Title
Non-convex Optimization Methods for Sparse and Low-rank Reconstruction

Permalink
https://escholarship.org/uc/item/31z2p19f

Author
Yin, Penghang

Publication Date
2016

Peer reviewed|Thesis/dissertation
Non-convex Optimization Methods for Sparse and Low-rank Reconstruction

DISSERTATION

submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Penghang Yin

Dissertation Committee:
Professor Jack Xin, Chair
Professor Hongkai Zhao
Professor Patrick Guidotti

2016
Dedication

To my family, teachers, friends and collaborators.
# Table of Contents

List of Figures v

List of Tables vi

Acknowledgements vii

Curriculum Vitae viii

Abstract of the Dissertation ix

1 Introduction 1

1.1 Compressed Sensing ............................................. 1

1.2 Phase Retrieval .................................................. 3

1.3 Preliminaries .................................................... 5

  1.3.1 Characterization of conditioning of measurement matrices ... 5

  1.3.2 Difference of convex functions algorithm .................... 6

  1.3.3 Alternating Direction Method of Multipliers ................. 7

2 Minimization of $\ell_1-\ell_2$ for Compressed Sensing 8

2.1 Theory of $\ell_{1-2}$ Minimization ................................ 9

  2.1.1 Exact and stable recovery ................................... 9

  2.1.2 Sparsity of local minimizers ................................ 15

2.2 Computational Approach ............................................. 18

  2.2.1 Convergence analysis ........................................ 19

  2.2.2 Solving the subproblem ...................................... 24

2.3 Hybrid Simulated Annealing ........................................ 26

2.4 Numerical Results ................................................ 31

  2.4.1 Selection of parameters ....................................... 33

  2.4.2 Exact recovery of sparse vectors ............................ 33

  2.4.3 Robust recovery in presence of noise ....................... 37

  2.4.4 MRI reconstruction ......................................... 38

3 Iterative $\ell_1$ Minimization for Non-convex Compressed Sensing 41

3.1 Iterative $\ell_1$ framework ...................................... 43

3.2 Recovery results .................................................. 46

3.3 Numerical experiments ............................................ 48
4 PhaseLiftOff: an Accurate and Stable Phase Retrieval Method Based on Difference of Trace and Frobenius Norms

4.1 Notations and Preliminaries. .................................................. 52
4.2 Exact and Stable Recovery Theory. ........................................ 55
  4.2.1 Equivalence. ................................................................. 55
  4.2.2 Exact and stable recovery under Gaussian measurements. ...... 57
  4.2.3 Computation of $\|A\|$. .................................................... 59
4.3 Algorithms. ................................................................. 60
  4.3.1 Convergence analysis. ..................................................... 61
  4.3.2 Solving the subproblem. .................................................. 65
  4.3.3 Real-valued, nonnegative signals. ...................................... 67
4.4 Numerical Experiments. ..................................................... 69
  4.4.1 Exact recovery from noise-free measurements. ....................... 69
  4.4.2 Robust recovery from noisy measurements. ......................... 70
  4.4.3 Phase retrieval in array imaging ..................................... 72

5 Conclusions and Future Work ................................................. 74

BIBLIOGRAPHY ................................................................. 75
List of Figures

Contours of three sparsity metrics. ........................................... 9
Relative error v.s. relative tolerance. ....................................... 34
Success rates using incoherent sensing matrix. .......................... 35
Success rates using randomly oversampled partial DCT matrices. .... 36
Left: Comparison of success rates of HSA algorithms. Right: Comparison of
CPU time. ............................................................................. 37
MRI reconstruction results. ....................................................... 40

\[ p'(0^+)|t| - p(|t|) \] is convex. ............................................. 43
Plots of success rates. ............................................................ 50

success rate v.s. number of measurements. ............................... 71
SNR of signal recovery v.s. noise level in measurement (in SNR dB). 72
Array imaging setup and results. ............................................. 72
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR of reconstruction (dB) under Gaussian measurements.</td>
<td>38</td>
</tr>
<tr>
<td>SNR of reconstruction (dB) using overampled DCT matrix.</td>
<td>38</td>
</tr>
<tr>
<td>Examples of sparse metrics and associated iterative $\ell_1$ scheme</td>
<td>45</td>
</tr>
<tr>
<td>Fixing $m = 4n$, $|A|$ is nearly linear in $n$ with $A$ being complex-valued Gaussian matrix.</td>
<td>60</td>
</tr>
</tbody>
</table>
Acknowledgments

I’d like to express my deepest gratitude to my advisor, Prof. Jack Xin, for continuous support of my Ph.D. study. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. His advice on both my research and career have been invaluable. He has set an excellent example to me as a successful mathematician and professor.

I am thankful to Prof. Hongkai Zhao and Prof. Patrick Guidotti for serving as my committee members. Special thanks to Prof. Long Chen and Chi Shing Sidney Tsang for their inspiring discussions on my research project. I’d also like to thank Department of Mathematics at University of California, Irvine for its generous financial support during the past four years.

Last but not least, I am greatly indebted to my family who have always been supportive of my pursuit of Ph.D. degree.

This work was partially supported by NSF grants DMS-0911277, DMS-0928427, DMS-1222507, and DMS-1522383.
Curriculum Vitae

Penghang Yin

B.S. in Mathematics, University of Science and Technology of China, 2010
M.A. in Mathematics, Arizona State University, 2012
Ph.D. in Applied Mathematics, University of California, Irvine, 2016
Abstract of the Dissertation

Non-convex Optimization Methods for Sparse and Low-rank Reconstruction

By

Penghang Yin

Doctor of Philosophy in Mathematics
University of California, Irvine, 2016
Professor Jack Xin, Chair

An algorithmic framework, based on the difference of convex functions algorithm, is proposed for minimizing difference of $\ell_1$ and $\ell_2$ norms ($\ell_{1-2}$ minimization) as well as a wide class of concave sparse metrics for compressed sensing problems. The resulting algorithm iterates a sequence of $\ell_1$ minimization problems. An exact sparse recovery theory is established to show that the proposed framework always improves on the basis pursuit ($\ell_1$ minimization) and inherits robustness from it. Numerical examples on success rates of sparse solution recovery illustrate further that, unlike most existing non-convex compressed sensing solvers in the literature, our method always out-performs basis pursuit, no matter how ill-conditioned the measurement matrix is.

As the counterpart of $\ell_{1-2}$ minimization for low-rank matrix recovery, we present a phase retrieval method via minimization of the difference of trace and Frobenius norms which we call PhaseLiftOff. The associated least squares minimization with this penalty as regularization is equivalent to the original rank-one least squares problem under a mild condition on the measurement noise. Numerical results show that PhaseLiftOff outperforms the convex PhaseLift and its non-convex variant (log-determinant regularization), and successfully recovers signals near the theoretical lower limit on the number of measurements without the noise.
Chapter 1

Introduction

1.1 Compressed Sensing

Compressed sensing (CS) has been a rapidly growing field of research in signal processing and mathematics stimulated by the foundational papers [16, 13, 31, 30] and related Bregman iteration methods [85, 51]. A fundamental issue in CS is to recover an $n$-dimensional vector $\vec{x}$ from $m \ll n$ measurements (the projection of $\vec{x}$ onto $m$ $n$-dimensional vectors), or in matrix form given $b = A\vec{x}$, where $A$ is the so-called $m \times n$ sensing (measurements) matrix. One can also view $\vec{x}$ as coefficients of a sparse linear representation of data $b$ in terms of redundant columns of matrix $A$ known as dictionary elements.

The conditioning of $A$ is related to its restricted isometry property (RIP) as well as the coherence (maximum of pairwise mutual angles) of the column vectors of $A$. Breakthrough results in CS have been established when $A$ is drawn from a Gaussian matrix ensemble or random row sampling without replacement from an orthogonal matrix (Fourier matrix), then $A$ is well-conditioned in the sense that if $\vec{x}$ is $s$-sparse ($s$ is much less than $n$), $m = O(s \log n)$ measurements suffice to recover $\vec{x}$ (the sparsest solution) with an overwhelming probability by $\ell_1$ minimization or the basis pursuit (BP) problem [13, 22]:

\[
\min_{\vec{x}} \|x\|_1 \quad \text{subject to} \quad Ax = b.
\]

In the above formulation, $\ell_1$ norm works as the convex relaxation of $\ell_0$ that counts the
nonzeros. Such a matrix $A$ has incoherent column vectors. On the other hand, if columns of $A$ are coherent enough, such as those arising in discretization of continuum imaging problems (radar and medical imaging) when the grid spacing is below the Rayleigh threshold [36], $\ell_1$ minimization may not give the sparsest solution [36, 82].

In the past decade, great efforts have been devoted to explore efficient and stable algorithms for solving BP problem and its associated $\ell_1$-regularized problem (also called lasso [72]):

$$
\min_x \frac{1}{2} \|Ax - b\|^2_2 + \lambda \|x\|_1,
$$

where $\lambda > 0$ is a free parameter. Bregman iterative method, now known to be equivalent to the augmented Lagrangian method, was proposed to solve the BP problem by Yin et al. [85]. There are many state-of-the-art algorithms available for lasso problem (1.2), such as the split Bregman [51], being equivalent to ADMM [8, 33], FPC [53], FISTA [4] among others [76, 81, 86, 77].

Non-convex (concave) functions, such as $\ell_p$ (quasi-)norm ($p < 1$) [18, 19] and log-det functional [17], have also been proposed as alternatives to $\ell_0$. Such non-Lipschitz continuous metrics usually require additional smoothing in minimization to avoid division by zero and to enhance sparsity. Besides, a general class of penalty functions satisfying unbiasedness, sparsity and continuity can be found in [35, 61]. While non-convex metrics are generally more challenging to minimize, they have advantages over the convex $\ell_1$ norm. Simply put, non-convex CS enables one to reconstruct the sparse signal of interest from substantially fewer measurements.

On the computational side, researchers have observed that under certain conditions on the sensing matrix $A$ (e.g. when columns of $A$ are sufficiently randomized), several non-convex CS solvers do produce solutions of better quality [17, 20, 59, 27, 45], even though none of them theoretically guarantees convergence to a global minimum. Algorithms that directly tackle the $\ell_0$ minimization include Compressive Sampling Matching Pursuit (CoSaMP) [66] which is a greedy method amongst variants of orthogonal matching pursuit (OMP) [73], iterative hard thresholding (IHT) algorithm [6, 5] and penalty decomposition method [64],
whereas iteratively reweighted least squares (IRLS) [27, 20, 59] and iteratively reweighted $\ell_1$ (IRL1) [24, 91, 17, 45] can be applied to minimize non-convex proxies of $\ell_0$.

1.2 Phase Retrieval

Phase retrieval has been a long standing problem in imaging sciences such as X-ray crystallography, electron microscopy, array imaging, optics, signal processing, [48, 42, 43, 55, 65] among others. It concerns with signal recovery when only the amplitude measurements (say of its Fourier transform) are available. Major recent advances have been made for phase retrieval by formulating it as a matrix completion and rank one minimization problem (PhaseLift) which is relaxed and solved as a convex trace (nuclear) norm minimization problem under sufficient measurement conditions [11, 21, 10, 9]; see also [2, 1] for related work. An alternative viable approach makes use of random masks in measurements to achieve uniqueness of solution with high probability [37, 38].

The phase retrieval problem aims to reconstruct an unknown signal $\hat{x} \in \mathbb{C}^n$ satisfying $m$ quadratic constraints

$$|\langle a_i, \hat{x} \rangle|^2 = b_i, \quad i = 1, \ldots, m,$$

where the bracket is inner product, $a_i \in \mathbb{C}^n$ and $b_i \in \mathbb{R}$. Letting $X = xx^* \in \mathbb{C}^{n \times n}$ be a rank-1 positive semidefinite matrix ($*$ is conjugate transpose), one can recast quadratic measurements as linear ones about $X$:

$$|\langle a_i, x \rangle|^2 = a_i^* X a_i, \quad i = 1, \ldots, m.$$

Thus we can define a linear operator $A$ uniquely determined by the measurement matrix $A = (a_1, \ldots, a_m) \in \mathbb{C}^{n \times m}$:

$$\mathbb{H}^{n \times n} \rightarrow \mathbb{R}^m$$

$$X \quad \mapsto \quad \text{diag}(A^* X A)$$

which maps Hermitian matrices into real-valued vectors. Denote $\hat{\hat{x}}^*$ by $\hat{X}$, and suppose $b = (b_1, \ldots, b_m)^T = A(\hat{X}) \in \mathbb{R}^m$ is the measurement vector. Then the phase retrieval
becomes the feasibility problem, being equivalent to a rank minimization problem:

$$\begin{align*}
\text{find } & \quad X \in \mathbb{C}^{n \times n} \\
\text{s.t.} & \quad \mathcal{A}(X) = b \\
& \quad X \succeq 0 \\
& \quad \text{rank}(X) = 1.
\end{align*}$$

(1.3)

To arrive at the original solution $\hat{x}$ to the phase retrieval problem, one needs to factorize the solution $\hat{X}$ of (1.3) as $\hat{x}\hat{x}^*$. It gives $\hat{x}$ up to multiplication by a constant scalar with unit modulus (a constant phase factor), because if $\hat{x}$ solves the phase retrieval problem, so does $c\hat{x}$, for any $c \in \mathbb{C}$ with $|c| = 1$. At least $3n - 2$ intensity measurements are necessary to guarantee uniqueness (up to a constant phase factor) of the solution to (1.3) [41], whereas $4n - 2$ generic measurements suffice for uniqueness with probability one [2].

Instead of (1.3), Candès et al. [9, 11] suggest solving the convex PhaseLift problem, namely minimizing the trace norm as a convex surrogate for the rank functional:

$$\min_{X \in \mathbb{C}^{n \times n}} \text{Tr}(X) \quad \text{s.t.} \quad \mathcal{A}(X) = b, \; X \succeq 0.$$  

It is shown in [10] that if each $a_i$ is Gaussian or uniformly sampled on the sphere, then with high probability, $m = O(n)$ measurements are sufficient to recover the ground truth $\hat{X}$ via PhaseLift. For the noisy case, the following variant is considered in [10]:

$$\min_{X \in \mathbb{C}^{n \times n}} \|\mathcal{A}(X) - b\|_1 \quad \text{s.t.} \quad X \succeq 0.$$  

In this case, $b = \mathcal{A}(\hat{X}) + e$ is contaminated by the additive noise $e \in \mathbb{R}^m$. Similarly, $m = O(n)$ measurements guarantee stable recovery in the sense that the solution $X^{\text{opt}}$ satisfies $\|X^{\text{opt}} - \hat{X}\|_F = O\left(\frac{\|e\|_1}{m}\right)$ with probability close to 1. On the computational side, the regularized trace-norm minimization is considered in [9, 11]:

$$\min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2 + \lambda\text{Tr}(X) \quad \text{s.t.} \quad X \succeq 0.$$  

(1.4)

If there is no noise, a tiny value of $\lambda$ would work well. However, when the measurements are
noisy, determining $\lambda$ requires extra work, such as employing the cross validation technique.

Besides PhaseLift and its nonconvex variant (log-determinant) proposed in [9], related formulations such as feasibility problem or weak PhaseLift [28] and PhaseCut [74] also lead to phase retrieval solutions under certain measurement conditions. PhaseCut is a convex relaxation where trace minimization is in the form $\min_U \text{Tr}(UM)$, where $M$ (resp., $U$) is a known (resp., unknown) positive semidefinite Hermitian matrix, and $\text{diag}(U) = 1$. The exact recovery (tightness) conditions for PhaseLift and PhaseCut are studied in [74] and references therein.

1.3 Preliminaries

1.3.1 Characterization of conditioning of measurement matrices

Restricted isometry property (RIP) introduced by Candès et al. [16] is one of the most used frameworks for CS, which characterizes matrices that are nearly orthonormal.

**Definition 1.3.1.** For each number $s$, $s$-restricted isometry constant of $A$ is the smallest $\delta_s \in (0, 1)$ such that

$$(1 - \delta_s) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_s) \|x\|_2^2$$

for all $x \in \mathbb{R}^n$ with sparsity of $s$. The matrix $A$ is said to satisfy the $s$-RIP with $\delta_s$.

Sensing matrices with small $\delta_s$ are suitable for reconstruction of sparse signals [16, 15]. It has been shown that with overwhelming probability, random Gaussian, random Bernoulli, and random partial Fourier matrices satisfy the RIP (with small restricted isometry constants) [16, 26, 68]. Given a deterministic matrix $A$, it is generally NP-hard however, to verify whether $A$ is a RIP matrix or not [3]. Another commonly used CS concept is the so-called mutual coherence or coherence [32] for short.

**Definition 1.3.2.** The coherence of a matrix $A$ is the maximum absolute value of the cross-correlations between the columns of $A$, namely

$$\mu(A) := \max_{i \neq j} \frac{|A_i^T A_j|}{\|A_i\|_2 \|A_j\|_2}.$$
Coherence is closely related to the RIP yet easy to examine. Specifically, a matrix satisfying some RIP tends to have small coherence or to be incoherent. Conversely, a highly coherent matrix is unlikely to possess small restricted isometry constants.

### 1.3.2 Difference of convex functions algorithm

The DCA is a descent method without line search developed by Tao and An [70, 71]. It addresses the problem of minimizing a function of the form \( f(x) = g(x) - h(x) \) on the space \( \mathbb{R}^n \), with \( g, h \) being lower semicontinuous proper convex functions:

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

\( g - h \) is called a DC decomposition of \( f \), while the convex functions \( g \) and \( h \) are DC components of \( f \). The DCA involves the construction of two sequences \( \{x^k\} \) and \( \{y^k\} \), the candidates for optimal solutions of primal and dual programs respectively. At the \((k+1)\)-th step, we choose a subgradient of \( h(x) \) at \( x^k \), namely \( y^k \in \partial h(x^k) \). We then linearize \( h \) at \( x^k \), which permits a convex upper envelope of \( f \). More precisely,

\[
f(x) = g(x) - h(x) \leq g(x) - (h(x^k) + \langle y^k, x - x^k \rangle), \quad \forall x \in \mathbb{R}^n
\]

with equality at \( x = x^k \).

By iteratively computing

\[
\begin{align*}
&y^k \in \partial h(x^k), \\
x^{k+1} = \arg\min_{x \in \mathbb{R}^n} g(x) - (h(x^k) + \langle y^k, x - x^k \rangle)
\end{align*}
\]

we have

\[
f(x^k) \geq g(x^{k+1}) - (h(x^k) + \langle y^k, x^{k+1} - x^k \rangle) \geq g(x^{k+1}) - h(x^{k+1}) = f(x^{k+1}).
\]

This generates a monotonically decreasing sequence \( \{f(x^k)\} \), leading to its convergence if \( f(x) \) is bounded from below.
1.3.3 Alternating Direction Method of Multipliers

Alternating Direction Method of Multipliers (ADMM) [7] solves problems in the form

$$\min_{x \in \mathbb{R}^n, z \in \mathbb{R}^m} f(x) + g(z) \quad \text{subject to} \quad Ax + Bz = c$$

with \(f\) and \(g\) being convex, \(A \in \mathbb{R}^{p \times n}\), \(B \in \mathbb{R}^{p \times m}\) and \(c \in \mathbb{R}^p\).

As in the method of multipliers, we form the augmented Lagrangian

$$L_\delta = f(x) + g(z) + \langle y, Ax + Bz - c \rangle + \frac{\delta}{2} \|Ax + Bz - c\|^2_2.$$ 

with \(\delta > 0\). ADMM consists of the following iterations:

$$x^{k+1} := \arg \min_x L_\delta(x, z^k, y^k)$$

$$z^{k+1} := \arg \min_z L_\delta(x^{k+1}, z, y^k)$$

$$y^{k+1} := y^k + \delta(Ax^{k+1} + Bz^{k+1} - c).$$
Chapter 2

Minimization of $\ell_{1-2}$ for Compressed Sensing

In this chapter, we study minimization of the non-convex yet Lipschitz continuous metric $\ell_{1-2}$, for sparse signal recovery and compare it with various CS solvers. $\ell_{1-2}$ was first addressed in [34] by Esser et al. in the context of nonnegative least squares problems and group sparsity with applications to spectroscopic imaging. A contour plot of $\ell_{1-2}$ metric can be seen in Fig. 2.1. Here we mainly discuss the constrained $\ell_{1-2}$ minimization problem

\begin{equation}
(2.1) \quad \min_{x \in \mathbb{R}^n} \|x\|_1 - \|x\|_2 \quad \text{subject to} \quad Ax = b
\end{equation}

and the unconstrained one

\begin{equation}
(2.2) \quad \min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda(\|x\|_1 - \|x\|_2),
\end{equation}

where $A \in \mathbb{R}^{m \times n}$ is an under-determined sensing matrix of full row rank and $b \in \mathbb{R}^m \setminus \{0\}$. We shall focus on the theoretical aspects such as sparsity of minimizers and convergence of minimization algorithms, and refer to the companion paper [62] for more extensive computational study with applications to imaging problems.

**Notations.** Let us fix some notations. For any $x, y \in \mathbb{R}^n$, $\langle x, y \rangle = x^T y$ is their inner product. $\text{supp}(x) := \{1 \leq i \leq n : x_i \neq 0\}$ denotes the support of $x$, and $\|x\|_0 := |\text{supp}(x)|$
Figure 2.1: Contours of three sparsity metrics.

is cardinality of supp(x). $B_r(x) = \{ y \in \mathbb{R}^n : \| y - x \|_2 < r \}$ denotes the $n$-dimensional Euclidean ball centered at $x$ with radius $r > 0$. Let $T \subseteq \{1, \ldots, n\}$ be an index set, and let $|T|$ be the cardinality of $T$. Moreover, for $A \in \mathbb{R}^{m \times n}$, $A_T \in \mathbb{R}^{m \times |T|}$ is the submatrix of $A$ with column indices $T$. $I_m$ is the identity matrix of dimension $m$. The $\text{sgn}(x)$ is the signum function defined as

$$\text{sgn}(x) := \begin{cases} 
1 & \text{if } x > 0, \\
-1 & \text{if } x < 0, \\
0 & \text{if } x = 0.
\end{cases}$$

2.1 Theory of $\ell_{1-2}$ Minimization

2.1.1 Exact and stable recovery

We have the following fundamental properties of the function $\|x\|_1 - \|x\|_2$, which will be frequently invoked later in the proofs.

**Lemma 2.1.1.** Suppose $x \in \mathbb{R}^n \setminus \{0\}$, $\Lambda = \text{supp}(x)$ and $\|x\|_0 = s$, then

(a) $(n - \sqrt{n}) \min_i |x_i| \leq \|x\|_1 - \|x\|_2 \leq (\sqrt{n} - 1)\|x\|_2$.

(b) $(s - \sqrt{s}) \min_{i \in \Lambda} |x_i| \leq \|x\|_1 - \|x\|_2 \leq (\sqrt{s} - 1)\|x\|_2$.

(c) $\|x\|_1 - \|x\|_2 = 0$ if and only if $s = 1$. 


Proof. (a) The upper bound is immediate from Cauchy-Schwarz inequality. To show the lower bound, without loss of generality, let us assume

$$|x_1| \geq |x_2| \geq \cdots \geq |x_n|.$$ 

Let $t = \lfloor \sqrt{n} \rfloor$, then we have

\begin{equation}
\lVert x \rVert_2 \leq \sum_{i=1}^{t} |x_i| + (\sqrt{n} - t)|x_{t+1}|.
\end{equation}

To see this, we square both sides

$$\sum_{i=1}^{n} |x_i|^2 \leq \sum_{i=1}^{t} |x_i|^2 + \sum_{i=1}^{t} \sum_{j=1}^{t} |x_i||x_j| + 2(\sqrt{n} - t)|x_{t+1}| \sum_{i=1}^{t} |x_i| + (\sqrt{n} - t)^2|x_{t+1}|^2$$

or equivalently,

$$\sum_{i=t+1}^{n} |x_i|^2 \leq \sum_{i=1}^{t} \sum_{j=1}^{t} |x_i||x_j| + 2(\sqrt{n} - t)|x_{t+1}| \sum_{i=1}^{t} |x_i| + (\sqrt{n} - t)^2|x_{t+1}|^2$$

Then (4.5) holds because

$$\sum_{i=t+1}^{n} |x_i|^2 \leq \sum_{i=t+1}^{n} |x_{t+1}|^2 = (n-t)|x_{t+1}|^2$$

and

$$\sum_{i=1}^{t} \sum_{j=1}^{t} |x_i||x_j| + 2(\sqrt{n} - t)|x_{t+1}| \sum_{i=1}^{t} |x_i| + (\sqrt{n} - t)^2|x_{t+1}|^2$$

$$\geq \sum_{i=1}^{t} \sum_{j=1}^{t} |x_{t+1}|^2 + 2(\sqrt{n} - t)|x_{t+1}| \sum_{i=1}^{t} |x_{t+1}| + (\sqrt{n} - t)^2|x_{t+1}|^2$$

$$= ((t^2 - t) + 2(\sqrt{nt} - t^2) + (\sqrt{n} - t)^2)|x_{t+1}|^2$$

$$= (n-t)|x_{t+1}|^2.$$
It follows from (4.5) that

\[ \|x\|_1 - \|x\|_2 \geq \|x\|_1 - (\sum_{i=1}^{t} |x_i| + (\sqrt{n} - t)|x_{t+1}|) = (t + 1 - \sqrt{n})|x_{t+1}| + \sum_{i=t+2}^{n} |x_i| \]

\[ \geq (t + 1 - \sqrt{n})|x_n| + \sum_{i=t+2}^{N} |x_n| = (n - \sqrt{n})|x_n| \]

\[ = (n - \sqrt{n}) \min_i |x_i|. \]

(b) Note that \( \|x\|_1 - \|x\|_2 = \|x_\Lambda\|_1 - \|x_\Lambda\|_2 \), (b) follows as we apply (a) to \( x_\Lambda \).

(c) If \( \|x\|_1 - \|x\|_2 = 0 \), then by (b)

\[ 0 = \|x\|_1 - \|x\|_2 \geq (s - \sqrt{s}) \min_{i \in \Lambda} |x_i| \]

So \( s - \sqrt{s} \leq 0 \) and thus \( s = 1 \).

The other direction is trivial.

A RIP based sufficient condition was derived in [15] for exact recovery of BP (1.1), here we derive an analogous condition for that of \( \ell_{1-2} \) minimization, demonstrating the capability of \( \ell_{1-2} \) to promote sparsity.

**Theorem 2.1.1.** Let \( \bar{x} \) be any vector with sparsity of \( s \) satisfying

\[ a(s) = \left( \frac{\sqrt{3s} - 1}{\sqrt{s} + 1} \right)^2 > 1, \]

and let \( b = A\bar{x} \), suppose \( A \) satisfies the following condition

\[ (2.4) \quad \delta_{3s} + a(s)\delta_{4s} < a(s) - 1, \]

then \( \bar{x} \) is the unique solution to (2.1).

**Proof.** The proof generally follows the lines of [15]. Let \( x \) be any feasible solution satisfying the constraint \( Ax = b \) yet with a smaller objective value, i.e.

\[ (2.5) \quad \|x\|_1 - \|x\|_2 \leq \|\bar{x}\|_1 - \|\bar{x}\|_2. \]
We write \( x = \bar{x} + v \) with \( v \in \ker(A) \), and want to show \( v = 0 \).
Let \( \Lambda = \text{supp}(\bar{x}) \), we further decompose \( v \) as \( v = v_{\Lambda} + v_{\Lambda^c} \). Then (2.5) becomes
\[
\| \bar{x} + v_{\Lambda} + v_{\Lambda^c} \|_1 - \| \bar{x} + v_{\Lambda} + v_{\Lambda^c} \|_2 \leq \| \bar{x} \|_1 - \| \bar{x} \|_2.
\]
On the other hand,
\[
\| \bar{x} + v_{\Lambda} + v_{\Lambda^c} \|_1 - \| \bar{x} + v_{\Lambda} + v_{\Lambda^c} \|_2 = \| \bar{x} + v_{\Lambda} \|_1 + \| v_{\Lambda^c} \|_1 - \| \bar{x} + v_{\Lambda} + v_{\Lambda^c} \|_2
\]
\[
\geq \| \bar{x} \|_1 - \| v_{\Lambda} \|_1 + \| v_{\Lambda^c} \|_1 - \| \bar{x} \|_2 - \| v_{\Lambda} \|_2 - \| v_{\Lambda^c} \|_2,
\]
So \( v \) must obey the following inequality constraint
\[
(2.6) \quad \| v_{\Lambda} \|_1 + \| v_{\Lambda} \|_2 \geq \| v_{\Lambda^c} \|_1 - \| v_{\Lambda^c} \|_2.
\]
Arrange the indices in \( \Lambda^c \) in order of decreasing magnitude of \( v_{\Lambda^c} \) and divide \( \Lambda^c \) into subsets of size \( 3s \). Then \( \Lambda^c = \Lambda_1 \cup \Lambda_2 \cup \cdots \cup \Lambda_l \), where each \( \Lambda_i \) contains \( m \) indices probably except \( \Lambda_l \). Denoting \( \Lambda_0 = \Lambda \cup \Lambda_1 \) and using the RIP of \( A \), we have
\[
0 = \| Av \|_2 = \| A_{\Lambda_0} v_{\Lambda_0} + \sum_{i=2}^l A_{\Lambda_i} v_{\Lambda_i} \|_2 \geq \| A_{\Lambda_0} v_{\Lambda_0} \|_2 - \| \sum_{i=2}^l A_{\Lambda_i} v_{\Lambda_i} \|_2
\]
\[
\geq \| A_{\Lambda_0} v_{\Lambda_0} \|_2 - \sum_{i=2}^l \| A_{\Lambda_i} v_{\Lambda_i} \|_2
\]
\[
(2.7) \quad \geq \sqrt{1 - \delta_{3s}} \| v_{\Lambda_0} \|_2 - \sqrt{1 + \delta_{3s}} \sum_{i=2}^l \| v_{\Lambda_i} \|_2
\]
Now we set an upper bound on \( \sum_{i=2}^l \| v_{\Lambda_i} \|_2 \). For each \( t \in \Lambda_i, i \geq 2 \),
\[
| v_t | \leq \min_{r \in \Lambda_{i-1}} | v_r | \leq \frac{\| v_{\Lambda_{i-1}} \|_1 - \| v_{\Lambda_{i-1}} \|_2}{3s - \sqrt{3s}},
\]
where the second inequality follows from Lemma 2.1.1(a). Then it follows that
\[
\| v_{\Lambda_i} \|_2 \leq \sqrt{3s} \frac{\| v_{\Lambda_{i-1}} \|_1 - \| v_{\Lambda_{i-1}} \|_2}{3s - \sqrt{3s}} = \frac{\| v_{\Lambda_{i-1}} \|_1 - \| v_{\Lambda_{i-1}} \|_2}{\sqrt{3s} - 1},
\]
12
and

\[(2.8) \quad \sum_{i=2}^{l} \|v_{\Lambda_i}\|_2 \leq \sum_{i=1}^{l-1} \frac{\|v_{\Lambda_i}\|_1 - \|v_{\Lambda_i}\|_2}{\sqrt{3s} - 1} \leq \frac{\sum_{i=1}^{l} \|v_{\Lambda_i}\|_1 - \sum_{i=1}^{l} \|v_{\Lambda_i}\|_2}{\sqrt{3s} - 1}.\]

Note that in (4.7)

\[\sum_{i=1}^{l} \|v_{\Lambda_i}\|_1 = \|v_{\Lambda^c}\|_1 \quad \text{and} \quad \sum_{i=1}^{l} \|v_{\Lambda_i}\|_2 \geq \sqrt{\sum_{i=1}^{l} \|v_{\Lambda_i}\|_2^2} = \|v_{\Lambda^c}\|_2.\]

Combining (4.7) and (4.6) gives

\[\sum_{i=2}^{l} \|v_{\Lambda_i}\|_2 \leq \frac{\|v_{\Lambda^c}\|_1 - \|v_{\Lambda^c}\|_2}{\sqrt{3s} - 1} \leq \frac{\|v_{\Lambda}\|_1 + \|v_{\Lambda}\|_2}{\sqrt{3s} - 1} \leq \frac{(\sqrt{s} + 1)\|v_{\Lambda}\|_2}{\sqrt{3s} - 1} = \frac{\|v_{\Lambda}\|_2}{\sqrt{a(s)}}.\]

So it follows from (4.14) that

\[0 \geq \sqrt{1 - \delta_{4s}} \|v_{\Lambda_0}\|_2 - \frac{\sqrt{1 + \delta_{3s}}}{\sqrt{a(s)}} \|v_{\Lambda}\|_2 \geq \sqrt{1 - \delta_{4s}} \|v_{\Lambda_0}\|_2 - \frac{\sqrt{1 + \delta_{3s}}}{\sqrt{a(s)}} \|v_{\Lambda_0}\|_2.\]

Since (2.4) amounts to

\[\sqrt{1 - \delta_{4s}} - \frac{\sqrt{1 + \delta_{3s}}}{\sqrt{a(s)}} > 0,\]

we have \(v_{\Lambda_0} = 0\). This implies \(v = 0\), which completes the proof. \(\square\)

**Remark 2.1.1.** (2.4) can be rewritten as

\[\delta_{3s} < a(s)(1 - \delta_{4s}) - 1.\]

Note that the RIP condition for exact recovery of BP derived in [15] reads

\[(2.9) \quad \delta_{3s} + 3\delta_{4s} < 2,\]

or equivalently

\[\delta_{3s} < 3(1 - \delta_{4s}) - 1.\]

The condition (2.4) required for \(\ell_{1-2}\) exact recovery appears more stringent than (2.9) for \(\ell_1\)
recovery since $a(s) < 3$ (and thus also stronger than the RIP for $\ell_p$ recovery with $0 < p < 1$ [18]). However, this does not mean $\ell_1$ norm is superior to $\ell_{1-2}$ in terms of sparsity promoting. On the contrary, in Section 2.4 it will be shown numerically that $\ell_{1-2}$ penalty consistently outperforms $\ell_1$. Besides possible technical issues lying in the proof (e.g. the estimate in (4.6) is in fact not sharp), another explanation can be that a RIP-based condition is just a sufficient condition to guarantee that a measurement matrix $A$ fits for exact reconstruction. It happens that two matrices have exactly the same performance and yet one satisfies RIP whereas the other does not [90].

**Remark 2.1.2.** The assumptions of Theorem 2.1 require $a(s) = \left( \frac{\sqrt{3s} - 1}{\sqrt{s} + 1} \right)^2 > 1$, which implies $s \geq 8$. Then a natural question is whether the uniqueness of $\bar{x}$ still holds for the case $1 \leq s \leq 7$. First of all, when $s = 1$, any minimizer of $\|x\|_1 - \|x\|_2$ other than $\bar{x}$ must be 1-sparse (and be a feasible solution of $Ax = b$). So the RIP condition to guarantee uniqueness is just $\delta_2 < 1$. When $s \geq 2$, we redefine $a(s)$ as $\left( \frac{\sqrt{3s} - 1}{\sqrt{s} + 1} \right)^2$. It is easy to check that $a(s) > 1$ for $s \geq 2$. By similar argument, we can show that the following RIP condition suffices for the uniqueness of $\bar{x}$:

$$\delta_{6s} + a(s)\delta_{7s} < a(s) - 1.$$ 

Similar to [14], we also establish the following stable recovery of $\ell_{1-2}$ when measurements are contaminated by noises.

**Theorem 2.1.2.** Under the assumptions of Theorem 2.1.1 except that $b = A\bar{x} + e$, where $e \in \mathbb{R}^m$ is any perturbation with $\|e\|_2 \leq \tau$, we have that the solution $x^{opt}$ to the following variant of problem (2.1)

$$\min_{x \in \mathbb{R}^n} \|x\|_1 - \|x\|_2 \quad \text{subject to} \quad \|Ax - b\|_2 \leq \tau$$

obeys $\|x^{opt} - \bar{x}\|_2 \leq C_s \tau$ for some constant $C_s > 0$ depending on $\delta_{3s}$ and $\delta_{4s}$.

**Proof.** Let $\Lambda$ be the support of $\bar{x}$ and $x^{opt} = \bar{x} + v$. Then starting from $\|x^{opt}\|_1 - \|x^{opt}\|_2 \leq$
\[ \| \bar{x} \|_1 - \| \bar{x} \|_2 \] and repeating the arguments in the proof of Theorem 2.1.1, we obtain

\[ \sum_{i=2}^{l} \| v_{\Lambda_i} \|_2 \leq \frac{\| v_{\Lambda_1} \|_2}{\sqrt{a(s)}} \]  

and

\[ \| Av \|_2 \geq \left( \sqrt{1 - \delta_{4s}} - \frac{\sqrt{1 + \delta_{3s}}}{\sqrt{a(s)}} \right) \| v_{\Lambda_0} \|_2. \]  

From (2.10) it follows that

\[ \| v \|_2 = \sqrt{\| v_{\Lambda_0} \|_2^2 + \sum_{i=2}^{l} \| v_{\Lambda_i} \|_2^2} \leq \sqrt{\| v_{\Lambda_0} \|_2^2 + \frac{\| v_{\Lambda_1} \|_2^2}{a(s)}} \leq \sqrt{1 + \frac{1}{a(s)}} \| v_{\Lambda_0} \|_2, \]

so (2.11) becomes

\[ \| Av \|_2 \geq \frac{\sqrt{a(s)(1 - \delta_{4s})} - \sqrt{1 + \delta_{3s}}}{\sqrt{1 + a(s)}} \| v \|_2. \]  

On the other hand, since \( \| A\bar{x} - b \|_2 \leq \tau \) and \( \| Ax^{\text{opt}} - b \|_2 \leq \tau \), by the triangular inequality,

\[ \| Av \|_2 = \| (Ax^{\text{opt}} - b) - (A\bar{x} - b) \|_2 \leq \| A\bar{x} - b \|_2 + \| Ax^{\text{opt}} - b \|_2 \leq 2\tau. \]

Combining (2.12) and (2.13), we have \( \| v \|_2 \leq C_s \tau \), where \( C_s := \frac{2\sqrt{1+a(s)}}{\sqrt{a(s)(1-\delta_{4s})}-\sqrt{1+\delta_{3s}}} > 0. \)  

**Remark 2.1.3.** The upper bound \( C_s \tau \) of the approximation error basically relies on how well the RIP condition (2.4) is satisfied. \( C_s \) is \( O(1) \) if \( \delta_{3s} \) and \( \delta_{4s} \) are small and \( a(s) \gg 1 \).

### 2.1.2 Sparsity of local minimizers

Next we shall prove that local minimizers of the problems (2.1) and (2.2) possess certain sparsity in the sense that they only extract linearly independent columns from the sensing matrix \( A \), whether \( A \) satisfies any RIP or not. In other words, minimizing \( \ell_1 - 2 \) will rule out redundant columns of \( A \). It is worth noting that similar results was proved in [15] for the \( \ell_p \) unconstrained problem using the second-order optimality condition. This demonstrates an
advantage of non-convex sparsity metrics over the convex $\ell_1$ norm.

**Theorem 2.1.3.** Let $x^*$ be a local minimizer of the constrained problem (2.1) and $\Lambda^* = \text{supp}(x^*)$, then $A_{\Lambda^*}$ is of full column rank, i.e. the columns of $A_{\Lambda^*}$ are linearly independent.

**Proof.** The proof simply uses the definition of local minimizer. Suppose the columns of $A_{\Lambda^*}$ are linearly dependent, then there exists $v \in \ker(A) \setminus \{0\}$ such that $\text{supp}(v) \subseteq \Lambda^*$. For any fixed neighborhood $B_r(x^*)$ of $x^*$, we scale $v$ so that

$$
\|v\|_2 < \min\{\min_{i \in \Lambda^*} |x^*_i|, r\}.
$$

Consider two feasible vectors in $B_r(x^*)$, $\hat{x} = x^* + v$ and $\bar{x} = x^* - v$. Since $\text{supp}(v) \subseteq \Lambda^*$, we have $\text{supp}(\hat{x}) \subseteq \Lambda^*$ and $\text{supp}(\bar{x}) \subseteq \Lambda^*$. Moreover,

$$
(x^* \pm v)_i = x^*_i \pm v_i = \text{sgn}(x^*_i)(|x^*_i| \pm \text{sgn}(x^*_i)v_i), \quad \forall i \in \Lambda^*.
$$

The above implies $\text{sgn}(\hat{x}_i) = \text{sgn}(\bar{x}_i) = \text{sgn}(x^*_i)$ for all $i \in \Lambda^*$, since

$$
|x^*_i| \pm \text{sgn}(x^*_i)v_i \geq |x^*_i| - |v_i| \geq \min_{i \in \Lambda^*} |x^*_i| - \|v\|_2 > 0, \quad \forall i \in \Lambda^*.
$$

In other words, $x^*$, $\hat{x}$ and $\bar{x}$ are located in the same orthant. It follows that

$$
(2.14) \quad \|x^*\|_1 = \frac{1}{2}\|\hat{x} + \bar{x}\|_1 = \frac{1}{2}\|\hat{x}\|_1 + \frac{1}{2}\|\bar{x}\|_1,
$$

and

$$
(2.15) \quad \|x^*\|_2 = \frac{1}{2}\|\hat{x} + \bar{x}\|_2 < \frac{1}{2}\|\hat{x}\|_2 + \frac{1}{2}\|\bar{x}\|_2.
$$

(2.14) holds since $\hat{x}$ and $\bar{x}$ are in the same orthant, and (2.15) holds because of the fact that $\hat{x}$ and $\bar{x}$ are not collinear since they both satisfy the linear constraint $Ax = b$. So

$$
\|x^*\|_1 - \|x^*\|_2 > \frac{1}{2}(\|\hat{x}\|_1 - \|\hat{x}\|_2 + \|\bar{x}\|_1 - \|\bar{x}\|_2)
\geq \min\{\|\hat{x}\|_1 - \|\hat{x}\|_2, \|\bar{x}\|_1 - \|\bar{x}\|_2\},
$$

16
which contradicts with the assumption that $x^*$ is a minimizer in $B_r(x^*)$. $\square$

Local minimizers of the unconstrained problem share the same property.

**Theorem 2.1.4.** Let $x^*$ be a local minimizer of the unconstrained problem (2.2) then the columns of $A_{\Lambda^*}$ are linearly independent.

**Proof.** We claim that $x^*$ is also a local minimizer of the following constrained problem

$$\min_{x \in \mathbb{R}^n} \|x\|_1 - \|x\|_2 \quad \text{subject to} \quad Ax = Ax^*.$$  

Suppose not, then $\forall r > 0$, there exists $\tilde{x} \in B_r(x^*)$ such that $A\tilde{x} = Ax^*$ and

$$\|\tilde{x}\|_1 - \|\tilde{x}\|_2 < \|x^*\|_1 - \|x^*\|_2.$$

This implies

$$\frac{1}{2} \|A\tilde{x} - b\|_2^2 + \lambda(\|\tilde{x}\|_1 - \|\tilde{x}\|_2) < \frac{1}{2} \|Ax^* - b\|_2^2 + \lambda(\|x^*\|_1 - \|x^*\|_2).$$

Thus for any $r > 0$, we actually find a $\tilde{x} \in B_r(x^*)$ yielding a smaller objective of (2.2) than $x^*$, which leads to a contradiction because $x^*$ is assumed to be a local minimizer. Thus the claim is validated.

Using the claim above and Theorem 2.1.3, we have that the columns of $A_{\Lambda^*}$ are linearly independent. $\square$

By Theorem 2.1.3 and 2.1.4, we readily conclude the following facts:

**Corollary 2.1.1.** (a) Suppose $x^*$ is a local minimizer of (2.1) or (2.2), since $\text{rank}(A) = m$, the sparsity of $x^*$ is at most $m$.

(b) If $x^*$ is a local minimizer of (2.1), then there is no such $x \in \mathbb{R}^n$ satisfying $Ax = b$ and $\text{supp}(x) \subseteq \Lambda^*$, i.e. it is impossible to find a feasible solution whose support is contained in $\text{supp}(x^*)$.

(c) Both the numbers of local minimizers of (2.1) and (2.2) are finite.
2.2 Computational Approach

Applying DCA introduced in Section 1.3, we decompose the objective in (2.2) as

\[
F(x) = \left( \frac{1}{2} \| Ax - b \|_2^2 + \lambda \| x \|_1 \right) - \lambda \| x \|_2.
\]

Note that \( \| x \|_2 \) is differentiable with gradient \( \frac{x}{\| x \|_2} \) for all \( x \neq 0 \) and that \( 0 \in \partial \| x \|_2 \) for \( x = 0 \), thus the strategy to iterate is as follows:

\[
x^{k+1} = \begin{cases} 
\arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \| Ax - b \|_2^2 + \lambda \| x \|_1 & \text{if } x^k = 0, \\
\arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \| Ax - b \|_2^2 + \lambda \| x \|_1 - \langle x, \lambda \frac{x^k}{\| x^k \|_2} \rangle & \text{otherwise.}
\end{cases}
\]

It will be shown in Proposition 4.3.1 that \( \| x^{k+1} - x^k \|_2 \to 0 \) as \( k \to \infty \), so a reasonable termination criterion for (4.10) can be

\[
\frac{\| x^{k+1} - x^k \|_2}{\max\{\| x^k \|_2, 1\}} < \epsilon
\]

for some given parameter \( \epsilon > 0 \).

The DCA in general does not guarantee a global minimum due to non-convex nature of the problem. One could in principle prove convergence to the global minimum by the branch and bound procedure (as done in [60]), but the cost is often too high. A good initial guess is therefore crucial for the performance of algorithm. The experiments in Section 2.4 will show that the DCA often produces a solution that is close to global minimizer when starting with \( x^0 = 0 \). The intuition behind our choice can be that the first step of (4.10) reduces to solving the unconstrained \( \ell_1 \) problem. So basically we are minimizing \( \ell_{1-2} \) on top of \( \ell_1 \), which possibly explains why we observed in the experiments that \( \ell_{1-2} \) regularization initialized by \( x^0 = 0 \) always outperforms \( \ell_1 \) regularization. Hereby we summarize DCA-\( \ell_{1-2} \) in Algorithm 1 below.
Algorithm 1 DCA-\(\ell_1-\ell_2\) for solving (2.2).

Define \(\epsilon > 0\) and set \(x^0 = 0, n = 0\).

for \(k = 0, 1, 2, \cdots, \text{Maxoit}\) do
  if \(x^k = 0\) then
    \(x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1\)
  else
    \(x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1 - \langle x, \lambda \frac{x^k}{\|x^k\|_2} \rangle\)
  end if
end for

2.2.1 Convergence analysis

Assuming each DCA iteration of (4.10) is solved accurately, we show that the sequence \(\{x^k\}\) is bounded and \(\|x^{k+1} - x^k\|_2 \to 0\), and limit points of \(\{x^k\}\) are stationary points of (2.2) satisfying the first-order optimality condition. Note that \(\text{ker}(A^TA)\) is non-trivial, so both the DC components in (4.9) only have weak convexity. As a result, the convergence of (4.10) is not covered by the standard convergence analysis for the DCA (e.g. Theorem 3.7 of [71]), because strong convexity is otherwise needed.

Lemma 2.2.1. For all \(\lambda > 0\), \(F(x) = \frac{1}{2} \|Ax - b\|_2^2 + \lambda(\|x\|_1 - \|x\|_2) \to \infty\) as \(\|x\|_2 \to \infty\), therefore \(F(x)\) is coercive in the sense that the level set \(\{x \in \mathbb{R}^n : F(x) \leq F(x^0)\}\) is bounded, \(\forall x^0 \in \mathbb{R}^n\).

Proof. It suffices to show that for any fixed \(x \in \mathbb{R}^n \setminus \{0\}\), \(F(cx) \to \infty\) as \(c \to \infty\).

\[
F(cx) = \frac{1}{2} \|cAx - b\|_2^2 + c\lambda(\|x\|_1 - \|x\|_2) \\
\geq \frac{1}{2} (c\|Ax\|_2 - \|b\|_2)^2 + c\lambda(\|x\|_1 - \|x\|_2)
\]

If \(Ax = 0\), i.e. \(x \in \text{ker}(A) \setminus \{0\}\), since \(\text{rank}(A) = m\), we have \(\|x\|_0 \geq m + 1 \geq 2\). Lemma 2.1.1(c) implies \(\|x\|_1 - \|x\|_2 > 0\), so

\[
F(cx) = \frac{1}{2} \|b\|_2^2 + c\lambda(\|x\|_1 - \|x\|_2) \to \infty \quad \text{as} \quad c \to \infty.
\]

If \(Ax \neq 0\), the claim follows as we notice that \(c\|Ax\|_2 - \|b\|_2 \to \infty\) as \(c \to \infty\). \(\square\)
Lemma 2.2.2. Let \( \{x^k\} \) be the sequence generated by the DCA (4.10). For all \( k \in \mathbb{N} \), we have

\[
F(x^k) - F(x^{k+1}) \geq \frac{1}{2} \|A(x^k - x^{k+1})\|_2^2 + \lambda(\|x^{k+1}\|_2 - \|x^k\|_2 - \langle y^k, x^{k+1} - x^k \rangle) \geq 0, \tag{2.19}
\]

where \( y^k \in \partial \|x^k\|_2 \).

Proof. A simple calculation shows

\[
F(x^k) - F(x^{k+1}) = \frac{1}{2} \|A(x^{k+1} - x^k)\|_2^2 + \langle A(x^k - x^{k+1}), Ax^{k+1} - b \rangle
+ \lambda(\|x^k\|_1 - \|x^{k+1}\|_1 - \|x^k\|_2 + \|x^{k+1}\|_2) \tag{2.20}
\]

Recall that \( x^{k+1} \) is the solution to the problem

\[
\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda(\|x\|_1 - \langle x, y^k \rangle)
\]

with \( y^k \in \partial \|x^k\|_2 \), then the first-order optimality condition holds at \( x^{k+1} \). More precisely, there exists \( w^{k+1} \in \partial \|x^{k+1}\|_1 \), such that

\[
A^T(Ax^{k+1} - b) + \lambda(w^{k+1} - y^k) = 0. \tag{2.21}
\]

Left multiplying (2.21) by \( (x^k - x^{k+1})^T \) gives

\[
\langle A(x^k - x^{k+1}), Ax^{k+1} - b \rangle + \lambda(\langle w^{k+1}, x^k \rangle - \|x^{k+1}\|_1) + \langle y^k, x^{k+1} - x^k \rangle = 0, \tag{2.22}
\]

where we used \( \langle w^{k+1}, x^k \rangle = \|x^{k+1}\|_1 \). Combining (2.20) and (2.22), we have

\[
F(x^k) - F(x^{k+1}) \geq \frac{1}{2} \|A(x^{k+1} - x^k)\|_2^2 + \lambda(\|x^{k+1}\|_2 - \|x^k\|_2 - \langle y^k, x^{k+1} - x^k \rangle)
\geq \frac{1}{2} \|A(x^{k+1} - x^k)\|_2^2 + \lambda(\|x^{k+1}\|_2 - \|x^k\|_2 - \langle y^k, x^{k+1} - x^k \rangle)
\geq 0.
\]
In the first inequality above, \( \|x^k\|_1 - \langle w^{k+1}, x^k \rangle \geq 0 \) since \( |w_i^{k+1}| \leq 1 \) for all \( 1 \leq i \leq n \), while the second one holds because \( y^k \in \partial \|x^k\|_2 \).

We now show the convergence results for Algorithm 1.

**Proposition 2.2.1.** Let \( \{x^k\} \) be the sequence of iterates generated by Algorithm 1, we have

(a) \( \{x^k\} \) is bounded.

(b) \( \|x^{k+1} - x^k\|_2 \to 0 \) as \( k \to \infty \).

(c) Any nonzero limit point \( x^* \) of \( \{x^k\} \) satisfies the first-order optimality condition

\[
\begin{align*}
0 & \in A^T(Ax^* - b) + \lambda (\partial \|x^*\|_1 - \frac{x^*}{\|x^*\|_2}),
\end{align*}
\]

which means \( x^* \) is a stationary point of (2.2).

**Proof.** (a) Using Lemma 2.2.1 and the fact that \( \{F(x^k)\} \) is monotonically decreasing, we have \( \{x^k\} \subseteq \{x \in \mathbb{R}^n : F(x) \leq F(x^0)\} \) is bounded.

(b) If \( x^1 = x^0 = 0 \), we then stop the algorithm producing the solution \( x^* = 0 \). Otherwise, it follows from (4.11) that

\[
F(0) - F(x^1) \geq \lambda \|x^1\|_2 > 0,
\]

so \( x^k \neq 0 \) whenever \( k \geq 1 \). In what follows, we assume \( x^k \neq 0 \) for all \( k \geq 1 \). Since \( \{F(x^k)\} \) is convergent, substituting \( y^k = \frac{x^k}{\|x^k\|_2} \) in (4.11), we must have

\[
\begin{align*}
\|A(x^k - x^{k+1})\|_2 & \to 0 \\
\|x^{k+1}\|_2 - \frac{\langle x^k, x^{k+1} \rangle}{\|x^k\|_2} & \to 0
\end{align*}
\]

We define \( c^k := \frac{(x^k, x^{k+1})}{\|x^k\|_2^2} \) and \( e^n := x^{k+1} - c^k x^k \), then it suffices to prove \( e^k \to 0 \) and \( c^k \to 1 \).

It is straightforward to check that

\[
\|e^k\|_2^2 = \|x^{k+1}\|_2^2 - \frac{(x^k, x^{k+1})^2}{\|x^k\|_2^2} \to 0,
\]
where we used (4.16). Then from (4.15) it follows that

\[ 0 = \lim_{k \to \infty} \|A(x^k - x^{k+1})\|_2 = \lim_{k \to \infty} \|A((c^k - 1)x^k - e^k)\|_2 = \lim_{k \to \infty} |c^k - 1| \|Ax^k\|_2. \]

If \( \lim_{k \to \infty} c^k - 1 \neq 0 \), then there exists a subsequence \( \{x^{k_j}\} \) such that \( Ax^{k_j} \to 0 \). So we have

\[ \lim_{k_j \to \infty} F(x^{k_j}) \geq \lim_{k_j \to \infty} \frac{1}{2} \|Ax^{k_j} - b\|_2^2 = \frac{1}{2} \|b\|_2^2 = F(x^0), \]

which is contradictory to the fact that

\[ F(x^{k_j}) \leq F(x^1) < F(x^0), \quad \forall k_j \geq 1. \]

Therefore \( c^k \to 1, \ e^k \to 0 \), and thus \( x^{k+1} - x^k \to 0 \) as \( k \to \infty \).

(c) Let \( \{x^{k_j}\} \) be a subsequence of \( \{x^k\} \) converging to \( x^* \neq 0 \), so the optimality condition at the \( k_j \)-th step of Algorithm 1 reads

\[ 0 \in A^T(Ax^{k_j} - b) + \lambda \partial \|x^{k_j}\|_1 - \lambda \frac{x^{k_j-1}}{\|x^{k_j-1}\|_2}; \]

or

\[ -(A^T(Ax^{k_j} - b) - \lambda \frac{x^{k_j-1}}{\|x^{k_j-1}\|_2}) \in \lambda \partial \|x^{k_j}\|_1. \tag{2.26} \]

Here \( \partial \|x\|_1 = \prod_{i=1}^n \text{SGN}(x_i) \subset \mathbb{R}^n \) with

\[ \text{SGN}(x_i) := \begin{cases} \{\text{sgn}(x_i)\} & \text{if } x_i \neq 0, \\ [-1, 1] & \text{otherwise}. \end{cases} \]

Since, by (b), \( x^{k_j} \to x^* \), we have, when \( k_j \) is sufficiently large, \( \text{supp}(x^*) \subseteq \text{supp}(x^{k_j}) \) and \( \text{sgn}(x^{k_j}_i) = \text{sgn}(x^*_i) \) for all \( i \in \text{supp}(x^*) \), which implies

\[ \partial \|x^{k_j}\|_1 \subseteq \partial \|x^*\|_1. \]
Then by (2.26), for large $k_j$ we have

$$(2.27) \quad -(A^T(Ax^{k_j} - b) - \lambda \frac{x^{k_j-1}_{x_j}}{\|x^{k_j-1}_j\|_2}) \in \lambda \partial \|x^*\|_1.$$ 

Moreover, since $x^*$ is away from $0$,

$$\lim_{k_j \to \infty} A^T(Ax^{k_j} - b) - \lambda \frac{x^{k_j}}{\|x^{k_j}\|_2} - \lambda \frac{x^{k_j} - 1}{\|x^{k_j-1}\|_2} = \lim_{k_j \to \infty} A^T(Ax^* - b) - \lambda \frac{x^*}{\|x^*\|_2}.$$ 

Let $k_j \to \infty$ in (2.27) and note that $\partial \|x^*\|_1$ is a closed set, then (2.23) follows. 

By choosing appropriate regularization parameters, we are able to control the sparsity of $x^*$.

**Theorem 2.2.1.** \(\forall s \in \mathbb{N}, \) there exists $\lambda_s > 0$ such that for any parameter $\lambda > \lambda_s$ in (2.2), we have $\|x^*\|_0 \leq s$, where $x^* \neq 0$ is a stationary point generated by Algorithm 1.

**Proof.** By the optimality condition (2.23), there exists $w^* \in \partial \|x^*\|_1$ satisfying

$$(2.28) \quad w^*_i = \begin{cases} \text{sgn}(x^*_i) & \text{if } i \in \text{supp}(x^*), \\ \in [-1, 1] & \text{otherwise.} \end{cases}$$

such that

$$-A^T(Ax^* - b) = \lambda (w^* - \frac{x^*}{\|x^*\|_2}).$$

Since, by (2.28) $\|w^*\|_2 \geq \sqrt{\|x^*\|_0}$, taking $\ell_2$ norm of both sides gives

$$(2.29) \quad \|A^T(Ax^* - b)\|_2 = \lambda \|w^* - \frac{x^*}{\|x^*\|_2}\|_2 \geq \lambda (\|w^*\|_2 - \|\frac{x^*}{\|x^*\|_2}\|_2) \geq \lambda (\sqrt{\|x^*\|_0} - 1).$$

On the other hand,

$$(2.30) \quad \|A^T(Ax^* - b)\|_2 \leq \|A\|_2 \|Ax^* - b\|_2 = \|A\|_2 \|Ax^* - b\|_2 \leq \|A\|_2 \|b\|_2,$$
where we used
\[ \frac{1}{2} \|Ax^* - b\|_2 \leq \frac{1}{2} \|Ax^* - b\|_2 + \lambda(\|x^*\|_1 - \|x^*\|_2) = F(x^*) \leq F(x^0) = \frac{1}{2} \|b\|_2. \]

Combining (2.29) and (2.30), we obtain
\[ \sqrt{\|x^*\|_0} \leq \frac{\|A\|_2 \|b\|_2}{\lambda} + 1 \quad \text{or} \quad \|x^*\|_0 \leq \left( \frac{\|A\|_2 \|b\|_2}{\lambda} + 1 \right)^2. \]

Moreover,
\[ \left( \frac{\|A\|_2 \|b\|_2}{\lambda} + 1 \right)^2 < s + 1 \quad \iff \quad \lambda > \lambda_s := \frac{\|A\|_2 \|b\|_2}{\sqrt{s + 1} - 1} \]

In other words, if \( \lambda > \lambda_s \), then \( \|x^*\|_0 < s + 1 \). \( \|x^*\|_0 \) and \( s \) are integers, so \( \|x^*\|_0 \leq s \).

### 2.2.2 Solving the subproblem

In each DCA iteration, it requires solving a \( \ell_1 \)-regularized convex subproblem of the following form

\[ (2.31) \min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \langle x, v \rangle + \lambda\|x\|_1, \]

where \( v \in \mathbb{R}^n \) is a constant vector. This problem can be done by the alternating direction method of multipliers (ADMM), a versatile algorithm first introduced in [52, 46]. A recent result on the \( O(1/n) \) convergence rate of ADMM was established in [56]. Just like the split Bregman [51], the trick of ADMM form is to decouple the coupling between the quadratic term and \( \ell_1 \) penalty in (2.31). Specifically, (2.31) can be reformulated as

\[ \min_{x, z \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \langle x, v \rangle + \lambda\|z\|_1 \quad \text{subject to} \quad x - z = 0, \]

We then form the augmented Lagrangian
\[ \mathcal{L}_\delta(x, z, y) = \frac{1}{2} \|Ax - b\|_2^2 + \langle x, v \rangle + \lambda\|z\|_1 + y^T(x - z) + \frac{\delta}{2} \|x - z\|_2^2, \]
where \( y \) is the Lagrange multiplier, \( \delta > 0 \) is the penalty parameter. ADMM consists of the iterations:

\[
\begin{aligned}
    x^{l+1} &= \arg \min_x \mathcal{L}_\delta(x, z^l, y^l) \\
    z^{l+1} &= \arg \min_z \mathcal{L}_\delta(x^{l+1}, z, y^l) \\
    y^{l+1} &= y^l + \delta(x^{l+1} - z^{l+1})
\end{aligned}
\]

The first two steps have closed-form solutions which are detailed in Algorithm 2. In the \( z \)-update step, \( \mathcal{S}(x, r) \) denotes the soft-thresholding operator given by

\[
(S(x, r))_i = \sgn(x_i) \max\{|x_i| - r, 0\}.
\]

The computational complexity of Algorithm 2 mainly lies in the \( x \)-update step. Since \( A \in \mathbb{R}^{m \times n} \) and \( b \in \mathbb{R}^m \), the computational complexity of \( A^T A + \delta I \) is \( O(mn^2 + n) = O(mn^2) \) and that of \( A^T b - v + \delta z^k - y^k \) is \( O(mn + n) = O(mn) \). Moreover, the inversion of matrix \( A^T A + \delta I \) requires \( O(n^3) \). Therefore, the computational complexity of Algorithm 2 per iteration is \( O(n^3 + mn^2 + mn) = O(n^3 + mn^2) \).

**Algorithm 2** ADMM for subproblem (2.31).

```
Define \( x^0, z^0 \) and \( u^0 \).
for \( l = 0, 1, 2, \cdots, \text{MAXit} \) do
    \( x^{l+1} = (A^T A + \delta I)^{-1}(A^T b - v + \delta z^l - y^l) \)
    \( z^{l+1} = \mathcal{S}(x^{l+1} + y^l/\delta, \lambda/\delta) \)
    \( y^{l+1} = y^l + \delta(x^{l+1} - z^{l+1}) \)
end for
```

According to [7], a stopping criterion of Algorithm 2 is given by:

\[
||r^l||_2 \leq \sqrt{n}\epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max\{||x^l||_2, ||z^l||_2\}, \quad ||s^l||_2 \leq \sqrt{n}\epsilon^{\text{abs}} + \epsilon^{\text{rel}}||y^l||_2,
\]

where \( r^l = x^l - z^l \), \( s^l = \delta(z^l - z^{l-1}) \) are primal and dual residuals respectively at the \( l \)-th iteration. \( \epsilon^{\text{abs}} > 0 \) is an absolute tolerance and \( \epsilon^{\text{rel}} > 0 \) a relative tolerance.
2.3 Hybrid Simulated Annealing

In this section, we employ a technique called simulated annealing (SA) to traverse local minima to reach a global solution. Combining the DCA with SA, we propose a hybrid simulated annealing DCA. There are many generic SA algorithms, see Kirkpatrick [58], Geman and Geman [47], Hajek [54], Gidas [49], and the references therein. In addition, this technique has many applications to image processing, such as Carnevali et al. [25].

The term ”annealing” is analogous to the cooling of a liquid or solid in a physical system. Consider the problem of minimizing the cost function $F(x)$. Simulated annealing algorithm begins with an initial solution and iteratively generates new ones, each of which is randomly selected among the ”neighborhood” of the previous state. If the new solution is better than the previous one, it is accepted; otherwise, it is accepted with certain probability. The probability of accepting a new state is given by $\exp(-\frac{F_{\text{new}}-F_{\text{curr}}}{T}) > \alpha$, where $\alpha$ is a random number between 0 and 1, and $T$ is a temperature parameter. The algorithm usually starts with a high temperature, and then gradually goes down to 0. The cooling must be slow enough so that the system does not get stuck into local minima of $F(x)$. The hybrid simulated annealing algorithm can be summarized as follows:

1. Choose an initial temperature $T$, an initial state $x_{\text{curr}}$, and evaluate $F(x_{\text{curr}})$.

2. Randomly determine a new state $x_{\text{new}}$, and run the DCA to get the near optimal solution $DCA(x_{\text{new}})$.

3. Evaluate $F(DCA(x_{\text{new}}))$. If $F(DCA(x_{\text{new}})) < F(x_{\text{curr}})$, accept $DCA(x_{\text{new}})$, i.e., $x_{\text{curr}} = DCA(x_{\text{new}})$; otherwise, accept $DCA(x_{\text{new}})$ if $\exp(-\frac{F(DCA(x_{\text{new}}))-F(x_{\text{curr}})}{T}) > \alpha$, where $\alpha$ is a random number between 0 and 1.

4. Repeat Steps 2 and 3 for some iterations with temperature $T$.

5. Lower $T$ according to the annealing schedule, and return to step 2. Continue this process until some criteria of convergence is satisfied.

There are two important aspects in implementing simulated annealing. One is how to lower the temperature $T$. Kirkpatrick [58] suggests that $T$ decays geometrically in the number
of cooling phases. Hajek [54] proves that if $T$ decreases at the rate of $\frac{d}{\log k}$, where $k$ is the number of iterations and $d$ is some certain constant, then the probability distribution for the algorithm converges to the set of global minimum points with probability one. In our algorithm, we follow Hajek’s [54] method by decreasing $T$ at the rate of $\frac{d}{\log k}$, with some constant $d$.

Another aspect is how to advance to a new state based on the current one in Step 2. One of the most common methods is to add random Gaussian noise, such as the method in [78]. We generate the Gaussian perturbation by the following probability density function

$$p(x, T_k) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(T_k + \beta)}} \exp(-\frac{x_i^2}{2(T_k + \beta)}),$$

where the temperature $T_k = \frac{T_0}{\log(k)}$, $\beta$ is a constant. We assume that there is only one iteration in each cooling scheme, i.e., the temperature $T_k$ decreases after each iteration. Then we have the following theorem.

**Theorem 2.3.1.** If we choose the $k$-th new state by the probability density function given in (2.1), then the hybrid simulated annealing algorithm converges to the global minimum $F^*$ in probability.

**Proof.** The proof is similar to Corollary 1 in [78]. We assume that there exists a bounded set $E \subset \mathbb{R}^n$ containing all the global minima. Let $r_i = \max_{x,y \in E} |y_i - x_i|$, $1 \leq i \leq n$, it suffices to show that there exist constants $C > 0$ and $k_0 > 0$, such that $\min_{x,y \in E} p(y - x, T_k) \geq C/k$, $\forall k > k_0$. Take $k_0 = 1$ and $C = \frac{1}{(2\pi)^n/(T_0 + \beta)^n/2 \exp(||r||^2/2\beta)}$, since $0 \leq T_k \leq T_0$, we have the following inequalities for all $k \geq 1$

$$\min_{x,y \in E} p(y - x, T_k) = \min_{x,y \in E} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(T_k + \beta)}} \exp(-\frac{(y_i - x_i)^2}{2(T_k + \beta)}) \geq \frac{1}{(2\pi)^n/(T_k + \beta)^n/2 \exp(||r||^2/2\beta)}$$

$$\geq \frac{1}{(2\pi)^n/(T_0 + \beta)^n/2 \exp(||r||^2/2\beta)} \geq C \geq C/k.$$
However, due to the presence of a large number of local minima, this Gaussian perturbation method would converge slowly. To overcome this difficulty, we propose a novel perturbation method. First, let us define the space $V = \{ x \in \mathbb{R}^n : x_i = 0, 1 \text{ or } -1 \}$ and $V_s = \{ x \in V : \|x\|_0 \leq s \}$. Note that given any $x \in \mathbb{R}^n$, $\text{sgn}(x) \in V$, where the signum function applies element wise. Define the mapping $W$ from $\mathbb{R}^n$ to $V$ by $W(x) = \text{sgn}(S(x, \nu))$ for some small number $\nu > 0$, where $S$ is the soft-thresholding operator. Denote the DCA function by $DCA(x)$ being output of DCA initialized by $x$. We randomly choose the new state $x_{\text{new}} \in V_s$ for some $0 < s < n$, such that $\|x_{\text{new}} - W(DCA(x_{\text{curr}}))\|_2 \leq \eta$.

The idea of this perturbation method is to keep the sparse properties of the current state and perturb inside the space $V_s$. Hence, we call this perturbation method the sparse perturbation method.

To prove the convergence to global minima of this hybrid simulated annealing with sparse perturbation, we shall work with the following assumption.

**Assumption 2.3.1.** If $x^*$ is a global minimizer of the cost function $F$ over $\mathbb{R}^n$, then there is a global minimizer of the cost function $J(x) := F(DCA(x))$ denoted by $x^*_J \in \arg \min_{x \in V} J(x)$, such that $x^* = DCA(x^*_J)$.

The above assumption says that a global minimizer of $F$ can be reached by a local DCA descent from a global minimizer of $J$ defined over a smaller set $V$ whose elements are vectors with components $0, \pm 1$. This assumption is akin to an interesting property of the Bregman iteration of $\ell_1$ minimization [84] where if the $n$-th step iteration gets the signs and support of an $\ell_1$ minimizer, the minimizer is reached at the $(n+1)$-th step. Though one could minimize $F$ directly as stated in Theorem 4.1, the passage to a global minimum of $F$ from that of $J$ via DCA is observed to be a short cut in our numerical experiments, largely because the global minima of $F$ in our problem are sparse. Under this assumption, we aim to show that the sequence $W(x_{\text{curr}})$ converges to a global minimizer of $J$ over space $V_s$. By our algorithm, for each state $x \in V$, we have a neighborhood of $x$, $N(x) \subset V$, where we generate the next state. We also assume that there is a transition probability matrix $Q$ such that $Q(x, y) > 0$ if and only if $y \in N(x)$.
We only need one iteration in each cooling scheme, because the temperature $T_k$ decreases after each iteration, and $\lim_{k \to \infty} T_k = 0$. Denote the sequence of states by $x^1_{\text{curr}}, x^2_{\text{curr}}, \cdots$, and the initial state by $x^0$. Define $y_k = W(x^k_{\text{curr}})$. Given $y^k = i$, a new potential next state $x^k_{\text{new}}$ is chosen from the neighborhood set $\mathbb{N}(i)$ with the conditional probability $P(x^k_{\text{new}} = j | y^k = i) = Q(i, j)$. Then we update the algorithm as follows. If $J(x^{k+1}_{\text{new}}) \leq J(y^k)$, $y^{k+1} = x^{k+1}_{\text{new}}$. If $J(x^{k+1}_{\text{new}}) > J(y^k)$,

\begin{equation}
(2.2) \quad y^{k+1} = \begin{cases} x^{k+1}_{\text{new}} & \text{with probability } \exp(-(J(x^{k+1}_{\text{new}}) - J(y^k))/T_k), \\ x^k_{\text{new}} & \text{otherwise.} \end{cases}
\end{equation}

In summary,

$$
P(y^{k+1} = j | y^k = i) = Q(i, j) \exp(-\frac{\max(J(j) - J(i))}{T_k}),$$

for $j \in \mathbb{N}(i) \subset \mathbb{V}$.

By the above updating method, SA algorithm is best understood as a nonhomogeneous Markov chain $y^k$ in which the transition matrix is dependent on the temperature $T_k$. Denote the set of the global minimizers of $J$ on $\mathbb{V}$ by $\mathbb{V}^*$. We aim to prove that

$$\lim_{k \to \infty} P(y^k \in \mathbb{V}^*) = 1.$$

To motivate the rationale behind the SA algorithm, we assume that the temperature $T_k$ is kept at constant value $T$. In addition, we assume that $y_k$ is irreducible, which means that for any two states $i, j \in \mathbb{V}$, we can choose a sequence of states $y^0 = i, y^1, \cdots, y^l = j$ for some $l \geq 1$ such that $y^{k+1} \in \mathbb{N}(y^k), 1 \leq k \leq l - 1$. We also assume that $Q$ is reversible, i.e., there is a distribution $a(i)$ on $\mathbb{V}$ such that $a(i)Q(i, j) = a(j)Q(j, i)$ for all $i, j \in \mathbb{V}$. One simple choice for $Q$ is

\begin{equation}
(2.3) \quad Q(i, j) = \begin{cases} \frac{1}{|\mathbb{N}(i)|} & \text{if } j \in \mathbb{N}(i), \\ 0, & \text{otherwise.} \end{cases}
\end{equation}
We then introduce the following lemma.

**Lemma 2.3.1.** Under the above assumptions, the state sequence \( \{y^k\} \) generated by SA algorithm satisfies

\[
\lim_{T \to 0} \lim_{k \to \infty} \mathbf{P}(y^k \in \mathbb{V}^*) = 1.
\]

**Proof.** Since the temperature \( T_k = T \) for all \( k \), the sequence \( y^k \) is a homogeneous Markov chain. Assume that its associated transition matrix is \( P_T \). Define a probability distribution by

\[
\pi_T(i) = \frac{a(i)}{Z_T} \exp\left(-\frac{J(i)}{T}\right),
\]

where \( Z_T = \sum_i a(i) \exp(-J(i)/T) \). A simple computation shows that \( \pi_T = \pi_T P_T \). So \( \pi_T \) is the invariant distribution for the Markov chain \( y_k \). By the reversibility of \( Q \) and the irreducibility of \( y^k \), the Markov ergodic convergence theorem implies that

\[
\mathbf{P}(y^k \in \mathbb{V}^*) = \sum_{i \in \mathbb{V}^*} \pi_T(i).
\]

Since \( J(i) > J(j) \) for all \( i \in \mathbb{V} \setminus \mathbb{V}^*, j \in \mathbb{V}^* \), and \( \lim_{T \to 0} \pi_T(i) = 0 \) for all \( i \in \mathbb{V} \setminus \mathbb{V}^* \), we have

\[
\lim_{T \to 0} \lim_{k \to \infty} \mathbf{P}(y^k \in \mathbb{V}^*) = 1.
\]

\( \square \)

To extend the convergence result to the case where \( \lim_{k \to \infty} T_k = 0 \), we introduce the following concept.

**Definition 2.3.1.** We say that the state \( i \) communicates with \( \mathbb{V}^* \) at height \( h \) if there exists a path in \( \mathbb{V} \) such that the largest value of \( J \) alone the path is \( J(i) + h \).

The main theorem is given as follows.

**Theorem 2.3.2.** Let \( d^* \) be the smallest number such that every \( i \in \mathbb{V} \) communicates with \( \mathbb{V} \) at height \( d^* \). Then the SA algorithm sequence \( y^k \) converges to the global minima set \( \mathbb{V}^* \) with probability one, if and only if

\[
\sum_{k=1}^{\infty} \exp\left(-d^*/T_k\right) = \infty.
\]
The detailed proof is given in [54]. By the theorem, we just need to choose \( T_k = d/ \log(k) \), where \( d \geq d^* \). To estimate \( d^* \), we can simply set \( d^* = |V_s| \), since \( V_s \) is a finite space.

2.4 Numerical Results

In this section, we present numerical experiments to demonstrate the efficiency of DCA-\( \ell_{1-2} \) method. We will compare it with the following state-of-the-art CS solvers:

- ADMM-lasso [8] that solves the lasso problem (1.2) by ADMM.
- A greedy method termed CoSaMP [66] which involves a sequence of support detections and least squares.
- The accelerated version of IHT (AIHT) [6] that solves
  \[
  \min_{x \in \mathbb{R}^n} \| Ax - b \|_2^2 \text{ subject to } \|x\|_0 \leq s.
  \]
  by hard thresholding iterations.
- An improved IRLS-\( \ell_p \) algorithm [59] that solves the unconstrained \( \ell_p \) problem with \( 0 < p < 1 \):
  \[
  (2.1) \quad \min_{x \in \mathbb{R}^n} \frac{1}{2} \| Ax - b \|_2^2 + \lambda \|x\|_p^p.
  \]
- Reweighted \( \ell_1 \) [17] which is at heart a non-convex CS solver based on IRL1 algorithm attempting to solve
  \[
  \min_{x \in \mathbb{R}^n} \sum_{i=1}^n \log(|x_i| + \varepsilon) \text{ subject to } Ax = b.
  \]
- Half thresholding [79] (Scheme 2) for \( \ell_{1/2} \) regularization, i.e., (2.1) with \( p = 0.5 \).

Note that all the proposed methods except ADMM-lasso are non-convex in nature.
**Sensing matrix for tests.** We will test the commonly used random Gaussian matrix, which is defined as

\[ A_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_m/m), \quad i = 1, \ldots, n \]

and random partial discrete cosine transform (DCT) matrix

\[ A_i = \frac{1}{\sqrt{m}} \cos(2i\pi \xi), \quad i = 1, \ldots, n \]

where \( \xi \in \mathbb{R}^m \sim \mathcal{U}([0, 1]^m) \) whose components are uniformly and independently sampled from [0,1]. These sensing matrices fit for CS, being incoherent and having small RIP constants with high probability.

We also test more ill-conditioned sensing matrix of significantly higher coherence. Specifically, a randomly oversampled partial DCT matrix \( A \) is defined as

\[ A_i = \frac{1}{\sqrt{m}} \cos(2i\pi \xi/F), \quad i = 1, \ldots, n \]

where \( \xi \in \mathbb{R}^m \sim \mathcal{U}([0, 1]^m) \) and \( F \in \mathbb{N} \) is the refinement factor. Actually it is the real part of the random partial Fourier matrix analyzed in [36]. The number \( F \) is closely related to the conditioning of \( A \) in the sense that \( \mu(A) \) tends to get larger as \( F \) increases. For \( A \in \mathbb{R}^{100 \times 2000} \), \( \mu(A) \) easily exceeds 0.99 when \( F = 10 \). This quantity is above 0.9999 when \( F \) increases to 20. Although \( A \) sampled in this way does not have good RIP by any means, it is still possible to recover the sparse vector \( \bar{x} \) provided its spikes are sufficiently separated. Specifically, we randomly select the elements of \( \text{supp}(\bar{x}) \) so as to satisfy the following condition:

\[ \min_{j,k \in \text{supp}(\bar{x})} |j - k| \geq L. \]

Here \( L \) is called the minimum separation. For traditional inversion methods to work, it is necessary for \( L \) to be at least 1 Rayleigh length (RL) which is unity in the frequency domain [36, 29]. In our case, the value of 1 RL is nothing but \( F \).
2.4.1 Selection of parameters

The regularization parameter $\lambda$ controls data fitting and sparsity of the solution. For noiseless case, a tiny value should be chosen. When measurements are noisy, a reasonable $\lambda$ should depend on the noise level. In this case, $\lambda$ needs to be tuned empirically (typically by cross-validation technique). Although our convergence result is established on the assumption that the sequence of subproblems is solved exactly by Algorithm 2, it suffices for practical use that the relative tolerance $\epsilon^{\text{rel}}$ and absolute tolerance $\epsilon^{\text{abs}}$ are adequately small. Fig. 2.2 shows that in the noiseless case, the relative error $\frac{\|x^*-\bar{x}\|_2}{\|\bar{x}\|_2}$ is linear in the tolerance at moderate sparsity level when $\epsilon^{\text{rel}} \leq 10^{-3}$ and $\epsilon^{\text{abs}} = 10^{-2}\epsilon^{\text{rel}}$. Here $\bar{x}$ is the test signal and $x^*$ is the recovered one by DCA-$\ell_{1-2}$ from the measurements $\hat{b} = A\bar{x}$. $\delta$ in Algorithm 2 should be well chosen, since sometimes the convergence can be sensitive to its value. [8] suggests to vary $\delta$ by iteration, aiming to stabilize the ratio between primal and dual residuals as they both go to zero. We adopt this strategy when having noise in measurements. More precisely,

$$
\delta^{l+1} = \begin{cases} 
2\delta^l & \text{if } \|r^l\|_2 > 10\|s^l\|_2, \\
\delta^l/2 & \text{if } 10\|r^l\|_2 < \|s^l\|_2, \\
\delta^l & \text{otherwise.} 
\end{cases}
$$

Recall that $r^l$ and $s^l$ are the primal and dual residuals respectively. In noiseless case where $\lambda$ is always set to a small number, it turns out that just taking $\delta = 10\lambda$ works well enough.

2.4.2 Exact recovery of sparse vectors

In the noiseless case, we compare the proposed methods in terms of success percentage and computational cost.

**Test on RIP matrix.** We carry out the experiments as follows. After sampling a sensing matrix $A \in \mathbb{R}^{m \times n}$, we generate a test signal $\bar{x} \in \mathbb{R}^n$ of sparsity $s$ supported on a random index set with i.i.d. Gaussian entries. We then compute the measurement $\hat{b} = A\bar{x}$ and apply each solver to produce a reconstruction $x^*$ of $\bar{x}$. The reconstruction is considered
as a success if the relative error
\[ \frac{\|x^* - \bar{x}\|_2}{\|\bar{x}\|_2} < 10^{-3}. \]

We run 100 independent realizations and record the corresponding success rates at various sparsity levels.

We chose \( \epsilon_{\text{abs}} = 10^{-7} \) and \( \epsilon_{\text{rel}} = 10^{-5} \) for DCA-\( \ell_1-2 \) in Algorithm 2. In the outer stopping criterion (2.18) in Algorithm 1, we set \( \epsilon = 10^{-2} \). \texttt{MAXoit} and \texttt{MAXit} are 10 and 5000 respectively. For ADMM-lasso, we let \( \lambda = 10^{-6}, \beta = 1, \rho = 10^{-5}, \epsilon_{\text{abs}} = 10^{-7}, \epsilon_{\text{rel}} = 10^{-5} \), and the maximum number of iterations \texttt{maxiter} = 5000. For CoSaMP, \texttt{maxiter} = 50, the tolerance \texttt{tol} = 10^{-8}. The \texttt{tol} for AIHT was 10^{-12}. For IRLS-\( \ell_p \), \( p = 0.5 \), \texttt{maxiter} = 1000, \texttt{tol} = 10^{-8}. For reweighted \( \ell_1 \), the smoothing parameter \( \varepsilon \) was adaptively updated as introduced in [17], and the outer stopping criterion adopted was the same as that of the DCA-\( \ell_{1-2} \). We solved its weighted \( \ell_1 \) minimization subproblems using the more efficient YALL1 solver (available at http://yall1.blogs.rice.edu/) instead of the default \( \ell_1 \)-MAGIC. The tolerance for YALL1 was set to 10^{-6}. For half thresholding, we let \texttt{maxiter} = 5000. In addition, CoSaMP, AIHT and half thresholding require an estimate on the sparsity of \( \bar{x} \), which we set to the \textit{ground truth}. All other settings of the algorithms were set to default ones.
Random Gaussian matrix  

Random partial DCT matrix  

Figure 2.3: Success rates using incoherent sensing matrix.

Fig. 2.3 depicts the success rates of the proposed methods with $m = 64$ and $n = 256$. For both Gaussian matrix and partial DCT matrix, IRLS-$\ell_p$ with $p = 0.5$ has the best performance, followed by reweighted $\ell_1$. DCA-$\ell_{1-2}$ is comparable to half thresholding and CoSaMP, which outperform both ADMM-lasso and AIHT.

**Test on highly coherent matrix.** Fixing the size of $A$ at 100-by-2000 and $L = 2F$, we repeat the experiment and present the results in Fig. 2.4 for $F = 10$ (left) and 20 (right). Note that in this example, the task of non-convex CS has become more challenging since ill-conditioning of the sensing matrix $A$ makes it much easier for the solvers to stall at spurious local minima. Here we do not take CoSaMP and AIHT into consideration in the comparison, because preliminary results show that even with $\bar{x}$ at low sparsity level, they do not work for matrix of large coherence at all (in terms of exact reconstruction). In this example, the DCA-$\ell_{1-2}$ is the best and provides robust performance regardless of large coherence of $A$. In contrast, the other non-convex solvers clearly encounter the trapping of local minima and perform worse than the convex ADMM-lasso. Moreover, by comparing the two plots in Fig. 2.4, one can tell that their reconstruction qualities suffer a decline as $A$ becomes more and more coherent.

The left plot of Fig. 2.5 shows the success rates for DCA-$\ell_{1-2}$ with/without aid of hybrid simulated annealing (HSA) methods. For each HSA method, we apply at most 100
Figure 2.4: Success rates using randomly oversampled partial DCT matrices.

iterations. The matrix size is 100 × 1500, $F = 20$, and the minimum separation $L = 2F$. We also compare the two different perturbation methods, referred to as HSA with Gaussian perturbations and HSA with sparse perturbations. Both of these HSA methods can improve the reconstruction capability of the plain DCA-ℓ₁−₂. However, HSA with sparse perturbations has the best performance. On the other hand, though the limit point of DCA-ℓ₁−₂ is not known theoretically to be a global minimum, in practice it is quite close. This can be seen from Fig. 2.5 where the additional improvement from the HSA is at most about 15% in the intermediate sparsity regime.

**Comparison of time efficiency under Gaussian measurements.** The comparison of CPU time using random Gaussian sensing matrix and non-convex CS solvers is presented in the right plot of Fig. 2.5. For each $n$, we fixe $m = n/4$ and $s = m/8$ and run 10 independent realizations. Parameters (mainly the tolerances) for the algorithms were tuned such that all resulting relative errors were roughly $10^{-5}$, for which except CoSaMP. CoSaMP stands out as its success relies on correct identification of the support of $\bar{x}$. Once the support is correctly identified, followed by least squares minimization, it naturally produces a solution of perfect accuracy with a tiny relative error close to the machine precision. It turns out that AIHT enjoy the best overall performances in terms of time consumption, being slightly faster than CoSaMP. But CoSaMP did provide substantially higher quality solutions in the
Figure 2.5: Left: Comparison of success rates of HSA algorithms. Right: Comparison of CPU time.

absence of noise. When $n > 1000$, DCA-$\ell_{1-2}$ is faster than the other regularization methods like IRLS-$\ell_p$ and half thresholding. This experiment was carried out on a laptop with 16GB RAM memory and 2.40 GHz Intel Core i7 CPU.

2.4.3 Robust recovery in presence of noise

In this example, we show robustness of DCA-$\ell_{1-2}$ in the noisy case. White Gaussian noise is added to the clean data $A\bar{x}$ to get contaminated measurements $b$, by calling $b = \text{awgn}(Ax, \text{snr})$ in MATLAB, where $\text{snr}$ corresponds to the value of signal-to-noise ratio (SNR) measured in dB. We then obtain the reconstruction $x^*$ using DCA-$\ell_{1-2}$, and compute the SNR of reconstruction given by

$$10 \log_{10} \frac{\|x^* - \bar{x}\|_2^2}{\|\bar{x}\|_2^2},$$

with $\frac{\|x^* - \bar{x}\|_2^2}{\|\bar{x}\|_2^2}$ being the relative mean squared error. Varying amount of the noise, we test DCA-$\ell_{1-2}$, ADMM-lasso, half thresholding, CoSaMP, and AIHT on both Gaussian matrix and the ill-conditioned oversampled DCT. For Gaussian measurements, we chose $n = 1024$, $m = 256$, and $s = 48$. For oversampled DCT, $n = 2000$, $m = 100$, $s = 15$, $F = 10$, and the minimum separation $L = 2F$. At each noise level, we run 50 times and record the average SNR of reconstruction (in dB).
Table 2.1: SNR of reconstruction (dB) under Gaussian measurements.

<table>
<thead>
<tr>
<th>snr (dB)</th>
<th>DCA-(\ell_{1-2})</th>
<th>ADMM-lasso</th>
<th>half thresholding</th>
<th>CoSaMP</th>
<th>AIHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>-29.2398</td>
<td>-27.6103</td>
<td>-37.3863</td>
<td>-32.8737</td>
<td>-36.9471</td>
</tr>
<tr>
<td>10</td>
<td>-3.6700</td>
<td>-3.1970</td>
<td>-1.4126</td>
<td>1.1522</td>
<td>-1.4048</td>
</tr>
</tbody>
</table>

Table 2.2: SNR of reconstruction (dB) using overampled DCT matrix.

<table>
<thead>
<tr>
<th>snr (dB)</th>
<th>DCA-(\ell_{1-2})</th>
<th>ADMM-lasso</th>
<th>half thresholding</th>
<th>AIHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>-35.2119</td>
<td>-25.9895</td>
<td>-3.8896</td>
<td>-3.8393</td>
</tr>
<tr>
<td>35</td>
<td>-17.0934</td>
<td>-12.2916</td>
<td>-4.2793</td>
<td>-3.7375</td>
</tr>
<tr>
<td>20</td>
<td>-3.1806</td>
<td>-3.0157</td>
<td>-2.6428</td>
<td>-0.8141</td>
</tr>
</tbody>
</table>

Table 2.1 shows the results under Gaussian measurements. Choosing an appropriate value of \(\lambda\) is necessary for both DCA-\(\ell_{1-2}\) and ADMM-lasso being well functional, for which we employ a "trial and error" strategy. CoSaMP and AIHT do not need such a parameter, whereas half thresholding embraces a self-adjusting \(\lambda\) during iterations. As a trade-off, however, they all require an estimate on the sparsity of \(\bar{x}\), for which we used the true value in the experiment. With this piece of crucial information, it then appears reasonable that they perform better than DCA-\(\ell_{1-2}\) and ADMM-lasso when there is not much noise, producing relatively smaller SNR of reconstruction. Table 2.2 shows the results for oversampled DCT with \(F = 10\). In this case, we do not display the result for CoSaMP since it yields huge errors. Half thresholding and AIHT are not robust as suggested by Table 2.2. In contrast, DCA-\(\ell_{1-2}\) and ADMM-lasso perform much better, still doing the job under moderate amount of noise. Nevertheless, due to the large coherence of \(A\), their performance have sunk compared to that in the Gaussian case. In either case, DCA-\(\ell_{1-2}\) consistently beats ADMM-lasso.

### 2.4.4 MRI reconstruction

We present a two-dimensional example of reconstructing Magnetic Resonance Imaging (MRI) from a limited number of projections. It is first introduced in [13] to demonstrate the success of compressed sensing. The signal/image is a Shepp-Logan phantom of size 256 × 256, as shown in Fig. 2.6. In this case, it is the gradient of the signal that is sparse, and therefore
the work [13] is to minimize $\ell_1$ norm of the gradient, or the so-called total variation (TV),

\[
\min \|\nabla u\|_1 \quad \text{subject to} \quad RFu = f
\]

where $F$ denotes the Fourier transform, $R$ is the sampling mask in the frequency space and $f$ is the data. It is claimed in [13] that 22 projections are necessary to have exact recovery, while we find 10 projections suffice by using the split Bregman method.

Our proposed $\ell_1-2$ on the gradient is expressed as

\[
\min |\partial_x u| + |\partial_y u| - \sqrt{|\partial_x u|^2 + |\partial_y u|^2} \quad \text{subject to} \quad RFu = f
\]

It is anisotropic TV. We apply the technique of DCA by linearizing the $\ell_2$ norm of the gradient,

\[
u^{k+1} = \arg \min_{u, d_x, d_y} |d_x| + |d_y| - \frac{(d_x, d_y)^T (\partial_x u^k, \partial_y u^k)}{\sqrt{|\partial_x u^k|^2 + |\partial_y u^k|^2}} + \frac{\mu}{2} \|RFu - f\|_2^2 + \frac{\lambda}{2} \|d_x - \partial_x u\|_2^2 + \frac{\lambda}{2} \|d_y - \partial_y u\|_2^2
\]

Let $(t_x, t_y) = (\partial_x u^k, \partial_y u^k)/\sqrt{|\partial_x u^k|^2 + |\partial_y u^k|^2}$ at the current step $u^k$. The subproblem to obtain a new solution $u^{k+1}$ can be solved by the split Bregman method, as detailed in Algorithm 3. Note that the matrix to be inverted in the algorithm is diagonal.

**Algorithm 3** The split Bregman method to solve (2.4)

Define $u = d_x = d_y = b_x = b_y = 0, z = f$ and MAXinner, MAXouter.

Let $D$ and $D^T$ be forward and backward difference operators respectively.

for 1 to MAXouter do
  for 1 to MAXinner do
    $u = (\mu R^T R - \lambda F \Delta F^T)^{-1} (\mu F^T R z + \lambda D_x^T (d_x - b_x) + \lambda D_y^T (d_y - b_y))$
    $d_x = S(D_x u + b_x + t_x/\lambda, 1/\lambda)$
    $d_y = S(D_y u + b_y + t_y/\lambda, 1/\lambda)$
    $b_x = b_x + D_x u - d_x$
    $b_y = b_y + D_y u - d_y$
  end for
  $z = z + f - RF u$
end for
Fig. 2.6 shows the exact recovery of 8 projections using the proposed method. We also compare with the classical filtered-back projection (FBP) and $\ell_1$ on the gradient or TV minimization, whose relative errors are 0.99 and 0.1 respectively. A similar work is reported in [18], where 10 projections are required for $\ell_p$ (p=0.5) on the gradient.
Chapter 3

Iterative $\ell_1$ Minimization for Non-convex Compressed Sensing

Inspired by the success of basis pursuit, researchers have began to investigate various non-convex CS models and algorithms. More and more empirical studies have shown that non-convex CS methods usually outperform basis pursuit when matrix $A$ is RIP-like, in the sense that they require fewer linear measurements to reconstruct signals of interest. Instead of minimizing $\ell_1$ norm, it is natural to consider minimization of non-convex (concave) sparse metrics, for instance, $\ell_q$ (quasi-)norm ($0 < q < 1$) [19, 20, 59], capped-$\ell_1$ [89, 63], and transformed-$\ell_1$ [61, 88]. Another category of CS methods in spirit rely on support detection of $\hat{x}$. To name a few, there are orthogonal matching pursuit (OMP) [73], iterative hard thresholding (IHT) [6], (re)weighted-$\ell_1$ scheme [17], iterative support detection (ISD) [75], and their variations [66, 92, 57].

On the other hand, it has been proved that even if $A$ is not RIP-like and contains highly correlated columns, basis pursuit still enables sparse recovery under certain conditions of $\hat{x}$ involving its support [12]. In this scenario, most of the existing non-convex CS methods, however, are not that robust to the conditioning of $A$, as suggested by [83]. Their success rates will drop as columns of $A$ become more and more correlated. In [83], based on the difference of convex functions algorithm (DCA) [70, 71], the authors propose DCA-$\ell_{1-2}$ for minimizing the difference of $\ell_1$ and $\ell_2$ norms [34, 82]. Extensive numerical experiments [83, 62, 63] show that DCA-$\ell_{1-2}$ algorithm consistently outperforms $\ell_1$ minimization, irrespective
of the conditioning of \( A \).

Stimulated by the empirical evidence found in [83, 62, 63], we propose a general DCA-based CS framework for the minimization of a class of concave sparse metrics. More precisely, we consider the reconstruction of a sparse vector \( \bar{x} \in \mathbb{R}^n \) by minimizing sparsity-promoting metrics:

\[
\min_{x \in \mathbb{R}^n} P(|x|) \quad \text{s.t.} \quad Ax = A\bar{x}.
\]

Throughout the paper, we assume that \( P(x) \) always takes the form \( \sum_{i=1}^{n} p(x_i) \) unless otherwise stated. To promote sparsity, \( p \) defined on \([0, +\infty)\) generally satisfies:

- \( p \) is continuous, concave and increasing.

- The right derivative \( p'(0+) \) exists with \( p'(0+) > 0 \).

A number of sparse metrics in the literature enjoy the above properties, including smoothly clipped absolute deviation (SCAD) [35], capped-\( \ell_1 \), transformed-\( \ell_1 \), and of course \( \ell_1 \) itself. Although \( \ell_q \) (\( q \in (0, 1) \)) and logarithm functional do not meet the second condition, their smoothed versions \( p(t) = (t + \varepsilon)^q \) and \( p(t) = \log(t + \varepsilon) \) are differentiable at zero. These proposed properties will be essential in the algorithm design as well as in the proof of main results.

Our proposed algorithm calls for solving a sequence of minimization subproblems. The objective of each subproblem is \( ||x||_1 \) plus a linear term, which is convex and tractable. We further validate robustness of this framework, by showing theoretically and numerically that it performs at least as well as basis pursuit in terms of uniform sparse recovery, independent of the conditioning of \( A \) and sparsity metric.

**Notations.** \( \mathbf{0} \in \mathbb{R}^n \) is the vector of zeros, and similar to \( \mathbf{1} \). \( \circ \) is Hadamard (entry-wise) product, meaning that \( x \circ y = \sum_{i} x_i y_i \). \( I_m \) is the identity matrix of dimension \( m \). For any function \( g \) on \( \mathbb{R}^n \), \( \nabla g(x) \in \partial g(x) \) is a subgradient of \( g \) at \( x \). For any set \( \Omega \subseteq \mathbb{R}^n \), \( \iota_{\Omega}(x) \) is given by

\[
\iota_{\Omega}(x) := \begin{cases} 
0 & \text{if } x \in \Omega, \\
\infty & \text{if } x \not\in \Omega.
\end{cases}
\]
3.1 Iterative $\ell_1$ framework

Our proposed iterative $\ell_1$ framework for solving (3.1) is built on $\ell_1$ minimization and DCA. Note that (3.1) can be equivalently written as

$$\min_{x \in \mathbb{R}^n} P(|x|) + \iota_{\{x : Ax = A\bar{x}\}}(x).$$

We then rewrite the above objective in DC decomposition form:

$$P(|x|) + \iota_{\{x : Ax = A\bar{x}\}}(x) = (p'(0+)\|x\|_1 + \iota_{\{x : Ax = A\bar{x}\}}(x)) - (p'(0+)\|x\|_1 - \sum_{i=1}^{n} p(|x_i|))$$

Clearly the first term on the right-hand side is convex in terms of $x$. We show below that the second term is also a convex function.

**Proposition 3.1.1.** $p'(0+)\|x\|_1 - \sum_{i=1}^{n} p(|x_i|)$ is convex in $x$.

**Proof.** For notational convenience, define $f(t) := p'(0+)t - p(t)$ on $[0, \infty)$. Since $p$ is concave on $[0, \infty)$, we have that $f$ is convex on $[0, \infty)$. We only need to show that $f(|\cdot|)$ is convex on $\mathbb{R}$, or equivalently, for all $t_1, t_2 \in \mathbb{R}$, $a \in (0, 1)$,

$$f(|at_1 + (1-a)t_2|) \leq af(|t_1|) + (1-a)f(|t_2|).$$

**Case 1.** If $t_1$ and $t_2$ have the same sign or one of them is 0. Since $f(|at_1 + (1-a)t_2|) =$
\[ f(a|t_1| + (1-a)|t_2|) \text{ and } f \text{ is convex on } [0, \infty), \text{ then the above inequality holds.} \]

**Case 2.** If \( t_1 \) and \( t_2 \) are of the opposite sign. By the concavity of \( p \) on \([0, \infty)\), we have

\[
p(t) \leq p(0) + p'(0+)t, \quad \forall t > 0,
\]

that is, \( f(t) \geq f(0) \) for all \( t > 0 \). Without loss of generality, we suppose \( a|t_1| \geq (1-a)|t_2| \).

Then

\[
f(|at_1 + (1-a)t_2|) = f(a|t_1| - (1-a)|t_2|)
\leq \frac{(1-a)(|t_1| + |t_2|)}{|t_1|} f(0) + \frac{a|t_1| - (1-a)|t_2|}{|t_1|} f(|t_1|)
\leq (1-a)f(|t_2|) + \frac{(1-a)|t_2|}{|t_1|} f(|t_1|) + \frac{a|t_1| - (1-a)|t_2|}{|t_1|} f(|t_1|)
= af(|t_1|) + (1-a)f(|t_2|)
\]

In the first inequality above, we used the convexity of \( f \) on \([0, \infty)\), whereas in the second one, we used the fact that \( f(t) \geq f(0) \) for \( t > 0 \).

At the \((k+1)\)th iteration, DCA calls for linearization of the second convex term at the current guess \( x^{(k)} \), and solving the resulting convex subproblem for \( x^{(k+1)} \). After converting back the linear constraint and removing the constant and the factor of \( p'(0+) \), we iterate:

\[
x^{(k+1)} = \arg \min_x \|x\|_1 - \langle R(x^{(k)}), x \rangle \quad \text{s.t.} \quad Ax = A\tilde{x},
\]

where

\[
R(x) := \operatorname{sgn}(x) \circ (1 - \frac{\nabla P(|x|)}{p'(0+)}) \in \partial(\| \cdot \|_1 - \frac{P(| \cdot |)}{p'(0+)})(x).
\]

Be aware that \( \nabla P(|x|) \in \partial P(\cdot)(|x|) \) denotes subgradient of \( P \) at \( |x| \) instead of subgradient of \( P(\cdot) \) at \( x \). In this way, the subproblem reduces to minimizing \( \|x\|_1 \) plus a linear term of \( x \), which can be efficiently solved by a variety of state-of-the-art algorithms for basis pursuit (with minor modifications). In Table 3.1, we list some non-convex metrics and the corresponding iterative \( \ell_1 \) algorithm.

For initialization, we take \( x^{(0)} = R(x^{(0)}) = 0 \), which is basically to solve \( \ell_1 \) minimization.
Table 3.1: Examples of sparse metrics and associated iterative \( \ell_1 \) scheme

<table>
<thead>
<tr>
<th>Sparse metric</th>
<th>( p(t) )</th>
<th>( p'(0+) )</th>
<th>( (R(x))_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capped-( \ell_1 )</td>
<td>( \min{t, \theta}, \theta &gt; 0 )</td>
<td>1</td>
<td>( \operatorname{sgn}(x_i)_{\overline{</td>
</tr>
<tr>
<td>Transformed-( \ell_1 )</td>
<td>( \frac{\theta + 1}{\theta} \frac{t}{t + \theta}, \theta &gt; 0 )</td>
<td>( \frac{\theta + 1}{\theta} )</td>
<td>( \operatorname{sgn}(x_i)(1 - \frac{\theta}{</td>
</tr>
<tr>
<td>Smoothed log</td>
<td>( \log(t + \varepsilon), \varepsilon &gt; 0 )</td>
<td>( \frac{1}{\varepsilon} )</td>
<td>( \operatorname{sgn}(x_i)(1 - \frac{\varepsilon}{</td>
</tr>
<tr>
<td>Smoothed ( \ell_q )</td>
<td>( (t + \varepsilon)^q, \varepsilon &gt; 0 )</td>
<td>( q\varepsilon^{-1} )</td>
<td>( \operatorname{sgn}(x_i)(1 - (\frac{\varepsilon}{</td>
</tr>
</tbody>
</table>

The proposed algorithm is thus summarized in Algorithm 4 below. Due to the descending property of DCA, Algorithm 4 produces a convergent sequence \( \{P(x^{(k)})\} \). Beyond that, we shall not prove any stronger convergence result in this paper. We refer the readers to [83], in which subsequential convergence of \( \{x^{(k)}\} \) is established for DCA-\( \ell_1 \).

**Algorithm 4** Iterative \( \ell_1 \) minimization

Initialize: \( x^{(0)} = 0 \).

for \( k = 1, 2, \ldots \) do

\( y^{(k)} = \operatorname{sgn}(x^{(k)}) \circ \left(1 - \frac{\nabla P(|Dx|)}{p'(0+)}\right) \)

\( x^{(k+1)} = \arg \min_x \|x\|_1 - \langle y^{(k)}, x \rangle \) s.t. \( Ax = A\bar{x} \)

end for

**Extensions.** Two natural extensions of (3.1) are regularized model:

\[
(3.3) \quad \min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda P(|Dx|),
\]

and denoising model:

\[
(3.4) \quad \min_{x \in \mathbb{R}^n} P(|Dx|) \text{ s.t. } \|Ax - b\|_2 \leq \sigma,
\]

where \( b \) is the measurement, \( D \) is a general matrix, and \( \lambda, \sigma > 0 \) are parameters. They find applications in magnetic resonance imaging (MRI) [13], total variation (TV) denoising [69] and so on. We can show that DC decomposition of \( P(|Dx|) \) is

\[
P(|Dx|) = p'(0+)\|Dx\|_1 - (p'(0+))\|Dx\|_1 - P(|Dx|)).
\]

The iterative \( \ell_1 \) frameworks are detailed in Algorithms 5 and 6 respectively.
Algorithm 5 Iterative $\ell_1$ regularization

Initialize: $x^{(0)} = 0$.

\begin{align*}
\text{for } k = 1, 2, \ldots & \text{ do} \\
& y^{(k)} = D^T \left( \text{sgn}(Dx^{(k)}) \circ \left( 1 - \frac{\nabla P(|Dx^{(k)}|)}{p'(0^+)} \right) \right) \\
& x^{(k+1)} = \arg \min_x \frac{1}{2} \|Ax - b\|_2^2 + \lambda p'(0^+) \left( \|Dx\|_1 - \langle y^{(k)}, x \rangle \right)
\end{align*}

end for

Algorithm 6 Iterative $\ell_1$ denoising

Initialize: $x^{(0)} = 0$.

\begin{align*}
\text{for } k = 1, 2, \ldots & \text{ do} \\
& y^{(k)} = D^T \left( \text{sgn}(Dx^{(k)}) \circ \left( 1 - \frac{\nabla P(|Dx^{(k)}|)}{p'(0^+)} \right) \right) \\
& x^{(k+1)} = \|Dx\|_1 - \langle y^{(k)}, x \rangle \quad \text{s.t.} \quad \|Ax - b\|_2 \leq \sigma
\end{align*}

end for

3.2 Recovery results

Although in general global minimum is not guaranteed in minimization, we can show that its performance is provably robust to the conditioning of measurement matrix $A$, by proving that it always tends to sharpen $\ell_1$ solution.

Let us take another look at the assumptions on $p$ which were crucial in the proof of Proposition 3.1.1. Since $p$ is concave and increasing on $[0, \infty)$, $0 \leq \frac{\nabla P(|x|)}{p'(0^+)} \leq 1$, $\forall x \in \mathbb{R}, 1 \leq i \leq n$, and thus $\|R(x)\|_\infty \leq 1$. Now we are ready to show the main results.

**Theorem 3.2.1** (Support-wise uniform recovery). Let $T \subseteq \{1, \ldots, n\}$ be an arbitrary but fixed index set. If basis pursuit uniquely recovers all $\bar{x}$ supported on $T$, so does (3.2).

**Proof.** By the assumption that basis pursuit uniquely recovers all $\bar{x}$ supported on $T$, and by the well-known null space property [44] for $\ell_1$ minimization, we must have

$$\|h_T\|_1 < \|h_{\ell^\infty}\|_1, \quad \forall h \in \text{Ker}(A) \setminus \{0\},$$

and $x^{(1)} = \bar{x}$ in (3.2). The 2\textsuperscript{nd} step of DCA reads

$$x^{(2)} = \arg \min_x \|x\|_1 - \langle R(\bar{x}), x \rangle \quad \text{s.t.} \quad Ax = A\bar{x}.$$
Let \( x^{(2)} = \bar{x} + h^{(2)} \), then

\[
\|\bar{x}\|_1 - \langle R(\bar{x}), \bar{x} \rangle \geq \|\bar{x} + h^{(2)}\|_1 - \langle R(\bar{x}), \bar{x} + h^{(2)} \rangle
\]

\[
\implies \|\bar{x}\|_1 - \langle R(\bar{x}), \bar{x} \rangle \geq \|\bar{x}\|_1 + \langle \text{sgn}(\bar{x}), h^{(2)}_T \rangle + \|h^{(2)}_T\|_1 - \langle R(\bar{x}), \bar{x} + h^{(2)}_T \rangle
\]

\[
\iff -\langle \text{sgn}(\bar{x}) - R(\bar{x}), h^{(2)}_T \rangle \geq \|h^{(2)}_T\|_1
\]

\[
\iff -\langle \text{sgn}(\bar{x}) \circ \frac{\nabla P(|\bar{x}|)}{p'(0+)} , h^{(2)}_T \rangle \geq \|h^{(2)}_T\|_1
\]

Since \( \|\nabla P(|\bar{x}|)\|_\infty \leq 1 \), we have \( \|h^{(2)}_T\|_1 \leq \|h^{(2)}_T\|_1 \). As a result, \( h^{(2)} \) must be \( 0 \). \( \square \)

If nonzero entries of \( \bar{x} \) have the same magnitude, a stronger result holds that (3.2) recovers any fixed signal whenever basis pursuit does.

**Theorem 3.2.2** (Recovery of equal-height signals). Let \( \bar{x} \) be a signal with equal-height peaks supported on \( T \), i.e. \( |x_i| = |x_j|, \forall i, j \in T \). If the basis pursuit uniquely recovers \( \bar{x} \), so does (3.2).

**Proof.** If basis pursuit uniquely recovers \( \bar{x} \), then for all \( h \in \text{Ker}(A) \setminus \{0\} \), \( \|\bar{x}\|_1 < \|\bar{x} + h\|_1 = \|\bar{x} + h_T\|_1 + \|h_T\|_1 \). This implies that for all \( h \in \text{Ker}(A) \setminus \{0\} \) and \( \|h\|_\infty \leq \min_{i \in T} |\bar{x}_i| \), \( \|\bar{x}\|_1 < \|\bar{x} + h_T\|_1 + \|h_T\|_1 = \|\bar{x}\|_1 + \langle \text{sgn}(\bar{x}), h_T \rangle + \|h_T\|_1 \). So for all \( h \in \text{Ker}(A) \setminus \{0\} \) and \( \|h\|_\infty \leq \min_{i \in T} |\bar{x}_i| \), we have \( -\langle \text{sgn}(\bar{x}), h_T \rangle < \|h_T\|_1 \).

Therefore,

(3.5) \quad \langle \overline{\text{sgn}(\bar{x})}, h_T \rangle < \|h_T\|_1, \forall h \in \text{Ker}(A) \setminus \{0\},

and also \( x^{(1)} = \bar{x} \).

We let \( x^{(2)} = \bar{x} + h^{(2)} \), and Suppose that \( h^{(2)} \neq 0 \). Repeating the argument in Theorem 3.2.1 and by (3.5), we arrive at

\[
-\langle \text{sgn}(\bar{x}) \circ \frac{\nabla P(|\bar{x}|)}{p'(0+)} , h^{(2)}_T \rangle \geq \|h^{(2)}_T\|_1 > -\langle \text{sgn}(\bar{x}), h^{(2)}_T \rangle.
\]

Since peaks of \( \bar{x} \) have equal height, \( \langle \frac{\nabla P(|\bar{x}|)}{p'(0+)} \rangle \in [0, 1) \) is a constant for all \( i \in T \). So \( -\langle \text{sgn}(\bar{x}) \circ \frac{\nabla P(|\bar{x}|)}{p'(0+)} , h^{(2)}_T \rangle \) is non-negative and less than \( -\langle \text{sgn}(\bar{x}), h^{(2)}_T \rangle \), which leads to a contradiction. \( \square \)
Remark 3.2.1. We did not make any assumption on the matrix $A$ in the above theorems.

Remark 3.2.2. It is easy to check that the recovery results can be extended to DCA-$\ell_{1-2}$ proposed in [83]. For DCA-$\ell_{1-2}$, $P(x) = \|x\|_1 - \|x\|_2$, which couples all the components together, and

$$R(x) = \begin{cases} \frac{x}{\|x\|_2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Theorem 3.2.1 provides a theoretical explanation for the experimental observations made in [83, 62, 63] that DCA-$\ell_{1-2}$ performs consistently better than $\ell_1$ minimization.

### 3.3 Numerical experiments

To demonstrate effectiveness of the proposed framework, we reconstruct sparse vector $\bar{x}$ using iterative $\ell_1$ algorithm (Algorithm 5 with $D = I_n$) for minimizing the regularized model (3.3) with smoothed $\ell_q$ norm (IL$_{1-\ell_q}$) and smoothed logarithm functional (IL$_{1-log}$), and compare them with two state-of-the-art non-convex CS algorithms, namely IRLS-$\ell_q$ [59] and IRL$_1$ [17]. Note that IRLS-$\ell_q$ and IRL$_1$ attempt to minimize $\ell_q$ and logarithm, respectively, and both involve a smoothing strategy in minimization. So it would be particularly interesting to compare IL$_{1-\ell_q}$ with IRLS-$\ell_q$, and IL$_1-log$ with IRL$_1$. $q = 0.5$ is chosen for IRLS-$\ell_q$ and IL$_1-\ell_q$ in all experiments. We shall also include ADMM-$\ell_1$ [8] for solving $\ell_1$ regularization (LASSO) in comparison.

Experiments are carried out as follows. We first sample a sensing matrix $A \in \mathbb{R}^{m \times n}$, and generate a test signal $\bar{x} \in \mathbb{R}^n$ of sparsity $s$ supported on a random index set with i.i.d. Gaussian entries. We then compute the measurement $A\bar{x}$ and apply each solver to produce a reconstruction $x^*$ of $\bar{x}$. The reconstruction is called a success if

$$\frac{\|x^* - \bar{x}\|_2}{\|\bar{x}\|_2} < 10^{-3}.$$ 

We run 100 independent realizations and record the corresponding success rates at different sparsity levels.

Matrices for testing. Again we test random Gaussian matrix with $m = 64$ and $n = 256$. 

48
We also use more ill-conditioned oversampled partial DCT matrix. We test at three coherence levels with $F = 5, 10, 15$. Note that $\mu(A) \approx 0.95$ for $F = 5$, $\mu(A) \approx 0.998$ for $F = 10$, and $\mu(A) \approx 0.9996$ for $F = 15$. Set $L = 2F$ in experiments.

**Algorithm implementation.** For ADMM-$\ell_1$, we let $\lambda = 10^{-6}$, $\beta = 1$, $\rho = 10^{-5}$, $\epsilon^{\text{abs}} = 10^{-7}$, $\epsilon^{\text{rel}} = 10^{-5}$, and the maximum number of iterations $\text{maxiter} = 5000$ [8, 83]. For IRLS-$\ell_q$, $\text{maxiter} = 1000$, $\text{tol} = 10^{-8}$. For reweighted $\ell_1$, the smoothing parameter $\varepsilon$ is adaptively updated as introduced in [17], and the outer iteration criterion is stopped if the relative error between two consecutive iterates is less than $10^{-2}$. The weighted $\ell_1$ minimization subproblems is solved by the YALL1 solver (available at http://yall1.blogs.rice.edu/). The tolerance for YALL1 was set to $10^{-6}$. All other settings of the algorithms are set to default ones.

For IL$_1$-$\ell_q$, we let $\lambda = 10^{-6}$, and the smoothing parameter $\varepsilon = \max\{\frac{|x^{(1)}|_d}{3}, 0.01\}$, where $x^{(1)}$ is the output from the first iteration, which is also the solution to LASSO. $|x|_d$ denotes the $d^{\text{th}}$ largest entry of $|x|$. We set $d$ to $\lfloor \frac{m}{4} \rfloor$. For IL$_1$-log, $\varepsilon = \max\{|x^{(1)}|_d, 0.01\}$. The subproblems are solved by alternating direction method of multipliers (ADMM), which is detailed Algorithm 2. The parameters for solving subproblems are the same as that for ADMM-$\ell_1$.

**Interpretation of results.** The plot of success rates is shown in Figure 3.2. When $A$ is Guassian, we see that all non-convex CS solvers are comparable and much better than ADMM-$\ell_1$, with IRLS-$\ell_q$ being the best. For oversampled DCT matrices, we see that the success rates of IRLS-$\ell_q$ and IRL$_1$ drop as $F$ increases, whereas the proposed IL$_1$-$\ell_q$ and IL$_1$-log are robust and consistently outperform ADMM-$\ell_1$. 

49
Figure 3.2: Plots of success rates.
Chapter 4

PhaseLiftOff: an Accurate and Stable Phase Retrieval Method Based on Difference of Trace and Frobenius Norms

In this chapter, we study a nonconvex Lipschitz continuous metric, the difference of trace and Frobenius norms, and show that its minimization characterizes the rank one solution exactly and serves as a new tool to solve the phase retrieval problem. We shall see that it is more accurate than trace norm or the heuristic log-determinant [39, 40], and performs the best when the number of measurements approaches the theoretical lower limit [2]. We shall call our method PhaseLiftOff, where Off is short for subtracting off Frobenius norm from the trace norm in PhaseLift [11, 9].

From the point of view of energy minimization, the phase retrieval problem is simply:

\[(4.1) \min_{X \in \mathbb{C}^{n \times n}} \|A(X) - b\|_2^2 \quad \text{s.t.} \quad X \succeq 0, \, \text{rank}(X) = 1.\]

This is a least squares-type model applicable to both noiseless and noisy cases. Our main contribution in this work is to reformulate the phase retrieval problem (4.1) as a nearly equivalent nonconvex optimization problem that can be efficiently solved by the so-called
difference of convex functions algorithm (DCA). Specifically, we propose to solve the following regularization problem:

\[(4.2) \quad \min_{X \in \mathbb{C}^{n \times n}} \varphi(X) := \frac{1}{2} \|A(X) - b\|_2^2 + \lambda(\text{Tr}(X) - \|X\|_F) \quad \text{s.t.} \quad X \succeq 0.\]

Recently the authors of [34, 83] have reported that minimizing the difference of $\ell_1$ and $\ell_2$ norms would promote sparsity when recovering a sparse vector from linear measurements. The $\ell_1 - \ell_2$ minimization is extremely favorable for the reconstruction of the 1-sparse vector $x$ because $\|x\|_1 - \|x\|_2$ attains the possible minimum value zero at such $x$. Note that when $X \succeq 0$, $\text{Tr}(X)$ is nothing but the $\ell_1$ norm of the vector $\sigma(X)$ formed by $X$’s singular values and $\|X\|_F$ the $\ell_2$ norm. Thus (4.2) is basically the counterpart of $\ell_1 - \ell_2$ minimization with nonnegativity constraint discussed in [34]. Similarly, $\text{Tr}(X) - \|X\|_F$ is minimized when $\sigma(X)$ is 1-sparse or equivalently rank($X$) = 1.

4.1 Notations and Preliminaries.

For any $X, Y \in \mathbb{C}^{n \times n}$, $\langle X, Y \rangle = \text{Tr}(X^*Y)$ is the inner product for matrices, which is a generalization of that for vectors. The Frobenius norm of $X$ is $\|X\|_F = \sqrt{\langle X, X \rangle}$, while $X \circ Y$ denotes the entry-wise product, namely $(X \circ Y)_{ij} = X_{ij}Y_{ij}, \forall i, j$. $\text{diag}(X) \in \mathbb{C}^n$ extracts the diagonal elements of $X$. The spectral norm of $X$ is $\|X\|_2$, while the nuclear norm of $X$ is $\|X\|_*$. We have the following elementary inequalities:

$$\|X\|_2 \leq \|X\|_F \leq \sqrt{\text{rank}(X)}\|X\|_2,$$

and

$$\|X\|_F \leq \|X\|_* \leq \sqrt{\text{rank}(X)}\|X\|_F.$$

For any vector $x \in \mathbb{R}^m$, $\|x\|_1$ and $\|x\|_2$ are the $\ell_1$ norm and $\ell_2$ norm respectively, while Diag($x$) $\in \mathbb{R}^{m \times m}$ is the diagonal matrix with $x$ on its diagonal.

We assume that $m \geq n$ and that $A$ is of full rank unless otherwise stated, i.e. rank($A$) = $n$. Recall that $\mathcal{A}(X) := \text{diag}(A^*XA)$ is a linear operator from $\mathbb{H}^{n \times n}$ to $\mathbb{R}^m$, then the adjoint
operator $\mathcal{A}^*$ is defined as $\mathcal{A}^*(x) := A\text{Diag}(x)A^* \in \mathbb{H}^{n \times n}$ for all $x \in \mathbb{R}^m$. Furthermore, the norms of $\mathcal{A}$ and $\mathcal{A}^*$ are given by

$$
\|\mathcal{A}\| := \sup_{X \in \mathbb{H}^{n \times n} \setminus \{0\}} \frac{\|\mathcal{A}(X)\|_2}{\|X\|_F}, \quad \|\mathcal{A}^*\| := \sup_{x \in \mathbb{R}^m \setminus \{0\}} \frac{\|\mathcal{A}^*(x)\|_F}{\|x\|_2}.
$$

Since $(\mathbb{H}^{n \times n}, \langle \cdot, \cdot \rangle)$ and $(\mathbb{R}^m, \langle \cdot, \cdot \rangle)$ are both Hilbert spaces, we have

$$
(4.3) \quad \|\mathcal{A}^*\|^2 = \|\mathcal{A}\|^2 = \|\mathcal{A}\mathcal{A}^*\|.
$$

The following lemma will be frequently used in the proofs.

**Lemma 4.1.1.** Suppose $X, Y \in \mathbb{C}^{n \times n}$ and $X, Y \succeq 0$, then

1. $\langle X, Y \rangle \geq 0$.

2. $\langle X, Y \rangle = 0 \iff XY = 0$.

3. $\|\mathcal{A}(X)\|_2 = 0 \iff X = 0$.

**Proof.** (1) Suppose $Y = U\Sigma U^*$ is the singular value decomposition (SVD), let $Y^{\frac{1}{2}} := U\Sigma^{\frac{1}{2}}U^* \succeq 0$, where the diagonal elements of $\Sigma^{\frac{1}{2}}$ are square roots of the singular values. Then we have $Y = Y^{\frac{1}{2}}Y^{\frac{1}{2}}$ and

$$
\langle X, Y \rangle = \text{Tr}(X^*Y) = \text{Tr}(XY) = \text{Tr}(Y^{\frac{1}{2}}XY^{\frac{1}{2}}) \geq 0.
$$

The last inequality holds because $Y^{\frac{1}{2}}XY^{\frac{1}{2}} \succeq 0$.

(2) ”$\Rightarrow$” Further assume $\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}$, where $\Sigma_1 \succ 0$, and let $Z = U^*XU \succeq 0$. By (1), we have $\text{Tr}(Y^{\frac{1}{2}}XY^{\frac{1}{2}}) = \langle X, Y \rangle = 0$, thus $Y^{\frac{1}{2}}XY^{\frac{1}{2}} = 0$. So

$$
0 = \Sigma^{\frac{1}{2}}U^*XU\Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}}Z\Sigma^{\frac{1}{2}} = \begin{pmatrix} \Sigma_1^{\frac{1}{2}} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{12}^* & Z_{22} \end{pmatrix} \begin{pmatrix} \Sigma_1^{\frac{1}{2}} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \Sigma_1^{\frac{1}{2}}Z_{11}\Sigma_1^{\frac{1}{2}} & 0 \\ 0 & 0 \end{pmatrix},
$$

then we have $\Sigma_1^{\frac{1}{2}}Z_{11}\Sigma_1^{\frac{1}{2}} = 0$ and $Z_{11} = 0$. Next we want to show $Z_{12} = 0$. Suppose $Z_{12} \neq 0$,
let us consider \( v_c = \begin{pmatrix} cZ_{12}w \\ w \end{pmatrix} \in \mathbb{C}^n \), where \( w \) is a fixed vector making \( Z_{12}w \) nonzero and \( c \in \mathbb{R} \). Then since \( Z \succeq 0 \), we have

\[
0 \leq v_c^* Zv_c = (cw^*Z_{12}, w^*) \begin{pmatrix} 0 & Z_{12} \\ Z_{12}^* & Z_{22} \end{pmatrix} \begin{pmatrix} cZ_{12}w \\ w \end{pmatrix} = 2c|Z_{12}w|^2 + w^*Z_{22}w, \quad \forall c \in \mathbb{R}.
\]

In above inequality, letting \( c \to -\infty \) leads to a contradiction. Therefore \( Z_{12} = 0 \). A simple computation gives \( U^*XU \Sigma = Z \Sigma = 0 \), and thus \( XY = XU \Sigma U^* = 0 \).

\( \Rightarrow \) If \( XY = 0 \), then \( \langle X, Y \rangle = \text{Tr}(X^*Y) = \text{Tr}(XY) = 0 \)

\( \leq \) Let \( X^{\frac{1}{2}} \succeq 0 \) such that \( X^{\frac{1}{2}}X^{\frac{1}{2}} = X \). Then

\[
0 = \|A(X)\|_2 = \|\text{diag}(A^*X^{\frac{1}{2}}X^{\frac{1}{2}}A)\|_2.
\]

So \( \text{diag}(A^*X^{\frac{1}{2}}X^{\frac{1}{2}}A) = 0 \) and thus \( 0 = \text{Tr}(A^*X^{\frac{1}{2}}X^{\frac{1}{2}}A) = \|A^*X^{\frac{1}{2}}\|_F^2 \). This together with \( \text{rank}(A) = n \leq m \) implies \( X^{\frac{1}{2}} = 0 \).

\( \Rightarrow \) Trivial. \( \square \)

**Karush-Kuhn-Tucker conditions.** Let us consider a first-order stationary point \( \tilde{X} \) of the minimization problem

\[
\min_{X \in \mathbb{C}^{n \times n}} f(X) \quad \text{s.t.} \quad X \succeq 0.
\]

Suppose \( f \) is differentiable at \( \tilde{X} \), then there exists \( \tilde{\Lambda} \in \mathbb{C}^{n \times n} \), such that the following Karush-Kuhn-Tucker (KKT) optimality conditions hold:

- Stationarity: \( \nabla f(\tilde{X}) = \tilde{\Lambda} \).
- Primal feasibility: \( \tilde{X} \succeq 0 \).
- Dual feasibility: \( \tilde{\Lambda} \succeq 0 \).
- Complementary slackness: \( \tilde{X} \tilde{\Lambda} = 0 \).

In order to make better use of the last condition, by Lemma 4.1.1 (2), we can express it as

- Complementary slackness: \( \langle \tilde{X}, \tilde{\Lambda} \rangle = 0 \).
4.2 Exact and Stable Recovery Theory.

In this section, we establish the PhaseLiftOff theory for exact and stable recovery of complex signals.

4.2.1 Equivalence.

We first develop mild conditions that guarantee the full equivalence between Phase Retrieval (4.1) and PhaseLiftOff (4.2).

**Theorem 4.2.1.** Let $\mathcal{A}$ be an arbitrary linear operator from $\mathbb{H}^{n \times n}$ to $\mathbb{R}^m$, and let $b = \mathcal{A}(\hat{X}) + e$. If $\|b\|_2 > \|e\|_2$ and $\lambda > \frac{\|\mathcal{A}\|\|e\|_2}{\sqrt{2} - 1}$, suppose $X^{opt}$ is a solution (global minimizer) to (4.2), then $\text{rank}(X^{opt}) = 1$. Moreover, minimization problems (4.1) and (4.2) are equivalent in the sense that they share the same set of solutions.

**Proof.** Let $X^{opt}$ be a solution to (4.2). Since $\varphi(\hat{X}) = \frac{1}{2}\|e\|_2^2 < \frac{1}{2}\|b\|_2^2 = \varphi(0)$, $X^{opt} \neq 0$. Suppose $\text{rank}(X^{opt}) = r \geq 1$, and let

$$X^{opt} = U\Sigma U^* = (U_1, U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} (U_1, U_2)^* = U_1\Sigma_1U_1^*$$

be the SVD, where $U_1 = (u_1, \ldots, u_r) \in \mathbb{C}^{n \times r}$, $U_2 = (u_{r+1}, \ldots, u_n) \in \mathbb{C}^{n \times (n-r)}$, and $\Sigma_1 = \text{Diag}((\sigma_1, \ldots, \sigma_r)) \in \mathbb{R}^{r \times r}$ with $X^{opt}$'s positive singular values on its diagonal.

Since $X^{opt}$ is a global minimizer, it is also a stationary point. This means KKT conditions must hold at $X^{opt}$, i.e., there exists $\Lambda \in \mathbb{C}^{n \times n}$ such that

$$\begin{align*}
\mathcal{A}^*(\mathcal{A}(X^{opt}) - b) + \lambda(I_n - \frac{X^{opt}}{\|X^{opt}\|_F}) &= \Lambda, \\
X^{opt} &\succeq 0, \quad \Lambda \succeq 0, \quad \langle X^{opt}, \Lambda \rangle = 0.
\end{align*}$$

(4.4)

Rewrite $I_n = UU^* = U_1U_1^* + U_2U_2^*$, then (4.4) becomes

$$\begin{align*}
\mathcal{A}^*(\mathcal{A}(X^{opt}) - b) &= \lambda(I_n - \frac{X^{opt}}{\|X^{opt}\|_F}) - \Lambda = \lambda U_1U_1^* + (\lambda U_2U_2^* - \Lambda) - \frac{\lambda X^{opt}}{\|X^{opt}\|_F}.
\end{align*}$$

55
Taking Frobenius norm of both sides above, we obtain

\[(4.5)\]
\[\|A^* (A(X^{opt}) - b)\|_F = \|\lambda U_1^* + (\lambda U_2^* - \Lambda) - \lambda \frac{X^{opt}}{\|X^{opt}\|_F}\|_F \geq \|\lambda U_1^* + (\lambda U_2^* - \Lambda)\|_F - \lambda.\]

Also, we have 0 = \langle X^{opt}, \Lambda \rangle = \langle U_1^* \Sigma^*_1, \Lambda \rangle = \sum_{i=1}^r \sigma_i \langle u_i^* u_i^*, \Lambda \rangle. But \langle u_i^* u_i^*, \Lambda \rangle \geq 0, since \Lambda \succeq 0 and u_i^* u_i^* \succeq 0. So \langle u_i^* u_i^*, \Lambda \rangle = 0 for 1 \leq i \leq m, and \langle U_1^*, \Lambda \rangle = \sum_{i=1}^r \langle u_i^* u_i^*, \Lambda \rangle = 0.

Moreover,
\[\langle U_1^* U_2^*\rangle = \sum_{i=1}^r \sum_{j=r+1}^n \langle u_i^* u_i^*, u_j^* u_j^*\rangle = \sum_{i=1}^r \sum_{j=r+1}^n \langle u_j^* u_i, u_j^* u_i\rangle = 0.\]

In a word, \(U_1^*\) is orthogonal to both \(U_2^*\) and \(\Lambda\). Then from Pythagorean theorem it follows that

\[\|\lambda U_1^* + (\lambda U_2^* - \Lambda)\|_F = \sqrt{\|\lambda U_1^*\|_F^2 + \|\lambda U_2^* - \Lambda\|_F^2} \geq \lambda \|U_1^*\|_F,\]

and thus (4.5) reduces to

\[(4.6)\]
\[\|A^* (A(X^{opt}) - b)\|_F \geq \|U_1^*\|_F - \lambda = \lambda (\sqrt{r} - 1).\]

On the other hand, since
\[\|A(X^{opt}) - b\|_2 \leq \sqrt{\|A(X^{opt}) - b\|_2^2 + 2\lambda \text{Tr}(X^{opt} - \|X^{opt}\|_F) = \sqrt{2\phi(X^{opt})}},\]

we have

\[(4.7)\]
\[\|A^* (A(X^{opt}) - b)\|_F \leq \|A^*\| \|A(X^{opt}) - b\|_2 = \|A\| \|A(X^{opt}) - b\|_2 \leq \|A\| \sqrt{2\phi(X^{opt})} \leq \|A\| \sqrt{2\phi(\hat{X})} = \|A\| \|e\|_2.\]
Combining (4.6) and (4.7) gives
\[ \lambda (\sqrt{r} - 1) \leq \|A\|\|e\|_2, \]
or equivalently
\[ r \leq \left( \frac{\|A\|\|e\|_2}{\lambda} + 1 \right)^2 < 2. \]

The last inequality above follows from the assumption \( \lambda > \frac{\|A\|\|e\|_2}{\sqrt{2} - 1} \). \( r \) is a natural number, so \( r = 1 \).

Note that \( \text{Tr}(X) - \|X\|_F \geq 0 \) for \( X \succeq 0 \) with equality when \( \text{rank}(X) = 1 \). It is not hard to see the equivalence between (4.1) and (4.2).

\[ \square \]

**Corollary 4.2.1.** In the absence of measurement noise, the equivalence between (4.1) and (4.2) holds for all \( \lambda > 0 \). In this sense, (4.2) is essentially a parameter-free model.

Theorem 4.2.1 claims that provided the noise in measurement is smaller than the measurement itself, all \( \lambda \) that exceed an explicit threshold would work equally well for (4.2) in theory. In contrast, the \( \lambda \) in (1.4) needs to be carefully chosen to balance the fidelity and penalty terms. Particularly in noiseless case, the \( \lambda \) in (4.2) acts like a ’fool-proof’ regularization parameter, and \( A(X) = b \) is always exact at the solution \( X^{\text{opt}} = \hat{X} \) whenever \( \lambda > 0 \), whereas a perfect reconstruction via solving (1.4) generally requires a dynamic \( \lambda \) that goes to 0.

**Remark 4.2.1.** Despite the tremendous room for \( \lambda \) values in view of Theorem 4.2.1, we should point out that in practice the choice of \( \lambda \) could be more subtle because

- The theoretical lower bound \( \frac{\|A\|\|e\|_2}{\sqrt{2} - 1} \) for \( \lambda \) may be too stringent, and a smaller \( \lambda \) could also be feasible.

- Choosing \( \lambda \) too large may reduce the mobility of the energy minimizing iterations due to trapping by local minima.

An efficient algorithm designed for PhaseLiftOff should be as insensitive as possible to the choice of \( \lambda \) when it is large enough.

**4.2.2 Exact and stable recovery under Gaussian measurements.**

In the framework of [11, 10], assuming \( a_i \)'s are i.i.d. complex-valued normally distributed random vectors, we establish the exact recovery and stability results for (4.2). Due to
the equivalence between (4.1) and (4.2) under the conditions stated in Theorem 4.2.1, it suffices to discuss the model (4.1) only. Similar to [10], \( m = O(n) \) measurements suffice to ensure exact recovery in noiseless case or stability in noisy case with probability close to 1. Although the required number of measurements for (4.1) and that for PhaseLift are both on the minimal order \( O(n) \), the scalar factor of the former is actually smaller, and so is the probability of failure.

**Theorem 4.2.2.** Suppose column vectors of \( A \) are i.i.d. complex-valued normally distributed. Fix \( \alpha \in (0, 1) \), there are constants \( \theta, \gamma > 0 \) such that if \( m > \theta [\alpha^{-2} \log \alpha^{-1}] n \), for any \( \hat{X} \), (4.1) is stable in the sense that its solution \( X_{\text{opt}} \) satisfies

\[
\| X_{\text{opt}} - \hat{X} \|_F \leq C_\alpha \frac{\| e \|_2}{\sqrt{m}}
\]

for some constant \( C_\alpha := \frac{\sqrt{2}}{(\sqrt{2} - 1)(1 - \alpha)} > 0 \) with probability at least \( 1 - 3e^{-\gamma m \alpha^2} \). In particular, when \( e = 0 \), the recovery is exact.

The proof is straightforward with the aid of Lemma 5.1 in [11]:

**Lemma 4.2.1 ([11]).** Under the assumption of Theorem 4.2.2, we have that \( A \) obeys the following property with probability at least \( 1 - 3e^{-\gamma m \alpha^2} \): for any Hermitian matrix \( X \) with rank(\( X \)) \( \leq 2 \),

\[
\frac{1}{m} \| A(X) \|_1 \geq 2(\sqrt{2} - 1)(1 - \alpha)\| X \|_2.
\]

**Proof of Theorem 4.2.2.**

Proof. Let \( X_{\text{opt}} = \hat{X} + H \), where \( \hat{X} \) satisfies \( A(\hat{X}) + e = b \), then \( H \) is Hermitian with rank(\( H \)) \( \leq 2 \). Since

\[
\| e \|_2 = \| A(\hat{X}) - b \|_2 \geq \| A(X_{\text{opt}}) - b \|_2 \geq \| A(X_{\text{opt}} - \hat{X}) \|_2 - \| A(\hat{X}) - b \|_2,
\]

we have \( \| A(H) \|_2 \leq 2\| e \|_2 \). Invoking Lemma 4.2.1 above, we further have

\[
\frac{1}{\sqrt{m}} \| A(H) \|_2 \geq \frac{1}{m} \| A(H) \|_1 \geq 2(\sqrt{2} - 1)(1 - \alpha) \| H \|_2 \geq \frac{2(\sqrt{2} - 1)(1 - \alpha)}{\sqrt{2}} \| H \|_F.
\]
Therefore,
\[ \|X^\text{opt} - \hat{X}\|_F = \|H\|_F \leq \frac{\sqrt{2}}{(\sqrt{2} - 1)(1 - \alpha)} \sqrt{m}. \]
The above inequality holds with probability at least \(1 - 3e^{-\gamma m \alpha^2}\).

\[ \square \]

### 4.2.3 Computation of \(\|A\|\).

The \(\|A\|\) in Theorem 4.2.1 can be actually computed. To do this, we first prove the following result:

**Lemma 4.2.2.** \(AA^*(x) = (A^*A \circ A^*A)x\), \(\forall x \in \mathbb{R}^m\), where the overline denotes complex conjugate.

**Proof.** By the definitions of \(A\) and \(A^*\), \(AA^*(x) = \text{diag}(A^*A\text{diag}(x)A^*A)\), then \(\forall 1 \leq i \leq m\), the \(i\)-th entry of \(AA^*(x)\) reads

\[
(AA^*(x))_i = (A^*A\text{diag}(x)A^*A)_{ii} = \sum_{j=1}^{m} x_j (A^*A)_{ij} (A^*A)_{ji}
\]

\[
= \sum_{j=1}^{m} x_j (A^*A)_{ij} (A^*A)_{ij} = \sum_{j=1}^{m} x_j (A^*A \circ A^*A)_{ij}
\]

\[
= ((A^*A \circ A^*A)x)_i.
\]

\[ \square \]

Hence, from Lemma 4.2.2 and (4.3) it follows that

\[ \|A\| = \sqrt{\|AA^*\|} = \sqrt{\|A^*A \circ A^*A\|_2}. \]

It would be interesting to see how fast \(\|A\|\) grows with dimensions \(n\) and \(m\) when \(A\) is a complex-valued random Gaussian matrix. In this setting, \(A\) enjoys approximate \(\ell_1\)-isometry properties as revealed by Lemma 3.1 of [11] (in complex case). Here we are most interested in the part that concerns the upper bound:
Lemma 4.2.3 ([11]). Suppose $A \in \mathbb{C}^{n \times n}$ is random Gaussian. Fix any $\delta > 0$ and assume $m \geq 16\delta^{-2}n$. Then with probability at least $1 - e^{-me^2/2}$, where $\delta/4 = \epsilon^2 + \epsilon$,

$$\frac{1}{m} \|A(X)\|_1 \leq (1 + \delta)\|X\|_*$$

holds for all $X \in \mathbb{C}^{n \times n}$.

Under assumptions of Lemma 4.2.3, with high probability we have

$$\frac{1}{\sqrt{m}} \|A(X)\|_2 \leq \frac{1}{m} \|A(X)\|_1 \leq (1 + \delta)\|X\|_* \leq (1 + \delta)\sqrt{n}\|X\|_F,$$

which implies $\|A\| = O(\sqrt{mn})$. For the phase retrieval problem to be well-posed, $m = O(n)$ is required; for instance, $m = 4n$ would be sufficient according to [2]. Then we expect that $\|A\|$ is on the order of $n$. This can be validated by a simple numerical experiment whose results are shown in Table 4.1 below.

<table>
<thead>
<tr>
<th>$n$</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|A|$</td>
<td>148</td>
<td>291</td>
<td>577</td>
<td>1149</td>
<td>2295</td>
<td>4584</td>
</tr>
</tbody>
</table>

Table 4.1: Fixing $m = 4n$, $\|A\|$ is nearly linear in $n$ with $A$ being complex-valued Gaussian matrix.

4.3 Algorithms.

We consider the computational aspects of the minimization problem (4.2). We apply the DCA to (4.2), where the objective naturally has the DC decomposition

$$(4.9) \quad \varphi(X) = \left(\frac{1}{2}\|A(X) - b\|_2^2 + \lambda \text{Tr}(X)\right) - \lambda\|X\|_F.$$

Since $\varphi(X) \geq 0$ for all $X \succeq 0$, the scheme

$$\begin{cases}
\Delta^k \in \partial\|X^k\|_F, \\
X^{k+1} = \arg \min_{X \in \mathbb{C}^{n \times n}} \left\{ \frac{1}{2}\|A(X) - b\|_2^2 + \lambda \text{Tr}(X) - \lambda(\|X^k\|_F + \langle \Delta^k, X - X^k \rangle) \right\} \text{ s.t. } X \succeq 0.
\end{cases}$$
yields a decreasing and convergent sequence \( \{ \varphi(X^k) \} \). Note that \( \|X\|_F \) is differentiable with gradient \( \frac{X}{\|X\|_F} \) at all \( X \neq 0 \) and that \( 0 \in \partial \|X\|_F \) at \( X = 0 \), by ignoring constants we iterate (4.10)

\[
X^{k+1} = \begin{cases} 
\arg \min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2} \|A(X) - b\|_2^2 + \lambda \text{Tr}(X) & \text{s.t. } X \succeq 0 \quad \text{if } X^k = 0, \\
\arg \min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2} \|A(X) - b\|_2^2 + \lambda \langle X, I_n - \frac{X^k}{\|X^k\|_F} \rangle & \text{s.t. } X \succeq 0 \quad \text{otherwise.}
\end{cases}
\]

Since \( X^k - X^{k-1} \to 0 \) as \( k \to \infty \) (Proposition 4.3.1 (2)), we stop the DCA when

\[
\frac{\|X^k - X^{k-1}\|_F}{\max\{\|X^k\|_F, 1\}} < \text{tol},
\]

for some given tolerance \( \text{tol} > 0 \). In practice the above iteration takes only a few steps to convergence. While the problem (4.2) is nonconvex, empirical studies have shown that the DCA usually produces a global minimizer with a good initialization. In particular, our initialization here is \( X^0 = 0 \), as suggested by the observations in [83]. This amounts to employing the (PhaseLift) solution of the regularized trace-norm minimization problem (1.4) as a start.

4.3.1 Convergence analysis.

We proceed to show that the sequence \( \{X^k\} \) is bounded and \( X^{k+1} - X^k \to 0 \), and limit points of \( \{X^k\} \) are stationary points of (4.2) satisfying KKT optimality conditions. Standard convergence results for the general DCA (e.g. Theorem 3.7 of [71]) take advantage of strong convexity of the DC components. However, the DC components in (4.9) only possess weak convexity as \( \ker(A^*A) \) is generally nontrivial. In this sense, our analysis below is novel.

**Lemma 4.3.1.** Suppose \( X \succeq 0 \), \( \varphi(X) \to \infty \) as \( X \to \infty \).

**Proof.** It suffices to show that for any fixed nonzero \( X \succeq 0 \), \( \varphi(cX) \to \infty \) as \( c \to \infty \).

\[
\varphi(cX) = \frac{1}{2} \|cA(X) - b\|_2^2 + c\lambda(\text{Tr}(X) - \|X\|_F) \geq \frac{1}{2}(c\|A(X)\|_2^2 - \|b\|_2^2).
\]

Since \( X \succeq 0 \) and is nonzero, by Lemma 4.1.1 (3), \( \|A(X)\|_2 > 0 \). Hence, \( c\|A(X)\|_2 - \|b\|_2 \to \infty \) as \( c \to \infty \), which completes the proof. \( \square \)
Lemma 4.3.2. Let \( \{X^k\} \) be the sequence generated by the DCA. For all \( k \in \mathbb{N} \), we have

\[
\varphi(X^k) - \varphi(X^{k+1}) \geq \frac{1}{2} \|A(X^k - X^{k+1})\|^2_2 + \lambda(\|X^{k+1}\|_F - \|X^k\|_F - \langle \Delta^k, X^{k+1} - X^k \rangle) \geq 0,
\]

where \( \Delta^k \in \partial\|X^k\|_F \).

Proof. We first calculate

\[
\varphi(X^k) - \varphi(X^{k+1}) = \frac{1}{2} \|A(X^k - X^{k+1})\|^2_2 + \langle A(X^k - X^{k+1}), A(X^{k+1}) - b \rangle + \lambda \text{Tr}(X^k - X^{k+1}) + \lambda(\|X^{k+1}\|_F - \|X^k\|_F).
\]

Recall that the \((k + 1)\)-th DCA iteration is to solve

\[
X^{k+1} = \arg \min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2} \|A(X) - b\|^2_2 + \lambda \langle X, I_n - \Delta^k \rangle \quad \text{s.t.} \quad X \succeq 0,
\]

where \( \Delta^k \in \partial\|X^k\|_F \). Then by the KKT conditions at \( X^{k+1} \), there exists \( \Lambda^{k+1} \) such that

\[
\mathcal{A}^*(\mathcal{A}(X^{k+1}) - b) + \lambda(I_n - \Delta^k) = \Lambda^{k+1},
\]

\( X^{k+1} \succeq 0, \Lambda^{k+1} \succeq 0, \langle \Lambda^{k+1}, X^{k+1} \rangle = 0. \)

Multiplying (4.13) by \( X^k - X^{k+1} \) (inner product) gives

\[
\langle \mathcal{A}(X^k - X^{k+1}), \mathcal{A}(X^{k+1}) - b \rangle + \lambda \text{Tr}(X^k - X^{k+1}) = \langle \Lambda^{k+1}, X^k \rangle - \lambda \langle \Delta^k, X^{k+1} - X^k \rangle.
\]

In (4.14), \( \langle \Delta^k, X^{k+1} - X^k \rangle \leq \|X^{k+1}\|_F - \|X^k\|_F \) since \( \Delta^k \in \partial\|X^k\|_F \), and \( \langle \Lambda^{k+1}, X^k \rangle \geq 0 \) by Lemma 4.1.1 (1). Combining (4.12) and (4.14) gives

\[
\varphi(X^k) - \varphi(X^{k+1}) \geq \frac{1}{2} \|A(X^k - X^{k+1})\|^2_2 + \lambda(\|X^{k+1}\|_F - \|X^k\|_F) + \langle \Lambda^{k+1}, X^k \rangle - \lambda \langle \Delta^k, X^{k+1} - X^k \rangle
\]

\[
\geq \frac{1}{2} \|A(X^k - X^{k+1})\|^2_2 + \lambda(\|X^{k+1}\|_F - \|X^k\|_F - \langle \Delta^k, X^{k+1} - X^k \rangle)
\]

\[
\geq 0.
\]
We are now in the position to prove convergence results of the DCA for solving the PhaseLiftOff problem (4.2).

**Proposition 4.3.1.** Let \( \{X^k\} \) be the sequence produced by the DCA starting with \( X^0 = 0 \).

1. \( \{X^k\} \) is bounded.
2. \( X^{k+1} - X^k \to 0 \) as \( k \to \infty \).
3. Any nonzero limit point \( \tilde{X} \) of the sequence \( \{X^k\} \) is a first-order stationary point, which means there exists \( \tilde{\Lambda} \), such that the following KKT conditions are satisfied:
   - **Stationarity:** \( A^*(A(\tilde{X}) - b) + \lambda (I_n - \frac{\tilde{X}}{\|\tilde{X}\|_F}) = \tilde{\Lambda} \).
   - **Primal feasibility:** \( \tilde{X} \succeq 0 \).
   - **Dual feasibility:** \( \tilde{\Lambda} \succeq 0 \).
   - **Complementary slackness:** \( \langle \tilde{X}, \tilde{\Lambda} \rangle = 0 \).

**Proof.** (1) By Lemma 4.3.1, the level set \( \Omega := \{X \in \mathbb{C}^{n \times n} : X \succeq 0, \varphi(X) \leq \varphi(0)\} \) is bounded. Since \( \{\varphi(X^k)\} \) is decreasing, \( \{X^k\} \subseteq \Omega \) is also bounded.

(2) Letting \( k = 0 \) and substituting \( \Delta^0 = 0 \) in (4.11), we obtain

\[
\varphi(0) - \varphi(X^1) \geq \frac{1}{2} \|A(X^1)\|_2^2 + \lambda \|X^1\|_F.
\]

If \( X^1 \neq 0 \), then \( \varphi(0) > \varphi(X^1) \geq \cdots \geq \varphi(X^k) \), so \( X^k \neq 0 \), \( \forall k \geq 1 \). Otherwise \( X^k \equiv 0 \).

Assuming \( X^k \neq 0 \), we show that \( X^{k+1} - X^k \to 0 \) as \( k \to \infty \) in what follows. Note that \( \{\varphi(X^k)\} \) is decreasing and convergent, and that \( \Delta^k = \frac{X^k}{\|X^k\|_F} \) when \( k \geq 1 \). Combining this with (4.11), we have the following key information about \( \{X^k\} \):

\[
\|A(X^k - X^{k+1})\|_2 \to 0 \quad (4.15)
\]

\[
\|X^{k+1}\|_F - \langle \frac{X^k}{\|X^k\|_F}, X^{k+1} \rangle \to 0. \quad (4.16)
\]
Define \( c^k := \frac{\langle X^k, X^{k+1} \rangle}{\|X^k\|_F} \geq 0 \) and \( E^k := X^{k+1} - c^k X^k \), then it suffices to prove \( E^k \to 0 \) and \( c^k \to 1 \). A simple computation shows

\[
\|E^k\|_F^2 = \|X^{k+1}\|_F^2 - \frac{\langle X^k, X^{k+1} \rangle^2}{\|X^k\|_F^2} \to 0,
\]

where (4.16) was used. Thus, from (4.15) it follows that

\[
0 = \lim_{k \to \infty} \|A(X^k - X^{k+1})\|_2 = \lim_{k \to \infty} \|A((c^k - 1)X^k - E^k)\|_2 = \lim_{k \to \infty} |c^k - 1| \|A(X^k)\|_2.
\]

Suppose \( \lim_{k \to \infty} c^k \neq 1 \), then there exists a subsequence \( \{X^{k_j}\} \) such that \( \|A(X^{k_j})\|_2 \to 0 \).

Since, by Lemma 4.1.1 (3), \( A(X) = 0 \iff X = 0 \) for \( X \succeq 0 \), we must have \( X^{k_j} \to 0 \) and \( \varphi(X^{k_j}) \to \varphi(0) \), which leads to a contradiction because

\[
\varphi(X^{k_j}) \leq \varphi(X^1) < \varphi(0).
\]

Therefore \( c^k \to 1 \) and \( X^{k+1} - X^k \to 0 \), as \( k \to \infty \).

(3) Let \( \{X^{k_j}\} \) be a subsequence of \( \{X^k\} \) converging to some limit point \( \tilde{X} \neq 0 \), then the optimality conditions at the \( k_j \)-th step read:

\[
A^*(A(X^{k_j}) - b) + \lambda(I_n - \frac{X^{k_j-1}}{\|X^{k_j-1}\|_F}) = \Lambda^{k_j},
\]

\[
X^{k_j} \succeq 0, \Lambda^{k_j} \succeq 0, \langle \Lambda^{k_j}, X^{k_j} \rangle = 0.
\]

Define

\[
\tilde{\Lambda} := \lim_{k_j \to \infty} \Lambda^{k_j}
\]

\[
= \lim_{k_j \to \infty} A^*(A(X^{k_j}) - b) + \lambda(I_n - \frac{X^{k_j-1}}{\|X^{k_j-1}\|_F})
\]

\[
= \lim_{k_j \to \infty} A^*(A(X^{k_j}) - b) + \lambda(I_n - \frac{X^{k_j}}{\|X^{k_j}\|_F}) + \lambda(\frac{X^{k_j}}{\|X^{k_j}\|_F} - \frac{X^{k_j-1}}{\|X^{k_j-1}\|_F})
\]

\[
= A^*(A(\tilde{X}) - b) + \lambda(I_n - \frac{\tilde{X}}{\|\tilde{X}\|_F}).
\]
In the last equality, we used \( \lim_{k_j \to \infty} X^{k_j} = \tilde{X} \neq 0 \) and \( X^{k_j} - X^{k_j-1} \to 0 \). Since \( X^{k_j} \succeq 0, \Lambda^{k_j} \succeq 0 \), their limits are \( \tilde{X} \succeq 0 \) and \( \tilde{\Lambda} \succeq 0 \). It remains to check that \( \langle \tilde{\Lambda}, \tilde{X} \rangle = 0 \). Using \( \langle \Lambda^{k_j}, X^{k_j} \rangle = 0 \), we have

\[
\langle \tilde{\Lambda}, \tilde{X} \rangle = \langle \tilde{\Lambda} - \Lambda^{k_j}, \tilde{X} - X^{k_j} \rangle + \langle \Lambda^{k_j}, \tilde{X} \rangle + \langle \tilde{\Lambda}, X^{k_j} \rangle.
\]

Let \( k_j \to \infty \) on the right hand side above, \( \langle \tilde{\Lambda}, \tilde{X} \rangle = 0 \).

**Remark 4.3.1.** In light of the proof of Proposition 4.3.1 (2), one can see that for all \( k \geq 1 \), either \( X^k \equiv 0 \) or \( \|X^k\|_F > \eta \) for some \( \eta > 0 \). A sufficient condition to ensure that the DCA does not yield \( \tilde{X} = 0 \) is as follows

\[
\frac{1}{2}\|b\|_2^2 > \frac{1}{2}\|c\|_2^2 + \lambda \text{Tr}(\tilde{X}) \iff \lambda < \frac{\|b\|_2^2 - \|c\|_2^2}{2 \text{Tr}(\tilde{X})},
\]

where \( \tilde{X} \) is the ground truth obeying \( b = A(\tilde{X}) + e \). The above condition would guarantee that \( X^1 \neq 0 \). Though the equivalence between (4.1) and (4.2) follows from Theorem 4.2.1 as long as \( \lambda \) is sufficiently large, in practice \( \lambda \) cannot get too large because the DCA iterations may stall at \( X^0 = 0 \).

### 4.3.2 Solving the subproblem.

At the \((k + 1)\)th DC iteration, one needs to solve a convex subproblem of the form:

\[
X^{k+1} = \arg \min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2} \|A(X) - b\|_2^2 + \langle X, W \rangle \quad \text{s.t.} \quad X \succeq 0.
\]

In our case, \( W = \lambda I_n \) or \( \lambda (I_n - \frac{X^k}{\|X^k\|_F}) \) is a known Hermitian matrix.

This problem can be treated as a weighted trace-norm regularization problem, which has been studied in [9]. The authors of [9] suggest using FISTA [4, 9] which is a variant of Nesterov’s accelerated gradient descent method [67]. An alternative choice is the alternating direction method of multipliers (ADMM). We only discuss the naive ADMM here, though this algorithm could be further accelerated by incorporating Nesterov’s idea [50]. To implement
ADMM, we introduce a dual variable $Y$ and form the augmented Lagrangian

$$L_\delta(X, Y, Z) = \frac{1}{2} \|A(X) - b\|^2_2 + \langle X, W \rangle + \langle Y, X - Z \rangle + \frac{\delta}{2} \|X - Z\|^2_F + g_\succcurlyeq(Z),$$

where

$$g_\succcurlyeq(Z) = \begin{cases} 0 & \text{if } Z \succeq 0, \\ \infty & \text{otherwise}. \end{cases}$$

ADMM consists of updates on both the primal and dual variables [7]:

$$\begin{align*}
X^{l+1} &= \arg \min_X L_\delta(X, Y^l, Z^l) \\
Z^{l+1} &= \arg \min_Z L_\delta(X^{l+1}, Y^l, Z) \\
Y^{l+1} &= Y^l + \frac{\delta}{\delta} (X^{l+1} - Z^{l+1})
\end{align*}$$

The first two steps have closed-form solutions, which are detailed in Algorithm 7. In the $X$-update step, one needs to know the expression of $(A^*A + \delta I_n)^{-1}$. The celebrated Woodbury formula implies

$$(A^*A + \delta I_n)^{-1} = \frac{1}{\delta} (I_n - A^* (A A^* + \delta I_m)^{-1} A).$$

By Lemma 4.2.2, $A A^* = A^* A \circ \overline{A} A$, so we have

$$A^*A + \delta I_n)^{-1}(X) = \frac{1}{\delta} (X - A^* ((A A^* + \delta I_m)^{-1} A(X))$$

$$= \frac{1}{\delta} (X - ADiag((A^* A \circ \overline{A} A + \delta I_m)^{-1} \text{diag}(A^* X A)) A^*)$$

In the $Z$-update step, $P_\succcurlyeq : \mathbb{H}^{n \times n} \to \mathbb{H}^{n \times n}$ represents the projection onto the positive semidef-
inite cone. More precisely, if \( X \) has the eigenvalue decomposition \( X = U \Sigma U^* \), then

\[
P_\geq(X) = U \max\{\Sigma, 0\} U^*.
\]

According to [7], the stopping criterion here is given by:

\[
\|R_l^t\|_F \leq n \epsilon_{\text{abs}} + \epsilon_{\text{rel}} \max\{\|X^t\|_F, \|Z^t\|_F\}, \quad \|S_l^t\|_F \leq n \epsilon_{\text{abs}} + \epsilon_{\text{rel}} \|Y^t\|_F,
\]

where \( R^t_l = X^t - Z^t \), \( S^t_l = \delta(Z^t - Z^{t-1}) \) are primal and dual residuals respectively at the \( l \)-th iteration. \( \epsilon_{\text{abs}}>0 \) is an absolute tolerance and \( \epsilon_{\text{rel}}>0 \) is a relative tolerance, and they are both algorithm parameters. \( \delta \) is typically fixed, but one can also adaptively update it during iterations following the rule in [7]; for instance,

\[
\delta^{l+1} = \begin{cases} 
2 \delta^l & \text{if } \|R^t_l\|_F > 10 \|S^t_l\|_F, \\
\delta^l/2 & \text{if } 10 \|R^t_l\|_F < \|S^t_l\|_F, \\
\delta^l & \text{otherwise.}
\end{cases}
\]

### 4.3.3 Real-valued, nonnegative signals.

If the signal is known to be real or nonnegative, we should add one more constraint to the complex PhaseLiftOff (4.2):

\[
\min_{X \in \mathbb{C}^{n \times n}} \varphi(X) \quad \text{s.t.} \quad X \succeq 0, \ X \in \Omega.
\]

Here \( \Omega = \mathbb{R}^{n \times n} \) (or resp., \( \mathbb{R}^+_{n \times n} \)), which means each entry of \( X \) is real (or resp., nonnegative). Thus we need to modify the DCA (4.10) accordingly:

\[
X^{k+1} = \begin{cases} 
\arg\min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2}\|A(X) - b\|^2 + \lambda \text{Tr}(X) \quad \text{s.t.} \quad X \succeq 0, \ X \in \Omega & \text{if } X^k = 0, \\
\arg\min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2}\|A(X) - b\|^2 + \lambda \langle X, I_n - \frac{X^k}{\|X^k\|_F} \rangle \quad \text{s.t.} \quad X \succeq 0, \ X \in \Omega & \text{otherwise.}
\end{cases}
\]
The above subproblem at each DCA iteration can also be solved by ADMM. Specifically, we want to solve the optimization problem of the following form:

\[
\min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2} \| A(X) - b \|_2^2 + \langle X, W \rangle \quad \text{s.t.} \quad X \succeq 0, \ X \in \Omega.
\]

In ADMM form, (4.21) is reformulated as

\[
\min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2} \| A(X) - b \|_2^2 + \langle X, W \rangle + g_\Omega(X) + g_{\geq}(Z) \quad \text{s.t.} \quad X - Z = 0,
\]

where \( g_{\geq}(Z) \) is the same as in (4.18), and

\[
g_\Omega(X) = \begin{cases} 
0 & \text{if } X \in \Omega, \\
\infty & \text{otherwise}.
\end{cases}
\]

Having defined the augmented Lagrangian

\[
L_\delta(X, Y, Z) = \frac{1}{2} \| A(X) - b \|_2^2 + \langle X, W \rangle + \langle Y, X - Z \rangle + \frac{\delta}{2} \| X - Z \|_F^2 + g_\Omega(X) + g_{\geq}(Z),
\]

we arrive at Algorithm 8 by alternately minimizing \( L_\delta \) with respect to \( X \), minimizing with respect to \( Z \), and updating the dual variable \( Y \). The operator \( \mathcal{P}_\Omega : \mathbb{H}^{n \times n} \to \Omega \) in Algorithm 8 represents the projection onto the set \( \Omega \). In particular, \( \mathcal{P}_\Omega(X) = \text{Re}(X) \) is the real part of \( X \) for \( \Omega = \mathbb{R}^{n \times n} \), whereas \( \mathcal{P}_\Omega(X) = \max\{\text{Re}(X), 0\} \) for \( \Omega = \mathbb{R}_+^{n \times n} \). Algorithm 8 is almost identical to Algorithm 7 except that an extra projection \( \mathcal{P}_\Omega \) is performed in the \( X \)-update step.
4.4 Numerical Experiments.

In this section, we report numerical results. Besides the proposed (4.2) and the regularized PhaseLift (1.4), we also discuss the following reweighting scheme from [9], which is an extension of reweighted $\ell_1$ algorithm in the regime of compressed sensing introduced in [17]:

\[
X^{k+1} = \arg \min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2} \|A(X) - b\|^2_2 + \lambda \langle W^k, X \rangle \quad \text{s.t.} \quad X \succeq 0,
\]

where $W^0 = I_n$ and $W^k = (X^k + \varepsilon I_n)^{-1}$ for $k \geq 1$ and for some $\varepsilon > 0$. The aim of this scheme is to provide more accurate solutions with lower rank than that of PhaseLift. Note that $W^k$ is exactly the gradient of $\log(\det(X + \varepsilon I_n))$ at $X^k$, the reweighting scheme is in essence an implementation of the DCA attempting to solve the nonconvex problem

\[
\min_{X \in \mathbb{C}^{n \times n}} \frac{1}{2} \|A(X) - b\|^2_2 + \lambda \log(\det(X + \varepsilon I_n)) \quad \text{s.t.} \quad X \succeq 0.
\]

Here the DC components are $\frac{1}{2} \|A(X) - b\|^2_2$ and $-\lambda \log(\det(X + \varepsilon I_n))$. We hereby remark that with positive semidefinite constraint, the DCA is basically equivalent to the reweighting scheme. In [9], (1.4) and the subproblem (4.22) of (4.23) are solved by FISTA. Here we solve them using ADMM (Algorithm 7) instead as we find it more efficient.

4.4.1 Exact recovery from noise-free measurements.

We set up a phase retrieval problem by 1) generating a random complex-valued signal $\hat{x}$ of length $n = 32$ whose real and imaginary parts are Gaussian, 2) sampling a Gaussian matrix $A \in \mathbb{C}^{n \times m}$ with $m = 60, 63, \ldots, 150$, and 3) computing the measurements $b = A(\hat{x}\hat{x}^*)$. We then solve (4.2), (1.4) and (4.23) to get approximations to $\hat{x}\hat{x}^*$. The ultimate goal of phase retrieval is to reconstruct the signal $\hat{x}$ rather than the rank-1 matrix $\hat{x}\hat{x}^*$. So given a solution $\tilde{X}$, we need to compute the relative mean squared error (rel. MSE) between $\tilde{x} = \sqrt{\sigma_1(\tilde{X})}u_1$ and $\hat{x}$ modulo a global phase term to measure the recovery quality, where $\sigma_1(\tilde{X})$ is the largest singular value (or eigenvalue) of $\tilde{X}$ and $u_1$ the corresponding unit-normed eigenvector. More
precisely, the rel. MSE is given by
\[
\min_{c \in \mathbb{C} : |c| = 1} \frac{||c \tilde{x} - \hat{x}||_2^2}{||\hat{x}||_2^2}.
\]
It is easy to show that its minimum occurs at
\[
\tilde{c} = \frac{\langle \tilde{x}, \hat{x} \rangle}{|\langle \tilde{x}, \hat{x} \rangle|}.
\]
A recovery is considered as a success if the rel. MSE is less than 10^{-6} (or equivalently, relative error < 10^{-3}). For each \( m = 60, 63, \ldots, 150 \), we repeat the above procedures 100 times and record the success rate for each model.

For (1.4), we set \( \lambda = 10^{-4}, \epsilon_{\text{rel}} = 10^{-5} \) and \( \epsilon_{\text{abs}} = 10^{-7} \) in its ADMM algorithm; for (4.2), \( \lambda = 10^{-4}, \epsilon_{\text{rel}} = 10^{-5}, \epsilon_{\text{abs}} = 10^{-7} \) and \( \text{tol} = 10^{-2} \); parameters for (4.23) are the same as those for (4.2) except that there is an additional parameter \( \varepsilon = 2 \). In addition, the maximum iteration set for all ADMM algorithms is 5000 and that for the DCA and the reweighting algorithm are both 10. All three methods start with the same initial point \( X^0 = 0 \).

The success rate v.s. number of measurements plot is shown in Figure 4.1. The result validates that nonconvex proxy for the rank functional gives significantly better recovery quality than the convex trace norm. A similar finding has been reported in the regime of compressed sensing [83]. We also observe that PhaseLiftOff outperforms log-det regularization. This is not surprising as the former always captures rank-1 solutions. In Figure 4.1, one can see that when the number of measurements is \( m \approx 3n = 96 \), solving our model by the DCA guarantees exact recovery with high probability. Recall that in theory [2] at least \( 3n - 2 \) measurements are needed to recover the signal exactly. This is an indication that the proposed method is likely to provide the optimal practical results one can hope for.

4.4.2 Robust recovery from noisy measurements.

We investigate how the proposed method performs in the presence of noise. The test signal \( \hat{x} \) is a Gaussian complex-valued signal of length \( n = 32 \). We sample \( m = 4n \) Gaussian measurement vectors in \( \mathbb{C}^n \) and compute the measurements \( b \in \mathbb{R}^m \), followed by adding
additive white Gaussian noise by means of the MATLAB function \texttt{awgn(b,snr)}. There are 6 noise levels varying from 5dB to 55dB. We then apply the DCA to achieve a reconstruction $\hat{X}$ and compute the signal-to-noise ratio (SNR) of reconstruction in dB defined as $-10 \log_{10}(\text{rel. MSE})$. The SNR of reconstruction for each noise level is finally averaged over 10 independent runs.

A crucial point to address here is how we set the value of $\lambda$. Theorem 4.2.1 predicts that provided the noise amount $\|e\|_2$ is known, when $\lambda > \frac{\|A\|\|e\|_2}{\sqrt{2}-1} \approx 2.414\|A\|\|e\|_2$, the PhaseLiftOff (4.2) is equivalent to the phase retrieval problem (4.1), and its solution is no longer related to $\lambda$. From computational perspective, however, $\lambda$ cannot be too large as the algorithm may often get stuck at a local solution. An extreme example is that if $\lambda$ is exceedingly large, the DCA will be trapped at the initial guess $X^0 = 0$. On the other hand, if $\lambda$ is too small, the reconstruction will be of course far from the ground truth as $A(X) = b$ tends to be enforced. But can we choose $\lambda$ that is less than $2.414\|A\|\|e\|_2$? The answer is yes, since this bound only provides a sufficient condition for equivalence.

Suppose the noise amount $\|e\|_2$ (or its estimate) is known, defining

$$\mu := \|A\|\|e\|_2 = \sqrt{\|A^*A \circ \widehat{A}^*\|_2\|e\|_2},$$

we try 4 different values of $\lambda$ in each single run. They are multiples of $\mu$, namely $0.01\mu$, $0.2\mu$, $2.5\mu$ and $50\mu$. The maximum outer and inner iterations are 10 and 5000 respectively.
The other parameters are $\epsilon_{\text{rel}} = 10^{-5}$, $\epsilon_{\text{abs}} = 10^{-7}$, $\text{tol} = 10^{-2}$. The reconstruction results are depicted in Figure 4.2. The two curves for $0.2\mu$ and $2.5\mu$ nearly coincide, and they are almost linear, which strongly suggest stable recoveries. In contrast, the algorithm with $0.01\mu$ and $50\mu$ performed poorly. Although $\lambda = 50\mu$ yields comparable reconstruction when there is little noise, the DCA clearly encounters local minima in the low SNR regime. On the other hand, $0.01\mu$ is too small. Summarizing these observations, we conclude that for the DCA method, a reasonable value for $\lambda$ lies in the interval (but not limited to) $[0.2\mu, 2.5\mu]$.

4.4.3 Phase retrieval in array imaging

Although the exact recovery theory is on Gaussian measurements, the algorithm works for non-Gaussian case, as shown in this array imaging example. Array imaging aims to determine
the locations and reflectivities of small distributed reflectors \( r \) by sending probing signals from a sensor array and recording the backscattered fields (left plot of Fig. 4.3). Phase or arrival time information in the time domain is essential, yet difficult to measure and record from the signals received at the array. This is known in single molecule detection with optical sensors and three-dimensional imaging of nanostructures. In [21], single frequency array imaging of localized scatterers is treated when only the intensity is recorded at the array. The ambient medium is homogeneous so that waves propagate coherently. The sensing (array response) matrix \( A \) depends on the Green’s function of the wave operator as well as the illumination vector (known). The goal is to recover \( r \) with a small number of illuminations. The Green’s function is approximated by neglecting multiple scattering when point objects are separated.

Suppose \( \{x_i\}_{i=1}^m \), \( \{y_i\}_{i=1}^n \) are locations of the sensors and the mesh points respectively. The response data \( b \) has the approximation

\[
\Pi_r(\omega)f(\omega) = \left( \sum_{i=1}^n r_i g(y_i, \omega) g^T(y_i, \omega) \right) f(\omega),
\]

where \( \Pi_r(\omega) \) approximates the array impulse response matrix, \( f(\omega) \) is the illumination vector,

\[
g(y, \omega) = \left( \frac{e^{i\kappa|x_1-y|}}{4\pi|x_1-y|}, \ldots, \frac{e^{i\kappa|x_m-y|}}{4\pi|x_m-y|} \right)^T, \quad \kappa = \frac{\omega}{c}
\]

is the Green’s function vector.

The authors of [21] studied phase retrieval by rank-1 lifting and trace norm relaxation (that is, PhaseLift). PhaseLiftOff can significantly improve the quality of recovery. We test 4 planar scatterers of reflectivities 1.0, 0.8, 0.7 and 0.5, within an image window of size \( 10 \times 10 \), using 21 transducers (5 wavelengths apart) with a distance 100 (all in units of wavelength) to the image window. The 5% independent Gaussian noise is added. In the middle plot of Fig. 4.3, the 4 recovered objects are blurry with 10 illuminations by PhaseLift. In the right plot, they are much more visible with only 3 illuminations by PhaseLiftOff.
Chapter 5

Conclusions and Future Work

We developed an iterative $\ell_1$ framework for a broad class of Lipschitz continuous non-convex sparsity promoting objectives, including those arising in statistics. The iterative $\ell_1$ algorithm is shown via theory and computation to improve on the $\ell_1$ minimization for CS problems independent of the conditioning of the sensing matrices. We introduced and analyzed a novel penalty (trace minus Frobenius norm) for phase retrieval in the PhaseLiftOff least squares regularization problem. We proved its equivalence with the original phase retrieval problem and stable recovery for noisy measurement at high probability. Numerical experiments showed that the PhaseLiftOff method outperforms PhaseLift for both Gaussian and physical sensing (non-Gaussian) measurements.

For future work, we shall apply the methods presented in the thesis to sparsity and rank constrained problems in imaging science, machine learning and portfolio optimization.
BIBLIOGRAPHY


79


[92] S. Zhou, N. Xiu, Y. Wang, L. Kong, and H. Qi A null-space-based weighted \( \ell_1 \) minimization approach to compressed sensing, Information and Inference, 2016: iaw002v1-iaw002.