Non-Convex Methods for Compressed Sensing and Low-Rank Matrix Problems

Gerosa, Daniele

2022

Document Version:
Publisher's PDF, also known as Version of record

– CENTRUM SCIENTIARUM MATHEMATICARUM –

Non-Convex Methods for Compressed Sensing and Low-Rank Matrix Problems

DANIELE GEROSA

Lund University
Faculty of Sciences
Centre for Mathematical Sciences
Mathematics

Non-Convex Methods for Compressed Sensing and Low-Rank Matrix Problems

Non-Convex Methods for Compressed Sensing and Low-Rank Matrix Problems

Daniele Gerosa

LUND
UNIVERSITY

DOCTORAL THESIS
Thesis advisor: Senior Lecturer Marcus Carlsson
Faculty opponent: Associate Professor Martin Skovgaard Andersen

To be publicly defended, by due permission of the Faculty of Science of Lund University, on Tuesday, the 26th of April 2022 at 15:00, in the Hörmander lecture hall, Sölvegatan 18A, Lund, for the Degree of Doctor of Philosophy in Mathematics.

Organization	Document name
LUND UNIVERSITY Centre for Mathematical Sciences Box 118 SE–221 00 LUND Sweden	**DOCTORAL DISSERTATION**
	Date of disputation 2022-04-26
	Sponsoring organization eSSENCE: the e-Science Collaboration (Project 4:3), Crafoord Foundation (grant n. 20190847)
Author(s) Daniele Gerosa	

Title and subtitle

Non-Convex Methods for Compressed Sensing and Low-Rank Matrix Problems

Abstract

In this thesis we study functionals of the type $\mathcal{K}_{f,A,\mathbf{b}}(\mathbf{x}) = \mathcal{Q}(f)(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2$, where A is a linear map, $\mathbf{b}$ a measurements vector and $\mathcal{Q}$ is a functional transform called *quadratic envelope*; this object is a very close relative of the *Lasry-Lions envelope* and its use is meant to regularize the functionals f. Carlsson and Olsson investigated in earlier works the connections between the functionals $\mathcal{K}_{f,A,\mathbf{b}}$ and their unregularized counterparts $f(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2$. For certain choices of f the penalty $\mathcal{Q}(f)(\cdot)$ acts as sparsifying agent and the minimization of $\mathcal{K}_{f,A,\mathbf{b}}(\mathbf{x})$ delivers sparse solutions to the linear system of equations $A\mathbf{x} = \mathbf{b}$. We prove existence and uniqueness results of the sparse (or low rank, since the functional f can have any Hilbert space as domain) global minimizer of $\mathcal{K}_{f,A,\mathbf{b}}(\mathbf{x})$ for some instances of f, under Restricted Isometry Property conditions on A. The theory is complemented with robustness results and a wide range of numerical experiments, both synthetic and from real world problems.

Key words

compressed sensing, low-rank matrix, phase retrieval, non-convex optimization

Classification system and/or index terms (if any)

Supplementary bibliographical information	Language English
ISSN and key title 1404-0034	ISBN 978-91-8039-088-0 (print) 978-91-8039-087-3 (pdf)
Recipient's notes	Number of pages 227 / Price
	Security classification

Non-Convex Methods for Compressed Sensing and Low-Rank Matrix Problems

Daniele Gerosa

LUND
UNIVERSITY

Funding information: The thesis work was supported by eSSENCE: the e-Science Collaboration (Project 4:3) as well as the Crafoord Foundation (grant n. 20190847)

Mathematics
Centre for Mathematical Sciences
Box 118
SE-22100 LUND
Sweden

Doctoral Theses in Mathematical Sciences 2022:3
ISSN: 1404-0034

ISBN: 978-91-8039-088-0 (print)
ISBN: 978-91-8039-087-3 (pdf)
LUNFMA-1044-2022

Printed in Sweden by Media-Tryck, Lund University, Lund 2022

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

En el Asia Menor o en Alejandría, en el segundo siglo de nuestra fe, cuando Basílides publicaba que el cosmos era una temeraria o malvada improvisación de ángeles deficientes, Nils Runeberg hubiera dirigido, con singular pasión intelectual, uno de los conventículos gnósticos. Dante le hubiera destinado, tal vez, un sepulcro de fuego; su nombre aumentaría los catálogos de heresiarcas menores, entre Satornilo y Carpócrates; algún fragmento de sus prédicas, exornado de injurias, perduraría en el apócrifo *Liber adversus omnes haereses* o habría perecido cuando el incendio de una biblioteca monástica devoró el último ejemplar del *Syntagma*. En cambio, Dios le deparó el siglo XX y la ciudad universitaria de Lund.

Tres versiones de Judas, Jorge Luis Borges

Acknowledgements

First and foremost I would like to thank my main supervisor Marcus Carlsson. Besides being a brilliant mathematician with a very strong geometric intuition and countless ideas, he is an *happy* human being (and I suspect and possibly theorize, as the Italian rock band Afterhours sang, that *the key of (his) happiness is disobedience*); his human and mathematical support throughout the whole PhD process has been, *ça va sans dire*, pivotal. None of the results present in this thesis could have been achieved without him. I would like then to thank Carl Olsson for his stellar guidance and never-ending patience, not to mention his powerful ideas and punctual encouragements; thank to him I could fully appreciate the theoretical and practical implications of our work. I would also like to thank Rajmund Mokso and Gerardina Carbone who helped me a lot, in particular at the beginning of my Swedish adventure.

I have at least one memory associated to each single member of the (extended) PhD gang, directly from my very first encounters with them. Bartosz and Douglas laughing at me when I told them that I wanted to learn teaching tricks from Marcus, Tien paying for one of my first lunches at Finn Inn, Dag swinging by my office to invite me to V. E. S. P. A, Adem talking about the Belgian Congo, Jens gently correcting a statement of mine about *my* university, Jonathan ranting about micromobility and shared electric scooters in Malmö. When I think about those early days I become quite nostalgic; I think we developed a sincere and serene friendship (that continued, continues and will hopefully continue outside the walls of the Matematikcentrum), and considering how usually tough is for an expat to integrate in the social tissue of a new country, I feel blessed to have met such a nice and friendly group of people.

I owe acknowledgements to other inhabitants of the Matematikcentrum, too: Alexandru, Anna-Maria, Dragi, Erik Wahlén, Eskil Rydhe, Kjell, Mats and Yacin. I felt particularly welcomed by all of them and I immensely valued the time that they chose to spend with me.

Thanks to my friends Nicola and Lorenzo. Despite the fact that I have not met them in 5 years, I somehow always felt their presence as they were just here nearby.

Grazie a mamma, papà e Ilaria per avermi, da lontano, supportato sempre.

Borges dice che nella biblioteca di Babele non ci sia un solo nonsenso che sia *assoluto*. Il *mio* segnale, criptico, che a volte somiglia ad un eco, dice così: "ti ha aspettato". Lo ha ripetuto continuamente, periodico, mentre tu, coi fatti, dimostravi quanto vera fosse questa verità. La pazienza è la più coraggiosa delle tue qualità, Dorotea; perché presuppone fiducia. Perché presuppone *visione*. Sono qui perché tu sei qui. Ti amo.

Popular science summary

R. Priemer, in his book *Introductory Signal Processing*, says that a *signal* is a function that conveys information about a phenomenon. A physical phenomenon is usually modelled using the mathematical language in one of its instances (i. e. branches), but oftentimes a signal is not directly accessible to human beings, i. e. it cannot be directly measured; this might be seen as a limitation of the model used, or as an intrinsic feature of the phenomenon. The simplest models are the linear ones where the measurements are assumed to be linear combinations of the underlying (discretized) ground truth. With an example: suppose that the signal $\mathbf{x}_0 = \begin{pmatrix} 1 & 0 & 0 & -2 & 0 & 0 \end{pmatrix}^t$ is "emitted" by some phenomenon, but not directly accessible for measurements; accessible are however the measurements

$$\mathbf{b} = \begin{pmatrix} -13 \\ 23 \\ -28 \end{pmatrix} = \underbrace{\begin{pmatrix} 7 & 9 & -5 & 10 & 10 & -8 \\ 9 & 3 & 1 & -7 & 0 & -2 \\ -8 & -8 & 10 & 10 & 6 & 9 \end{pmatrix}}_{=A} \begin{pmatrix} 1 \\ 0 \\ 0 \\ -2 \\ 0 \\ 0 \end{pmatrix}$$

with A a known matrix sometimes called *sensing matrix*. Thus, having $\mathbf{b}$ and A, we would like to retrieve $\mathbf{x}_0$. This is of course not possible in our example: the matrix A has a 3-dimensional kernel and thus $\mathbf{x}_0$ is not the only solution to the linear system of equations $A\mathbf{x} = \mathbf{b}$. Things change and become possible when we look for an $\mathbf{x}_0$ with *a lot of zeros* (which is actually the case in our toy example) and when A is assumed to have some structural property called *Restricted Isometry Property*. This is essentially what this thesis is about: looking for solutions with a lot of zeros to under-determined linear systems of equations.

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

Paper I **An unbiased approach to compressed sensing**

Marcus Carlsson, **Daniele Gerosa** and Carl Olsson
Inverse Problems, 36(11) 115014, 2020

Paper II **An unbiased approach to low rank recovery**

Marcus Carlsson, **Daniele Gerosa** and Carl Olsson
Preprint (submitted), 2019

Paper III **Bias versus non-convexity in compressed sensing**

Daniele Gerosa, Marcus Carlsson and Carl Olsson
Journal of Mathematical Imaging and Vision (online), 1-16, 2022

Paper IV **Relaxations for non-separable cardinality/rank penalties**

Carl Olsson, **Daniele Gerosa** and Marcus Carlsson
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 162-171, 2021

Paper V **On phase retrieval via matrix completion and the estimation of low rank PSD matrices**
Marcus Carlsson and **Daniele Gerosa**
Inverse Problems, 36(1) 015006, 2020

Author contributions

Remarks: the order of the authors was chosen to be alphabetical in Papers I, II and V, while it reflects the amount of work done in Papers III and IV. Chronological order:

Paper I → Paper III → Paper II → Paper V → Paper III (reworked) → Paper IV.

Paper I. I did all the work pertaining **Sections 6** and **2.4** and I proved some intermediate results (**8.1**, **4.7**, **5.4**). I contributed to the development and writing of the other sections.

Paper II. I did most of the work in **Sections 2**, **4** (except **4.5**), **5**, **6**, 7 and **8.2**. The remaining sections were developed and written as joint effort.

Paper III. I did all the work pertaining **Sections 2**, **5** and all the numerical experiments in **Sections 7.1, 7.2** and **7.3**. The remaining sections were developed and written as joint effort.

Paper IV. I proved **Theorem 3.1** (and its generalization, that was omitted for space constraints; it can be found in the **Miscellaneous** section) and I did all the work pertaining **Section 4.1**. I contributed to **Sections 3.2** and **3.3** with discussions, observations and insights.

Paper V. I designed the paper and I did most of the work. **Sections 3** and **4.1** were a joint effort with M. Carlsson.

Contents

Preface

Thesis structure

This thesis contains 5 scientific papers. Papers **I**, **II**, **III** and **IV** pertain to a subfield of signal processing known as *compressed sensing*, *compressive sensing* or *compressive sampling*. Throughout this entire work we will use the first terminology. Paper **V** deals with an (instance of an) inverse problem called the *Fourier phase retrieval problem* arising (for example) in X-ray cristallography and coherent diffractive imaging; in this paper the phase retrieval problem is considered from a "low-rank perspective" and thus, ultimately, from a compressed sensing perspective.

A Miscellaneous section will come after the five papers; we will present there few unpublished observations. The structure of this preface is the following: there will be two main sections, one dedicated to compressed sensing and one to the phase retrieval problem. In each section classical theory and glimpses of state of the art will be outlined; our results, together with their relation with the state of the art, will then be discussed.

The papers, in their latest versions, will come after this preface.

Compressed Sensing

The Nyquist-Shannon theorem states that a 1-dimensional absolutely integrable signal f with Fourier transform $\hat{f}$ supported in some interval $[-M, M] \subseteq \mathbb{R}$ can be completely reconstructed via equispaced samples of f of length $1/2M$. This well-known signal processing result somehow suggests the idea that a *sufficiently large* number of measurements is needed in order to retrieve an unknown signal. In some sense this was already pretty clear in basic Linear Algebra: in the simplest possible scenario, when measurements $\mathbf{b} \in \mathbb{C}^m$ are obtained as linear combinations of an underlying signal (to be reconstructed) $\mathbf{x}_0 = (x_1, \dots, x_n) \in \mathbb{C}^n$ so that $\mathbf{b} = A\mathbf{x}_0$ for a sensing matrix $A \in \mathbb{M}_{m\times n}(\mathbb{C})$, there is no chance of recovering $\mathbf{x}_0$ if A has a non-trivial kernel (so, the measurements $\mathbf{b}$ are not enough). However things dramatically changes if more hypothesis are added on A, m and on the number of non-zero entries of $\mathbf{x}_0$ (this property is called *sparsity*). Compressed sensing bases its power on some sort of paradigmatic assumption, that signals can often be represented using a basis that make them *sparse*, i. e. with a lot of zeros; compressed sensing can thus be seen as the mathematical theory that clarifies under which circumstances the *underdetermined* linear system of equations

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

has a *unique* solution, given a sparsity constraint on $\mathbf{x}$.

From a merely mathematical standpoint this does not really come as a surprise: paraphrasing Ennio De Giorgi, the more hypothesis are added the stronger the theoretical conclusions will be; nonetheless from a more engineering point of view the idea of an undetermined linear system with a unique solution is thrilling, simply because the paradigm

less measurements = more resources saved

holds in the real world.

Classical vector theory. The possibly most intuitive approach to find a sparse solution to (1) would be to solve the problem

$$\min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{x}\|_{\ell^0} \quad \text{subject to } A\mathbf{x} = \mathbf{b} \tag{2}$$

where $\|\mathbf{x}\|_{\ell^0} := \operatorname{card}(\mathbf{x}) = \operatorname{card}(\{x_j \,:\, x_j \neq 0\})$; however the map $\mathbf{x} \mapsto \|\mathbf{x}\|_{\ell^0}$ is discontinuous and thus any optimization-based approach is deemed to fail. One could also attack (2) using combinatorial techniques, but Natarajan showed in [48] that the problem is NP-hard. For $p \in (0, 1)$ the non-convex ℓ^p quasi-norms have been proposed as alternatives to the cardinality with various degree of success, motivating the approach with the simple observation that $\|\mathbf{x}\|_{\ell^p}^p \to \|\mathbf{x}\|_{\ell^0}$ as $p \to 0^+$ for all $\mathbf{x}$.

The notorious turning point was the ℓ^1 norm: its convexity and almost everywhere differentiability made it an appealing candidate and ultimately it became widely used in the whole signal processing community. From an intuitive/geometric point of view it is rather clear why the solution to

$$\min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{x}\|_{\ell^1} \quad \text{subject to } A\mathbf{x} = \mathbf{b} \tag{3}$$

should be, roughly speaking, sparse: Figure 1 illustrates pretty well this property of the ℓ^1 diamond-shaped balls. The ℓ^1 ball of minimal radius (bold blue) intersects the linear subspace $A\mathbf{x} - \mathbf{b} = \mathbf{0}$ (bold red) at one of its vertices (black dot), which is sparse. The dashed blue diamonds are ℓ^1 balls of bigger radii. This simple observation led mathematicians to develop a rather strong and juicy theory revolving around the ℓ^1 norm; we will try here to outline it. We begin by recalling the (definition of) *null-space property*:

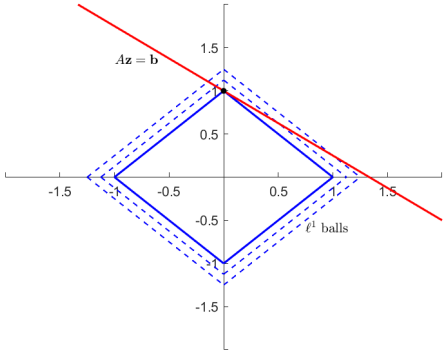

Figure 1: 2-D geometric idea explaining why a solution to (3) is likely sparse.

Definition. A matrix $A \in \mathbb{M}_{m\times n}(\mathbb{C})$ is said to satisfy the *null-space property* relative to a set $S \subset \{1, \dots, n\}$ if

$$\|\mathbf{x}_S\|_{\ell^1} < \|\mathbf{x}_{S^c}\|_{\ell^1} \quad \forall\, \mathbf{x} \in \ker(A) \setminus \{\mathbf{0}\}. \tag{4}$$

(Here the vector $\mathbf{x}_S$ is obtained from $\mathbf{x}$ by means of setting $x_j = 0$ if $j \in S^c$, being S^c the set-theoretic complement of S.) Consequently we say that A satisfies the *null-space property of order s* if the condition (4) holds for *all* sets S with $\text{card}(S) \le s$. We also say that A satisfies the *robust* null-space property if the inequality in (4) holds in the form $\|\mathbf{x}_S\|_{\ell^1} \le \rho\|\mathbf{x}_{S^c}\|_{\ell^1} + \tau\|A\mathbf{x}\|$ $\forall\, \mathbf{x} \in \mathbb{C}^n$, for a $\rho \in (0, 1)$ and $\tau \ge 0$.

The null-space property is fundamental to establish uniqueness of the solution to (3) and to its non-convex relative with the ℓ^p quasi-norms instead. The following results follow closely the formulation of [28], Chapter 4:

Theorem. *For a matrix $A \in \mathbb{M}_{m\times n}(\mathbb{C})$, every s-sparse vector $\mathbf{x}_0 \in \mathbb{C}^n$ is the* unique *solution to* (3) *with $\mathbf{b} = A\mathbf{x}_0$ if and only if A satisfies the null-space property of order s.*

The above theorem is quite interesting because, under the null-space property, it builds a bridge between (3) and (2): in particular it is easy to see that if $\mathbf{x}$ is s-sparse and it solves (3), then it solves (2) too (assume that $\mathbf{z}$ solves (2); then $\|\mathbf{z}\|_{\ell^0} \le \|\mathbf{x}\|_{\ell^0}$, thus also $\mathbf{z}$ is s-sparse and $\mathbf{b} = A\mathbf{z}$. But the solution to (3) is unique, thus $\mathbf{x} = \mathbf{z}$). There are ℓ^p versions of

the null-space property and of the theorem above, see Theorem 4.9 in [28]; for redundancy reasons we omit them here.

The robust null-space property is the key condition on A to solve a relaxed version of (3), where noisy measurements are allowed:

$$\min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{x}\|_{\ell^1} \quad \text{subject to } \|A\mathbf{x}-\mathbf{b}\|_{\ell^2} \leq \epsilon \tag{5}$$

for some positive tolerance ϵ and $\mathbf{b} = A\mathbf{x}_0 + \mathbf{e}$. Problem (5) is often called *quadratically constrained basis pursuit*. The following result holds (Theorem 4.19 in [28]):

Theorem. *Suppose that a matrix $A \in \mathbb{M}_{m\times n}(\mathbb{C})$ satisfies the robust null-space property of order s with constants $\rho \in (0,1)$ and $\tau > 0$. Then, for any $\mathbf{x}_0 \in \mathbb{C}^n$, a solution $\tilde{\mathbf{x}}$ to (5) with $\mathbf{b} = A\mathbf{x}_0 + \mathbf{e}$ and $\|\mathbf{e}\|_{\ell^2} \leq \epsilon$ satisfies the following inequality:*

$$\|\mathbf{x}_0 - \tilde{\mathbf{x}}\|_{\ell^1} \leq \frac{2(1+\rho)}{1-\rho}\sigma_s(\mathbf{x}_0)_1 + \frac{4\tau}{1-\rho}\epsilon, \tag{6}$$

where $\sigma_s(\mathbf{x}_0)_1 := \inf_{\|\mathbf{x}\|_{\ell^0}\leq s} \|\mathbf{x}_0 - \mathbf{x}\|_{\ell^1}$.

Thus, while the solution to (5) is not necessarily $\mathbf{x}_0$, the ℓ^1-mismatch between the two can be controlled from above.

The issue with the null-space property is that it seems somehow artificial and rather elusive, often difficult to check. The paper by Candès and Tao [17], where the notion of *Restricted Isometry Property* (RIP) was introduced for the first time, tried to overcome to this issue:

Definition. The *restricted isometry constant* δ_k of order k of a matrix $A \in \mathbb{M}_{m\times n}(\mathbb{C})$ is the smallest number $\delta \in (0,1)$ such that

$$(1-\delta)\|\mathbf{x}\|_{\ell^2}^2 \leq \|A\mathbf{x}\|_{\ell^2}^2 \leq (1+\delta)\|\mathbf{x}\|_{\ell^2}^2$$

for all vectors $\mathbf{x} \in \mathbb{C}^n$ with $\text{card}(\mathbf{x}) \leq k$.

The subtle interconnections between the restricted isometry property and the null-space property have been investigated for instance in [12], but in some sense RIP is stronger than the null-space property, as the following lemma proved by Candès in [14] states:

Lemma. *Suppose that a matrix $A \in \mathbb{M}_{m\times n}(\mathbb{R})$ satisfies the RIP of order $2s$ with constant δ_{2s}. Then it satisfies the robust null-space property with constants $\tau = 0$ and $\rho = \sqrt{2}\delta_{2s}/(1-\delta_{2s})$.*

The cornerstone of this compressed sensing theory is most likely the theorem proved by Candès, Romberg and Tao in [16] (Theorem 1), where they show that, if the sensing matrix obeys some RIP condition, then (5) has solution "very close to" $\mathbf{x}_0$, and the ℓ^2 error can be bounded by a quantity *linear* in the noise magnitude:

Theorem. *Let A and $S \subset \{1, \dots, n\}$ with $|S| \leq s$ be such that $\delta_{3s} + 3\delta_{4s} < 2$. Then for any signal $\mathbf{x}_0 \in \mathbb{R}^n$ with supp$(\mathbf{x}_0) \subset S$ and any noise $\mathbf{e}$ with $\|\mathbf{e}\|_{\ell^2} \leq \epsilon$ the solution $\tilde{\mathbf{x}}$ to (5) obeys*

$$\|\mathbf{x}_0 - \tilde{\mathbf{x}}\|_{\ell^2} \leq \frac{4/\sqrt{3}}{\sqrt{1-\delta_{4s}} - \sqrt{1/3}\sqrt{1+\delta_{3s}}}\epsilon. \tag{7}$$

The above estimate and condition on the RIP constant were subsequently improved in [14] where $\delta_{2s} < \sqrt{2} - 1$ was enough to obtain $\|\mathbf{x}_0 - \tilde{\mathbf{x}}\|_{\ell^2} \leq \frac{4\sqrt{1+\delta_{2s}}}{1-(1+\sqrt{2})\delta_{2s}}\epsilon$. In [13] similar and possibly improved results were proved, which however are more intricate since they depend on another structural constant of the matrix A, i. e. the so called (s, s')-*restricted orthogonality constant* $\theta_{s,s'}$; they proved that if $\mathbf{x}_0$ is s-sparse and $\delta_s + \sqrt{s}\theta_{s,1} < 1$ then

$$\|\mathbf{x}_0 - \tilde{\mathbf{x}}\|_{\ell^2} \leq \frac{2\sqrt{1+\delta_s}\sqrt{1+s}}{1-\delta_s-\sqrt{s}\theta_{s,1}}\epsilon. \tag{8}$$

Before concluding this subsection we want to recall that the quadratically constrained basis pursuit (5) is strongly connected to the *basis pursuit denoising* problem, that means solving

$$\min_{\mathbf{x}\in\mathbb{C}^n} \lambda\|\mathbf{x}\|_{\ell^1} + \|A\mathbf{x} - \mathbf{b}\|_{\ell^2}^2 \tag{9}$$

for some $\lambda > 0$; in particular if, for a fixed $\lambda > 0$, $\tilde{\mathbf{x}}$ is a minimizer of (9), then there exists a $\epsilon = \epsilon_{\tilde{\mathbf{x}}}$ such that $\tilde{\mathbf{x}}$ solves (5) with $\epsilon_{\tilde{\mathbf{x}}}$. This fact, together with the rather strong theoretical properties listed above and the broad variety of algorithmic tools developed to compute the solution[1], made the ℓ^1 method a popular approach to tackle compressed sensing or low rank problems (see next section).

Shrinking bias and non-convex alternatives. As explained in the previous section, the ℓ^1 norm method enjoys appealing theoretical properties as well as a wide range of powerful optimization routines developed to solve the minimization problem associated to it. However this method comes with a *shrinking bias*, since small values are set to zero but bigger ones are shrunk by a factor depending on λ, generating a certain degree of distortion in the reconstruction. This phenomenon was described in [24] via numerical evidences and a simple inspection of the 1-D case: if we indeed consider the simple minimization problem $\min_{y\in\mathbb{R}}(y-x)^2/2 + \lambda|y|$

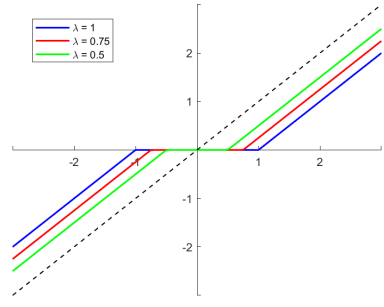

Figure 2: Soft-thresholding for different values of λ.

[1](9) can indeed be solved for example using the Chambolle-Pock algorithm [21], the Forward-Backward Splitting algorithm [41], the adaptive inverse scale space method [11] and the homotopy method [22] (where the last two seem to work only in the real-valued case, though).

we see that the solution (called *soft-thresholding* - see
Figure 2) is 0 for small values of x and it is pushed down (with respect to the line $y = x$) for "bigger" values of x; here is where the bias is introduced. This weakness of the ℓ^1 method opened the gate to several non-convex alternatives where the functional in (9) is replaced with

$$p(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|_{\ell^2}^2,$$

being now p a possibly non-convex function (that in jargon is called *penalty*).

In [24] the authors claim that "a good penalty function should result in an estimator with three properties: (nearly) *unbiasedness*[2], [...] *sparsity*, [...] *continuity*". The hard-thresholding [8] is sparsity-inducing and unbiased, but it is discontinuous; the ℓ^p quasi-norms, $p \in (0, 1)$, are continuous and sparsity-inducing, but they still suffer from a (weaker) form of shrinking bias. Their ([24]) proposed non-convex penality - called Smoothly Clipped Absolute Deviation (SCAD) - enjoys all the three properties; moreover it is proved that SCAD has the *oracle property*, i.e. it generates a sequence of *sparse* local minimizers (of some penalized likelihood function) that is $\sqrt{n}$-consistent (this means that the distance between the estimation and the ground truth decays, in probability, as $1/\sqrt{n}$, being n the dimension of the model). The authors also conjecture that the oracle property does *not* hold for the ℓ^1 penalty.

Several other non-convex alternatives were proposed to tackle the sparsity problem: the Adaptive LASSO [64], further studied in [18], the Exponential-Type Penalty (ETP) [29], the Non-negative Garrote (that first appeared as thresholding rule only in [10]), the Minimax Concave Penalty (MCP) introduced by Zhang in [63] (which was presented as nearly unbiased, but it is actually unbiased), the Huber loss [37]. Most of them come with some sort of statistical justification - asymptotic oracle property or consistency - but there seems to be no clear winner, also because these penalties are *separable*, i. e. they are of the type

$$p(\mathbf{x}) = \sum_{i=1}^{n} p_i(x_i);$$

this is a limitation, because they impose a choice between a bias for large values or a 0 gradient for large values, which possibly implies the existence of many stationary points and stagnation of gradient-based algorithms. To understand this better, we inspect a $1-$D case: consider indeed the problem

$$\min_x r(|x|) + (x - b)^2$$

for some unknown penalty r. The solution is either 0 or $x = b - \text{sign}(x)r'(|x|)/2$ and thus the derivative r' needs to be 0 to recover $x = b$ when b is large. Applied to each p_i, this simple observation explains why separability is a rather undesirable property.

[2]The notion of *near unbiasedness* does not seem to have a universal characterization. Some papers call nearly unbiased an estimator that keeps sufficiently large values unchanged; in other papers if the estimator's bias decays when the problem dimension increases.

Separable non-convex penalties that preserve the convexity of the objective functional were constructed for instance in [44][49], where however the focus seems to be concentrated more on algorithm development rather than on the theoretical understanding of which type of sparse minimum the functional has, or what is its relation to the ground truth or to the oracle solution.

Interesting results about *non-separable* penalties were proved in the noise-free scenario by Wipf and co-authors in [61], where some conditions for the uniqueness of the sparse global minimizer are given (Theorem 8); remarkable are also the results by Selesnick and co-workers [53][54][38], who constructed different non-convex non-separable penalties that preserve the convexity of the objective functional, nevertheless under rather strong structural assumptions on the sensing matrix A and with again little said about the nature of the sparse minimizer.

It is worth to mention here the ratio between ℓ^1 and ℓ^2 norms $\|\mathbf{x}\|_{\ell^1}/\|\mathbf{x}\|_{\ell^2}$, that has gained a significant attention in the recent years. The rationale is heuristic; there are evidences that the method might outperform the classic ℓ^1 in some problem instances [51] and it is also parameter independent. In [62] it is shown that the solution to (2) is a local minimizer of $\|\mathbf{x}\|_{\ell^1}/\|\mathbf{x}\|_{\ell^2}$ constrained to $A\mathbf{x} = \mathbf{b}$ under some null-space property of A. In [56] the proximal operator of $(\|\mathbf{x}\|_{\ell^1}/\|\mathbf{x}\|_{\ell^2})^+$ is computed, and an ADMM-based algorithm to solve the problem

$$\min_{\mathbf{x}\geq\mathbf{0}} \frac{\|\mathbf{x}\|_{\ell^1}}{\|\mathbf{x}\|_{\ell^2}} + \|A\mathbf{x}-\mathbf{b}\|^2$$

is proposed. This penalty function is however neither convex nor concave, and it is actually not even globally continuous; moreover the theory backing this method still seems rather modest.

In the previous pages we tried to outline theory and challenges of the modern compressed sensing. On one side we have the convex "world", mainly represented by the ℓ^1 penalty: a very rich theory with its limitations, sometimes even severe. On the other side the non-convex one, which is more a constellation of fairly disjoint objects. Our journey began here, with an attempt to create a synthesis of these two sides.

Low-rank matrices. There are several problem instances / applications where a low-rank matrix belonging to a prescribed linear, convex or non-convex set is sought. Examples are the Rigid Structure from Motion problem [57], the Non-Rigid Structure from Motion problem [9], Clustering and Classification [42], Computer Algebra [58], Compressive Hyperspectral Imaging [30][59] to mention a few. Typically the problem can be casted as

$$\min_{X\in\mathbb{M}_{m\times n}(\mathbb{C})} \operatorname{rank}(X) \quad \text{subject to } X\in\mathcal{C}.$$

Very often $\mathcal{C} = \{X \in \mathbb{M}_{m\times n}(\mathbb{C}) \,:\, \mathcal{A}(X) = M\}$, where $\mathcal{A}$ is a linear operator and M are measurements, often corrupted by noise or incomplete due to missing data. Sometimes

additional information on the rank of the sought matrix is available, either an estimate thereof or its exact value (as in the lifted phase retrieval problem, see [19]); in that case the low-rank problem can be recasted as

$$\text{find } X \in \mathbb{M}_{m\times n}(\mathbb{C}) \quad \text{subject to } X \in \mathcal{C} \text{ and } \operatorname{rank}(X) \leq r.$$

As in the vector case, the operator $X \mapsto \operatorname{rank}(X)$ is discontinuous and thus convex and non-convex relaxations have been proposed to replace it. The first (heuristic) penalty proposed was the trace norm $\operatorname{Tr}(X)$ [46] that however works only with square symmetric matrices; the nuclear norm $\|X\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(X)$ [25], being $\sigma_i(X)$ the singular values of X, pretty much imposed itself as standard approach, playing a role similar to the one played by the ℓ^1 norm in the vector case. The rationale behind using the nuclear norm is that $\|X\|_*$ is the convex envelope of $\operatorname{rank}(X)$ over the set $\{X \,:\, \sigma_1(X) \leq 1\}$ [25]; moreover it is possible to show that the two problems

$$\min_{X\in\mathbb{M}_{m\times n}} \|X\|_* \quad \text{subject to } \mathcal{A}(X) = M$$

and

$$\min_{X\in\mathbb{M}_{m\times n},\, Y\in\mathbb{H}_{m\times m},\, Z\in\mathbb{H}_{n\times n}} \operatorname{tr}(Y) + \operatorname{tr}(Z) \quad \text{subject to } \begin{cases} \mathcal{A}(X) = M \\ \begin{pmatrix} Y & X \\ X^* & Z \end{pmatrix} \succcurlyeq 0 \end{cases}$$

are equivalent, and thus the nuclear norm minimization problem enjoys all the numerical benefits of semidefinite programs. In a similar fashion to the vector case, uniqueness and robustness results are proved under the restricted isometry property (for matrices) [50] or rank null-space property [52].

Mirroring again the vector theory, the nuclear norm suffers from a shrinking bias; moreover, as noted in [32], there is usually no μ for which the solution to

$$\min_X \mu\|X\|_* + \|X - M\|_F^2 \quad \text{subject to } \operatorname{rank}(X) \leq r$$

is the projection of M onto the manifold of rank $\leq r$ matrices (that would be the solution to $\min_{\operatorname{rank}(X)\leq r} \|X-M\|_F^2$, cfr. Eckart-Young theorem), which is an undesirable property. For these very reasons several non-convex alternatives were proposed [39][47][36][32].

Our contributions

This thesis deals with functionals of the type

$$\mathcal{Q}_\gamma(f)(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2$$

where $\mathcal{Q}_\gamma$ is a general transform called *quadratic envelope* and studied in depth in [20], $f : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is a functional (virtually any) on a separable Hilbert space $\mathcal{V}$ and $A : \mathcal{V} \to \mathcal{W}$ is a linear operator. The quadratic envelope of f is defined as

$$\mathcal{Q}_\gamma(f)(\mathbf{x}) = \sup_{\alpha\in\mathbb{R},\, \mathbf{y}\in\mathcal{V}} \left\{\alpha - \frac{\gamma}{2}\|\mathbf{x} - \mathbf{y}\| \,:\, \alpha - \frac{\gamma}{2}\|\cdot - \mathbf{y}\|^2 \leq f\right\}$$

and it is possible to prove (Theorem 3.1 in [20]) that

$$\mathcal{Q}_\gamma(f)(\mathbf{x}) = \left(f + \frac{\gamma}{2}\|\cdot - \mathbf{d}\|^2\right)^{**}(\mathbf{x}) - \frac{\gamma}{2}\|\mathbf{x} - \mathbf{d}\|^2$$

where * is the Fenchel conjugate. Even though $\mathcal{Q}_\gamma(f)(\mathbf{x})$ might not be necessarily convex, it enjoys a remarkable set of good properties: it is continuous in the interior of its domain (if f is lower semi-continuous) and if f is semi-algebraic and $\mathcal{V}$ is finite-dimensional, then $\mathcal{Q}_\gamma(f)$ is semi-algebraic too. Moreover $\mathcal{Q}_\gamma(f)(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2$ has possibly fewer local minimizers than $f(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2$ and the *global* minimizers of the latter are not moved by regularizing f. This construction was already more concretely studied in [40] with $f = \mu\,\mathrm{rank}(X)$, in the context of low-rank matrices.

We detail now the contributions of each single manuscript:

Unbiasedness. In Paper I we study the functionals

$$\mathcal{K}_{\mu,\mathrm{reg}}(\mathbf{x}) = \mathcal{Q}_2(\mu\mathrm{card})(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|_{\ell^2}^2 \tag{10}$$

and

$$\mathcal{K}_{k,\mathrm{reg}}(\mathbf{x}) = \mathcal{Q}_2(\iota_{P_k})(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|_{\ell^2}^2 \tag{11}$$

where $\iota_{P_k}(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathrm{card}(\mathbf{x}) \leq k \\ \infty & \text{else.} \end{cases}$ Both penalties have an explicit formula [20] [1] and while $\mathcal{Q}_2(\mu\mathrm{card})(\mathbf{x})$ is separable (and coincides with the MCP introduced in [63]), $\mathcal{Q}_2(\iota_{P_k})(\mathbf{x})$ is non-separable and a completely new mathematical object. If the measurements $\mathbf{b}$ are noisy, i. e. are of the type $\mathbf{b} = A\mathbf{x}_0 + \mathbf{e}$ for some ground truth $\mathbf{x}_0$ and noise $\mathbf{e}$, it is impossible to retrieve $\mathbf{x}_0$ and the best that one could hope for is the so-called *oracle solution*, i. e. the solution to $A_S\mathbf{x} = \mathbf{b}$ in the least square sense, being $S = \mathrm{supp}(\mathbf{x}_0)$ and A_S the matrix obtained by removing (or setting to zero) the columns of indices in S^c from

A; essentially it is the solution to $A\mathbf{x} = \mathbf{b}$ *as if the support of* $\mathbf{x}_0$ *was a priori known.* Our main results essentially state that, under some conditions on the noise and on the sensing matrix A, the oracle solution is the *unique* global minimizer of both (10) and (11). Before stating the two main theorems we recall that the Lower Restricted Isometry Property (LRIP) constant δ_k^- of order k of a matrix A [7] is defined as

$$\delta_k^- = 1 - \inf \left\{ \frac{\|A\mathbf{x}\|_{\ell^2}^2}{\|\mathbf{x}\|_{\ell^2}^2} \, : \, \mathbf{x} \neq 0, \ \mathrm{card}(\mathbf{x}) \leq k \right\};$$

we moreover introduce the notation, for a matrix A, $\|A\|_{\infty,\mathrm{col}} = \max_i \|a_i\|_2$, being a_i the i-th column of A.

We are now ready to state the main theorems of Paper **I**:

Theorem (4.9). *Suppose that* $\mathbf{b} = A\mathbf{x}_0 + \mathbf{e}$ *where* A *is an* $m \times n$ *matrix with* $\|A\|_{\infty,\mathrm{col}} \leq 1$ *and set* $\mathrm{card}(\mathbf{x}_0) = k$. *Let* $N \geq 2k$, *assume that* $\|\mathbf{e}\|_{\ell^2} \leq (1 - \delta_k^-)\sqrt{\mu}$ *and*

$$|\mathbf{x}_{0,j}| > \left(\frac{1}{1 - \delta_k^-} + 1 \right) \sqrt{\mu}, \quad j \in \mathrm{supp}(\mathbf{x}_0).$$

Then the oracle solution $\mathbf{x}' = \mathbf{x}_{\mathrm{or}}$ *is the unique global minimum of* (10) *as well as* $\mu \mathrm{card}(\mu) + \|A\mathbf{x} - \mathbf{b}\|_{\ell^2}^2$, *with the property that* $\mathrm{supp}(\mathbf{x}') = \mathrm{supp}(\mathbf{x}_0)$, *that*

$$\|\mathbf{x}' - \mathbf{x}_0\|_{\ell^2} \leq \frac{\|\mathbf{e}\|_{\ell^2}}{\sqrt{1 - \delta_k^-}} \tag{12}$$

and that $\mathrm{card}(\mathbf{x}'') > N - k$ *for any other stationary point* $\mathbf{x}''$ *of* (10).

A similar result holds for (11):

Theorem (5.5). *Suppose that* $n \geq m + k + 2$ *(or* $n \geq 2m + k + 2$ *in the complex case) and that* $\|A\|_{\infty,\mathrm{col}} < 1$. *If* $\mathbf{e} \neq 0$ *and*

$$|\mathbf{x}_{0,j}| > \left(\frac{\|\mathbf{e}\|_{\ell^2}}{\sqrt{1 - \delta_k^-}} + \frac{2\|\mathbf{e}\|_{\ell^2}}{\sqrt{1 - \delta_{2k}^-}} \right), \quad j \in \mathrm{supp}(\mathbf{x}_0).$$

Then the oracle solution is a global minimum of (11) *and*

$$\|\mathbf{x}' - \mathbf{x}_0\|_{\ell^2} \leq \frac{\|\mathbf{e}\|_{\ell^2}}{\sqrt{1 - \delta_k^-}}.$$

We summarize here the main points of strength of our approach:

- both penalties / methods are *unbiased* because

$$\mathbb{E}(\mathbf{x}_{\text{or}}) - \mathbf{x}_0 = \mathbb{E}((A_S^* A_S)^\dagger A_S^* \mathbf{b}) - \mathbf{x}_0 = \mathbf{x}_0 - \mathbb{E}(\mathbf{e}) - \mathbf{x}_0 = 0$$

  as long as the noise has zero mean (with † we indicate the Moore-Penrose inverse). For the sake of completeness we must underline that ours is not the first proof of unbiasedness of MCP appearing in literature; Corollary 1 in [43] is for instance very similar in its conclusions to our Theorem 4.9, even though it is very hard to compare the premises: the assumptions on the sensing matrix A made in [43] are similar and / or comparable to LRIP (a property called *restricted strong convexity* is used instead), but there are further hypothesis on the covariance matrix of A that, together with a plethora of constants the magnitude of which is not clearly stated, seem rather hard to check in practice. Moreover Corollary 1 in [43] states that, under the aforementioned hypothesis the MCP has a *unique* stationary point while it was numerically observed both in Paper **I** and in [55] that, under the assumptions of Theorem 4.9, MCP has *a lot* of stationary points. We thus believe that our framework is more general, cleaner and somehow simpler (it does not involve asymptotic / probabilistic tools, but only machinery borrowed from classical convex analysis theory).

- the penalty $\mathcal{Q}_2(\iota_{P_k})$ is unbiased, non-separable, parameter independent and a completely new mathematical object; to the best of our knowledge no other non-separable penalty was shown to have such strong properties;

- we use LRIP instead of RIP, which is weaker; moreover there are no restrictions on δ_k^- other than < 1. In addition the estimate (12) is sharper than its convex counterparts (7) and (8), as the following picture shows:

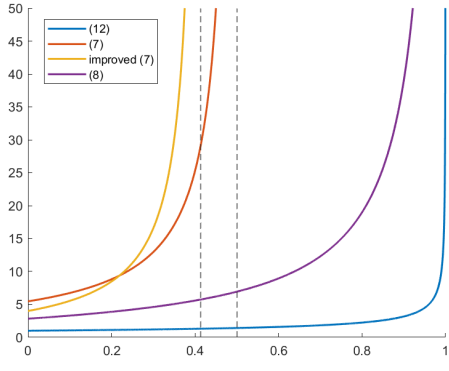

Figure 3: Noise coefficients comparison. δ_k on the x-axis and $C(\delta_k)$ on the y-axis.

The estimates (12), (7) and (8) are of the type $\|\mathbf{x}' - \mathbf{x}_0\|_{\ell^2} \leq C\|\mathbf{e}\|_{\ell^2}$ where C depends on δ_k or higher order RIP constants. In (8) we set $s = 1$ and $\theta_{s,1} = 0$; our

method is vastly better, even considering that the constants in (7) and improved (7) are functions of δ_{2k}, δ_{3k} and δ_{4k}, and $\delta_k \leq \delta_{2k} \leq \delta_{3k} \leq \delta_{4k}$ holds.

Escape routes and stationary points suppression. In Papers III and IV we tried to mitigate the effect that non-convexity might have (for instance on the optimization routines). Specifically, in Paper III we consider the functional

$$r_{\mu,\lambda}(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|_{\ell^2}^2 = \mathcal{Q}_2(\mu\text{card} + \lambda\|\cdot\|_{\ell^1})(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|_{\ell^2}^2; \tag{13}$$

the idea behind adding the (small) perturbation $\lambda\|\mathbf{x}\|_{\ell^1}$ is that it might help to mitigate the non-convexity of the original functional, by the price of introducing a small bias. This might be considered as a crossover method between those introduced in Paper I and the classical LASSO. We show there that actually $\mathcal{Q}_2(f+\lambda\|\cdot\|_{\ell^1})(\mathbf{x}) = \mathcal{Q}_2(f)(\mathbf{x}) + \lambda\|\mathbf{x}\|_{\ell^1}$, and moreover

$$\text{prox}_{\mathcal{Q}_2(f+\lambda\|\cdot\|_{\ell^1})/\rho}(\mathbf{y}) = \text{prox}_{\mathcal{Q}_2(f)/\rho}(\text{prox}_{\lambda\|\cdot\|_{\ell^1}}(\mathbf{y})), \quad \mathbf{y} \in \mathbb{R}^n$$

for any lower semi-continuous sign-invariant function $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(\mathbf{0}) = 0$ and $f(\mathbf{x}+\mathbf{y}) \geq f(\mathbf{x})$ for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+$.

In addition we prove that the λ-oracle solution, i. e. the solution to

$$\mathbf{x}_\lambda = \text{argmin}\, \lambda\|\mathbf{x}\|_{\ell^1} + \|A_S\mathbf{x} - \mathbf{b}\|_{\ell^2}^2$$

with $\text{supp}(\mathbf{x}_0) = S$, is a stationary point of (13) (Theorem 3.1) and that, under some assumptions on A, μ, λ and $\mathbf{x}_\lambda$, the other stationary points of (13) are non-sparse (Theorem 4.2). In conclusion, Figure 4 heuristically shows the correctness of our intuition: a very ill-posed problem was (randomly) generated and a cloud of starting points for the minimization routine was drawn. The small ℓ^1 term introduced in (13) dramatically enlarges the *convergence basin*, i. e. the set of starting points that lead to a "satisfactory" reconstruction. This means that either (13) suppresses some stationary points of (10) or that at least it allows the optimization algorithm (in our case, the Forward-Backward Splitting) to escape them.

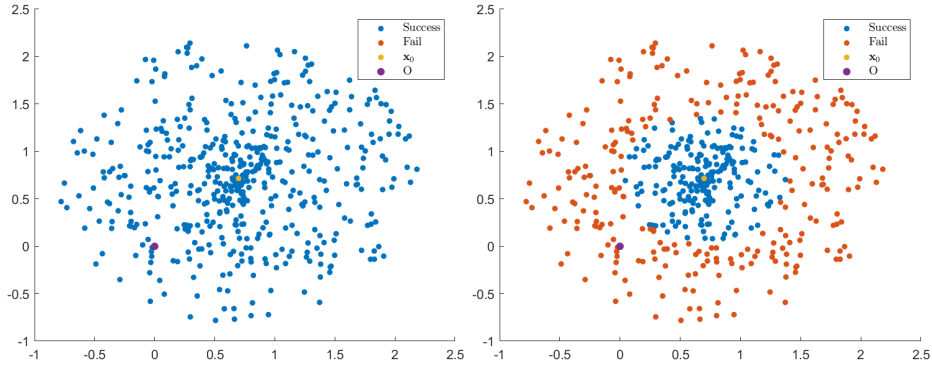

Figure 4: Converge basin of (13) (left) and (10) (right). A small λ helps with the convergence.

In Paper I we observed that the separability of $\mathcal{Q}_2(\mu\text{card})(\mathbf{x})$ makes (10) prone to have many dense local minima; this fact was not supposed to come as a surprise, but it actually came as such when we observed that the minimization routines tended to get stuck when the point $(A^*A)^\dagger A^*\mathbf{b}$ was used as starting point. With the aim of circumventing this issue, in Paper IV we thus investigated a further generalization of $\mathcal{Q}_2(\mu\text{card})(\mathbf{x})$. Specifically, given the weights function

$$G(k) = \sum_{i=0}^{k} g_i$$

where $0 = g_0 \leq g_1 \leq \cdots \leq g_k \leq \infty$, we considered the functional

$$\mathcal{Q}_2(G(\text{card}))(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|_{\ell^2}^2. \tag{14}$$

The functions $\text{card}(\mathbf{x})$ and $\iota_{P_k}(\mathbf{x})$ are particular cases of $G(\text{card})(\mathbf{x})$, for when $g_i = 1$ for all i and $g_i = 0$ for $i \leq k$ and $g_i = \infty$ for $i > k$ respectively, and this is already a natural reason to study (14). We show that those dense local minima can be suppressed by choosing for instance $g_i = i^2$ for $i \leq k_{\max}$ and $g_i = \infty$ else, being $k_{\max}$ a certain index "tolerance" (that might depend on the available information on the problem); when the g_i are different the penalty $G(\text{card})(\mathbf{x})$ is non-separable and this makes it effectively capable to penalize the cardinality. The paper is complemented with strong theoretical evidences on existence and uniqueness of sparse stationary points of the functional (14). Numerical experiments, both synthetic and from real world applications (Non-Rigid Structure from Motion problem) demonstrate robustness and superior (with respect to the competitor methods) performances of the proposed approach.

Matrix theory. Paper II mirrors Paper I and it shows the flexibility of the theory introduced, which creates a unitary theoretical framework for these type of problems.

We here consider the functionals

$$\mathcal{Q}_2(\mu\text{rank})(X) + \|\mathcal{A}(X) - \mathbf{b}\|_{\ell^2}^2 \tag{15}$$

and

$$\mathcal{Q}_2(\iota_{R_k})(X) + \|\mathcal{A}(X) - \mathbf{b}\|_{\ell^2}^2 \tag{16}$$

and we study their properties. Section 4 provides the mathematical tools and observations to re-use the vector theory developed in Paper **1**: the idea is that $\text{rank}(X) = \text{card}(\sigma(X))$ and $\iota_{R_k}(X) = \iota_k(\sigma(X))$ where $\sigma(X)$ is the vector of the singular values of X; this hints how vector and matrix theory are connected.

Since the vector oracle solution does not have a matrix counterpart, we consider the best rank k solution X_B to $\mathcal{A}(X) = \mathbf{b}$ and we show that under some assumptions on the ground truth X_0, the noise level and the RIP constant of the operator $\mathcal{A}$, results resembling those in Paper **1** hold:

Theorem. *Suppose that* $\mathbf{b} = \mathcal{A}(X_0) + \mathbf{e}$ *where* $\|\mathcal{A}\| < 1$ *and* $\text{rank}(X_0) = K$. *Assume that* $\|\mathbf{e}\|_{\ell^2} \leq (1 - \delta_{2K}^-)^{3/2}\sqrt{\mu}/3$ *and*

$$\sigma_K(X_0) > \left[\frac{1}{1 - \delta_{2K}^-} + (1 - \delta_{2K}^-)\right]\sqrt{\mu}.$$

Then the best rank K *solution* X_B *is unique and equals the (also unique) global minimum of* (15). *Moreover*

$$\|X_0 - X_B\|_F \leq 2\mathbf{e}/\sqrt{1 - \delta_{2K}^-}$$

and any other stationary point X *of* (15) *has* $\text{rank}(X) > K$.

The following theorem holds for (16):

Theorem. *Suppose that* $\mathbf{b} = \mathcal{A}(X_0) + \mathbf{e}$ *where* $\|\mathcal{A}\| < 1$ *and that* $\delta_{2K}^- < 1/2$ *and* $\text{rank}(X_0) = K$. *Assume that*

$$\sigma_K(X_0) > \left[\frac{5}{(1 - \delta_{2K}^-)^{3/2}}\right]\|\mathbf{e}\|_{\ell^2}.$$

Then the best rank K *solution* X_B *is the global minimum of* (16)*, and there are no other local minima. Moreover*

$$\|X_0 - X_B\|_F \leq 2\mathbf{e}/\sqrt{1 - \delta_{2K}^-}$$

holds.

The Fourier phase retrieval problem

Problem and uniqueness. The (noiseless) Fourier phase retrieval problem is a famous inverse problem that arises in X-ray crystallography and Coherent Diffractive Imaging (CDI) [26]. It consists in reconstructing a function $g : X \to \mathbb{R}$ given the modulus of its Fourier transform

$$|\mathcal{F}(g)(\xi)| = \left| \int_{-\infty}^{\infty} g(\xi) e^{-2\pi i \xi \cdot \mathbf{x}} \, d\mathbf{x} \right|$$

together with some additional constraints on its support.

Figure 5: Lena (left) and the modulus of its 2-D Fourier transform (right, in logarithmic scale)

In applications the phase retrieval problem is usually stated as *discretized*: introducing the pure oscillatory exponential functions $f_{\mathbf{k}}(\mathbf{n}) = e^{2\pi i \frac{\mathbf{k}\cdot\mathbf{n}}{m}}$ with $\mathbf{k} \in \{0, \dots, m-1\}^d$ and $\mathbf{n} \in \{0, \dots, n-1\}^d$, the discretized phase retrieval problem amounts to finding $\mathbf{x} \in \otimes_{j=1}^{d} \mathbb{C}^n$ given

$$b_{\mathbf{k}} = |\langle \mathbf{x}, f_{\mathbf{k}} \rangle|^2 \tag{17}$$

plus possible geometrical constraints on $\mathbf{x}$. Typically the measurements b are not "pure", but corrupted by Gaussian and Poisson noise. When $d = 2$ the latter can be written in a better known form:

$$b_{(k_1,k_2)} = \left| \sum_{n_1=0}^{m-1} \sum_{n_2=0}^{m-1} \mathbf{x}_{(k_1,k_2)} e^{-\frac{-2\pi i (k_1 n_1 + k_2 n_2)}{m}} \right|$$

which is, indeed, the modulus of the 2-D discrete Fourier transform (DFT).

The uniqueness of the solution to the Fourier phase retrieval problem roughly depends on the dimension d. First of all one needs to consider the so-called *trivial ambiguities*:

indeed given a signal $\mathbf{x} \in \otimes_{j=1}^{d} \mathbb{C}^n$, the signals $\mathbf{x}(\mathbf{n} + \mathbf{n}_0)$ (spatial shift), $\mathbf{x}e^{i\theta_0}$ (global phase shift) and $\overline{\mathbf{x}(-\mathbf{n})}$ (conjugate inversion) with $\theta_0 \in \mathbb{R}$ and $\mathbf{n}_0 \in \{0, \dots, n-1\}^d$ fixed have the same Fourier transform magnitude of $\mathbf{x}$. Therefore uniqueness has to be intended as modulo these ambiguities. The uniqueness problem for $d = 1$ was investigated in [35], where it was shown that, given a one-dimensional compactly supported complex signal $\mathbf{x}$, it is possible to construct non-trivial signals $\mathbf{y}$ with the same autocorrelation of $\mathbf{x}$ (and therefore with the same Fourier transform magnitude). In [5] it was showed that not even the assumption that $\mathbf{x} \in \mathbb{R}_+^d$ is enough to ensure uniqueness.

When $d \geq 2$ Hayes showed in [33] (Theorem 7) that two signals $\mathbf{x}$ and $\mathbf{y}$ in $\otimes_{j=1}^{d} \mathbb{C}^n$ with the same Fourier magnitudes and the same support are essentially the same (i.e. the same up to trivial ambiguities) if $m_i \geq 2n_i - 1$ for all $i = 1, \dots, d$ and the z-transform $\sum_{\mathbf{n}} \mathbf{x}(\mathbf{n})\mathbf{z}^{-\mathbf{n}}$ has at most one irreducible non-symmetric factor. Hayes and McClellan showed in [34] that if $d \geq 2$ then the set of complex coefficients of irreducible polynomials in d variables is isomorphic to a subset of $\mathbb{R}^{2\alpha(k,d)}$ of Lebesgue zero-measure, and thus the set of compactly supported signals that cannot be uniquely identified (modulo the aforementioned trivial ambiguities) by the magnitude of their Fourier transform is in some sense negligible. This gives some sort of well-posedness of the Fourier phase retrieval problem when $d \geq 2$.

Algorithms. The very first family of algorithms, based on the idea of alternatingly adjust the support and the Fourier modulus, was proposed and developed by Fienup (one intensity measurement), Gerchberg and Saxton (two intensity measurements) respectively in [26] and [31].

The idea of projecting back and forth is old (for affine spaces it goes back at least to von Neumann [60]) and rather elementary: given two non-empty closed convex sets $C, D \subseteq \mathbb{R}^n$ one would like to find a point $x \in C \cap D$. It is possible to show [3] that the sequence $\{\mathbf{x}_n\}_{n \geq 1} \subseteq \mathbb{R}^n$ defined recursively by

$$\mathbf{x}_{n+1} = \mathcal{P}_C(\mathcal{P}_D(\mathbf{x}_n)) \tag{18}$$

converges to a point $x \in C \cap D$ (if the intersection is nonempty); $\mathcal{P}_C(\mathbf{x})$ is the projection of $\mathbf{x}$ on C.

Fienup applied this principle to the Fourier phase retrieval problem with $C = \{g : \text{supp}(g) \subseteq S\}$ ($S \subseteq \mathbb{R}^2$ prescribed support) and $D = \{g : |\mathcal{F}(g)(\mathbf{x})|^2 = b(\mathbf{x})\}$ ($b(\mathbf{x})$ given as measurement); however in this case D is not convex and the convergence is not guaranteed anymore. Also, the convergence of this Error Reduction (ER) scheme tends to be slow. Fienup subsequently refined and analysed a new method called Hybrid Input-Output (HIO) [26][27] where he introduced an additional step as an attempt to overcome the speed limitations of Error Reduction. The Hybrid Input-Output can be written [26] as

$$\mathbf{x}_{n+1} = \begin{cases} \mathcal{P}_D(\mathbf{x}_n)(\mathbf{r}) & \text{if } \mathbf{r} \in S \\ (I - \beta\mathcal{P}_D)(\mathbf{x}_n)(\mathbf{r}) & \text{otherwise} \end{cases} \tag{19}$$

being S the support (or an estimate thereof) of the sought ground truth signal. Other proposed schemes are the Hybrid Projection-Reflection (HPR) [4]

$$\mathbf{x}_{n+1} = \frac{1}{2}[\mathcal{R}_C(\mathcal{R}_D + (1-\beta)\mathcal{P}_D) + I + (1-\beta)\mathcal{P}_D](\mathbf{x}_n)$$

where $\mathcal{R}_C = 2\mathcal{P}_C - I$ and $\mathcal{R}_D = 2\mathcal{P}_D - I$ are the reflections and β is a parameter, usually belonging to $[0.5, 1]$, the Difference Map (DM) [23]

$$\mathbf{x}_{n+1} = \{I + \beta\mathcal{P}_C[(1+\gamma)\mathcal{P}_D - \gamma I] - \beta\mathcal{P}_D[(1+\gamma)\mathcal{P}_C - \gamma I]\}(\mathbf{x}_n)$$

and so forth; see [45] for a more comprehensive overview.

Despite the significant number of algorithms developed throughout the years to tackle the Fourier phase retrieval reconstruction in dimension > 1, the problem remains rather elusive and a robust universal approach with convergence guarantees is, to the best of our knowledge, still lacking [6]. Input-Output schemes remain the most popular approach, even though they turn out to be (empirically) successful only when coupled with more advanced min-max / saddle points techniques [45] (for example for the steplength selection).

PhaseLift approach. The PhaseLift approach was introduced and popularized by Candès and co-workers in [19][15]. They observed that $|\langle \mathbf{x}, f_{\mathbf{k}}\rangle|^2 = \langle \mathbf{x}\otimes\overline{\mathbf{x}}, f_{\mathbf{k}}\otimes\overline{f_{\mathbf{k}}}\rangle$ where $f_{\mathbf{k}}(\mathbf{n}) = e^{-2\pi i \mathbf{k}\cdot\mathbf{n}/m}$ and therefore the Fourier phase retrieval problem as stated in (17) boils down to find a rank 1 positive semidefinite linear operator $\mathbf{X}$ on $\otimes_{j=1}^{d}\mathbb{C}^n$ such that

$$\langle \mathbf{X}, f_{\mathbf{k}}\otimes\overline{f_{\mathbf{k}}}\rangle = b_{\mathbf{k}}, \qquad \mathbf{k}\in\{0,\dots,m-1\}^d. \tag{20}$$

The problem becomes then linear, but the price to be paid is that the dimension is squared. If we denote with T the space of positive semidefinite operators on $\otimes_{j=1}^{d}\mathbb{C}^n$ we can define the linear operator $\mathcal{A}: T \to \otimes_{j=1}^{d}\mathbb{C}^m$ by

$$\mathbf{X} \mapsto (\langle \mathbf{X}, f_{\mathbf{k}}\otimes\overline{f_{\mathbf{k}}}\rangle)_{\mathbf{k}} \tag{21}$$

and solve the lifted Fourier phase retrieval problem (20) using an optimization-based approach, i.e.

$$\min \|\mathcal{A}(\mathbf{X}) - \mathbf{b}\|^2 \text{ subject to } \operatorname{rank}(\mathbf{X}) = 1, \mathbf{X}\succcurlyeq 0. \tag{22}$$

Due to the non-convexity of the constraint $\operatorname{rank}(\mathbf{X}) = 1$, Candès and co-workers proposed to minimize the (convex) functional

$$\lambda\|\mathbf{X}\|_* + \|\mathcal{A}(\mathbf{X}) - \mathbf{b}\|^2$$

where $\|\cdot\|_*$ is the nuclear norm, known for being sparsity inducing (cfr. previous section) and λ is a parameter that trades between sparsity and data fit.

We note that the PhaseLift approach is computationally heavy and rather unfeasible for large scale problems; the reason is intrinsic: an image of 512×512 pixels, when "lifted", is represented by a $512^2\times 512^2 = 262144\times 262144$ matrix, which would require, if dense, approximatively 150 GigaBytes of RAM.

Our contributions

The (Fourier) phase retrieval problem in its PhaseLift form is a good example of a problem where the rank of the sought ground truth is known to be *exactly* equal to 1.

There are (at least) two operations that are used in practice with the aim of stabilizing this problem: the first is *oversampling* (that is some sense is a mere mathematical trickery), the other is *masking* (that would need a concrete physical counterpart). Oversampling consists in using intermediate frequencies in the Discrete Fourier transform calculation; in practice this simply means that $m_i > n_i$ for all i and it is usually achieved by simply considering the underlying ground truth as zero-padded along each dimension. Masking is mostly achieved on situ by means of partially covering the sample from the sensing beam, so that the actual specimen becomes $\mathbf{x}_0 \chi_{C_i}(\mathbf{n})$, being χ_{C_i} the indicator function of some set C_i. Multiple masks might be used for improved performances.

In Paper **v** we firstly establish a result on oversampling: we show that the rank of the operator (21) equals $\min\{|S|, (2n-1)^d\}$, being S the support of the ground truth $\mathbf{X}_0$ and $|S|$ its cardinality; this essentially says that oversampling does not necessarily add *linearly independent* conditions. We then propose an algorithm to tackle the 1-D and 2-D Fourier phase retrieval problem (with random masks), heavily relying on the tools developed in Paper **i** and Paper **ii**. We consider and minimize the functional

$$\mathcal{Q}_2(\iota_{R_1^+})(\mathbf{X}) + \|\mathcal{A}(\mathbf{X}) - \mathbf{b}\|^2$$

with

$$\iota_{R_1^+}(\mathbf{X}) = \begin{cases} 0 & \text{if } \operatorname{rank}(\mathbf{X}) \leq 1 \text{ and } \mathbf{X} \succcurlyeq 0 \\ \infty & \text{else} \end{cases}$$

and $\mathbf{b}$ as in (20). The minimization is done by means of the Forward-Backward Splitting algorithm; we provide an explicit formula for the proximal operator of the penalty function involved, together with a fast and optimized procedure to calculate the gradient step using the Fast Fourier Transform. Our experiments with the 1-D problem show that our method outperforms the standard approach based on the nuclear norm in a noisy scenario, and it has comparable performances to the re-weighted nuclear norm, with the advantage of being able to retrieve matrices of *exactly* rank 1. Additional experiments are conducted to investigate the role of oversampling, and they ultimately confirm our theoretical analysis.

Possible future directions

In this thesis we propose a rather rich theory that somehow spontaneously developed itself, especially after the foundations in Paper **i** were laid. We tried to address and answer to most of the natural questions we came across, but few of them remain open. We outline here few possible future research directions / problems:

- If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is lower semicontinuous and semialgebraic, then so is $\mathcal{Q}_\gamma(f)$ [20]; and thus the latter has the local Kurdyka-Łojasiewicz property under which the Forward-Backward Splitting algorithm is proved to be convergent to a stationary point [2]. Nonetheless it might be interesting to develop a tailor-made algorithm that fully uses the topological advantages of the transform $\mathcal{Q}$, in the same fashion as [8].

- In Paper IV we proved that oracle-type solutions are stationary points of (14) under some conditions (both in the vector and the matrix case). In the vector case they are also local minima under essentially the same conditions (cfr. Miscellaneous section). It's unclear to us if this is also true in the matrix scenario.

References

[1] F. Andersson, M. Carlsson, and C. Olsson. Convex envelopes for fixed rank approximation. *Optimization Letters*, 11(8):1783–1795, 2017.

[2] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137:91–129, 2013.

[3] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, 2010.

[4] H. H. Bauschke, P. L. Combettes, and R. Luke. Hybrid projection-reflection method for phase retrieval. *Journal of the Optical Society of America A*, 20(6):1025–1034, 2003.

[5] R. Beinert. Non-negativity constraints in the one-dimensional discrete-time phase retrieval problem. *Information and Inference: A Journal of the IMA*, 6(2):213–224, 2017.

[6] T. Bendory, R. Beinert, and Y. C. Eldar. Fourier Phase Retrieval: Uniqueness and Algorithms. *Compressed Sensing and its Applications. Applied and Numerical Harmonic Analysis*, pages 55–91, 2018.

[7] J. D. Blanchard, C. Cartis, and J. Tanner. Compressed Sensing: How Sharp Is the Restricted Isometry Property? *SIAM Review*, 53(1):105–125, 2011.

[8] T. Blumensath and M. E. Davis. Iterative Thresholding for Sparse Approximations. *Journal of Fourier Analysis and Applications*, 14:629–654, 2008.

[9] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. *Proceedings of IEEE Conference in Computer Vision and Pattern Recognition*, 2000.

[10] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[11] M. Burger, M. Moeller, M. Benning, and S. Osher. AN ADAPTIVE INVERSE SCALE SPACE METHOD FOR COMPRESSED SENSING. *Mathematics of Computation*, 82(281):269–299, 2013.

[12] J. Cahill, X. Chen, and R. Wang. The gap between the null space property and the restricted isometry property. *Linear Algebra and its Applications*, 501:363–375, 2016.

[13] T. T. Cai, L. Wang, and G. Xu. Shifting Inequality and Recovery of Sparse Signals. *IEEE Transactions on Signal Processing*, 58(3):1300–1308, 2009.

[14] E. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9–10):589–592, 2008.

[15] E. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase Retrieval via Matrix Completion. *SIAM Review*, 6(1):225–251, 2015.

[16] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[17] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[18] E. Candès, M. Wakin, and S. Boyd. Enhancing Sparsity by Reweighted ℓ^1 Minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2008.

[19] E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[20] M. Carlsson. On Convex Envelopes and Regularization of Non-convex Functionals Without Moving Global Minima. *Journal of Optimization Theory and Applications*, 183:66–84, 2019.

[21] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40:120–145, 2011.

[22] D. L. Donoho and Y. Tsaig. Fast Solution of ℓ^1-Norm Minimization Problems When the Solution May Be Sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, 2008.

[23] V. Elser. Phase retrieval by iterated projections. *Journal of the Optical Society of America A*, 20(1):40–55, 2003.

[24] J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[25] M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. *Proceedings of the American Control Conference*, 2001.

[26] J. R. Fienup. Reconstruction of an object from the modulus of its fourier transform. *Applied Optics*, 3(1):27–29, 1978.

[27] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982.

[28] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[29] C. Gao, N. Wang, Q. Yu, and Z. Zhang. A Feasible Nonconvex Relaxation Approach to Feature Selection. *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 25(1), 2011.

[30] T. Gelvez, H. Rueda, and H. Arguello. Joint sparse and low rank recovery algorithm for compressive hyperspectral imaging. *Applied Optics*, 56(24):6785–6795, 2017.

[31] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[32] C. Grussler, A. Rantzer, and P. Giselsson. Low-Rank Optimization With Convex Constraints. *IEEE Transactions on Automatic Control*, 63(11):4000–4007, 2018.

[33] M. H. Hayes. The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 30(2):140–154, 1982.

[34] M. H. Hayes and J. H. McLellan. Reducible polynomials in more than one variable. *Proceeding of the IEEE*, 70(2):197–198, 1982.

[35] E. Hofstetter. Construction of time-limited functions with specified autocorrelation functions. *IEEE Transactions on Information Theory*, 10(2):119–126, 1964.

[36] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and Accurate Matrix Completion via Truncated Nuclear Norm Regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2117–2130, 2012.

[37] Peter J. Huber. Robust Estimation of a Location Parameter. *Annals of Statistics*, 53(1):73–101, 1964.

[38] A. Lanza, S. Morigi, I. Selesnick, and F. Sgallari. Nonconvex nonsmooth optimization via convex-nonconvex majorization-minimization. *Numerische Mathematik*, 136:343–381, 2017.

[39] V. Larsson and C. Olsson. Convex Low-Rank Approximation. *International Journal of Computer Vision*, 120:194–214, 2016.

[40] V. Larsson, C. Olsson, E. Bylow, and E. Kahl. Rank Minimization with Structured Data Patterns. *European Conference on Computer Vision*, pages 250–265, 2014.

[41] P. L. Lions and B. Mercier. Splitting Algorithms for the Sum of Two Nonlinear Operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[42] G. Liu, Z. Lin, and Y. Yu. Robust Subspace Segmentation by Low-Rank Representation. *IEEE Transactions on Cybernetics*, 44(8):1432–1445, 2013.

[43] P.-L. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *Annals of Statistics*, 45(6):2455–2482, 2017.

[44] M. Malek-Mohammadi, C. R. Rojas, and B. Wahlberg. A Class of Nonconvex Penalties Preserving Overall Convexity in Optimization-Based Mean Filtering. *IEEE Transactions on Signal Processing*, 64(24):6650–6664, 2016.

[45] S. Marchesini. Invited article: A unified evaluation of iterative projection algorithms for phase retrieval. *Review of Scientific Instruments*, 78, 2007.

[46] M. Mesbahi. On the semi-definite programming solution of the least order dynamic output feedback synthesis. *Proceedings of the 38th IEEE Conference on Decision and Control*, 1999.

[47] K. Mohan and M. Fazel. Iterative reweighted algorithms for matrix rank minimization. *The Journal of Machine Learning Research*, 13(1):3441–3473, 2012.

[48] B. K. Natarajan. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[49] M. Nikolova. Estimation of binary images by minimizing convex criteria. *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, 1998.

[50] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. A simplified approach to recovery conditions for low rank matrices. *2011 IEEE International Symposium on Information Theory Proceedings*, 2011.

[51] Y. Rahimi, C. Wang, H. Dong, and Y. Lou. A scale-invariant approach for sparse signal recovery. *SIAM Journal on Scientific Computing*, 41(6):3649–3672, 2019.

[52] B. Recht, M. Fazel, and P. P. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, 2010.

[53] I. Selesnick. Sparse Regularization via Convex Analysis. *IEEE Transactions in Signal Processing*, 65(17):4481–4494, 2017.

[54] I. Selesnick and İ Bayram. Enhanced sparsity by non-separable regularization. *IEEE Transactions on Signal Processing*, 64(9):2298–2313, 2016.

[55] E. Soubies, L. LeBlanc-Féraud, and G. Aubert. A Continuous Exact ℓ_0 Penalty (CEL0) for Least Squares Regularized Problem. *SIAM Journal on Imaging Sciences*, 8(3):1607–1639, 2015.

[56] M. Tao and X.-P. Zhang. A unified study on ℓ^1 over ℓ^2 minimization. *ArXiv preprint: 2108.01269v1*, 2021.

[57] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[58] K. Usevich and I. Markovsky. Variable projection methods for approximate (greatest) common divisor computations. *Theoretical Computer Science*, 681:176–198, 2013.

[59] H. Vargas and H. Arguello. A Low-Rank Model for Compressive Spectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):9888–9899, 2019.

[60] J. von Neumann. On rings of operators. reduction theory. *Annals of Mathematics*, 50:401–485, 1949.

[61] D. P. Wipf, B. D. Rao, and S. Nagarajan. Latent Variable Bayesian Models for Promoting Sparsity. *IEEE Transactions on Information Theory*, 57(9):6236–6255, 2011.

[62] Y. Xu, A. Narayan, T. Hoang, and C. G. Webster. Analysis of the ratio of ℓ^1 and ℓ^2 norms in compressed sensing. *Applied and Computational Harmonic Analysis*, 55:486–511, 2021.

[63] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[64] H. Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 100(476):1418–1429, 2006.

Miscellaneous

In this short section we collect an handful of unpublished observations that we believe significant enough to appear in this thesis. We prove a generalization of Theorem 3.1 in Paper IV, i. e. that oracle-type solutions are local minimizer (and not only stationary points) for the functional $\mathcal{Q}(G(\mathrm{card}))(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2$ under some conditions on the coefficients g_i (Theorem C); in order to do that, a more detailed study of the Frechét subdifferential $\hat{\partial}\mathcal{Q}(G(\mathrm{card}))(\mathbf{x})$ (Theorem B) structure is carried out.

The oracle solution is a local minimizer of $\mathcal{Q}(G(\mathrm{card}))(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2$.

This note generalizes Theorem 3.1 in Paper IV. Recall that $G : \mathbb{N} \to \mathbb{R} \cup \{\infty\}$ is a function defined by

$$G(k) = \sum_{i=1}^{k} g_i$$

with $0 \leq g_1 \leq g_2 \leq \cdots \leq g_n \leq \infty$ integers. We set $f(\mathbf{x}) \coloneqq G(\mathrm{card}(\mathbf{x}))$, $k \coloneqq \#S$ and $p \coloneqq \#S^c$ so that $n = k + p$. With $\|\cdot\|$ we mean the ℓ^2-norm. The operator $\widetilde{\cdot}$ is such that $|\widetilde{\mathbf{x}}|$ is ordered decreasingly.

Lemma A. *There exists a non-empty set $V \subseteq \mathbb{C}^n$ where*

$$f(\mathbf{y}) = \mathcal{Q}(f)(\mathbf{y}) \quad \forall\, \mathbf{y} \in V.$$

Proof. By definition of the quadratic envelope $\mathcal{Q}$ we have that

$$f(\mathbf{y}) - \|\mathbf{x} - \mathbf{y}\|^2 \overset{(*)}{\leq} f(\mathbf{x}) \quad \forall\, \mathbf{x} \in \mathbb{R}^n \Longrightarrow f(\mathbf{y}) = \mathcal{Q}(f)(\mathbf{y}).$$

If we define

$$\bar{j} \coloneqq \begin{cases} \max\{j \,:\, g_j < \infty\} & \text{if } g_n < \infty \\ g_n & \text{if } g_n = \infty \end{cases}$$

then we need to assume $l \coloneqq \mathrm{card}(\mathbf{y}) \leq \bar{j}$. Now $(*)$ means $\sum_{j=1}^{l} g_j - \|\mathbf{x} - \mathbf{y}\|^2 \leq \sum_{j=1}^{\mathrm{card}(\mathbf{x})} g_j$ which is trivially satisfied if $\mathrm{card}(\mathbf{x}) \geq l$. If $0 \leq \mathrm{card}(\mathbf{x}) < l$ we get $\sum_{j=\mathrm{card}(\mathbf{x})+1}^{l} g_j \leq \|\mathbf{x} - \mathbf{y}\|^2$ and in order for this to be true for all cardinalities of $\mathbf{x}$ we need conditions on $\mathbf{y}$ of the type

$$\begin{cases} \|\mathbf{y}\|^2 \geq \sum_{j=1}^{l} g_j \\ \|\mathbf{y}\|^2 - |\widetilde{\mathbf{y}}|_1^2 \geq \sum_{j=2}^{l} g_j \\ \qquad \vdots \\ |\widetilde{\mathbf{y}}|_l^2 \geq g_l; \end{cases}$$

since the g_j are increasing, the last condition is enough to define the set

$$V \coloneqq \{\mathbf{y} \in \mathbb{C}^n,\ \mathrm{card}(\mathbf{y}) = l\ :\ |\widetilde{\mathbf{y}}|_l \geq \sqrt{g_l}\}.$$

□

Theorem B (Subdifferential structure). *Assume* $|\widetilde{\mathbf{x}}_{or}|_k > \sqrt{g_k}$. *Then*

$$\begin{cases} \mathbf{z}_j = 0 & if\, j \in S \\ |\mathbf{z}_j| \leq 2\min\{\sqrt{g_{k+1}}, |\widetilde{\mathbf{x}}|_k\} & if\, j \in S^c \end{cases} \Longrightarrow \mathbf{z} \in \widehat{\partial}\mathcal{Q}(f)(\mathbf{x}_{or}).$$

Proof. We want to show that

$$L = \liminf_{\|\mathbf{y}\|\to 0} \frac{\mathcal{Q}(f)(\mathbf{x}_{or}+\mathbf{y}) - \mathcal{Q}(f)(\mathbf{x}_{or}) - \langle \mathbf{z}, \mathbf{y}\rangle}{\|\mathbf{y}\|} \geq 0.$$

According to [1] we have that

$$\mathcal{Q}(f)(\mathbf{y}) = \sup_{\mathbf{x},\alpha}\{\alpha - \|\mathbf{x}-\mathbf{y}\|^2\ :\ \alpha - \|\mathbf{x}-\mathbf{v}\|^2 \leq f(\mathbf{v})\ \forall\, \mathbf{v}\};$$

the idea is thus to carefully select a quadratic lower bound for $\mathcal{Q}(f)$ for estimating the fraction in L from below. For brevity we introduce $q(\alpha, \mathbf{x}, \mathbf{v}) \coloneqq \alpha - \|\mathbf{x}-\mathbf{v}\|^2$.

We claim now that $q(\bar{\alpha}, \overline{\mathbf{w}}, \mathbf{v}) \leq f(\mathbf{v})$ for all $\mathbf{v} \in \mathbb{R}^n$, where $\bar{\alpha} = \sum_{j=1}^k g_j + \gamma$ and $\overline{\mathbf{w}} \in \mathbb{R}^n$ is such that $\overline{\mathbf{w}}_S = \mathbf{x}_{or}$, $|\overline{\mathbf{w}}_j| \leq \min\{\sqrt{g_{k+1}}, |\widetilde{\mathbf{x}}|_k\}$ for $j \in S^c$ and $\|\overline{\mathbf{w}}_{S^c}\|^2 = \gamma \leq p\min\{\sqrt{g_{k+1}}, |\widetilde{\mathbf{x}}|_k\}^2$. We prove this claim:

- First of all notice that

  $$q(\bar{\alpha}, \overline{\mathbf{w}}, \mathbf{x}_{or}) = \sum_{j=1}^k g_j + \gamma - \|\overline{\mathbf{w}} - \mathbf{x}_{or}\|^2 = \sum_{j=1}^k g_j + \gamma - \|\overline{\mathbf{w}}_{S^c}\|^2 = f(\mathbf{x}_{or})$$

  and $q(\bar{\alpha}, \overline{\mathbf{w}}, \mathbf{0}) = \sum_{j=1}^k g_j - \|\overline{\mathbf{w}}_S\|^2 \leq 0$.

- For any $0 < v \coloneqq \mathrm{card}(\mathbf{v}) < k$ the claim amounts to show that $\sum_{j=v+1}^k g_j + \gamma \leq \|\overline{\mathbf{w}} - \mathbf{v}\|^2$, and this is clearly true because $\min_{\mathbf{v},\mathrm{card}(\mathbf{v})=v} \|\overline{\mathbf{w}} - \mathbf{v}\|^2 = \|\overline{\mathbf{w}}_{S^c}\|^2 + \|\mathbf{u}_v\|^2$ where $\mathbf{u}_v$ is equal to $\mathbf{x}_{or}$ without its v biggest (in modulus) components.

- For $\mathbf{v}$ with $\mathrm{card}(\mathbf{v}) = k$ it's clear.

- For any $v \coloneqq \mathrm{card}(\mathbf{v}) > k$ the claim amounts to show that $-\sum_{j=k+1}^v g_j + \gamma \leq \|\overline{\mathbf{w}} - \mathbf{v}\|^2$ and this is again true because $\min_{\mathbf{v},\mathrm{card}(\mathbf{v})=v} \|\overline{\mathbf{w}} - \mathbf{v}\|^2 = \|\mathbf{u}_v\|^2$ where

$\mathbf{u}_v$ is now a vector obtained from $\overline{\mathbf{w}}$ by removing/subtracting $\mathbf{x}_{or}$ and the biggest $v-k$ components of $\overline{\mathbf{w}}_{S^c}$. Now

$$\sum_{j=k+1}^{v} g_j + \|\mathbf{u}\|^2 \geq (v-k)g_{k+1} + \|\mathbf{u}\|^2 \geq \|\overline{\mathbf{w}}_{S^c}\|^2 = \gamma$$

and this concludes the proof of the claim.

From the claim it follows that $q(\bar{\alpha}, \overline{\mathbf{w}}, \mathbf{x}_{or} + \mathbf{y}) \leq \mathcal{Q}(f)(\mathbf{x}_{or} + \mathbf{y})$; also notice that $\mathcal{Q}(f)(\mathbf{x}_{or}) = f(\mathbf{x}_{or})$. Therefore

$$\frac{\mathcal{Q}(f)(\mathbf{x}_{or}+\mathbf{y}) - \mathcal{Q}(f)(\mathbf{x}_{or}) - \langle \mathbf{z}, \mathbf{y}\rangle}{\|\mathbf{y}\|} \geq \frac{q(\bar{\alpha}, \overline{\mathbf{w}}, \mathbf{x}_{or}+\mathbf{y}) - f(\mathbf{x}_{or}) - \langle \mathbf{z}, \mathbf{y}\rangle}{\|\mathbf{y}\|} := A.$$

But

$$A = \frac{\gamma - \|\mathbf{y} - \overline{\mathbf{w}}_{S^c}\|^2 - \langle \mathbf{z}, \mathbf{y}\rangle}{\|\mathbf{y}\|} = \frac{\gamma - \|\mathbf{y}\|^2 - \|\overline{\mathbf{w}}_{S^c}\|^2 + \langle 2\overline{\mathbf{w}}_{S^c} - \mathbf{z}, \mathbf{y}\rangle}{\|\mathbf{y}\|}$$

and if $2\overline{\mathbf{w}}_{S^c} - \mathbf{z} = \mathbf{0}$ we simply have $A = -\|\mathbf{y}\|$; in this case $L \geq 0$. The structure of $\mathbf{z}$ follows from the arbitrariness of $\overline{\mathbf{w}}_{S^c}$.

□

We can prove now that the oracle solution $\mathbf{x}_{or}$ is a local minimum for $\mathcal{G}$ under moderate noise. As usual measured data $\mathbf{b}$ are of the type $A\mathbf{x}_0 + \mathbf{e}$ for some underlying ground truth $\mathbf{x}_0$ and some noise $\mathbf{e}$. We assume here that A has some restricted isometry property. Define $M := 2\min\{\sqrt{g_{k+1}}, |\widetilde{\mathbf{x}}|_k\}$.

Theorem C. *Assume that $|\widetilde{\mathbf{x}}_{or}|_k > \sqrt{g_k}$ and that $\|\mathbf{e}\| \leq M/2$. Then the oracle solution $\mathbf{x}_{or}$ is a local minimizer for $\mathcal{G}$.*

Proof. In the computations of the proof of Theorem B we constructed a quadratic lower bound for $\mathcal{Q}(f)$; consider now a perturbation $\mathbf{y} = \mathbf{y}_S + \mathbf{y}_{S^c}$ with $\mathbf{y}_S$ "fixed" such that $\min_j |\mathbf{x}_{or,j} + \mathbf{y}_{S,j}| > \sqrt{g_k}$. Now we redefine $\overline{\mathbf{w}}$ in the following way: $\overline{\mathbf{w}}_S = \mathbf{x}_{or} + \mathbf{y}_S$ and $\overline{\mathbf{w}}_j = A^*(A\mathbf{x}_{or} - \mathbf{b})_j + \beta e^{i\phi_j}$ for $j \in S^c$; if $\mathbf{y}_S$ is small enough in norm and, say, $|\overline{\mathbf{w}}_j| \leq M/2$ ($j \in S^c$), then for all $\phi_j \in (0, 2\pi]$ we have $q(\overline{\alpha}, \overline{\mathbf{w}}, \mathbf{v}) \leq f(\mathbf{v})$ for all $\mathbf{v} \in \mathbb{C}^n$ with $\overline{\alpha} = \sum_{j=1}^{k} g_j + \|\overline{\mathbf{w}}_{S^c}\|^2$.

We need to prove now that $\mathcal{G}(\mathbf{x}_{or} + \mathbf{y}) \geq \mathcal{G}(\mathbf{x}_{or})$ for small $\mathbf{y}$. Notice that

$$\begin{aligned}\|A(\mathbf{x}_{or}+\mathbf{y}) - \mathbf{b}\|^2 &= \|A\mathbf{x}_{or} - \mathbf{b}\|^2 + \|A\mathbf{y}\|^2 + 2\langle A\mathbf{y}, A\mathbf{x}_{or} - \mathbf{b}\rangle \\ &= \|A\mathbf{x}_{or} - \mathbf{b}\|^2 + \|A\mathbf{y}\|^2 + 2\langle \mathbf{y}, A^*(A\mathbf{x}_{or} - \mathbf{b})\rangle;\end{aligned}$$

moreover $\langle \mathbf{y}, A^*(A\mathbf{x}_{or}-\mathbf{b})\rangle_j = 0$ if $j \in S$, so we have $\langle \mathbf{y}, A^*(A\mathbf{x}_{or}-\mathbf{b})\rangle = \langle \mathbf{y}_{S^c}, A^*(A\mathbf{x}_{or}-\mathbf{b})\rangle$. Therefore if we now set $\phi_j = \arg(\mathbf{y}_{S^c,j})$, we do get

$$\begin{aligned}\mathcal{Q}(f)(\mathbf{x}_{or}+\mathbf{y}_S+\mathbf{y}_{S^c}) &\geq q(\overline{\alpha}, \overline{\mathbf{w}}, \mathbf{x}_{or}+\mathbf{y}) \\ &= \sum_{j=1}^{k} g_j - \|\mathbf{y}_{S^c} - \overline{\mathbf{w}}_{S^c}\|^2 \\ &= \sum_{j=1}^{k} g_j - \|\mathbf{y}_{S^c}\|^2 + 2\langle \mathbf{y}_{S^c}, A^*(A\mathbf{x}_{or}-\mathbf{b})\rangle + \beta\|\mathbf{y}_{S^c}\|_{\ell^1};\end{aligned}$$

using $\mathcal{Q}(f)(\mathbf{x}_{or}+\mathbf{y}_S) = \sum_{j=1}^{k} g_j$ we get that

$$\mathcal{G}(\mathbf{x}_{or}+\mathbf{y}) - \mathcal{G}(\mathbf{x}_{or}) \geq -\|\mathbf{y}_{S^c}\|^2 + \beta\|\mathbf{y}_{S^c}\|_{\ell^1} + \|A\mathbf{y}\|^2$$

and the RHS is strictly positive if $\mathbf{y}$ is small enough.

□

References

[1] M. Carlsson, *On convex envelopes and regularization of non-convex functionals without moving global minima*, Journal of Optimization Theory and Applications, 183 (2019), pp. 66–84.

Printed by Media-Tryck, Lund 2022 NORDIC SWAN ECOLABEL 3041 0903

Doctoral Theses in Mathematical Sciences 2022:3

ISBN 978-91-8039-088-0
ISSN 1404-0034