mathrm{dist}(A_i x|\{-b_i, b_i\}) \tag{5.7.1}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m_+$ and $|Ax|$ means absolute value function apply to each component of vector Ax. Denote the circle $\{a \in \mathbb{C}|\, \|a\| = x\}$ by C_x. In practice b is the observation and x is the signal we hope to recover. We used IRLS algorithm to (5.7.1) as an application. In this example we consider the image recover problem from MNIST data set. The size of the image matrix is $28 \times 28 = 784$. We flatten the image matrix to a vector of size $n = 784$. For the measurement matrix A, we make the each entry sampled independently from $N(0, 1)$. Let the number of the measurements, ℓ, be 4 times of the signal vector, namely $\ell := 4 \times 28 \times 28 = 3136$.

In the first experiment, we consider robust phase retrieval 1 without noise. We sample a image matrix and flatten it. Let the vector be x_* and $b = |Ax_*|$. We use the initialization algorithm in [76]. First initialize $\epsilon_0 = 0.1$. For each sub-step, if $\|x_{k+1} - x_k\| < 10^{-5}$ and $\epsilon > 10^{-12}$, let $\epsilon_{k+1} = 0.9\epsilon_k$; and $\epsilon_{k+1} = \epsilon_k$ otherwise. We only do the reducing update for $\epsilon > 10^{-12}$ since an extremely small ϵ will lead to instability of the experiments. We plot the objective $J(x_k)$ versus the number of iterations.

In the second experiment, we consider robust phase retrieval 1 with sparse noise. we also sample an image as the vector we want to recover. Instead of letting $b = |Ax|$, we let $b = ||Ax_*| + e_s|$, where the error vector $e_s \in \mathbb{R}^{3136}$ is sparse. In particular, the first 100 entries of e_s are sampled independently from $N(0, 1)$ and the rest of entries are zeros. We keep the same initialization algorithm and ϵ updating strategy as in the first experiment. Since in the second chapter of the thesis, we show that when the measurement matrix is i.i.d. $N(0, 1)$ and the error is sparse, x_* should still be the global minimizer. Therefore in this case we plot the objective difference $J(x_k) - J(x_*)$ versus the number of iterations.

In the third experiment, we consider robust phase retrieval 1 with Gaussian noise. We use the same image as in the first and second example. In this case, $b = b = ||Ax| + 0.5e_g|$, where $e_g \in \mathbb{R}^{3136}$ are sampled independently from Gaussian distribution $N(0, 1)$. The same initialization algorithm is used. Meanwhile, we keep ϵ unchanged during the algorithm, in order the track the change of the subdifferentials. We plot $\|u(x_k)\|$ versus the number of iterations plot in this case. Note $T(x_k) \leq \|u(x_k)\|^2$.

5.7.2 Robust Phase Retrieval 2

A more popular formalization of phase retrieval problem is to minimize

$$\min_{x\in\mathbb{C}^n} J'(x) := \big\||Ax|^2 - b\big\|_1 \tag{5.7.2}$$

We call this robust phase retrieval 2. For each row A_i of matrix A $(1 \leq i \leq m)$, note

$$\begin{aligned} ||A_ix|^2 - b_i| &= \sqrt{b_i}||A_ix| - \sqrt{b_i}| + |A_ix||A_ix| - b_i| \\ &= \text{dist}(\sqrt{b_i}A_ix|\{-\sqrt{b_i}, \sqrt{b_i}\}) + \text{dist}(A_ix|\{0\})\text{dist}(A_ix|\{b_i, -b_i\}) \end{aligned}$$

Therefore Problem (5.7.2) fits the IRLS formalization (5.1.1). Similar as robust phase retrieval 1, we consider the noiseless case, the case where sparse noise presents and the case with Gaussian noise. Instead of minimizing Problem (5.7.2), we minimize

$$\min_x J(x) := \big\|(Ax)^2 - b^2\big\|_1$$

for convenience of illustration.

In experiment 4, vector x_* is also sampled from MXNET dataset. Let $b = |Ax_*|$. The ϵ updating strategy is the same as of experiment 1. We plot the objective $J(x_k)$ versus the number of iterations.

In experiment 5, the observation $b = ||Ax_*| + e_s|$ where the definition of e_s is the same as experiment 2. We plot the objective $J(x_k) - J(x_*)$ versus the number of iterations.

In experiment 6, the observation $b = ||Ax_*| + 0.5e_g|$ where e_g is the i.i.d. standard normal noise.

For all the experiments from 4 to 6, we also do the initialization according to [76]. We plot $u(x_k)$ versus the number of iterations. Note $u(x_k)$ belongs to the limiting subdifferential of the approximation problem evaluating at x_k.

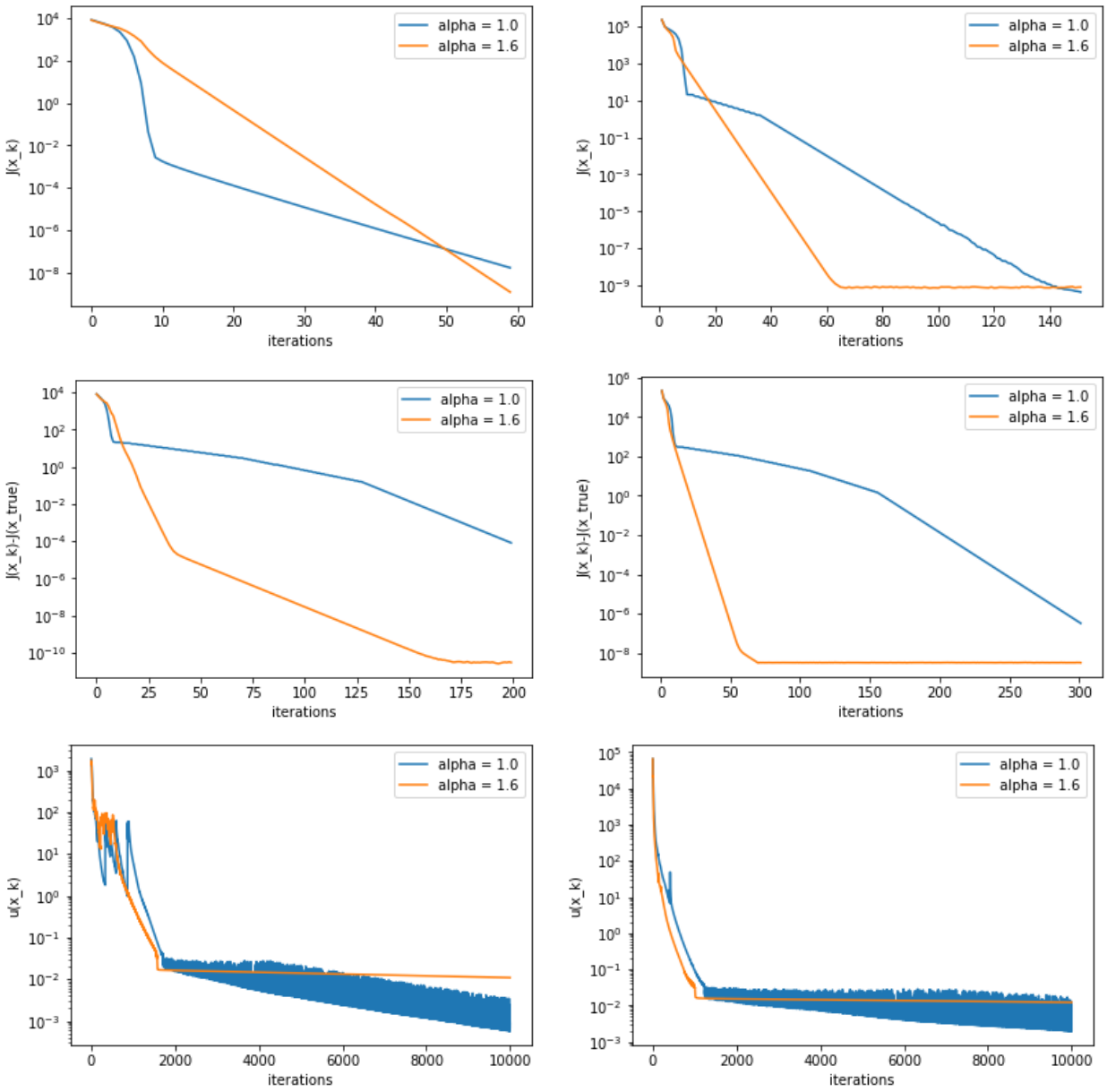

Figure 5.5: Experiments in the first row represents the noiseless case. Top left plot is experiment 1, while top right is experiment 4. The second row denotes the case where sparse noise presents. Middle left is experiment 2 and Middle right is experiment 5. The last row are the experiments where Gaussian noise presents. Top left is experiment 1. Bottom left is experiment 3 and bottom right is experiment 6 with $\epsilon = 10^{-3}$.

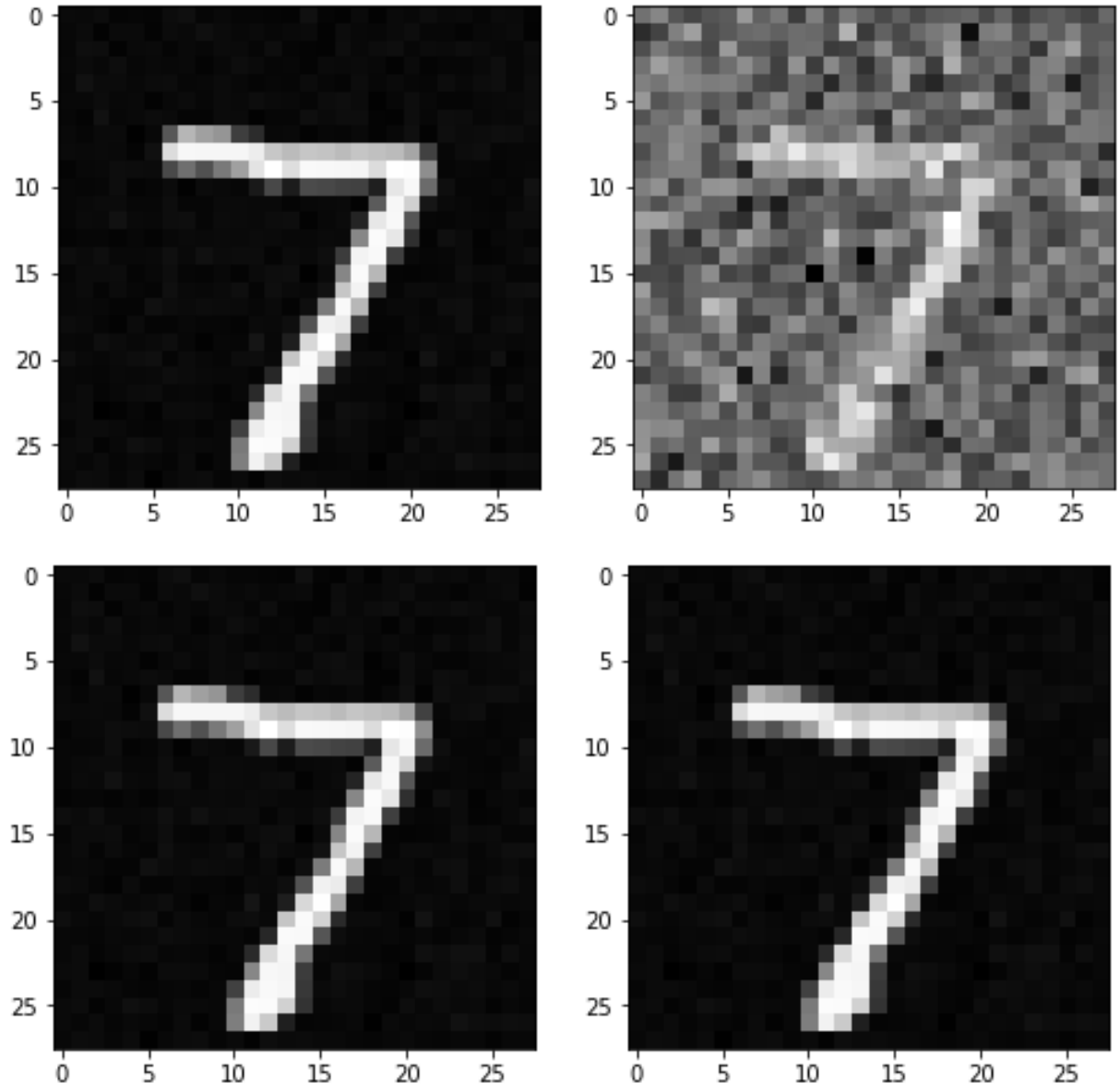

Figure 5.6: The top left figure is the original image. The top right is the image of initialization. The bottom left figure is the recovery image of IRLS with $\alpha = 1$ for experiment 1. The bottom is the recovery image of IRLS with $\alpha = 1$ for experiment 4.

5.7.3 Nesterov's Chebyshev-Rosenbrock Functions

Nesterov considered the following non-smooth function:

$$\begin{aligned}\min_{x\in\mathbb{R}^n} f(x) &= \frac{1}{4}|x_1-1| + \sum_{i=1}^{n-1}|x_{i+1}-2|x_i|+1| \\ &= \frac{1}{4}\text{dist}(x_1|C^0) + \sum_{i=1}^{n-1}[\text{dist}((\sqrt{2}x_{i+1}, 2\sqrt{2}x_i)|C^1) + \text{dist}((2x_{i+1}, 4x_i)|C^2)],\end{aligned} \tag{5.7.3}$$

where $C^0 := \{1\}$, $C^1 = \{(x,y)|\sqrt{2}x+\sqrt{2}y+1=0 \text{ or } \sqrt{2}x-\sqrt{2}y+1=0\}$ and $C = \{(x,y)|x \geq -\frac{1}{2} \text{ or } y=0\}$. The second equality of (5.7.3) is by

$$\begin{aligned}|a-2|b|+1| &= \begin{cases} ||a+1|-2|b|| & a \geq -1 \\ ||a+1|-2|b|| + 2\min\{|a+1|, 2|b|\} & a < -1 \end{cases} \\ &= \sqrt{2}\text{dist}((a,2b)|\tilde{C}_1) + 2\text{dist}((a,2b)|\tilde{C}_2)\end{aligned}$$

where $\tilde{C}_1 = \{(x,y)|x+y+1=0 \text{ or } x-y+1=0\}$ and $\tilde{C}_2 = \{(x,y)|x \geq -1 \text{ or } y=0\}$. Another non-smooth variation, also raised by Nesterov, is:

$$\begin{aligned}\tilde{f}(x) &= \frac{1}{4}(x_1-1)^2 + \sum_{i=1}^{n-1}|x_{i+1}-2x_i^2+1| \\ &= \frac{1}{4}(x_1-1)^2 + \sum_{i=1}^{m}\text{dist}((2\sqrt{2}x_{i+1}, 2\sqrt{2}x_i, 2\sqrt{2}x_{i+1})|C^3)\text{dist}((x_{i+1},x_i)|\{(-\frac{7}{8},0)\}) \\ &+ \sum_{i=1}^{m}\text{dist}((2\sqrt{2}x_{i+1}, 2\sqrt{2}x_i, 2\sqrt{2}x_{i+1})|C^3)\text{dist}(x_{i+1}|\{-\frac{9}{8}\})\end{aligned}$$

We know

$$\begin{aligned}|a-2b^2+1| &= 2|b^2+(a+\frac{7}{8})^2-(a+\frac{9}{8})^2| \\ &= 2|\sqrt{b^2+(a+\frac{7}{8})^2}+|a+\frac{9}{8}||\,|\sqrt{b^2+(a+\frac{7}{8})^2}-|a+\frac{9}{8}|| \\ &= 2\sqrt{2}[\text{dist}((a,b)|\{(-\frac{7}{8},0)\}) + \text{dist}(a|\{-\frac{9}{8}\})]\text{dist}((a,b,a)|C_3),\end{aligned}$$

where the cone $C_3 = \{(x,y,z)|(x+\frac{7}{8})^2+y^2=(z+\frac{9}{8})^2\}$. The minimizers of both problems are $\bar{x} = (1,1,...,1)^T$. We follow the initialization in the doctoral thesis of A. Engle. We have

two random initializations. We either sample each component uniformly on $[-0.2, 0.2]$ or sample each component uniformly on $[0.5, 1.5]$. The former initialization is relatively farther form the global minimizer. We call the former one the hard initialization and the latter on the easy initialization.

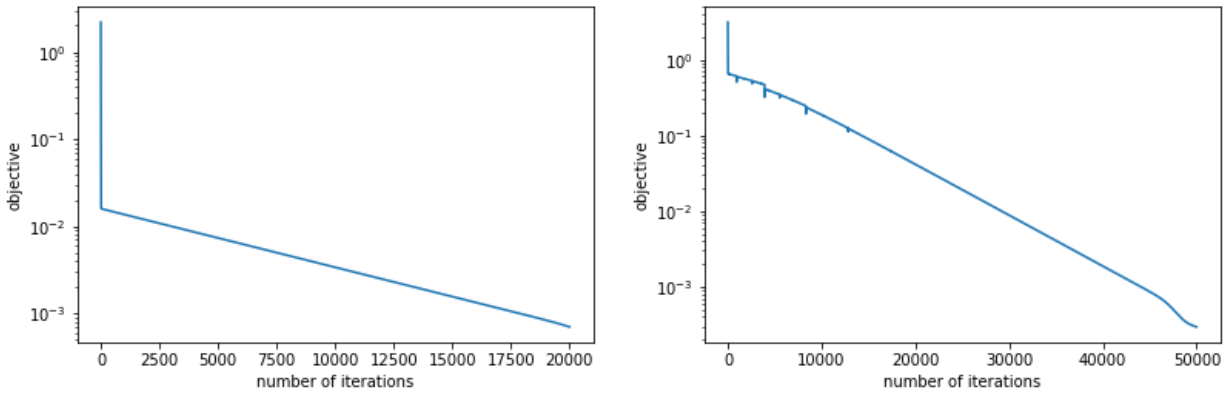

Figure 5.7: Let $n = 5$ for the first Nesterov's function. The left side is objective value vs iterations for easy initialization . The right side is objective value vs iterations for hard initialization. In the IRLS algorithm, $\alpha = 1.6$. We first initialization $\epsilon = 1$. We then update ϵ_k by $\epsilon_{k+1} = 0.999\epsilon_k$ if $\|x_{k+1} - x_k\| < 10^{-5}$, and $\epsilon_{k+1} = \epsilon_k$ otherwise.

BIBLIOGRAPHY

[1] A. Aravkin, J.V. Burke, D. Drusvyatskyi, M. Friedlander, and S. Roy. Level-set methods for convex optimization,. *Mathematical Programming, Series B*, 174.

[2] Aleksandr Aravkin, James Burke, and Daiwei He. On the global minimizers of real robust phase retrieval with sparse noise. *arXiv preprint arXiv:1905.10358*, 2019.

[3] Aleksandr Aravkin, James V Burke, and Daiwei He. Iteratively re-weighted least squares for non-convex optimization. Technical report, University of Washington, Preprint, 2019.

[4] Aleksandr Y Aravkin, James V Burke, and Daiwei He. Irls for sparse recovery revisited: Examples of failure and a remedy. *arXiv preprint arXiv:1910.07095*, 2019.

[5] Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.

[6] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[7] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[8] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[9] James Burke and Sien Deng. Weak sharp minima revisited part i: basic theory. *Control and Cybernetics*, 31:439–469, 2002.

[10] James V Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33(3):260–279, 1985.

[11] James V Burke, Frank E Curtis, Hao Wang, and Jiashan Wang. Iterative reweighted linear least squares for exact penalty subproblems on product sets. *SIAM Journal on Optimization*, 25(1):261–294, 2015.

[12] James V Burke and Sien Deng. Weak sharp minima revisited, part ii: application to linear regularity and error bounds. *Mathematical programming*, 104(2-3):235–261, 2005.

[13] James V Burke and Sien Deng. Weak sharp minima revisited, part iii: Error bounds for differentiable convex inclusions. *Mathematical Programming*, 116(1-2):37–56, 2009.

[14] James V Burke and Michael C Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

[15] James V Burke and Michael C Ferris. A gauss—newton method for convex composite optimization. *Mathematical Programming*, 71(2):179–194, 1995.

[16] J.V. Burke, F. Curtis, H. Wang, and J. Wang. Iteratively reweighted linear least squares for exact penalty subproblems on product sets. *SIAM J. Optim.*, 25:261–294, 2015.

[17] Emmanuel Candes, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *arXiv preprint math/0409186*, 2004.

[18] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[19] Emmanuel J Candès and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

[20] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[21] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[22] Emmanuel J Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[23] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[24] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

[25] Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3869–3872. IEEE, 2008.

[26] Jinghui Chen, Lingxiao Wang, Xiao Zhang, and Quanquan Gu. Robust wirtinger flow for phase retrieval with arbitrary corruption. *arXiv preprint arXiv:1704.06256*, 2017.

[27] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[28] Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.

[29] Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[30] Francis H Clarke, RJ Stern, and PR Wolenski. Proximal smoothness and the lower-c2 property. *J. Convex Anal*, 2(1-2):117–144, 1995.

[31] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best -term approximation. *Journal of the American mathematical society*, 22(1):211–231, 2009.

[32] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 63(1):1–38, 2010.

[33] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017.

[34] Laurent Demanet and Paul Hand. Stable optimizationless recovery from phaseless linear measurements. *Journal of Fourier Analysis and Applications*, 20(1):199–221, 2014.

[35] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[36] David L Donoho et al. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[37] David L Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE transactions on information theory*, 47(7):2845–2862, 2001.

[38] David L Donoho and Benjamin F Logan. Signal recovery and the large sieve. *SIAM Journal on Applied Mathematics*, 52(2):577–591, 1992.

[39] David L Donoho and Philip B Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989.

[40] Dmitriy Drusvyatskiy, Alexander D Ioffe, and Adrian S Lewis. Generic minimizing behavior in semialgebraic optimization. *SIAM Journal on Optimization*, 26(1):513–534, 2016.

[41] John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *arXiv preprint arXiv:1705.02356*, 2017.

[42] Yonina C Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[43] Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

[44] Massimo Fornasier, Holger Rauhut, and Rachel Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011.

[45] Daniel Gabay and Bertrand Mercier. *A dual algorithm for the solution of non linear variational problems via finite element approximation*. Institut de recherche d'informatique et d'automatique, 1975.

[46] Mariano Giaquinta and Giuseppe Modica. *Mathematical analysis: foundations and advanced techniques for functions of several variables*. Springer Science & Business Media, 2011.

[47] Irina F Gorodnitsky and Bhaskar D Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Transactions on signal processing*, 45(3):600–616, 1997.

[48] Rémi Gribonval and Morten Nielsen. *Sparse representations in unions of bases*. PhD thesis, INRIA, 2002.

[49] Rémi Gribonval and Morten Nielsen. Sparse representations in unions of bases. *IEEE transactions on Information theory*, 49(12):3320–3325, 2003.

[50] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[51] Ming-Jun Lai, Yangyang Xu, and Wotao Yin. Improved iteratively reweighted least squares for unconstrained smoothed \ell_q minimization. *SIAM Journal on Numerical Analysis*, 51(2):927–957, 2013.

[52] Charles Lawrence Lawson. Contribution to the theory of linear least maximum approximation. *Ph. D. dissertation, Univ. Calif.*, 1961.

[53] Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

[54] Qiuying Lin. *Sparsity and Nonconvex Nonsmooth Optimization*. PhD thesis, University of Washington, Seattle, WA, 2009.

[55] Canyi Lu, Zhouchen Lin, and Shuicheng Yan. Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Transactions on Image Processing*, 24(2):646–654, 2014.

[56] Canyi Lu, Zhouchen Lin, and Shuicheng Yan. Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Transactions on Image Processing*, 24(2):646–654, 2015.

[57] D Russell Luke, James V Burke, and Richard G Lyon. Optical wavefront reconstruction: Theory and numerical methods. *SIAM review*, 44(2):169–224, 2002.

[58] Michael Lustig, Juan M Santos, Jin-Hyung Lee, David L Donoho, and John M Pauly. Application of compressed sensing for rapid mr imaging. *SPARS,(Rennes, France)*, 2005.

[59] Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *Journal of Machine Learning Research*, 13(Nov):3441–3473, 2012.

[60] Boris S Mordukhovich. *Variational Analysis and Applications*. Springer, 2018.

[61] Michael R Osborne, Brett Presnell, and Berwin A Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

[62] Michael Robert Osborne. *Finite algorithms in optimization and data analysis*. Wiley New York, 1985.

[63] Dianne P. OLeary. Robust regression computation using iteratively reweighted least squares. *SIAM Journal on Matrix Analysis and Applications*, 11(3):466–480, 1990.

[64] Panos M Pardalos and Stephen A Vavasis. Quadratic programming with one negative eigenvalue is np-hard. *Journal of Global Optimization*, 1(1):15–22, 1991.

[65] René Poliquin and R Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838, 1996.

[66] Holger Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.

[67] R Tyrrell Rockafellar. Favorable classes of lipschitz continuous functions in subgradient optimization. 1981.

[68] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[69] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[70] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(8):1025–1045, 2008.

[71] Shriram Sarvotham, Dror Baron, Michael Wakin, Marco F Duarte, and Richard G Baraniuk. Distributed compressed sensing of jointly sparse signals. In *Asilomar conference on signals, systems, and computers*, pages 1537–1541, 2005.

[72] Dharmpal Takhar, Jason N Laska, Michael B Wakin, Marco F Duarte, Dror Baron, Shriram Sarvotham, Kevin F Kelly, and Richard G Baraniuk. A new compressive imaging camera architecture using optical-domain compression. In *Computational Imaging IV*, volume 6065, page 606509. International Society for Optics and Photonics, 2006.

[73] Yan Shuo Tan and Roman Vershynin. Phase retrieval via randomized kaczmarz: Theoretical guarantees. *arXiv preprint arXiv:1706.09993*, 2017.

[74] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[75] Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[76] Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2018.

[77] Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

[78] Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Provable non-convex phase retrieval with outliers: Median truncatedwirtinger flow. In *International conference on machine learning*, pages 1022–1031, 2016.

[79] Huishuai Zhang and Yingbin Liang. Reshaped wirtinger flow for solving quadratic system of equations. In *Advances in Neural Information Processing Systems*, pages 2622–2630, 2016.

[80] Huishuai Zhang, Yi Zhou, Yingbin Liang, and Yuejie Chi. A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. *The Journal of Machine Learning Research*, 18(1):5164–5198, 2017.

[81] Peng Zheng and Aleksandr Aravkin. Relax-and-split method for nonsmooth nonconvex problems. *arXiv preprint arXiv:1802.02654*, 2018.