An unbiased approach to compressed sensing

Citation for the original published paper (version of record):
Carlsson, M., Gerosa, D., Olsson, C. (2020). An unbiased approach to compressed sensing. Inverse Problems, 36(11). http://dx.doi.org/10.1088/1361-6420/abbd7f

An unbiased approach to compressed sensing

OPEN ACCESS
IOP Publishing
Inverse Problems
Inverse Problems **36** (2020) 115014 (38pp)
https://doi.org/10.1088/1361-6420/abbd7f

An unbiased approach to compressed sensing

Marcus Carlsson[1], Daniele Gerosa[1,*] and Carl Olsson[1,2]

[1] Centre for Mathematical Sciences, Lund University, Lund, Sweden
[2] Electrical Engineering, Chalmers University of Technology, Göteborg, Sweden

E-mail: mc,gerosa,calle@maths.lth.se, daniele.gerosa@math.lu.se and caols@chalmers.se

Received 27 May 2020, revised 22 September 2020
Accepted for publication 1 October 2020
Published 23 October 2020

Abstract

In compressed sensing a sparse vector is approximately retrieved from an underdetermined equation system $Ax = b$. Exact retrieval would mean solving a large combinatorial problem which is well known to be NP-hard. For b of the form $Ax_0 + \epsilon$, where x_0 is the ground truth and ϵ is noise, the 'oracle solution' is the one you get if you *a priori* know the support of x_0, and is the best solution one could hope for. We provide a non-convex functional whose global minimum is the oracle solution, with the property that any other local minimizer necessarily has high cardinality. We provide estimates of the type $\|\hat{x} - x_0\|_2 \leqslant C\|\epsilon\|_2$ with constants C that are significantly lower than for competing methods or theorems, and our theory relies on soft assumptions on the matrix A, in comparison with standard results in the field. The framework also allows to incorporate *a priori* information on the cardinality of the sought vector. In this case we show that despite being non-convex, our cost functional has no spurious local minima and the global minima is again the oracle solution, thereby providing the first method which is guaranteed to find this point for reasonable levels of noise, without resorting to combinatorial methods.

Keywords: compressed sensing, regularization, non-convex optimization, non-smooth optimization

(Some figures may appear in colour only in the online journal)

*Author to whom any correspondence should be addressed.

1. Introduction

1.1. Background

We consider the classical compressed sensing problem of minimizing the cardinality card $(x) = \|x\|_0$ of an approximate solution to an underdetermined equation system $Ax = b$, i.e.

$$\underset{x:\|Ax-b\|_2<\eta}{\operatorname{argmin}} \ \operatorname{card}(x), \tag{1}$$

where $\eta > 0$ is some allowed tolerance of the error and x_0 lies in $\mathbb{R}^n$ or $\mathbb{C}^n$. Problem (1) is NP-hard [36] and a popular approach is to replace card(x) with the convex function $\|x\|_1$, i.e.

$$\underset{x:\|Ax-b\|_2<\eta}{\operatorname{argmin}} \ \|x\|_1. \tag{2}$$

This method goes back (at least) to the 70s (see the introduction of [37] for a nice historical overview) but received increasing attention in the late 90s due to the work by Chen *et al* [21] on what they called *basis pursuit*, which amounts to solving

$$\operatorname{argmin}\left\{\lambda\|x\|_1 + \frac{1}{2}\|Ax-b\|_2^2\right\} \tag{3}$$

for a suitable choice of parameter λ, playing the role of η in (2). In fact, (3) is the dual problem of (2) in the sense that for each η there is a λ such that the solution of (2) and (3) coincides. The method received massive attention after the works of Donoho, Candés and coworkers in the early 2000, and the term compressed sensing was coined. In [14], Candés *et al* proved the surprising result that, given a k-sparse vector x_0 and a measurement

$$b = Ax_0 + \epsilon, \tag{4}$$

where ϵ is Gaussian noise, solving (2) yields (for a suitable choice of η) a vector $\hat{x}$ that satisfies

$$\|\hat{x} - x_0\|_2 < C_k\|\epsilon\|_2, \tag{5}$$

where C_k is a constant. Arguing that it is impossible to beat a linear dependence on the noise (even knowing the true support of x_0 *a priori*), the estimate (5) led the authors to conclude that 'no other method can significantly outperform this'. The result holds given certain assumptions on the matrix A, related to the restricted isometry property (RIP) of A, which in a separate publication (theorem 1.5, [15]) was shown to hold with 'overwhelming probability'.

These results give the impression that the theory is more or less complete and that improvements only can be marginal. However, what is not so well known is that the mentioned results usually do not apply to regular applications of the framework. For example, that 'statement $A(n)$' holds with 'overwhelming probability' only entails that the probability of $A(n)$ being false decays exponentially with the size n of the application, hence a statement can hold with 'overwhelming probability' and at the same time be false for most moderately sized applications (in some applications the dimension of the signals, in our case n, is modest. See for instance [29]. For a more extensive survey on compressed sensing applications, see [42]). In addition other assumptions need to be fulfilled for the 'overwhelming probability'-results to kick in, for example theorem 1.5 in [15] requires (according to the text below the theorem) that k/n is of magnitude 10^{-4}, which rules out most applications independent of whether n is large or not. While the theory has been improved since 2005, the main problem that the results often do not apply to standard applied settings, remains. For example the recent works

[1, 2, 12] provide *asymptotic* theorems about when compressed sensing works in concrete setups. Moreover, in [47] it was even shown that the method fails with probability tending to 1, as a function of n, if the ratio of k, n and m (the amount of measurements) is held fixed.

In addition to all this, whereas very strong recovery results were reported e.g. in [13, 20, 23] for the case of *exact data* $b = Ax_0$, in the presence of noise the method gives a well known bias (see e.g. [26, 35]). The ℓ^1 term not only has the (desired) effect of forcing many entries in x to 0, but also the (undesired) effect of diminishing the size of the non-zero entries. This is clearly visible even in the one-dimensional situation; the function $\mathbb{R} \ni x \mapsto \lambda|x| + \frac{1}{2}|x - x_0|^2$ has its minimum shifted toward 0 from the sought point x_0. This has led to a large amount of non-convex suggestions to replace the ℓ^1-penalty, see e.g. [4, 7–10, 37, 20, 26–28, 33–35, 41, 48, 50, 53, 54]. However, among these there is no clear winner and still ℓ^1-methods seems to be the standard choice among engineers, maybe also due to its simplicity. A fairly well-known non-convex alternative is the minimax concave penalty (MCP) by Zhang, which was coined *nearly unbiased* since the results in [52] imply that the method does find the *oracle solution* with probability tending to one under the assumptions of that paper. The 'oracle solution' is sort of the holy grail of compressed sensing, and aside from Zhang's work and this publication, there seems to be no reliable methods (with proofs) of how to find it.

1.2. Quadratic envelopes

In this paper we analyze two different methods to find the oracle solution, one which actually coincides with Zhang's MCP-penalty and a more intricate (and reliable) one that assumes *a priori* knowledge of the sparsity level k. In fact, these two are the tip of an iceberg of possible methods based on the 'quadratic envelope', which we now introduce. Consider the general problem of minimizing

$$\mathcal{K}(x) = f(x) + \|Ax - b\|_2^2 \tag{6}$$

where f is some non-convex penalty and x is a vector in some linear space, not necessarily $\mathbb{R}^n$. The standard non-convex example mentioned in most introductions to papers on compressed sensing is $f(x) = \mu\,\mathrm{card}(x)$ for some trade off parameter $\mu > 0$. However, if the desired cardinality k is known *a priori*, we can take f to be the indicator function ι_{P_k} of the set $P_k = \{x : \mathrm{card}(x) \leqslant k\}$ in which case (6) reduces to

$$\operatorname*{argmin}_{\mathrm{card}(x)\leqslant k} \|Ax - b\|_2. \tag{7}$$

In [17] *quadratic envelope* $\mathcal{Q}_2(f)$ was introduced, where $\mathcal{Q}_2$ is the *quadratic biconjugate* and $f : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ can be any functional on a separable Hilbert space $\mathcal{V}$; apart from the name, this transform was introduced already in [16] and goes back to the work of Larsson, Olsson [32]. It is defined as

$$\mathcal{Q}_2(f)(x) = \sup_{\alpha\in\mathbb{R},\ y\in\mathcal{V}} \{\alpha - \|x - y\|^2 :\ \alpha - \|\cdot - y\|^2 \leqslant f\} \tag{8}$$

see figure 1 (taken from [17]) for an illustration. An explicit form for $\mathcal{Q}_2$ is not always possible, but it is for the two functions that this paper examines, see (24) and (51). The quadratic envelope has also the property that $\mathcal{Q}_2(f)(x) + \|x\|_2^2$ is the lower semi-continuous convex envelope of $f(x) + \|x\|_2^2$. The relationship between

$$\mathcal{K}_{\mathrm{reg}}(x) = \mathcal{Q}_2(f)(x) + \|Ax - b\|_2^2 \tag{9}$$

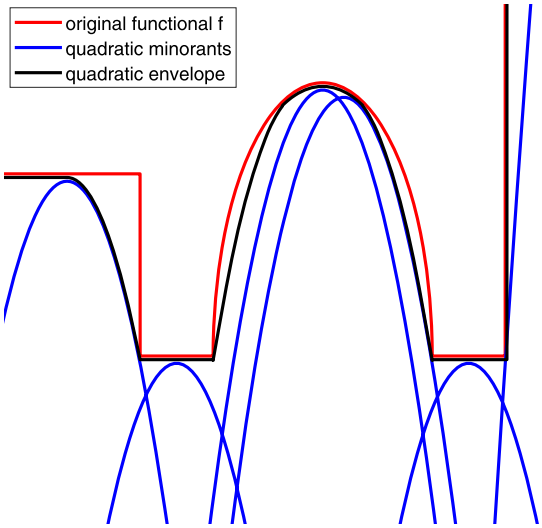

Figure 1. Illustration of a non-convex function f (red) and its quadratic envelope $Q_2(f)$ (black). The black graph lies slightly below for illustration only.

and the original functional in (6) was investigated in [17]. The inequality $\mathcal{K}_{\text{reg}} \leqslant \mathcal{K}$ is immediate by (8) and will be used throughout.

Given that $\|A\|_{\text{op}} < 1$ (which always can be achieved by rescaling), the main result of [17] is that the set of local minimizers to (9) is a subset of the local minimizers of (6), and most importantly that the global minimizers coincide[3].

In the particular case of $f(x) = \mu\,\text{card}(x)$, which is the first instance considered in this paper, the functional (9) has previously been introduced by Zhang [52] under the name MCP and independently by Soubies *et al* [46] under the name *CE*ℓ0. It also shows up in earlier publications, for example (2.4) in [26], but it seems like [52] is the first comprehensive performance study and [46] the first publication where the connection with convex envelopes appears. For this choice of f, the value of the contributions of the present paper is mainly theoretical, which goes much beyond what was previously known. In particular we show that the global minimizer with the MCP-penalty (i.e. $\mathcal{Q}_2(\text{card})$) is the oracle solution (for an appropriate range of the parameter μ); hence it follows that the MCP is actually *unbiased*, not merely *nearly unbiased* as claimed in [52].

To clarify what we mean by this, we note that it is easy to prove that the error in the oracle solution depends linearly on the noise, and hence the expectation of the error will be zero as long as the expectation of the noise is zero. In this sense any method finding the oracle solutions will be unbiased, which justifies the title of the paper.

The second penalty under consideration in this paper, $\mathcal{Q}_2(\iota_{P_k})$, is a new object that has only appeared previously in earlier publications by the authors of the present article. It also has the capacity of finding the oracle solution and the benefit that it does not rely on an appropriate parameter choice μ, as long as the model order is known. In contrast to the MCP-penalty (and most other previously studied sparsity priors) it is not separable but assigns a penalty that

[3] For the functionals considered in this paper the condition $\|A\|_{\text{op}} < 1$ can be substantially relaxed, as we will explain further below.

depends on the number of non-zero elements. This leads to significant differences with respect to the optimization landscape and the distribution of stationary points that we will study in this paper. In this article we provide theoretical results of the type (5) for the two concrete functionals $\mathcal{Q}_2(\text{card})$ and $\mathcal{Q}_2(\iota_{P_k})$. A more extensive discussion of previous results concerning MCP/$CE\ell 0$ is found in section 2.1, as well as other related results on non-convex optimization.

1.3. Contributions

A clear drawback with non-convex optimization schemes is that algorithms are bound to get stuck in local minima, and in concrete situations it is hard to determine whether this is the case or not. In the present article we give simple conditions which imply that the global minima of (9) for $f(x) = \mu\,\text{card}(x)$ *is the oracle solution*, and moreover that any local minima necessarily has a high cardinality unless it is the global minima. Hence, if a sparse local minima is found one can be sure that it is the oracle solution. In the case of $f = \iota_{P_k}$ we take this one step further and give conditions under which (9) has a unique local minimizer, which hence must be the oracle solution and also the solution to the original problem (7).

To be more precise, when the 'measurement' b has the form $b = Ax_0 + \epsilon$ and x_0 is a sparse vector, we significantly improve the state of the art in compressed sensing in a number of ways. Firstly, the conditions on A hold in greater generality, in the sense that our counterpart to conditions such as 'small RIP-values' or 'small mutual coherence' (see e.g. [30]) hold to a much greater extent than existing theory for other approaches such as ℓ^1-minimization or iterative hard thresholding (IHT). Secondly, since the global minimizer of our functionals is the oracle solution, we obtain an estimate corresponding to (5) where the involved constants are significantly smaller than C_k (or other constants with a similar role found in the references). Thirdly, we show numerically that forward–backward splitting (FBS) finds this in scenarios when competitors fail, thereby providing novel robust completely unbiased algorithms for compressed sensing (at least in the setting when A has normalized Gaussian random columns).

In section 2 we present highlights from the theory, show some numerical results and compare with the traditional ℓ^1-method (3). In section 2.1 we give a brief review of the field. The remainder of the paper, sections 3–5, are devoted to developing the theory.

2. Main results and innovations

Again, we will investigate minimizers of (9) for the two penalties $\mathcal{Q}_2(\mu\ \text{card})$ and $\mathcal{Q}_2(\iota_{P_k})$. We present key findings in sections 2.2 and 2.3. First we give a brief review of the field.

2.1. Brief review of related results

Needless to say, we are not the first group to address the shortcomings of traditional ℓ^1-minimization by use of non-convex penalties. In fact, even before the birth of compressed sensing, the shortcomings of ℓ^1-techniques were debated and non-convex alternatives were suggested, we refer to [26] for an overview of early publications on this issue. Moreover, shortly after publishing the celebrated result (5), Candés, Wakin and Boyd suggested an improvement called 'reweighted ℓ^1-minimization' [37] which also became a big success. They provide a theoretical understanding of this algorithm as minimizing the non-convex functional

$$f(x) = \sum_j \log(\epsilon + |x_j|)$$

where ϵ is a parameter chosen by the user. Figure 2 shows the functions $\text{card}(x)$, $|x|$ and $\log(0.1 + |x|) - \log(0.1)$ as well as $\mathcal{Q}_2(\text{card})$. As is clear to see, $\log(0.1 + |x|) - \log(0.1)$

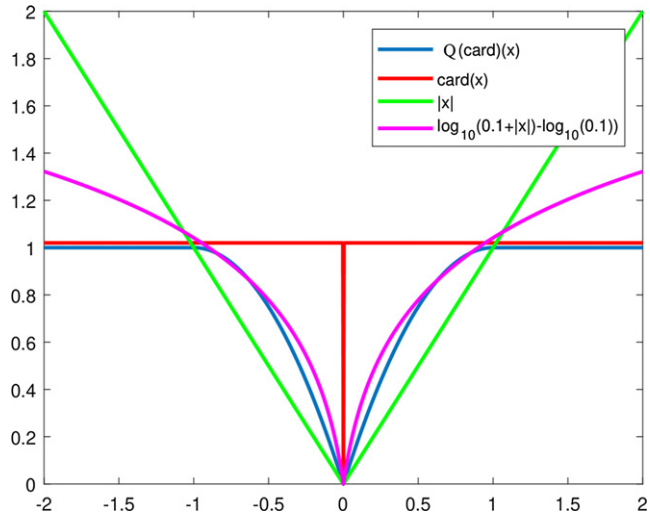

Figure 2. Illustration of penalties.

is closer to card(x) than $|x|$, which may explain the better performance by reweighted ℓ^1-minimization reported in [37, 23]. The functional $\mathcal{Q}_2$(card) is even closer to card(x), and while this certainly is one reason behind the superior theoretical results reported in this paper, there is still the issue of getting stuck in stationary points. In [46] the authors provide a macro algorithm to avoid non-local minima. In the same vein, Zhang [52] proposes to iteratively update relevant parameters to reach the desired global minima with higher probability.

Favorable results for $\mathcal{Q}_2(\mu$ card)/MCP were reported in the recent paper [34], which compares the use of MCP with ℓ^1 and reweighted ℓ^1 (called LSP in [34]) as well as SCAD (introduced in [26] which has similar performance as MCP). The numerical results in this paper seems also to reconfirm this, despite not employing any algorithm ensuring that we do not converge to an undesired stationary point.

The first theoretical justification of using MCP/$\mathcal{Q}_2(\mu$ card) is corollary 1 of [52], which roughly speaking contains an algorithm which finds the global minimum of MCP with high probability, and shows that the probability that this differs from the oracle solution is low. The result is based on very technical assumptions involving constants $c_*, c^*, d^*, d^0, \gamma, \sigma, w^0, \beta_*$ and $\tilde{p}_1$, and so it seems hard to verify if this result applies in a concrete situation.

A more recent theoretical justification to support the use of MCP is given in [34] which, under a number of assumptions, prove that (9) with $\mathcal{Q}_2(\mu$ card) does have the oracle solution as a unique stationary point with high probability, and provide an estimate of the type (5), see corollary 1. However, as with the results of Zhang, this result relies on a number of constants whose values are difficult to estimate, so it is hard to know when exactly the theorem applies. In addition we note that in many practical cases the MCP formulation has local minima, see our experimental evaluation, indicating that the assumptions made to ensure uniqueness are very restrictive. We believe that the corresponding theory in the present paper is much more transparent, with conditions that are more general and comparatively easy to verify, as well as stronger conclusions. We postpone further discussion of this to section 2.5.

The papers [4, 7, 8, 38, 39] considers (6) for the cases $f(x) = \text{card}(x)$ as well as $f(x) = \iota_{P_k}(x)$, and [4] show in particular that the FBS-algorithm applied to (6) converges to a stationary point, but a further analysis of this point is not present. In fact, it seems to us that

these papers fail to recognize that the oracle solution often is the global minimizer of both (10) and (15), which follows from the results of this paper (see corollaries 2.1 and 2.3 respectively).

Many other non-convex penalties have been proposed over the years [9, 10, 37, 20, 26–28, 33–35, 41, 48, 50, 52–54], and we make no attempt to review them here. The introduction of [34] contains a recent overview. A common denominator seems to be that the penalty function is separable, i.e. has the form $p(x) = \sum_j p_j(x_j)$ where p_j are functions on $\mathbb{R}$ (except the recent contribution [48]). The penalty $\mathcal{Q}_2(\iota_{P_k})$ is not of this form. In fact, $\mathcal{Q}_2(\iota_{P_k})$ is the simplest of a vast field of possible penalties introduced in [32] that can be more tailormade to the problem at hand, neither of which is separable.

2.2. Sparse recovery via $\mathcal{Q}_2(\mu\ \mathrm{card})$

We return to the first problem of minimizing (9) for $f = \mu\,\mathrm{card}(x)$ i.e.

$$\mathcal{K}_\mu(x) := \mu\ \mathrm{card}(x) + \|Ax - b\|_2^2 \tag{10}$$

where the parameter μ controls the tradeoff between sparsity and data-fit. Motivated by section 1.2 we propose to regularize $\mathcal{K}_\mu$ with

$$\mathcal{K}_{\mu,\mathrm{reg}}(x) = \mathcal{Q}_2(\mu\ \mathrm{card})(x) + \|Ax - b\|_2^2. \tag{11}$$

The graph of $\mathcal{Q}_2(\mathrm{card})$ is depicted in figure 2.

We will study uniqueness of sparse minimizers of both (10) and (11), in the sense that we give concrete conditions such that if there exists one local minimizer x' of (11) with the property that $\mathrm{card}(x') \ll m$ (in a manner to be made precise), then

- x' is automatically a global minimizer and also a solution to (10)
- Any other stationary point x'' of (11) satisfies $\mathrm{card}(x'') \gg \mathrm{card}(x')$.

To state our results, we remind the reader that A satisfies a RIP for integer k, if any k columns of A behaves approximately as an isometry, in the sense that

$$(1 - \delta_k)\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1 + \delta_k)\|x\|_2^2$$

for all k-sparse vectors $x \in \mathbb{R}^n$, $k \in \mathbb{N}$, and some constant $0 \leqslant \delta_k < 1$. Classical results from compressed sensing literature usually require that the numbers δ_k are small, something which we have found is hard to fulfill in practice. For example, the famous estimate (5) holds under the assumption that $\delta_{3k} + 3\delta_{4k} < 2$. This condition was later improved to the simpler estimate

$$\delta_{2k} < \sqrt{2} - 1 \approx 0.4, \tag{12}$$

(see [11]) which is the estimate currently reproduced in textbooks on the subject, such as [30]. Our numerical evaluation (see section 2.4) shows that this condition is usually not satisfied for a Gaussian random matrix A (with normalized columns) of size 100×200 (a common size for many applications), except for $k = 1$. The statement that RIP holds with overwhelming probability [15] is therefore somewhat misleading.

We base the theory of this paper on the lower restricted isometry property (LRIP), basically constituting the lower estimate of the RIP (introduced in [6]). More precisely, we define

$$1 - \delta_k^- = \inf\left\{\frac{\|Ax\|_2^2}{\|x\|_2^2} :\ x \neq 0,\ \mathrm{card}(x) \leqslant k\right\} \tag{13}$$

for $k = 1 \ldots n$. We say that A satisfies LRIP with respect to the property $P_k = \{x : \mathrm{card}(x) \leqslant k\}$ if $\delta_k^- < 1$. In other words A is LRIP with respect to this property if and only if any k chosen

columns of A are linearly independent. Clearly $\delta_k^- \leqslant \delta_k$ and for Gaussian matrices inequality typically holds, which is further discussed in section 2.4.

To give the reader an early insight into key findings, we state a simplified version of the main result of section 4, theorem 4.9 (for the particular case $N = 2k$).

Corollary 2.1. *Suppose that A has columns in the unit ball of $\mathbb{R}^n$ or $\mathbb{C}^n$, that $b = Ax_0 + \epsilon$ and set* $\mathrm{card}(x_0) = k$. *Assume that the noise is small enough that the open interval*

$$\left(\frac{\|\epsilon\|_2}{1-\delta_{2k}^-}, \frac{(1-\delta_{2k}^-)\min_{j\in supp\, x_0}|x_{0,j}|}{2}\right)$$

is non-empty. Then for any μ with $\sqrt{\mu}$ in the above interval, we have that

(*a*) *Then there exists a unique global minimum x' to $\mathcal{K}_{\mu,reg}$ as well as $\mathcal{K}_\mu$, and it is the oracle solution.*

(*b*) *We have that* $\mathrm{supp}\, x' = \mathrm{supp}\, x_0$

(*c*)

$$\|x' - x_0\|_2 \leqslant \frac{\|\epsilon\|_2}{\sqrt{1-\delta_k^-}},$$

(*d*) $\mathrm{card}(x'') > k$ *for any other stationary point x'' of $\mathcal{K}_{\mu,reg}$.*

Moreover, if the above estimates hold for some $N \gg 2k$ we can state that x'' has cardinality higher than $N - k$. In other words, either the algorithm finds the oracle solution or one with substantially higher cardinality. Although the theorem gives conditions on how to pick μ, the involved quantities are generally not exactly known. For some matrix families good estimates exist e.g. [6]. For other problems one has to proceed by trial and error (as with all to us known CS-methods). However, in our experience, the method is very robust and finds the oracle solution for a range of μ-values, as opposed to e.g. traditional ℓ^1-minimization (3) which gives a different solution for each λ.

Note that the conditions on 'noise' ϵ and 'ground truth' x_0 are very natural; if the noise is too large or if the non-zero entries of x_0 are too small, there is no hope of correctly retrieving the support. Also note the absence of a condition forcing δ_{2k}^- to be 'small', in sharp contrast to other results in the field such as (12) or $\delta_{3k} < 1/\sqrt{32}$ in [7] (conditions that are very hard to satisfy, see section 2.4). On the contrary, as long as $\delta_{2k}^- < 1$, corollary 2.1 holds, and in order for it to apply for some μ one needs that the signal to noise ratio, measured as

$$\mathrm{SNR} = \frac{\min_{j\in \mathrm{supp}\, x_0}|x_{0,j}|}{\|\epsilon\|_2}, \tag{14}$$

has to be sufficiently large, (more precisely larger than $\frac{2}{(1-\delta_{2k}^-)^2}$, for then the interval in the corollary is non-void).

2.3. Sparse recovery via $\mathcal{Q}_2(\iota_{P_k})$.

We now discuss the situation when the model order, i.e. the amount k of non-zero entries, is known. This problem is also known as the k-sparse problem and studied e.g. in [7]. For simplicity we restrict attention to $\mathbb{R}^n$, corresponding results for $\mathbb{C}^n$ are similar but the assumptions on A are slightly more technical (see section 5.1). As pointed out earlier the NP-hard problem (7) can be written

$$\mathcal{K}_k(x) = \iota_{P_k}(x) + \|Ax - b\|_2^2 \tag{15}$$

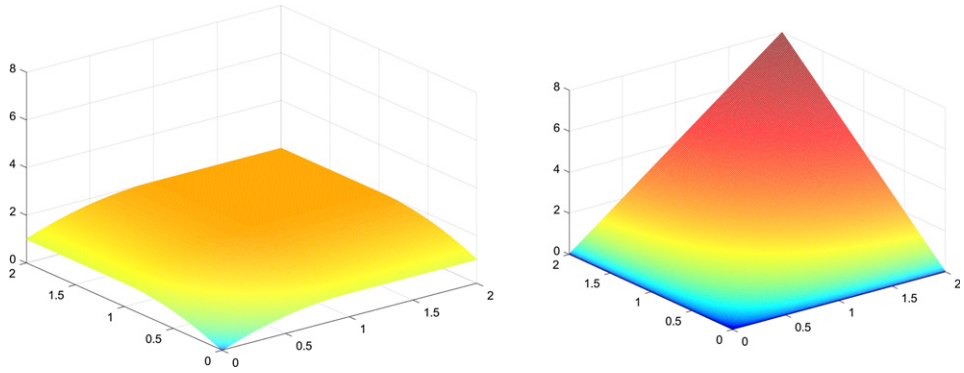

Figure 3. Two dimensional illustrations of the functions $\mathcal{Q}_2(\text{card})$ (left) and $\mathcal{Q}_2(\iota_{P_1})$ (right).

(where the subindex k separates the notation from (10)) which we regularize with

$$\mathcal{K}_{k,\text{reg}}(x) = \mathcal{Q}_2(\iota_{P_k})(x) + \|Ax - b\|_2^2. \tag{16}$$

Figure 3 shows $\mathcal{Q}_2(\iota_{P_1})$ as a function of two variables (in the positive quadrant). The penalty assigned is zero for all vectors with no more than one non-zero variable. For comparison we also plot $\mathcal{Q}_2(\text{card})$. Note that $\mathcal{Q}_2(\text{card})$ is constant in the region where both variables are larger than $\sqrt{\mu}$. This shape makes it likely that $\mathcal{K}_{\text{reg}}$ has local minimizers of high rank. In contrast $\mathcal{Q}_2(\iota_{P_1})$ has large gradients in this area which as we shall se makes it possible to exclude such stationary points for $K_{k,\text{reg}}$.

We first present a result where b is not necessarily given by $Ax_0 + \epsilon$.

Corollary 2.2. *Let A have columns in the unit ball such that no pair is orthogonal, and assume that $n \geqslant m + k + 2$. Any local minimizer x' of $\mathcal{K}_{k,reg}$ then satisfies* $\text{card}(x') \leqslant k$. *Moreover, set $z' = (I - A^*A)x' + A^*b$, let $\tilde{z}'$ contain the elements of z' sorted by decreasing magnitude, and assume that*

$$|\tilde{z}'_{k+1}| < (1 - 2\delta^-_{2k})|\tilde{z}'_k|. \tag{17}$$

Then x' is the unique global minimum of $\mathcal{K}_k$ and $\mathcal{K}_{k,reg}$.

A similar result also holds in the situation of the previous section. The interesting point to note is that there is a simple verifiable condition on whether a solution to (7) has been found, given that some estimate of δ^-_{2k} is available (see e.g. theorem 9.26 of [30] or [6]).

Corollary 2.2 is a combination of theorems 5.1 and 5.4. We now consider the case when $b = Ax_0 + \epsilon$ and we wish to retrieve x_0, where $\text{card}(x_0) = k$. By theorem 5.5, we have (for A as in the previous corollary);

Corollary 2.3. *Assume the SNR (as measured in (14)) is greater than $\frac{3}{\sqrt{1-\delta^-_{2k}}}$. Then the oracle solution is a unique global minimizer x' to $\mathcal{K}_{k,reg}$ with* $\text{supp}\,(x') = \text{supp}\,(x_0)$ *and moreover it satisfies $\|Ax' - b\|_2 \leqslant \|\epsilon\|_2$ and*

$$\|x' - x_0\|_2 \leqslant \frac{\|\epsilon\|_2}{\sqrt{1 - \delta^-_k}}.$$

Finally, if the SNR is also greater than

$$\left(\frac{1}{1-2\delta_{2k}^{-}}+\frac{1}{\sqrt{1-\delta_{k}^{-}}}\right),$$

Corollary 5.6 says that $\mathcal{K}_{k,\mathrm{reg}}$ has no local minimizers (except for the oracle solution). An interesting point to note is that minimizing $\mathcal{K}_k$ can be seen as finding one among $\binom{n}{k}$ possible minimizers (see the proof of theorem 5.5). However, $\binom{n}{k}$ is typically a large number, for example if $k=10$ and $n=1000$ it is around 2×10^{23}. The above corollary states that all but one of these, the relevant one, disappears when regularizing with $\mathcal{Q}_2(\iota_{P_k})$, which is rather amazing, in the authors humble opinion. However, this clearly demands that $\delta_{2k}^{-}<0.5$, to be compared with the state of the art assumption $\delta_{2k}<\sqrt{2}-1\approx 0.4$ for when standard compressed sensing results kick in [30]. In which situations is it likely to assume that either of these hold? We try to shed some light on this in the next section.

2.4. On the size of RIP/LRIP-constants

RIP-values are notoriously difficult to estimate, which makes it hard to compare theorems in compressed sensing. For example, the currently best known estimate for (5) was proven in [11], and is reproduced in textbooks such as [30]. It says that $C_k=8.5$ if $\delta_{2k}=0.2$, but how likely is that to happen? In [30] very intricate estimates in this direction are give in theorem 9.27, which claims that the $2k$-RIP constant δ_{2k} of a random Gaussian matrix $A/\sqrt{m}$ is[4]

$$\leqslant 2\left(1+\frac{1}{\sqrt{2\ \ln(e\cdot n/2k)}}\right)\eta+\left(1+\frac{1}{\sqrt{2\ \ln(e\cdot n/2k)}}\right)^{2}\eta^{2}$$

with probability $1-\epsilon$ if

$$m\geqslant 2\eta^{-2}(2k\ \ln(e\cdot n/2k)+\ln(2\epsilon^{-1})).$$

Now let us suppose we are interested in a very sparse signal, $k=10$, and $n=1000$. Then

$$\left(1+\frac{1}{\sqrt{2\ \ln(e\cdot 1000/20)}}\right)\approx 1.32$$

and $\left(1+\frac{1}{\sqrt{2\ \ln(e\cdot 1000/20)}}\right)^{2}\approx 1.74$. The equation $1.74\eta^2+2\times 1.32\eta=c$ gives the positive solution

$$\eta(c)\approx(\sqrt{1.74c+1.32^2}-1.32)/1.74.$$

For $c=0.2$, $\eta\approx 0.072$. Therefore we would need $m\geqslant 37\,878$, independently on the probability degree ϵ; this is absurd since we would like $m\ll n=1000$.

Of course, there is the possibility that the estimates for δ_{2k} are poor and that the reality is different. To test this we computed values of δ_j and δ_j^{-} for matrices of various size. The test matrices where generated by first drawing elements from i.i.d Gaussian distributions and then

[4] Constructing the matrix like this gives expected value of the column norms equal to 1, so is very similar to normalizing the columns, as done in the examples of this paper.

Table 1. $m = 25, n = 50$.

j	2	3	4	5	6
δ_j:	0.66	1.04	1.36	1.65	1.88
δ_j^-	0.66	0.79	0.87	0.92	0.95

Table 2. $m = 100, n = 200$.

j	2	3	4
δ_j	0.39	0.61	0.80
δ_j^-	0.39	0.52	0.62

Table 3. $m = 250, n = 500$.

j	2	3	4
δ_j	0.28	0.43	0.55
δ_j^-	0.28	0.38	0.44

normalizing the resulting columns. Note that this gives a matrix with columns drawn from a uniform distribution on the sphere. The results presented below where averaged over 5 trials.

From table 1 we can note several interesting things. For example, δ_j^- is usually a bit smaller than δ_j, and whereas the latter can become larger than 1 the former cannot, by definition. In fact, by the definition it is easy to see that $\delta_j^- = 1$ if and only if there are j linearly dependent columns in the matrix. This means that with probability 1, we always have $\delta_j^- < 1$ for $j \leqslant m$ whereas $\delta_j^- = 1$ for all $j > m$. In particular, corollary 2.1 is applicable with probability 1 whenever $k \leqslant m/2$.

The second thing to note is that we do not present very many values, which is related to the computational time. If we were interested in computing δ_{20} for a matrix with $n = 1000$, as discussed initially, we would need to perform $\binom{1000}{20} \approx 4 \times 10^{41}$ SVD's. In fact, even computing δ_7 for $n = 50$ requires around 10^9 SVD's, (which is not impossible but we skipped it since the numbers are very poor anyway). For this reason, the typical sizes of δ_j's remain a mystery, which likely is a reason behind the widespread belief that these numbers often are decent. To shed some light for larger matrices, we now compute for j up to 4 and $m = 100$ as well as 250 (with $n = 2m$) (tables 2 and 3).

The most striking thing to note is that the numbers are still terribly poor, even for $m = 250$. It certainly came as a surprise to the authors that none of the classical results on compressed sensing applies in the 250×500 setting, unless $k = 1$ and in this case the constant C_k is approximately 14 (based on our five trials average). Here a strength of the results of this paper becomes apparent, because even for an extremely poor value like $\delta_k^- = 0.95$ we have that the constant in the error estimate $\|x' - x_0\|_2 \leqslant \frac{1}{\sqrt{1-\delta_k^-}}\|\epsilon\|_2$ equals 4.5, almost half the value you get for C_k when $\delta_{2k} = 0.2$ in [11], as reported initially.

In fact, despite the difficulty in estimating the constants, it is not impossible to compare the quality of estimates. If we set $f_C(x) = \frac{4\sqrt{1+x}}{1-(1+\sqrt{2})x}$ and $f_{\mathrm{CGO}}(x) = \frac{1}{\sqrt{1-x}}$ then the constant C_k in (5), as defined in [11], is given by $f_C(\delta_{2k})$ whereas the corresponding constant in corollaries 2.1 and 2.3 is given by $f_{\mathrm{CGO}}(\delta_k^-)$. The functions f_C and f_{CGO} are displayed in figure 4. Clearly the latter constant is vastly better by just comparing these graphs, and this conclusion is further strengthened by noting that $\delta_k^- \leqslant \delta_k \leqslant \delta_{2k}$.

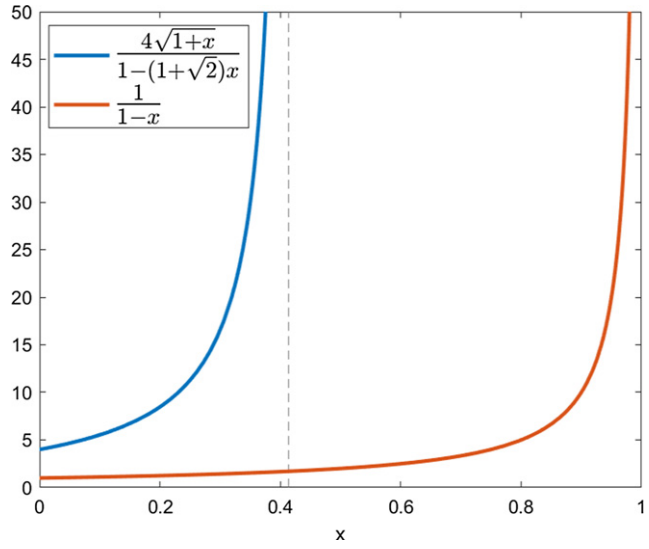

Figure 4. f_{CGO} in red and f_C in blue.

A fourth thing to note from the tables is that the δ_k's do decrease with m, as predicted by the theory, and once we hit $m = 250$ all reported numbers are below 0.5, the requirement for corollary 2.3 to kick in. However, note that once corollary 2.3 applies the error estimate immediately gets a very favorable constant, since $1/\sqrt{1-0.5} = \sqrt{2} \approx 1.4$. On the other hand, the ℓ^1-results by [11] still only applies for $k = 1$ and then the constant C_k equals 13.96.

How much better does it then get in the asymptotic regime? The best estimates of this we have found is in [6], which gives advanced probabilistic estimates as well as extensive numerical evaluations using sophisticated methods to estimate RIP/LRIP-values. In particular figure 2.3 and 2.4 are enlightening, where it is shown that for $\frac{m}{n} = 0.5$, one needs to have k well below 1% of m to have any hope of achieving $\delta_{2k} = 0.4$, which is what is required in (12). More precisely, following [6] we need $\mathcal{L}(0.5, \frac{2k}{m})$ and $\mathcal{U}(0.5, \frac{2k}{m})$[5] to be below 0.4, which happens around $k/m \approx 1.5 \times 10^{-3}$. It is also clear that RIP-values are consistently higher with a notable difference. Based on this, it seems safe to conclude that for a large amount of settings where ℓ^1-methods are used, there is very limited theoretical evidence for their applicability, at best.

2.5. What's in a theorem?

The so called 'oracle solution', i.e. the one you would get if an oracle told you the true support S of x_0 and you were to solve the (overdetermined) equations system $A_S x = b$ where A_S denotes the $m \times k$ matrix whose columns are those with indices in S (and then expand x to $\mathbb{R}^n$ by inserting zeroes off S). This is clearly the best possible solution one could hope for (as argued also in [14]).

[5] $\mathcal{L}$ and $\mathcal{U}$ are the asymptotic RIP bounds. Informally speaking, and here we quote [6] verbatim, 'for large matrices from the Gaussian ensemble, it is overwhelmingly unlikely that the RIP asymmetric constants $L(k, m, n)$ and $U(k, m, n)$ will be greater than $\mathcal{L}(\delta, \rho)$ and $\mathcal{U}(\delta, \rho)$'. δ and ρ are such that $n/N \to \delta$ and $k/n \to \rho$ as $n \to \infty$. $L(k, m, n)$ is what we called δ_k^- for an $m \times n$ matrix and $U(k, m, n)$ is its natural upper-counterpart.

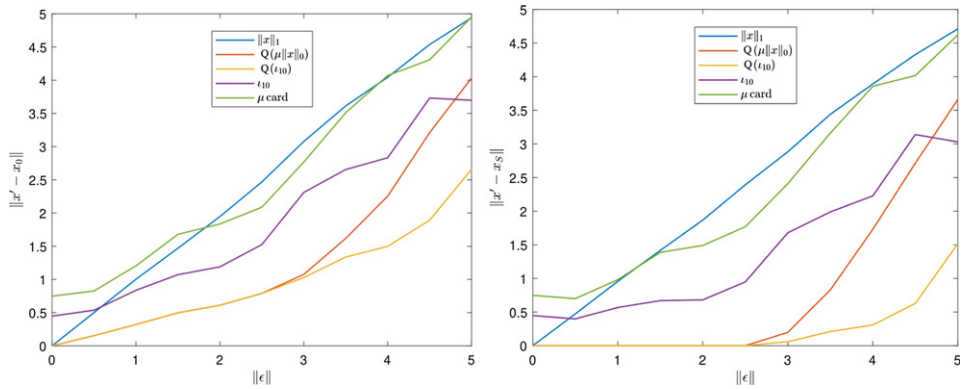

Figure 5. $\|x' - x_0\|_2$ (left) and $\|x' - x_S\|_2$ (right) versus $\|\epsilon\|_2$ for the 5 methods (3), (10), (11) and (15), (16) minimized using with FBS. The methods based on $\mathcal{Q}_2$(card) and $\mathcal{Q}_2(\iota_{P_k})$ work perfectly down to SNR ≈ 4.

Corollaries 2.1 and 2.3 claim that the oracle solution is a unique global minimizer of the respective functional, not necessarily that a given algorithm will find this global minimizer. In our experience, working either with FBS or ADMM, the algorithms do find the oracle solution in very difficult scenarios when one initializes at zero[6], but we do not have a proof for this. We can prove that FBS converges to a stationary point and that the stationary points in corollary 2.3 are not local minima, except for the oracle solution, see section 6.

What is the value of these observations and how do they compare with the existing literature? For example, [8] studies the minimization of (15) itself (which, if we apply FBS, leads to IHT for k-sparsity, denoted IHT$_k$), and it actually guarantees that IHT$_k$ converges to within $5\|\epsilon\|$ of the oracle solution. They do not prove that they have found the oracle solution, but combined with corollary 2.3 it follows that this is indeed the case (for SNR's such that the corollary applies). On first sight this is a much stronger conclusion, since they actually prove that their algorithm avoids unwanted stationary points. However, the method performs much worse in practice, see figure 5. The difference lies in the fact that [8] assumes that $\delta_{3k} < \frac{1}{\sqrt{32}} \approx 0.18$, whereas corollary 2.3 applies as long as $\delta_{2k}^- < 0.5$, which is much more easy to fulfill in practice.

The strength of a result not only in the conclusion, but in how much one needs to assume. For example, there are many papers giving conditions under which minimization of (3) or non-convex alternatives find the true support. If we have a method that would find a vector x' with the correct support S (with a bias or not), we can always get this unbiased solution by simply discarding x' and follow the above procedure to get the oracle solution. Therefore the issue of finding the support is maybe more central than having a good estimate of $\|x' - x_0\|_2$. Conditions under which LASSO finds the correct support are given e.g. in [47] and for a more general class of non-convex penalties in [34]. In both cases however, the theorems involve constants whose size is unknown, and their applicability cannot be verified in a concrete problem instance. To be more concrete, the latter paper does have a result claiming that MCP finds the oracle solution with given probability, but apart from involving conditions that are very difficult to verify, the conclusion contains the statement that MCP has a unique stationary point. This is rarely true

[6] Initializing $\mathcal{K}_{\text{reg}}$ at the least squares solution is not good. As evidenced by our numerical evaluation in section 6 there seems to be many local minima nearby.

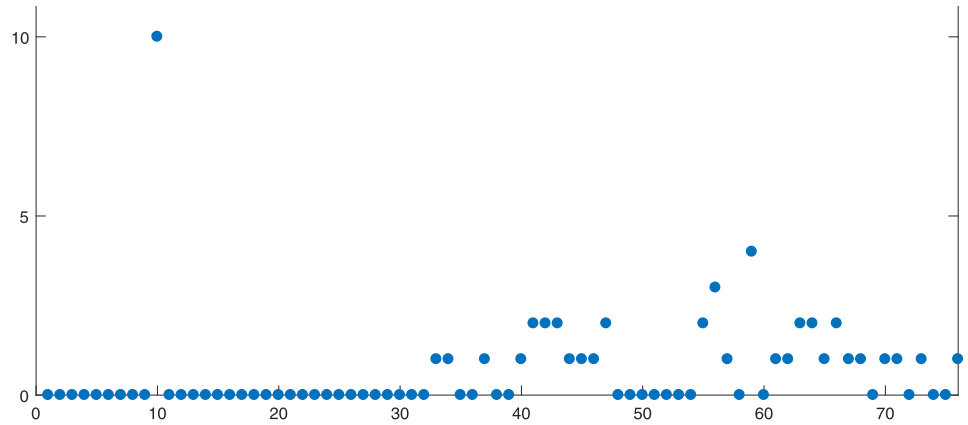

Figure 6. Histogram of cardinality for 50 trials of (11) with $\|\epsilon\|_2 = 2.5$.

(see e.g. [46] or figure 6 below) and hence it follows that these conditions are substantially more strict than ours. Concrete theorems proving that LASSO finds the true support, such as those found in [24] are rather weak, see [23] which studies this topic in depth.

We therefore believe that our framework is more generally applicable even for the already well-studied MCP penalty, and that the relative simplicity and verifiability of the assumptions makes the theory attractive, in particular combined with superior performance numerically, at least in the standard synthetic setting (see section 6). Moreover, the penalty $\mathcal{Q}_2(\iota_{P_k})$ is not separable, and the underlying ideas of this work are based on the quadratic envelope as a regularizer and extends to a whole class of more advanced sparsifying penalties, of which $\mathcal{Q}_2(\iota_{P_k})$ is merely one. We leave further extensions for future work, but remark that the whole machinery developed here can also be lifted to low rank matrix problems, see [18].

3. Uniqueness of sparse stationary points

We now turn to the heart of the matter, namely uniqueness of sparse minimizers of $\mathcal{K}_{\text{reg}}$ as defined in (9), where f can be any function with values in $\mathbb{R} \cup \{\infty\}$. We say that x is a stationary point of a given function g if

$$\liminf_{\substack{y \to x \\ y \neq 0}} \frac{g(x+y) - g(x)}{\|y\|} \geqslant 0. \tag{18}$$

If g is a sum of a convex function g_{c} and a differentiable function g_{d} and we work in $\mathbb{R}^n$, it is not hard to see that x is a stationary point if and only if $-\nabla g_{\text{d}}(x) \in \partial g_{\text{c}}(x)$ where ∂g_{c} denotes the usual subdifferential used in convex analysis, and ∇g_{d} the standard gradient. The same is true in the complex case, i.e. when working over $\mathbb{C}^n$, upon suitable modification of the concept of subdifferential and gradient. For convenience we provide the details in appendix A.1.

Set

$$\mathcal{G}(x) = \frac{1}{2}\mathcal{Q}_2(f)(x) + \frac{1}{2}\|x\|_2^2, \tag{19}$$

i.e. $2\mathcal{G}$ the l.s.c. convex envelope of $f(x) + \|x\|_2^2$. We have

$$\mathcal{K}_{\text{reg}}(x) = 2\mathcal{G}(x) - \|x\|_2^2 + \|Ax - b\|_2^2 \tag{20}$$

which upon differentiation yields that x' is a stationary point of $\mathcal{K}_{\text{reg}}$ if and only if

$$(I - A^*A)x' + A^*b \in \partial\mathcal{G}(x'). \tag{21}$$

since $\nabla\left(\|Ax - b\|_2^2 - \|x\|_2^2\right) = 2A^*(Ax - b) - 2x$, see the appendix for details. Given any x, we therefore associate with it a new point z via

$$z = (I - A^*A)x + A^*b. \tag{22}$$

This point will play a key role in this paper, in fact, it has already appeared in corollary 2.2. Suppose now that x' and x'' are two *sparse* stationary points in the sense that $x'' - x' \in P_N$ for some N less than m.

Proposition 3.1. *Let x' and x'' be distinct stationary points of $\mathcal{K}_{reg}$ such that $x'' - x' \in P_N$. Then*

$$\text{Re}\,\langle z'' - z', x'' - x'\rangle \leqslant \delta_N^-\|x'' - x'\|_2^2. \tag{23}$$

The above proposition will mainly be used backwards, i.e. we will show that (23) does not hold and thereby conclude that $x'' - x' \notin P_N$.

Proof. We have

$$z'' - z' = (I - A^*A)x'' + A^*b - (I - A^*A)x' - A^*b = (I - A^*A)(x'' - x'),$$

so taking the scalar product with $x'' - x'$ gives

$$\text{Re}\,\langle z'' - z', x'' - x'\rangle = \|x'' - x'\|_2^2 - \|A(x'' - x')\|_2^2 \leqslant \delta_N^-\|x'' - x'\|_2^2,$$

as desired. Note that it is not necessary to take the real part, but we leave it since scalar products in general can be complex numbers. □

As we shall see, the point z' has a decisive influence on the coming sections. To begin with, it has the following interesting property.

Proposition 3.2. *A point x' is a stationary point of $\mathcal{K}_{reg}$ if and only if it solves the convex problem*

$$x' \in \operatorname*{argmin}_x \mathcal{Q}_2(f)(x) + \|x - z'\|_2^2.$$

Note the absence of A in the above formula, which in particular implies that $\mathcal{Q}_2(f)(x) + \|x - z'\|_2^2$ is the convex envelope of $f(x) + \|x - z'\|_2^2$.

Proof. As noted in (21), x' is a stationary point of $\mathcal{K}_{\text{reg}}$ if and only if $z' \in \partial\mathcal{G}(x')$. By the same token, x' is a stationary point of

$$\mathcal{Q}_2(f)(x) + \|x - z'\|_2^2 = 2\mathcal{G}(x) - 2\,\text{Re}\,\langle x, z'\rangle + \|z'\|_2^2$$

if and only if $z' \in \partial\mathcal{G}(x')$, and since the functional is convex (and clearly has a well defined minimum) the stationary points coincide with the set of minimizers. □

4. The sparsity problem

We return to the sparsity problem, and consider $f(x) = \mu\,\mathrm{card}(x)$ where μ is a parameter and $\mathrm{card}(x)$ is the number of non-zero entries in the vector x. In this case we have,

$$\mathcal{Q}_2(\mu\ \mathrm{card})(x) = \sum_{j=1}^{n} \mu - \left(\max\{\sqrt{\mu} - |x_j|, 0\}\right)^2. \tag{24}$$

To recapitulate, we want to minimize (10), i.e.

$$\mathcal{K}_\mu(x) = \mu\ \mathrm{card}(x) + \|Ax - b\|_2^2 \tag{25}$$

which we replace by (11), i.e.

$$\mathcal{K}_{\mu,\mathrm{reg}}(x) = \mathcal{Q}_2(\mu\ \mathrm{card})(x) + \|Ax - b\|_2^2. \tag{26}$$

4.1. Equality of minimizers for $\mathcal{K}_\mu$ and $\mathcal{K}_{\mu,reg}$

As noted by Aubert, Blanc-Feraud and Soubies (see theorems 4.5 and 4.8 in [46]), $\mathcal{K}_{\mu,\mathrm{reg}}$ has the same global minima and potentially fewer local minima than $\mathcal{K}_\mu$ if

$$\|A\|_{\infty,\mathrm{col}} = \sup_i \|a_i\|_2 \leqslant 1, \tag{27}$$

where a_i denotes the columns of A. Below we (essentially) reproduce their statement in the terminology of this paper. A proof is included in the appendix for completeness.

Theorem 4.1. *If $\|A\|_{\infty,\mathrm{col}} < 1$, then any local minimizer of $\mathcal{K}_{\mu,reg}$ is a local minimizer of $\mathcal{K}_\mu$, and the (nonempty) set of global minimizers coincide. If merely $\|A\|_{\infty,\mathrm{col}} = 1$, then any global minimizer of $\mathcal{K}_{\mu,reg}$ which is not a global minimizer of $\mathcal{K}_\mu$, belongs to a connected component of global minimizers which includes at least two global minima of $\mathcal{K}_\mu$.*

4.2. On the uniqueness of sparse stationary points

Next we take a closer look at the structure of the stationary points. Given N such that $\delta_N^- < 1$, we will show that under certain assumptions the difference between two stationary points always has at least N elements. Hence if we find a stationary point with less than $N/2$ elements then we can be sure that this is the sparsest one. The main theorem reads as follows:

Theorem 4.2. *Let x' be a stationary point of $\mathcal{K}_{\mu,reg}$, let z' be given by (22), and assume that*

$$|z_i'| \notin \left[(1 - \delta_N^-)\sqrt{\mu}, \frac{\sqrt{\mu}}{1 - \delta_N^-}\right] \tag{28}$$

for all $i \in \{1, \ldots, n\}$. If x'' is another stationary point of $\mathcal{K}_{\mu,reg}$ then

$$\mathrm{card}(x'' - x') > N.$$

Note that we allow $\delta_N^- < 0$ in the above theorem, in which case the condition on z' is automatically satisfied. The proof depends on a sequence of lemmas, and is given at the end of

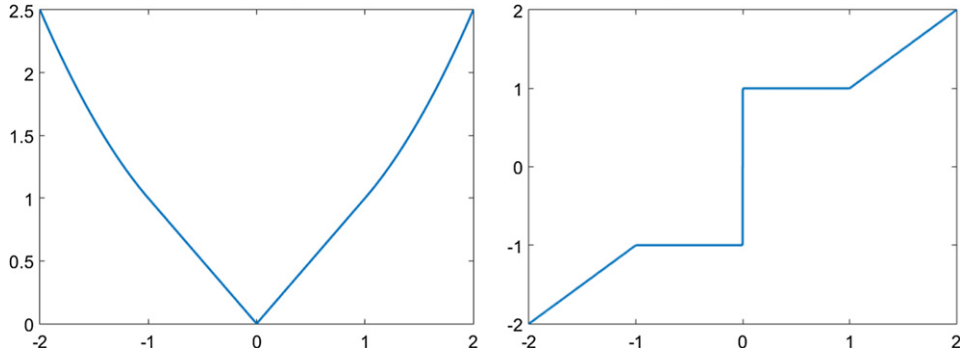

Figure 7. The function $g(x)$ (left) and its sub-differential $\partial g(x)$ (right), for $\mu = 1$. Note that the sub-differential contains a unique element everywhere except at $x = 0$.

the section. Clearly, we will rely on proposition 3.1, which requires an investigation of the functional $\mathcal{G}$ (19) and in particular its sub-differential. Introducing the function g as

$$g(x) = \begin{cases} \dfrac{\mu + |x|^2}{2} & |x| \geqslant \sqrt{\mu} \\ \sqrt{\mu}|x| & 0 \leqslant |x| \leqslant \sqrt{\mu} \end{cases}. \tag{29}$$

we get

$$\mathcal{G}(x) = \sum_{j=1}^{n} g(x_j). \tag{30}$$

Its sub-differential is given by

$$\partial g(x) = \begin{cases} \{x\} & |x| \geqslant \sqrt{\mu} \\ \{\sqrt{\mu}\dfrac{x}{|x|}\} & 0 < |x| \leqslant \sqrt{\mu} \\ \sqrt{\mu}\ \mathbb{D} & x = 0 \end{cases} \tag{31}$$

where $\mathbb{D}$ is the closed unit disc in $\mathbb{C}$ or, if working over $\mathbb{R}$, $\mathbb{D} = [-1, 1]$. In the remainder we suppose for concreteness that we work over $\mathbb{C}$ (but show the real case in pictures). Note that the sub-differential consists of a single point for each $x \neq 0$. Figure 7 illustrates g and its sub-differential.

The following two results establish a bound on the sub-gradients of $\mathcal{G}$. We begin with some one-dimensional estimates of g.

Lemma 4.3. *Assume that $z_0 \in \partial g(x_0)$ and $\delta_N^- > 0$. If*

$$|z_0| > \frac{\sqrt{\mu}}{1 - \delta_N^-} \tag{32}$$

then for any x_1, z_1 with $z_1 \in \partial g(x_1)$ and $x_1 \neq x_0$, we have

$$\mathrm{Re}(z_1 - z_0)\overline{(x_1 - x_0)} > \delta_N^- |x_1 - x_0|^2. \tag{33}$$

Proof. By rotational symmetry (i.e. $\partial g(\mathrm{e}^{\mathrm{i}\phi}x) = \mathrm{e}^{\mathrm{i}\phi}\partial g(x)$), it is no restriction to assume that $z_0 > 0$. By $\frac{1}{1-\delta_N^-} > 1$ and (32), we see that $z_0 > \sqrt{\mu}$, and hence the identity $z_0 \in \partial g(x_0)$ and (31) together imply that $z_0 = x_0$ and in particular that $x_0 \in \mathbb{R}$ and

$$x_0 > \frac{\sqrt{\mu}}{1-\delta_N^-}. \tag{34}$$

To prove the result we now minimize the quotient

$$\frac{\mathsf{Re}(z_1 - z_0)\overline{(x_1 - x_0)}}{|x_1 - x_0|^2} \tag{35}$$

and show that it is larger than δ_N^-. There are three cases to consider; $x_1 = 0$, $0 < |x_1| < \sqrt{\mu}$ and $|x_1| \geqslant \sqrt{\mu}$. The latter case is easy since then $z_1 - z_0 = x_1 - x_0$ and since $\delta_N^- < 1$, the desired conclusion is immediate.

For the two other cases we first show that z_1 and x_1 can be assumed to be real. If $x_1 = 0$ the above quotient is equivalent to $1 - \mathsf{Re}(z_1/x_0)$ over $z_1 \in \sqrt{\mu}\mathbb{D}$, since $x_0 = z_0$ is real and positive, which is clearly minimized for the real value $z_1 = \sqrt{\mu}$.

For the middle case, z_1 and x_1 have the same angle with $\mathbb{R}$. We first hold the radii fixed and only consider the angle as an argument. Recall $z_0 = x_0$ and set $R = |z_1|/|x_1|$. Then we have

$$\begin{aligned}\mathsf{Re}(z_1 - z_0)\overline{(x_1 - x_0)} &= R|x_1|^2 - (R+1)\mathsf{Re}\; x_1\overline{x_0} + |x_0|^2 \\ &= \frac{1}{2}\left((R-1)|x_1|^2 + (R+1)|x_1 - x_0|^2 + (1-R)|x_0|^2\right).\end{aligned}$$

So the quotient (35) only depends on $|x_1 - x_0|^2$ (for fixed radii), which shows that the quotient is minimized when x_1 is real (which then automatically applies to z_1 as well).

Summarizing the above we may thus assume that x_1 and z_1 are real and $x_1 \in [-\sqrt{\mu}, \sqrt{\mu}]$, which simplifies the quotient (35) to $\frac{x_0 - z_1}{x_0 - x_1}$. We now hold x_1, z_1 fixed and consider x_0 as the variable. Recall that $|z_1| \geqslant |x_1|$. If these are negative we immediately get that the quotient is $\geqslant 1 > \delta_N^-$ and the proof is done. In the positive case, the quotient is minimized when x_0 is as small as possible (since $z_1 \geqslant x_1$). By (34) we hence conclude that the minimum of (35) is strictly greater than $\frac{\frac{\sqrt{\mu}}{1-\delta_N^-} - z_1}{\frac{\sqrt{\mu}}{1-\delta_N^-} - x_1}$. The minimum of this is in its turn clearly attained at $x_1 = 0$ and $z_1 = \sqrt{\mu}$. Summing up, we have that

$$\frac{\mathsf{Re}(z_1 - z_0)\overline{(x_1 - x_0)}}{|x_1 - x_0|^2} > \frac{\frac{\sqrt{\mu}}{1-\delta_N^-} - z_1}{\frac{\sqrt{\mu}}{1-\delta_N^-} - x_1} \geqslant \frac{\frac{\sqrt{\mu}}{1-\delta_N^-} - \sqrt{\mu}}{\frac{\sqrt{\mu}}{1-\delta_N^-}} = \delta_N^-.$$

□

Lemma 4.4. *Assume that $z_0 \in \partial g(x_0)$ and $\delta_N^- > 0$. If*

$$|z_0| < (1 - \delta_N^-)\sqrt{\mu} \tag{36}$$

then for any x_1, z_1 with $z_1 \in \partial g(x_1)$, $x_1 \neq x_0$, we have

$$\mathsf{Re}(z_1 - z_0)\overline{(x_1 - x_0)} > \delta_N^-|x_1 - x_0|^2. \tag{37}$$

Proof. The proof is similar to the previous lemma. We first note that $x_0 = 0$, $x_1 \neq 0$ and that z_0 may be assumed to be in $(0, (1-\delta_N^-)\sqrt{\mu})$ by rotational symmetry. For a fixed radius $R = \frac{|z_1|}{|x_1|}$ the quotient

$$\frac{\mathrm{Re}(z_1 - z_0)\overline{(x_1 - x_0)}}{|x_1 - x_0|^2} = \frac{\mathrm{Re}(z_1 - z_0)\overline{x_1}}{|x_1|^2} = R - z_0 \frac{\mathrm{Re}\ x_1}{|x_1|^2}$$

is smallest when x_1 is real valued and positive (which then also applies to z_1 which by (31) equals $\max(x_1, \sqrt{\mu})$). The expression in question then becomes $R - \frac{z_0}{x_1}$, which is minimized by maximizing z_0. For any choice of x_1, (36) implies that our expression is strictly bigger than

$$\frac{z_1 - (1-\delta_N^-)\sqrt{\mu}}{x_1} = \frac{\max(x_1, \sqrt{\mu}) - (1-\delta_N^-)\sqrt{\mu}}{x_1}.$$

Basic calculus shows that the minimum of this quantity is attained at $x_1 = \sqrt{\mu}$ and equals δ_N^-, as desired. □

We are now ready to prove theorem 4.2.

Proof of theorem 4.2. By proposition 3.1 it suffices to verify

$$\mathrm{Re}\langle z'' - z', x'' - x' \rangle > \delta_N^- \|x'' - x'\|_2^2, \quad x'' \neq x'. \tag{38}$$

The claim will follow by contradiction. Suppose first that $\delta_N^- > 0$. Since $\partial \mathcal{G}(x) = \sum_{j=1}^n \partial g(x_j)$, lemmas 4.3 and 4.4 imply that

$$\mathrm{Re}(z_i'' - z_i')\overline{(x_i'' - x_i')} > \delta_N^- |x_i'' - x_i'|^2,$$

for all i with $x_i'' - x_i' \neq 0$. Since $x_i'' - x_i' = 0$ gives $(z_i'' - z_i')\overline{(x_i'' - x_i')} = 0$ summing over i gives the result.

Suppose now that $\delta_N^- < 0$. By (38) it suffices to prove that $\mathrm{Re}\langle z'' - z', x'' - x' \rangle \geqslant 0$ for all $x'' \neq x'$. Fix i in $\{1, \ldots, n\}$. By rotational symmetry it is easy to see that we can assume that $x_i', z_i' \geqslant 0$. Moreover, for fixed values of $|z_i''|$ and $|x_i''|$ (but variable complex phase) it is easy to see that $\mathrm{Re}(z_i'' - z_i')(x_i'' - x_i')$ achieves min when these are also real, i.e. we can assume that $x_i'', z_i'' \in \mathbb{R}$. Since the graph of ∂g is non-decreasing it follows that $(z_i'' - z_i')(x_i'' - x_i') \geqslant 0$ for all i, as desired.

It remains to consider the case when $\delta_N^- = 0$, and as above we reach a contradiction if we prove that $\mathrm{Re}\langle z'' - z', x'' - x' \rangle > 0$. Again we can assume that $x_i', z_i' \geqslant 0$ and that $x_i'', z_i'' \in \mathbb{R}$. Then (28) implies that $z_i' \neq \sqrt{\mu}$ for all $1 \leqslant i \leqslant n$, which via $z_i' \in \partial g(x_i')$ also implies that $x_i' \notin (0, \sqrt{\mu}]$. If $x'' \neq x'$ we must have $x_i'' \neq x_i'$ for some i. Using that $z_i'' \in \partial g(x_i'')$, examination of (31) yields that also $z_i'' \neq z_i'$. With this at hand we see that the left-hand side of (38) is strictly positive, whereas the right equals 0, which again is a contradiction. □

4.3. Conditions on global minimality

Theorem 4.5. *Let A satisfy $\|A\|_{\infty,\mathrm{col}} \leqslant 1$, let x' be a stationary point of $\mathcal{K}_{\mu,reg}$ and let z' be given by (22). Assume that*

$$|z_i'| \notin \left[(1-\delta_N^-)\sqrt{\mu}, \frac{1}{1-\delta_N^-}\sqrt{\mu}\right], \quad 1 \leqslant i \leqslant n. \tag{39}$$

If

$$2\ \mu\ card(x') + \|Ax' - b\|_2^2 < \mu N + \mu, \tag{40}$$

then x' is the unique global minimum of $\mathcal{K}_\mu$ and $\mathcal{K}_{\mu,reg}$.

Obviously, it is desirable to pick N as large as possible, which is limited by (39) and the fact that δ_N^- increases with N. Also note that $\delta_N^- \geqslant 0$ since $1 - \delta_1^- \leqslant \min_i\{\|a_i\|_2^2\} = \|A\|_{\infty,\mathrm{col}}^2 \leqslant 1$ so $\delta_1^- \geqslant 0$.

Proof. Set $k = \mathrm{card}(x')$ and assume that x' is not the unique global minimizer of $\mathcal{K}_{\mu,\mathrm{reg}}$. Let x'' be another. Either $\mathcal{K}_{\mu,\mathrm{reg}}(x) = \mathcal{K}_\mu(x)$ for $x = x''$ or x'' is part of a connected component of global minimizers to $\mathcal{K}_{\mu,\mathrm{reg}}$ including two points that satisfy the equation, by theorem 4.1. Since one of these must be different from x', we may assume that $\mathcal{K}_{\mu,\mathrm{reg}}(x'') = \mathcal{K}_\mu(x'')$. Theorem 4.2 then shows that $\mathrm{card}(x'') \geqslant N - k + 1$. Since $\mathcal{K}_\mu(x'') = \mathcal{K}_{\mu,\mathrm{reg}}(x'')$ and $\mathcal{K}_\mu(x') \geqslant \mathcal{K}_{\mu,\mathrm{reg}}(x')$ it follows from (40) that

$$\begin{aligned}\mathcal{K}_{\mu,\mathrm{reg}}(x'') - \mathcal{K}_{\mu,\mathrm{reg}}(x') &\geqslant \mathcal{K}_\mu(x'') - \mathcal{K}_\mu(x') \\ &\geqslant \mu(N - k + 1) - (\mu k + \|Ax' - b\|_2^2) > 0.\end{aligned}$$

This is a contradiction, and hence x' must be the unique global minimizer of $\mathcal{K}_{\mu,\mathrm{reg}}$. By theorem 4.1 it then follows that x' is also unique minimizer of $\mathcal{K}_\mu$. □

4.4. Finding the oracle solution

In this final subsection we return to the compressed sensing problem of retrieving a sparse vector x_0 given corrupted measurements $b = Ax_0 + \epsilon$, where ϵ is noise and x_0 is sparse. More precisely we set $S = \mathrm{supp}\, x_0$ where we assume that $\#S = k$ is much smaller than m—the amount of rows in A (i.e. number of measurements). Here $\#S$ denotes the amount of elements in S and the noise can be of any type, our theory only relies on knowledge of $\|\epsilon\|$.

We let $x_{0,j}$ denote the elements of the vector x_0. Let A_S denote the matrix obtained from A by setting columns outside of S to 0, and let x_{or} denote the least squares solution to $A_S x_{\mathrm{or}} = b$. Note that this is the so called 'oracle solution' discussed in the introduction, which can also be written $x_{\mathrm{or}} = (A_S^* A_S)^\dagger A_S^* b$ where $(A_S^* A_S)^\dagger$ denotes the Moore–Penrose inverse.

Our first result collects some general observations about the oracle solution.

Proposition 4.6. *Let A satisfy $\|A\|_{\infty,\mathrm{col}} \leqslant 1$ and let $c > 0$. If*

$$|x_{0,j}| > c + \frac{\|\epsilon\|_2}{\sqrt{1 - \delta_k^-}}$$

for all $j \in S$ then the oracle solution $x' = x_{\mathrm{or}}$ satisfies $\mathrm{supp}\,(x') = \mathrm{supp}\,(x_0)$. We also have $|x'_j| > c, j \in S$, $\|Ax' - b\|_2 \leqslant \|\epsilon\|_2$, and

$$\|x' - x_0\|_2 \leqslant \frac{\|\epsilon\|_2}{\sqrt{1 - \delta_k^-}}.$$

Proof. Consider the equation $A_S x = Ax_0 + \epsilon$ and note that $Ax_0 = A_S x_0$. The least squares solution is obtained by applying $(A_S^* A_S)^\dagger A_S^*$ which gives the solution

$$x' = x_0 + (A_S^* A_S)^\dagger A_S^* \epsilon = x_0 + \eta,$$

where we set $(A_S^* A_S)^\dagger A_S^* \epsilon = \eta$. By construction of the Moore–Penrose inverse, $\mathrm{supp}\,\eta \subset S$, and hence

$$A\eta = A_S \eta = P_{\mathrm{Ran} A_S} \epsilon,$$

where $P_{\mathrm{Ran}A_S}$ denotes the orthogonal projection onto the range of A_S. In particular,

$$\|\eta\|_2 \leqslant \frac{\|A_S\eta\|_2}{\sqrt{1-\delta_k^-}} = \frac{\|P_{\mathrm{Ran}A_S}\epsilon\|_2}{\sqrt{1-\delta_k^-}} \leqslant \frac{\|\epsilon\|_2}{\sqrt{1-\delta_k^-}},$$

which establishes the final inequality in the proposition. Also $\|\eta\|_\infty \leqslant \|\eta\|_2$ which implies

$$|x'_j| \geqslant |x_{0,j}| - |\eta_j| > c + \frac{\|\epsilon\|_2}{\sqrt{1-\delta_k^-}} - \frac{\|\epsilon\|_2}{\sqrt{1-\delta_k^-}} = c, \quad j \in S. \tag{41}$$

This also gives $\operatorname{supp} x' = \operatorname{supp} x_0$ since by construction we clearly have

$$\operatorname{supp} x' \subset \operatorname{supp} x_0 \cup \operatorname{supp} \eta \subset S.$$

Finally, consider $Ax' - b$, which equals

$$\begin{aligned} Ax' - b &= A_S x' - b = A_S x_0 + A_S(A_S^*A_S)^\dagger A_S^*\epsilon - (A_S x_0 + \epsilon) \\ &= (P_{\mathrm{Ran}A_S} - I)\epsilon = -P_{(\mathrm{Ran}A_S)^\perp}\epsilon \end{aligned} \tag{42}$$

and hence $\|Ax' - b\|_2 \leqslant \|\epsilon\|_2$. □

The below proposition shows that the oracle solution is under mild assumptions a local minimizer of $\mathcal{K}_{\mu,\mathrm{reg}}$, which we denote by x' for notational consistency.

Proposition 4.7. *Let A satisfy $\|A\|_{\infty,\mathrm{col}} \leqslant 1$. If $\|\epsilon\|_2 < \sqrt{\mu}$ and*

$$|x_{0,j}| > \sqrt{\mu} + \frac{\|\epsilon\|_2}{\sqrt{1-\delta_k^-}}$$

for all $j \in S$ then the oracle solution $x' = x_{\mathrm{or}}$ is a strict local minimum to $\mathcal{K}_{\mu,reg}$ with $\operatorname{supp}(x') = \operatorname{supp}(x_0)$. We also have $|x'_j| > \sqrt{\mu}$, $j \in S$, $\|Ax' - b\|_2 \leqslant \|\epsilon\|_2$, and

$$\|x' - x_0\|_2 \leqslant \frac{\|\epsilon\|_2}{\sqrt{1-\delta_k^-}}.$$

Proof. All inequalities follow by applying proposition 4.6 with $c = \sqrt{\mu}$, so it remains to prove that x' is a local minimum of $\mathcal{K}_{\mu,\mathrm{reg}} = \mathcal{Q}_2(\mu\ \mathrm{card}) + \|Ax - b\|_2^2$. To this end, consider $\mathcal{K}_{\mu,\mathrm{reg}}(x' + v)$. Since $|x'_j| > \sqrt{\mu}$ for $j \in S$, the term $\mathcal{Q}_2(\mu\ \mathrm{card})$ (see (24)) is constant for the corresponding indices of v, as long as v is small. For v in a neighborhood of 0 we get

$$\begin{aligned} \mathcal{K}_\mu(x' + v) = &\sum_{j \in S^c} \left(2\sqrt{\mu}|v_j| - |v_j|^2\right) + 2\ \mathrm{Re}\,\langle v, A^*(Ax' - b)\rangle \\ &+ \|Av\|_2^2 + \mathcal{K}_{\mu,\mathrm{reg}}(x'). \end{aligned}$$

Since x' solves the least squares problem posed initially, the vector $A_S^*(Ax' - b) = A_S^*(A_S x' - b)$ must be 0. With this in mind the above expression simplifies to

$$2\left(\sum_{j \in S^c} \sqrt{\mu}|v_j| + \mathrm{Re}\left(v_j \langle a_j, Ax' - b\rangle\right)\right) - \sum_{j \in S^c} |v_j|^2 + \|Av\|_2^2 + \mathcal{K}_{\mu,\mathrm{reg}}(x'). \tag{43}$$

By the Cauchy–Schwartz inequality and (42) we have

$$|\langle a_j, Ax' - b\rangle| \leqslant \|a_j\|_2\|\epsilon\|_2 < \|A\|_{\infty,\mathrm{col}}\sqrt{\mu} \leqslant \sqrt{\mu}.$$

It follows that the term $\sum_{j\in S^c}\sqrt{\mu}|v_j| + \mathsf{Re}\left(v_j\langle a_j, Ax' - b\rangle\right)$ in (43) can be estimated from below by

$$\sum_{j\in S^c}|v_j|(\underbrace{\sqrt{\mu} - |\langle a_j, Ax' - b\rangle|}_{:=\alpha_j}) \geqslant \alpha\sum_{j\in S^c}|v_j|$$

where $\alpha = \min_j\{\alpha_j\} > 0$ for all j. Hence

$$2\left(\sum_{j\in S^c}\sqrt{\mu}|v_j| + \mathsf{Re}\left(v_j\langle a_j, Ax' - b\rangle\right)\right) - \sum_{j\in S^c}|v_j|^2 > 0 \tag{44}$$

for v in a neighborhood of 0, as long as $\sum_{j\in S^c}|v_j|^2 \neq 0$. To have $\mathcal{K}_{\mu,\mathrm{reg}}(x' + v) \leqslant \mathcal{K}_{\mu,\mathrm{reg}}(x')$, (43) shows that we need the terms in (44) to be zero, or equivalently $\mathrm{supp}\, v \subset S$. But then (43) reduces to $\|Av\|_2^2 + \mathcal{K}_{\mu,\mathrm{reg}}(x')$, and since $\delta_k^- < 1$ it follows that $\|Av\|_2^2 > 0$ unless $v = 0$. In other words, x' is a strict local minimizer. □

In the above proposition, there is nothing said as to whether x' is a global minimum or not. To get further, let z' correspond to x' via (22). We need conditions such that (39) holds for z', i.e.

$$|z_i'| \notin \left[(1 - \delta_N^-)\sqrt{\mu}, \frac{\sqrt{\mu}}{1 - \delta_N^-}\right]. \tag{45}$$

We remind the reader that N is a number which preferably is a bit larger than $2k$, where k is the cardinality of x_0.

Proposition 4.8. *Let A satisfy $\|A\|_{\infty,\mathrm{col}} \leqslant 1$. If $\|\epsilon\|_2 < (1 - \delta_N^-)\sqrt{\mu}$ and*

$$|x_{0,j}| > \frac{\sqrt{\mu}}{1 - \delta_N^-} + \frac{(1 - \delta_N^-)\sqrt{\mu}}{\sqrt{1 - \delta_k^-}}, \quad j \in S, \tag{46}$$

then (45) holds.

Proof. Using (42) we get

$$z' = (I - A^*A)x' + A^*b = x' - A^*(Ax' - b) = x' + A^*P_{(\mathrm{Ran}A_S)^\perp}\epsilon. \tag{47}$$

Since $A^*P_{(\mathrm{Ran}A_S)^\perp}$ is 0 on rows with index $j \in S$ (being a scalar product of a vector in $\mathrm{Ran}A_S$ and another in its orthogonal complement), we see that $z_j' = x_j'$ for such j. Combining this with the final estimate of proposition 4.6, we see that

$$|z_j'| \geqslant |x_{0,j}| - |x_{0,j} - x_j'| > \frac{1}{1 - \delta_N^-}\sqrt{\mu}, \quad j \in S$$

holds as a consequence of (46). For the remaining z_j', (i.e. $j \in S^c$), we have $x_j' = 0$ so (47) implies

$$|z'_j| = |(A^* P_{(\mathrm{Ran} A_S)^\perp} \epsilon)_j| = \left| \left\langle P_{(\mathrm{Ran} A_S)^\perp} \epsilon, a_j \right\rangle \right|$$
$$\leqslant \|A\|_{\infty,\mathrm{col}} \|\epsilon\|_2 \leqslant \|\epsilon\|_2 < (1 - \delta_N^-)\sqrt{\mu}, \tag{48}$$

which establishes (45). □

Putting all the results together and combining with simple estimates, we finally get

Theorem 4.9. *Suppose that $b = Ax_0 + \epsilon$ where A is an $m \times n$-matrix with $\|A\|_{\infty,\mathrm{col}} \leqslant 1$ and set $\mathrm{card}(x_0) = k$. Let $N \geqslant 2k$ and assume that $\|\epsilon\|_2 < (1 - \delta_N^-)\sqrt{\mu}$ and*

$$|x_{0,j}| > \left(\frac{1}{1 - \delta_N^-} + 1 \right) \sqrt{\mu}, \quad j \in \mathit{supp}\, x_0.$$

Then the oracle solution $x' = x_{\mathrm{or}}$ is a unique global minimum to $\mathcal{K}_{\mu,reg}$ as well as $\mathcal{K}_\mu$, with the property that $\mathrm{supp}\, x' = \mathrm{supp}\, x_0$, that

$$\|x' - x_0\|_2 \leqslant \frac{\|\epsilon\|_2}{\sqrt{1 - \delta_k^-}},$$

and that $\mathrm{card}(x'') > N - k$ for any other stationary point x'' of $\mathcal{K}_{\mu,reg}$.

Proof. All the statements follow by theorems 4.2, 4.5 and proposition 4.7, so we just need to check that these apply. Note that $\sqrt{1 - \delta_N^-} \leqslant \sqrt{1 - \delta_k^-} \leqslant \|A\|_{\infty,\mathrm{col}} \leqslant 1$ which will be used repeatedly.

We begin to verify that proposition 4.7 applies, which is easy by noting that $\|\epsilon\|_2 \leqslant (1 - \delta_N^-)\sqrt{\mu} < \sqrt{\mu}$ and

$$\sqrt{\mu} + \frac{\|\epsilon\|_2}{\sqrt{1 - \delta_k^-}} \leqslant \frac{\sqrt{\mu}}{1 - \delta_N^-} + \frac{(1 - \delta_N^-)\sqrt{\mu}}{\sqrt{1 - \delta_k^-}} \leqslant \frac{\sqrt{\mu}}{1 - \delta_N^-} + \sqrt{\mu} < |x_{0,j}|.$$

Now, to verify that theorem 4.2 applies we need to check the condition (45), which follows if we show that proposition 4.8 applies. This is almost immediate since the estimate on $\|\epsilon\|_2$ is satisfied by assumption and (46) follows by noting that $\frac{1 - \delta_N^-}{\sqrt{1 - \delta_k^-}} \leqslant 1$. By this we also get the first condition of theorem 4.5 for free. We are done once we also verify (40). To this end, note that $\|Ax' - b\|_2 \leqslant \|\epsilon\|_2 < (1 - \delta_N^-)\sqrt{\mu}$ by proposition 4.7, so (40) holds if $2\mu k + (1 - \delta_N^-)^2 \mu \leqslant \mu N + \mu$, which is clearly the case since $N \geqslant 2k$. □

As a final remark, a simpler statement is found by setting $N = 2k$, which gives the loosest conditions to verify. We spelled this out in corollary 2.1, where we also simplified further by replacing $\frac{1}{1 - \delta_N^-} + 1$ by $\frac{2}{1 - \delta_N^-}$, for aesthetic reasons.

5. Known model order; the *k*-sparsity problem

Let $P_k = \{x : \mathrm{card}(x) \leqslant k\}$ where x is a vector in $\mathbb{C}^n$ or $\mathbb{R}^n$. Set $f(x) = \iota_{P_k}(x)$ and note that the problem

$$\operatorname*{argmin}_{\mathrm{card}(x) \leqslant k} \|Ax - b\|_2 \tag{49}$$

is equivalent to finding the minimum of

$$\mathcal{K}_k(x) = \iota_{P_k}(x) + \|Ax - b\|_2^2, \tag{50}$$

(where we put a subindex k to distinguish from $\mathcal{K}_\mu$ in the previous section)[7]. Again, we will approach this problem by using

$$\mathcal{K}_{k,\mathrm{reg}}(x) = \mathcal{Q}_2(\iota_{P_k})(x) + \|Ax - b\|_2^2.$$

This is in some ways much simpler than the situation in the previous sections, for example all local minimizers of $\mathcal{K}_k$ are clearly in P_k. On the other hand, $\mathcal{Q}_2(\iota_{P_k})$ turns out to be rather complicated. We recapitulate the essentials, which follows by adapting the computations in [3] (for matrices) to the vector setting. Define $\tilde{x}$ to be the vector x resorted so that $(|\tilde{x}_j|)_{j=1}^d$ is a decreasing sequence. Then

$$\mathcal{Q}_2(\iota_{P_k})(x) = \frac{1}{k_*}\left(\sum_{j>k-k_*} |\tilde{x}_j|\right)^2 - \sum_{j>k-k_*} |\tilde{x}_j|^2 \tag{51}$$

where k_* is the largest value of $l \in \{1, \ldots, k\}$ for which the non-increasing sequence

$$s(l) = \left(\sum_{j>k-l} |\tilde{x}_j|\right) - l|\tilde{x}_{k+1-l}| \tag{52}$$

is non-negative (note that it clearly is non-negative for $l = 1$). For any given vector x this is clearly computable, although one has to go through a number of cases, but the good thing is that there is an efficient way to implement the corresponding proximal operator (discussed in section 6.2) so in practice this is of little importance. Although it is not very clear from the above expression, $\mathcal{Q}_2(\iota_{P_k})$ is known to be continuous (see e.g. proposition 3.2 in [17]), and this will be used without comment below. We first show that the global minima of $\mathcal{K}_{k,\mathrm{reg}}$ and $\mathcal{K}_k$ coincide.

5.1. Equality of minimizers for $\mathcal{K}_k$ and $\mathcal{K}_{k,reg}$

As before A is a matrix of size $m \times n$, which we need to impose some additional conditions on. The theory in the entire section 5 assumes that

(a) $n \geqslant m + k + 2$ (when working over the reals) whereas $n \geqslant 2m + k + 2$ when working in $\mathbb{C}^n$.

(b) Either $\|A\|_{\infty,\mathrm{col}} < 1$ or $\|A\|_{\infty,\mathrm{col}} \leqslant 1$ and all possible scalar products $\langle a_i, a_j \rangle$ are non-zero.

The equivalent of theorem 4.1 now reads.

Theorem 5.1. *Under assumption (a) and (b) all local minimzers of $\mathcal{K}_{k,reg}$ lie in P_k (and hence are minimizers to $\mathcal{K}_k$). In particular the global minimizers exist and coincide.*

We note that the conclusion is the same as that of theorem 5.1 in [17], which holds for almost any penalty f. However, that proof assumes that $\|A\| < 1$ which is unnecessarily strong in the present setting. For example it would rule out all Gaussian random matrices with normalized columns. The proof of theorem 5.1 is given in appendix A.3.

[7] Admittedly, the notation is not perfect since if k and μ equal the same integer, then the two symbols become the same, but we hope the reader can live with this.

5.2. On the uniqueness of sparse stationary points

We now give a condition, similar to (28) in section 4.2, to ensure that a sparse stationary point is unique, in the sense that other stationary points must have higher cardinality.

Theorem 5.2. *Let x' be a stationary point of $\mathcal{K}_{k,reg}$ with cardinality k, let z' be given by* (22)*, and assume that*

$$|\tilde{z}'_{k+1}| < (1 - 2\delta^-_{2k})|\tilde{z}'_k|. \tag{53}$$

If x'' is another stationary point of $\mathcal{K}_{k,reg}$ then $\operatorname{card}(x'') > k$.

Again, we allow $\delta^-_{2k} < 0$ in the above theorem, in which case the condition on z is automatically satisfied. We begin with a lemma. Recall $\mathcal{G}$ given by (19), i.e. $\frac{1}{2}\mathcal{Q}_2(\iota_{P_k})(x) + \frac{1}{2}\|x\|_2^2$ in the present case. We need an expression for $\partial\mathcal{G}(x)$ for $x \in P_k$.

Lemma 5.3. *If $x \in P_k$ then $z \in \partial\mathcal{G}(x)$ if and only if $z_j = x_j$ for $j \in \operatorname{supp} x$ and $z_j \in |\tilde{x}_k|\mathbb{D}$ for all other j.*

Proof. Since $\mathcal{Q}_2(\iota_{P_k}) + \|x\|_2^2$ is the l.s.c. convex envelope of $\iota_{P_k} + \|x\|_2^2$, we have that $\mathcal{G}(x) = \frac{1}{2}\mathcal{Q}_2(\iota_{P_k}) + \frac{1}{2}\|x\|_2^2$ is the double Fenchel conjugate of $\frac{1}{2}\iota_{P_k} + \frac{1}{2}\|x\|_2^2$. The Fenchel conjugate of the latter is easily computed to

$$\mathcal{G}^*(y) = \frac{1}{2}\sum_{j=1}^{k} |\tilde{y}_j|^2.$$

By the well-known identity $z \in \partial\mathcal{G}(x) \Leftrightarrow x \in \partial\mathcal{G}^*(z)$ (see e.g. proposition 16.9 in [5]) we have $z \in \partial\mathcal{G}(x)$ if and only if

$$\mathcal{G}^*(w) \geqslant \mathcal{G}^*(z) + \langle x, w - z\rangle,$$

for all w which means that

$$z = \operatorname*{argmax}_z \operatorname{Re}\langle x, z\rangle - \frac{1}{2}\sum_{j=1}^{k} |\tilde{z}_j|^2. \tag{54}$$

By standard results on reordering of sequences (see e.g. chapter 1 in [45]), the maximum is attained for a z which is ordered in the same way as x. In other words we can choose a permutation π such that $|x(\pi(j))| = |\tilde{x}_j|$ and $|z(\pi(j))| = |\tilde{z}_j|$ holds for all j. This in turn implies that $\langle x, z\rangle = \sum_{j=1}^{n} x(\pi(j))\overline{z(\pi(j))}$. Combined with $x(\pi(j)) = 0$ for $j > k$, we see that (54) turns into

$$z = \frac{1}{2}\operatorname*{argmax}_z -\sum_{j=1}^{k} |x_j(\pi(j)) - z_j(\pi(j))|^2. \tag{55}$$

The lemma now easily follows. □

Proof of theorem 5.2. If $\operatorname{card}(x'') \leqslant k$ we clearly have $x'' - x' \in P_{2k}$ and both z' and z'' have the structure stipulated in lemma 5.3. Let $I' = \operatorname{supp} x'$ and $I'' = \operatorname{supp} x''$. Then

$\mathrm{Re}\,\langle z'' - z', x'' - x'\rangle$ can be written

$$\mathrm{Re}\left(\sum_{\substack{i\in I'\\ i\in I''}} |x_i'' - x_i'|^2 + \sum_{\substack{i\in I'\\ i\notin I''}} (x_i' - z_i'')\overline{x_i'} + \sum_{\substack{i\notin I'\\ i\in I''}} (x_i'' - z_i')\overline{x_i''}\right). \tag{56}$$

As before we want to reach a contradiction to proposition 3.1, i.e. we want to prove $\mathrm{Re}\,\langle z'' - z', x'' - x'\rangle > \delta_{2k}^-\|x'' - x'\|_2^2$. Note that

$$\|x'' - x'\|_2^2 = \sum_{\substack{i\in I'\\ i\in I''}} |x_i'' - x_i'|^2 + \sum_{\substack{i\in I'\\ i\notin I''}} |x_i'|^2 + \sum_{\substack{i\notin I'\\ i\in I''}} |x_i''|^2, \tag{57}$$

that the first term in (56) and (57) are the same, and that $\delta_{2k}^- < 1$. Since the second and third sums have the same number of terms it suffices to show that

$$\mathrm{Re}(x_i' - z_i'')\overline{x_i'} + (x_j'' - z_j')\overline{x_j''} > \delta_{2k}^-(|x_i'|^2 + |x_j''|^2), \tag{58}$$

for any pair $i \in I'$, $i \notin I''$ and $j \notin I'$, $j \in I''$. This in turn will follow upon showing that

$$z_i''\overline{x_i'} + z_j'\overline{x_j''} \leqslant |z_i''||x_i'| + |z_j'||x_j''| < (1 - \delta_{2k}^-)(|x_i'|^2 + |x_j''|^2).$$

Since $i \notin I''$ and $j \in I''$ we have $|z_i''| \leqslant |z_j''|$ by lemma 5.3, as well as that $z_j'' = x_j''$. Turning to z_j' we can say more due to assumption (53). More precisely, since $i \in I'$ and $j \notin I'$ we have $|z_j'| < (1 - 2\delta_{2k}^-)|z_i'| = (1 - 2\delta_{2k}^-)|x_i'|$, where again lemma 5.3 was used in the last identity. Summing up we have

$$\begin{aligned}|z_i''||x_i'| + |z_j'||x_j''| < |x_j''||x_i'| + (1 - 2\delta_{2k}^-)|x_i'||x_j''| &= 2(1 - \delta_{2k}^-)|x_i'||x_j''|\\ &\leqslant (1 - \delta_{2k}^-)(|x_i'|^2 + |x_j''|^2),\end{aligned}$$

as desired. □

5.3. Conditions on global minimality

The statements in this section are actually quite a bit stronger than the corresponding ones in section 4.3. On the other hand, the condition (53) entails that we must have $\delta_{2k}^- < 1/2$, which limits the applicability.

Theorem 5.4. *Let A satisfy (a) and (b) and let $x' \in P_k$ be a stationary point of $\mathcal{K}_{k,reg}$. Let z' be given by (22) and assume that (53) applies. Then x' is a unique global minimizer of $\mathcal{K}_k$ and $\mathcal{K}_{k,reg}$, and $\mathcal{K}_{k,reg}$ has no other local minimizers either.*

Proof. By theorem 5.1 there exists $x'' \in P_k$ which is a global minimizer for both $\mathcal{K}_k$ and $\mathcal{K}_{k,\mathrm{reg}}$. Clearly x'' is then a stationary point, so if $x' \neq x''$ this would contradict theorem 5.2, so we must have $x' = x''$. The same argument works for the local minimizers. □

5.4. Finding the oracle solution

We now assume that b is of the form $Ax_0 + \epsilon$ where ϵ is noise and x_0 is sparse. More precisely we set $S = \operatorname{supp} x_0$ where we assume that $\#S = k$. As before let A_S denote the matrix obtained from A by setting columns outside of S to 0.

In this case, theorem 5.1 is strong enough so that we do not need any longer argument to establish that x_{or} is a global minimizer, since all local minimizers of $\mathcal{K}_{k,\text{reg}}$ are to be found in P_k. We obtain the following result.

Theorem 5.5. *Let A satisfy (a) and (b). If $\epsilon \neq 0$ and*

$$\min_{j\in S}|x_{0,j}| > \left(\frac{\|\epsilon\|_2}{\sqrt{1-\delta_k^-}} + \frac{2\|\epsilon\|_2}{\sqrt{1-\delta_{2k}^-}}\right),$$

then the estimates of proposition 4.6 applies and the oracle solution is a global minimum of $\mathcal{K}_k$ and $\mathcal{K}_{k,reg}$.

Proof. Proposition 4.6 immediately applies with $c = \frac{2\|\epsilon\|_2}{\sqrt{1-\delta_k^-}}$. Let $J \subset \{1,\dots,n\}$ have cardinality k and consider the problem

$$x_J = \operatorname*{argmin}_x \|A_J x - b\|^2.$$

Searching over J gives rise to (at most) $\binom{n}{k}$ points (since $\delta_k^- < 1$), among which the minimizers of $\mathcal{K}_k$ are found (see lemma 8.1 for more details). By theorem 5.1 a subset of these are the local minimizers of $\mathcal{K}_{k,\text{reg}}$, and the global minimizer must be the one that gives the lowest value for $\|A_J x - b\|^2$. With this notation we have $x_{\text{or}} = x_S$ and the estimates of proposition 4.6 gives $\|Ax_S - b\|_2 \leqslant \|\epsilon\|_2$ and

$$|x_{S,j}| > \frac{2\|\epsilon\|_2}{\sqrt{1-\delta_k^-}}, \quad j \in S.$$

If $J \neq S$ is another set with cardinality k then x_J and x_S must differ in at least one coordinate, so

$$\|x_S - x_J\|_2 > \frac{2\|\epsilon\|_2}{\sqrt{1-\delta_{2k}^-}}, \quad j \in S.$$

But then

$$\begin{aligned}\|Ax_J - b\|_2 &= \|A(x_J - x_S) + Ax_S - b\|_2 \\ &\geqslant \sqrt{1-\delta_{2k}^-}\|x_S - x_J\|_2 - \|\epsilon\|_2 > \|\epsilon\|_2.\end{aligned}$$

Thus x_S is the one with the lowest value for $\|Ax_J - b\|_2$, which was to be shown. □

When minimizing $\mathcal{K}_{k,\mathrm{reg}}$ in practice, it would of course be good to know if there are local minima where one can get stuck. To rule out this possibility, we need unfortunately to assume that $\delta^-_{2k} < 1/2$.

Corollary 5.6. *If in addition to what is assumed in theorem 5.5 we have*

$$\min_{j\in S}|x_{0,j}| > \left(\frac{1}{1-2\delta^-_{2k}} + \frac{1}{\sqrt{1-\delta^-_k}}\right)\|\epsilon\|_2,$$

then the there are no local minimizers of $\mathcal{K}_{k,reg}$ except the oracle solution.

Proof. Theorem 5.5 clearly ensures that $x' = x_{\mathrm{or}}$ is a stationary point. The desired result follows from theorem 5.4 once we verify that (53) applies for z' given by (22). We need to check that $|\tilde{z}'_{k+1}| < (1-2\delta^-_{2k})|\tilde{z}'_k|$. Note that $|\tilde{z}'_{k+1}| \leqslant \|\epsilon\|_2$ by the same estimate as (48). Moreover, since $z' \in \partial\mathcal{G}(x')$, lemma 5.3 implies that $|\tilde{z}'_k| = |\tilde{x}'_k|$ so it suffices to show that $\|\epsilon\|_2 < (1-2\delta^-_{2k})|\tilde{x}'_k|$. This in turn holds by applying proposition 4.6 with $c = \frac{\|\epsilon\|^2}{1-2\delta^-_{2k}}$, and the proof is complete. □

6. Experimental evaluation

In this section we present an experiment designed to validate our main theoretical results. Our goal is to verify that both the proposed methods are able to recover the oracle solution when the signal to noise level is sufficiently large. For both our formulation we need to specify a parameter; μ in case of $\mathcal{K}_{\mu,\mathrm{reg}}$ and k for $\mathcal{K}_{k,\mathrm{reg}}$. Since we are working with synthetic data generated by $b = Ax_0 + \epsilon$, with a known vector x_0 we can set μ so that the non-zero elements are large enough to be preserved, and k so that $k = \text{ supp }(x_0)$ (see section 6.1 for a more detailed description).

In realistic settings where x_0 is unknown selecting parameters is more difficult and requires a precise definition of what constitutes a good solution. This could be based on application specific prior information about the support or size of the elements. If the size of the correct support is assumed to be known, then $\mathcal{K}_{k,\mathrm{reg}}$ is the convenient choice. On the other hand formulations able to directly specify the sought cardinality are uncommon. Therefore soft penalties such as $\lambda\|\cdot\|_1$ or $\mathcal{Q}_2(\mu \text{ card})$ are often utilized by searching over the parameter until a suitable cardinality solution is found. The ℓ_1-norm has been used in this way for a number of practical applications e.g. face recognition [51], subspace clustering [25], non rigid structure from motion [31] and outlier detection [40], diffraction imaging [44], MRI tomography [43] to name a few.

In [23, 19] solutions of a given cardinality was recovered using $\mathcal{Q}_2(\mu \text{ card})$ by searching over μ. Note however that while it is one-dimensional, the search criterion is not guaranteed to be unimodal and it is not clear over what range nor at what density one needs to sample in order not to miss the sought solution.

6.1. Numerical recovery results

In [37] astonishing results are shown in the noise free case. For example in figure 2 (of that paper) we see how $k = 130$ non-zero entries are recovered using a matrix A of size $m \times n = 256 \times 512$, (which incidentally is close to the theoretical bound $2k \leqslant m$ in the present

paper[8]). However, in the presence of noise, performance seems to drop drastically. In figure 7 (of the same paper) we see an example where performance is evaluated with $k = 8, m = 72$ and $n = 256$.

Here we will present numerical results for the case of $k = 10, m = 100$ and $n = 200$. We use a matrix A with Gaussian randomly generated columns, which are subsequently normalized, and solve problems (3), (11) and (16) for $b = Ax_0 + \epsilon$ for different levels of noise $\|\epsilon\|_2$ between 0 and 5. The vector x_0 has random entries between 2 and 4 in magnitude, and a total magnitude $\|x_0\|_2 = 11$. To solve the optimization problems we use FBS which is known to converge to a stationary point (by [4] in combination with section 2.4 of [16] or section 6 of [17]).

We compare with ℓ^1-minimization (3) as well as two forms of IHT, which arise when applying FBS to the unregularized problems (10) and (15). In the first case the proximal operator will simply threshold at $\sqrt{\mu}$ and in the latter threshold by keeping only the k largest entries. Convergence of such algorithms are proven e.g. in section 5 of [4], and convergence of the latter has also been shown in [8] when $\delta_{3k} < 0.18$. For this reason, we also included graphs for the result of minimizing (10) and (15) (labeled μ card and ι_{10} in the plots). Each point on the respective curves is an average over 50 trials, where we have used 1000 iterations and with a step-size parameter of $0.9/\|A\|^2$, which is close to the upper theoretical bound given in [4] (which coincides with the bound for the convex case, see e.g. [22]).

To set the parameter λ for the ℓ^1-problem (3) we used the formula

$$\lambda = \frac{\|\epsilon\|_2}{\sqrt{n}}\sqrt{2\ \log(n)}$$

corresponding to the recommendations in section 5.2 of [21]. For (10), (11) we used $\mu = 1$ and k was set to 10 for (15), (16), which we motivate as follows:

If the value of δ_{2k}^- is near 0, then the conditions in corollary 2.1 hold given that $2\sqrt{\mu} \lesssim \min\{|x_{0,j}| : |x_{0,j}| \neq 0\}$ where the latter in our case is 2.05 and $\|\epsilon\|_2 \leqslant \sqrt{\mu}$, whereas the conditions in corollary 2.3 hold as long as $3\|\epsilon\|_2 \lesssim 2.05$. In both cases, the estimate for $\|x' - x_0\|_2$ reads $\|x' - x_0\|_2 \lesssim \|\epsilon\|_2$ which is supposed to hold at least for $\|\epsilon\|_2 \lesssim 2/3$. Despite the fact that $\delta_{2k}^- \approx 0$ is quite unlikely (as we saw in section 2.4), the graph in figure 5 (left) indicates that the reality looks even better. Both algorithms find the oracle solution in 100% of the trial for $\|\epsilon\|_2$ up to 2.5, and the true bound (for this particular example) seems to be $\|x' - x_0\|_2 \lesssim \frac{1}{3}\|\epsilon\|_2$ for both (11) and (16), whereas the true constant for ℓ^1 is around 1 (despite $C_{10} = \infty$ as seen in section 2.4, as δ_{20} with high likelihood is greater than 0.4 [6]).

The first version of IHT, i.e. minimization of the unregularized functional (10), is similar to ℓ^1 in performance, whereas (15) is slightly better. Comparing with (11) and (16) the benefits of using the quadratic envelope are undeniable. Note that all 3 methods work for noise-levels much greater than stipulated by the theory. We also remark that, rather surprisingly, there is no major difference between (11) and (16) for moderate noise levels. However, both these methods are designed to find the oracle solution x_{or}, not x_0, so to evaluate this performance we include in figure 5 (right) also the graph of $\|x' - x_{\text{or}}\|_2$ versus $\|\epsilon\|_2$. From this we deduce that both work perfectly until $\|\epsilon\|_2 = 2.5$, but that (11) deteriorates substantially faster beyond this point. In other words, in this example both methods based on $\mathcal{Q}_2(\mu\ \text{card})$ and $\mathcal{Q}_2(\iota_{P_{10}})$ work as expected down to SNR around 4. In [23] a much more thorough comparison between (3), the two methods considered here, and other popular techniques such as reweighted ℓ^1 [37] and Huber-fitting [48] is carried out. This paper also optimizes over hyperparameters, as opposed

[8] Note indeed that the condition $\delta_{2k}^- < 1$ is equivalent to any $2k$ columns of A being linearly independent, which holds with probability 1 for Gaussian random matrices as long as $2k \leqslant m$.

to fixing them *a priori* as in this section. We refrain from similar experiments here since this is a theoretical paper and the above experiment was designed to illustrate and validate the theory, not to compare optimal performance of algorithms.

Another issue that we have not discussed is the starting point. We have used 0 for all examples above, and (a bit surprisingly) this seems to work better than using the least squares solution x_{LS} of $Ax = b$, which seems to have many local minima near it when we use $\mathcal{Q}_2(\mu \text{ card})$. This is clearly seen in our final graph where we plot a histogram of the cardinality of x' over 50 trials with the noise level $\|\epsilon\|_2 = 2.5$, using $\mathcal{Q}_2(\text{card})$ and x_{LS} as starting point. Concerning $\mathcal{Q}_2(\text{card})$ it is interesting to note the following dichotomy, either the cardinality is around 10, or substantially larger, as predicted by theorem 4.2. For this noise level and starting point x_{LS}, $\mathcal{Q}_2(\iota_{10})$ still works perfectly, which is why its performance is excluded; the histogram hits 50 at $k = 10$, in accordance with corollary 5.6. Combined with figure 5, this underlines that when k is known, $\mathcal{Q}_2(\iota_{10})$ is the best penalty.

6.2. Implementation technicalities

Basically anywhere there is a method involving a sparsity inducing $\|x\|_1$-term, it can be easily replaced with $\mathcal{Q}_\gamma(\mu \text{ card})$ or $\mathcal{Q}_\gamma(\iota_{P_k})$ if the model order is known. We encourage the reader to try these on his or her particular problem, and to facilitate this we here discuss briefly some implementational aspects and parameter choices. Code for evaluation of the corresponding proximal operators is available at the following GitHub repository:

https://github.com/Marcus-Carlsson/Quadratic-Envelopes.

First of all we note that it is often customary to put a factor $1/2$ in front of the quadratic term in (9) and moreover the quadratic envelope depends on a parameter γ which we have throughout kept fixed at 2. A more general version of (9) would be

$$\mathcal{Q}_\gamma(f)(x) + \frac{1}{2}\|Ax - b\|_2^2, \quad \gamma > 0. \tag{59}$$

To pass between various normalizations, we note that given any $\alpha > 0$ one has

$$\alpha\mathcal{Q}_\gamma(f) = \mathcal{Q}_{\alpha\gamma}(\alpha f),$$

so in particular (11) is equivalent with $\mathcal{Q}_1(\frac{\mu}{2} \text{ card}) + \frac{1}{2}\|Ax - b\|_2^2$ and (16) with $\mathcal{Q}_1(\iota_{P_k}) + \frac{1}{2}\|Ax - b\|_2^2$, and the entire paper could as well have been written in this setting.

In order for the global minima of $f(x) + \frac{1}{2}\|Ax - b\|_2^2$ to not move when switching to (59), the general theory of [17] states that γ should be less than $\|A\|^2$. In practice, this is too conservative. Reformulated in the general context (59), the condition $\|A\|_{\infty,\mathrm{col}} \leqslant 1$ turns into

$$\|A\|_{\infty,\mathrm{col}} \leqslant \sqrt{\gamma},$$

so by this we should set $\gamma = \sqrt{\|A\|_{\infty,\mathrm{col}}^2}$ in general. This is a much more realistic estimate in practice, but still it is given by a theoretical upper bound. We recall that γ equals the maximum negative curvature of $\mathcal{Q}_\gamma(f)(x)$, and hence lowering the value of γ makes the penalty 'less non-convex', intuitively speaking. We have found that, for the problems considered in this paper, values of γ as low as 0.3 give better performance (i.e. less chance of getting stuck in local minima), while still maintaining the property of finding the oracle solution. With that said, optimal parameter choices will be investigated elsewhere.

Concerning algorithms to minimize (59), we have found no significant difference between ADMM and FBS. The latter is guaranteed to converge to a stationary point when applied to

(9) (under mild assumptions). This follows by the main result of [4] combined with section 6 of [17]. ADMM on the other hand, to our best knowledge, still lacks a proof of convergence at least for the non-separable penalty $\mathcal{Q}_\gamma(\iota_{P_k})$, (the separable case, which does apply to (11), is considered in [49]).

7. Conclusions

With the wealth of papers analyzing sparsity-inducing penalties, is there a need for yet another one? The existing literature can be divided into two groups, either the results are asymptotic in nature (hence say little in a concrete setting) or they assume that δ_{2k} (or some analogous quantity) is sufficiently small. As argued in section 2.4, for the case $m = 2n$, this forces the sparsity k to be well below than 1% of n to achieve $\delta_{2k} \approx 0.4$ or less. On the other hand, in many concrete applications k is substantially larger, and so there is a vast regime where there is *no theoretical support* for that either ℓ^1-minimization (3) or IHT gets anywhere near the ground truth.

The majority of our results, on the other hand, applies as long as any $2k$ columns of A are linearly independent, for then $\delta_{2k}^- < 1$, with the natural catch that if δ_{2k}^- is poor then a large SNR is needed. This is a significant theoretical improvement; if we are in the range $k/n > 0.01$, then ℓ^1-minimization (3) is *convex* and therefore e.g. FBS applied to it is guaranteed to converge to some point x_1', but by the results of [4, 17], the same is true for (11) and (16), it just may happen that the convergence point x_2' is not the global minimum. However, whereas there is *no support* for the hypothesis that x_1' is anywhere near ground truth, the theorems of this paper states that *if* x_2' is the global minimum, then it is the oracle solution which is the best possible outcome (and else it may not be near ground truth, just like x_1'). This gives the two methods studied here a significant theoretical advantage over ℓ^1-minimization, (or IHT or reweighted ℓ^1 as well for that matter). Combined with the numerical section which demonstrates superior performance in the entire range, this paper challenges the ℓ^1-penalty as the penalty of choice for compressed sensing and sparsity based methods in general.

Finally, this paper studies design of sparsity inducing functionals, *not algorithms to find their global minima or stationary points.* We prove that the global minima, under verifiable conditions, is the oracle solution. The fact that both ADMM and FBS (with 0 as starting point) seems to converge to the global minima is a numerical observation whose proof we leave as an open question.

Appendix A

A.1. Appendix to section 3

While it is possible to deal with gradients and subdifferentials in $\mathbb{C}^n$ by simply identifying it with $\mathbb{R}^{2n}$ in the canonical way, the calculus becomes more intuitive if avoid this step. Instead, we say that a function $g_{\mathrm{d}} : \mathbb{C}^n \to \mathbb{R}$ is differentiable at a point x is there is a vector $v \in \mathbb{C}^n$ such that

$$\lim_{\|y\| \to 0+} \frac{g_{\mathrm{d}}(x+y) - g_{\mathrm{d}}(x) - \mathsf{Re}\,\langle y, v\rangle}{\|y\|} = 0. \tag{60}$$

In this case we write $v = \nabla g_{\mathrm{d}}(x)$. For example, consider the function $g_{\mathrm{d}}(x) = \|Ax - b\|^2$. Upon noting that $g_{\mathrm{d}}(x+y) = \|Ax - b\|^2 + 2\,\mathsf{Re}\,\langle y, A^*(Ax-b)\rangle + \|Ay\|^2$, it readily follows

that $\nabla g_{\mathrm{d}}(x) = 2A^*(Ax - b)$. Similarly, if g_{c} is convex and v is a vector such that

$$g_{\mathrm{c}}(x + y) - g_{\mathrm{c}}(x) - \mathsf{Re}\,\langle y, v\rangle \geqslant 0$$

for all y, we say that v is in the subdifferential of g_{c} which we denote by $v \in \partial g_{\mathrm{c}}(x)$.

Let us establish the claim following (18), i.e. that a function g of the type $g_{\mathrm{c}} + g_{\mathrm{d}}$ for functions as above has a stationary point at x if and only if $-\nabla g_{\mathrm{d}}(x) \in \partial g_{\mathrm{c}}(x)$. The condition (18) for stationarity translates to

$$0 \leqslant \liminf_{\substack{y\to x\\ y\neq 0}} \frac{g(x+y) - g(x)}{\|y\|} = \liminf_{\substack{y\to x\\ y\neq 0}} \frac{g_{\mathrm{c}}(x+y) - g_{\mathrm{c}}(x) + \mathsf{Re}\,\langle y, \nabla g_{\mathrm{d}}(x)\rangle}{\|y\|}.$$

To see this, just add and subtract $\mathsf{Re}\,\langle y, \nabla g_{\mathrm{d}}(x)\rangle$ to the numerator and invoke (60). It immediately follows that if $-\nabla g_{\mathrm{d}}(x) \in \partial g_{\mathrm{c}}(x)$ holds then x is stationary. Conversely, suppose that x is stationary. If there exists a y such that $g_{\mathrm{c}}(x + y) - g_{\mathrm{c}}(x) + \mathsf{Re}\,\langle y, \nabla g_{\mathrm{d}}(x)\rangle < 0$, then for $t \in [0, 1]$ we have by convexity that $g_{\mathrm{c}}(x + ty) \leqslant tg_{\mathrm{c}}(x + y) + (1 - t)g_{\mathrm{c}}(x)$ so

$$g_{\mathrm{c}}(x + ty) - g_{\mathrm{c}}(x) + \mathsf{Re}\,\langle ty, \nabla g_{\mathrm{d}}(x)\rangle \leqslant t(g_{\mathrm{c}}(x + y) - g_{\mathrm{c}}(x) + \mathsf{Re}\,\langle y, \nabla g_{\mathrm{d}}(x)\rangle)$$

by which it follows that the above lim inf must also be < 0, a contradiction.

A.2. Appendix to section 4.1

The full statement of theorem 4.1 follows by combining the below four lemmas. For concreteness assume that we work over $\mathbb{C}^n$.

Lemma 8.1. *Without any restriction on A, the functional $\mathcal{K}_\mu$ attains its infimum.*

Proof. Fix $1 \leqslant j \leqslant n$ and consider submatrices $A(:, J)$ of A with m rows and j columns, $J \subseteq \{1, \ldots, n\}$ and $\#J = j$; J determines which columns of A are selected. Now for each fixed J, the minimum of $\|A(:, J)x - b\|_2^2$ is attained and can be computed by solving the normal equations. Let c_j be a corresponding vector in $\mathbb{C}^n$ with zeroes off J, such that $\|Ac_j - b\|_2^2$ equals the minimum in question. Among the $\{c_J\}_{\#J=j}$ we denote by c_j one that satisfies

$$\|Ac_j - b\|_2^2 = \min_{\#J=j} \min_{x\in\mathbb{C}^j} \|A(:, J)x - b\|_2^2.$$

If $I = \inf\ \mathcal{K}_\mu(x)$ we can select a sequence $x_i \in \mathbb{C}^n$ such that $\mathcal{K}_\mu(x_i) \to I$. By construction it must be

$$\mathcal{K}_\mu(c_{\mathrm{card}(x_i)}) \leqslant \mathcal{K}_\mu(x_i).$$

Since $\mathcal{K}_\mu(x_i)$ is arbitrarily close to I and the c_j are finite, it must exist a $\bar{j}$—at least one—such that $\mathcal{K}_\mu(c_{\bar{j}}) = I$. □

Lemma 8.2. *If $\|A\|_{\infty,\mathrm{col}} \leqslant 1$, the functional $\mathcal{K}_{\mu,reg}$ attains its infimum, which equals that of $\mathcal{K}_\mu$.*

Proof. In the light of the basic inequality $\mathcal{K}_{\mu,\mathrm{reg}} \leqslant \mathcal{K}_\mu$ and the previous lemma, the two infima can only be different if there exists a point x_0 such that $\mathcal{K}_{\mu,\mathrm{reg}}(x_0) < \inf\ \mathcal{K}_\mu$. We prove by contradiction that this is impossible. In particular $\mathcal{K}_{\mu,\mathrm{reg}}(x_0) < \mathcal{K}_\mu(x_0)$, which implies that $\mathcal{Q}_2(\mu\ \mathrm{card})(x_0) < \mu\ \mathrm{card}(x_0)$ since the quadratic terms are the same. This in turn implies (by

(24)) that there must be some index j such that the corresponding value in $\mathcal{Q}_2(\mu\ \mathrm{card})(x_0)$ is different from $\mu\,\mathrm{card}(x_{0,j})$, which happens if and only if

$$0 < |x_{0,j}| < \sqrt{\mu}. \tag{61}$$

Let e_j equal 1 in coordinate j and zero elsewhere and consider

$$t \mapsto \mathcal{K}_{\mu,\mathrm{reg}}(x_0 + t\frac{x_{0,j}}{|x_{0,j}|}e_j)$$

for real t such that $0 < |x_{0,j}| + t < \sqrt{\mu}$. This must be a quadratic polynomial, again by inspection of (24), which also gives that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{K}_{\mu,\mathrm{reg}}(x_0 + t\frac{x_{0,j}}{|x_{0,j}|}e_j)\bigg|_{t=0} = -2 + 2\|a_j\|_2^2 \leqslant 0. \tag{62}$$

Hence this quadratic polynomial attains its minimum over the stated range at an endpoint.

It follows that we can redefine $x_{0,j}$ to equal either 0 or $\sqrt{\mu}$, so that the resulting point x_1 satisfies $\mathcal{K}_{\mu,\mathrm{reg}}(x_1) \leqslant \mathcal{K}_\mu(x_0)$. We can now continue like this for another index j such that (61) holds (if it exists), and this process must terminate after finitely many steps N. Denoting the resulting point by x_N, we see that it satisfies $\mathcal{K}_\mu(x_N) = \mathcal{K}_{\mu,\mathrm{reg}}(x_N) < \inf\ \mathcal{K}_\mu$, a contradiction. Hence $\inf\ \mathcal{K}_\mu = \inf\ \mathcal{K}_{\mu,\mathrm{reg}}$.

Let x_0 be a point where the first infimum is attained. Then $\mathcal{K}_{\mu,\mathrm{reg}}(x_0) \leqslant \mathcal{K}_\mu(x_0)$ so we must have identity and hence the infimum of $\mathcal{K}_{\mu,\mathrm{reg}}$ is also attained. □

Lemma 8.3. *Let $\|A\|_{\infty,\mathrm{col}} \leqslant 1$ and let x_0 be a global minima of $\mathcal{K}_{\mu,reg}$ which is not a global minima for $\mathcal{K}_\mu$. Then it belongs to a connected set of global minima of $\mathcal{K}_{\mu,reg}$ including at least two global minima of $\mathcal{K}_\mu$.*

Proof. By repetition of the previous proof we conclude that the first and second derivative of $\mathcal{K}_{\mu,\mathrm{reg}}(x_0 + te_j)$ must be equal to 0, so the quadratic polynomial is constant in the range $0 < |x_{0,j}| + t < \sqrt{\mu}$. Setting t to be one for the endpoints gives two new global minimizers x_1 with either $x_{1,j} = 0$ or $|x_{1,j}| = \sqrt{\mu}$. Either x_1 is a minimizer of $\mathcal{K}_\mu$ or we can continue the process with another subindex. The result now easily follows. □

Lemma 8.4. *Let $\|A\|_{\infty,\mathrm{col}} < 1$, then any local minima of $\mathcal{K}_{\mu,reg}$ is a local minima of $\mathcal{K}_\mu$. In particular, the sets of global minimizers coincide.*

Proof. Let x_0 be a local minimizer of $\mathcal{K}_{\mu,\mathrm{reg}}$ but not of $\mathcal{K}_\mu$. We again repeat the arguments in lemma 8.2, but this time we get strict inequality in (62), which is impossible. Hence such minimizers do not exist.

If now x_0 is a global minimizer to $\mathcal{K}_{\mu,\mathrm{reg}}$ then it is a local minimizer of $\mathcal{K}_\mu$, which in the light of $\mathcal{K}_\mu \geqslant \mathcal{K}_{\mu,\mathrm{reg}}$ means that it is a global minimizer, and the proof is complete. □

A.3. Appendix to section 5.1

The proof will follow after a collection of minor results.

Proposition 8.5. *For any $m + 2$ vectors $v_1, \ldots, v_{m+2}$ in $\mathbb{R}^m$, we can always pick two such that $\langle v_i, v_j\rangle \geqslant 0$.*

We note that the proposition is sharp since the $m+1$ vortices of a simplex in $\mathbb{R}^m$ do have negative scalar products.

Proof. This follows from a simple induction argument. It is indeed clear in $\mathbb{R}$. Suppose now we take $m+2$ vectors v_i such that $\langle v_i, v_j\rangle < 0$ if $i \neq j$. If V is the hyperplane perpendicular to v_{m+2}, the projections $v_1', \dots, v_{m+1}'$ of $v_1, \dots, v_{m+1}$ on V must also have negative scalar products (since the projections onto v_{m+2} always are in the same direction, opposite that of v_{m+2}, so $\langle v_i', v_j'\rangle < \langle v_i, v_j\rangle$). Since V is an $(m-1)$-dimensional vector space, the desired result is immediate by induction. □

Recall that $a_1, \dots, a_n$ denote the columns of A.

Lemma 8.6. *Let $T \subset \{1, \dots, n\}$ have cardinality $\#T \geqslant n-k$ and consider $\{a_j\}_{j\in T}$. Under assumption (a) and (b), we can pick indices $i, j \in T$ such that $\|a_i - a_j\|^2 < 2$.*

Proof. We first consider the real case $\mathbb{R}^m$. Then $\#T \geqslant m+2$ by (a) and since $\|a_i - a_j\|^2 = \|a_i\|^2 - 2\,\mathrm{Re}\,\langle a_i, a_j\rangle + \|a_j\|^2$, the result is immediate by (b) and proposition 8.5. Finally, since $\mathbb{C}^m$ is isomorphic with $\mathbb{R}^{2m}$, the corresponding result in the complex case follows analogously, since now $\#T \geqslant 2m+2$ by (a). □

Armed with the above statements we can now start to characterize global minimizers of $\mathcal{K}_{k,\mathrm{reg}}$, which is annoyingly difficult. It is even difficult to prove that they exist, so as a first step we shall restrict attention to a closed ball. Recall that $\mathbb{D}$ denotes either the unit disc in $\mathbb{C}$ or, if we work over the reals, the interval $[-1, 1]$.

Lemma 8.7. *There exists an $R_0 > 0$ such that for any $R > R_0$, any global minimum x' of $\mathcal{K}_{k,reg}$ restricted to $(R\mathbb{D})^n$ must satisfy*

$$|\tilde{x}_{k+1}'| \leqslant \frac{R}{2}.$$

Proof. Introduce

$$U = \left\{ x \neq 0 : \ |\tilde{x}_{k+1}| \geqslant \frac{1}{2}|\tilde{x}_1| \right\}.$$

We first note that $\mathcal{Q}_2(\iota_{P_k})(x) > 0$ for all $x \in P_k^c$, which follows by the definition (see (8)), so in particular this holds for all $x \in U$. Define

$$\alpha = \inf\left\{ \mathcal{Q}_2(\iota_{P_k})(x) : \ x \in U, \ \|x\|_2 = 1 \right\}. \tag{63}$$

Since we are minimizing a continuous (non-zero) positive function over a compact set, $\alpha > 0$. Let us write $s = s_x$ for the function defined in (52), when there is a need to make the dependence on x clear. The function s is radially dependent, i.e. $s_{tx} = ts_x$ for $t \in \mathbb{R}$, and hence k_* is radially independent. Looking at the expression for $\mathcal{Q}_2(\iota_{P_k})$ we see that

$$\mathcal{Q}_2(\iota_{P_k})(tx) = t^2 \mathcal{Q}_2(\iota_{P_k})(x) \quad t \in \mathbb{R}.$$

Note that $\mathcal{K}_k(0) = \mathcal{K}_{k,\mathrm{reg}}(0) = \|b\|_2^2$ so the global minimum of $\mathcal{K}_{k,\mathrm{reg}}$ is less than or equal to this. Let R_0 be such that $\alpha(R_0/2)^2 > \|b\|_2^2$. If $x \in U$ satisfies $\|x\|_2 > (R_0/2)$, then

$$\mathcal{K}_{k,\mathrm{reg}}(x) \geqslant \mathcal{Q}_2(\iota_{P_k})(x) \geqslant \alpha\|x\|_2^2 > \|b\|_2^2$$

so it follows that such a point is no global minimizer (at least not on any set containing 0).

Now let R and x' be as stated in the lemma. If $|\tilde{x}'_{k+1}| > R/2$ then clearly $\|x'\|_2 \geqslant R_0/2$ so x' cannot be in U. But then $\frac{1}{2}|\tilde{x}'_1| > |\tilde{x}'_{k+1}| > R/2$ which means that $|\tilde{x}'_1|$ is outside of $R\mathbb{D}$. This is impossible, so the proof is complete. □

We define an angle of a complex number z to be any number α_z such that $z = |z|\mathrm{e}^{\mathrm{i}\alpha_z}$. While this is unique modulo 2π for $z \neq 0$, it can be any number for $z = 0$. Recall that $e_1, \ldots, e_n$ denotes the canonical basis in $\mathbb{R}^n$ (or $\mathbb{C}^n$).

Lemma 8.8. *Let x be any vector and let $p, q \in \{1, \ldots, n\}$ be different indices such that $|x_p| \leqslant |\tilde{x}_{k+1}|$ and $0 < |x_q| \leqslant |\tilde{x}_{k+1}|$ holds. Fix corresponding angles α_p and α_q and set*

$$x(t) = x + t\ \mathrm{e}^{\mathrm{i}\alpha_q}\ \mathrm{e}_p - t\ e^{i\alpha_q}\ e_q.$$

Then $\mathcal{Q}_2(\iota_{P_k})(x(t))$ is twice differentiable at 0 and

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{Q}_2(\iota_{P_k})(x(t))\bigg|_{t=0} = -4.$$

Proof. This is relatively easy to see in the case when $|x_p|$ and $|x_q|$ are strictly less than $|\tilde{x}_k|$, so we first assume this. Then the two points where t show up in the sequence $\tilde{x}(t)$ are beyond k, assuming t is kept small enough. For any $l \geqslant 1$ we then have that

$$\sum_{j>k-l} |\tilde{x}_j(t)| = \sum_{j>k-l} |\tilde{x}_j|, \tag{64}$$

because the left-hand side includes one term like $|\tilde{x}_p| + t$ and one term like $|\tilde{x}_q| - t$, which therefore cancel out. (This is were we used $|x_q| > 0$). Looking at the expression (52) which is used to determine k_*, we see that all the values $s_{x(t)}(l)$ are unaffected by small t, and hence k_* is unaffected by t (as long as it is small enough). Now, the first part of the expression (51) for $\mathcal{Q}_2(\iota_{P_k})(x(t))$ also contain (64) (for the particular value $l = k_*$), and hence this is constant. The second part equals

$$-\sum_{j>k-k_*} |\tilde{x}_j(t)|^2 = -\sum_{j>k-k_*} |\tilde{x}_j|^2 - 2|x_p|t + 2|x_q|t - 2t^2,$$

whose second derivative at 0 equals -4, as was to be shown.

Now assume that $|x_p|$ or $|x_q|$ (or both) equals $|\tilde{x}_k|$. The conclusion will follow as above, once we verify that (i)k_* is invariant for small t and, (ii) both terms with t in them appear in $\{|\tilde{x}_j(t)|\}_{j>k-k_*}$.

To see (i), let a be the largest integer such that $|\tilde{x}_{k+1-a}| = |\tilde{x}_k|$ and note that $s_x(1) > 0$ since we have assumed $|\tilde{x}_{k+1}| > 0$. Moreover, by inspection of (52) we have that $s_x(l) = s_x(1)$ for all $1 \leqslant l \leqslant a$, so $k_* \geqslant a$. By this it follows, if we write $k_*(t)$ for the k_* associated with $x(t)$, that we also have $k_*(t) \geqslant a$ for small t, by continuity. Moreover both terms with t's show up in $\{|\tilde{x}_j(t)|\}_{j>k-l}$ for all $l \geqslant a$, so for such l we have that $s_{x(t)}(l)$ is unaffected by small t's by the same cancellation effects as in (64). By this we finally conclude that $k_*(t)$ is constant in a neighborhood of 0, i.e. (i). Since we also know $k_* \geqslant a$, (ii) follows as well by what was written above. The proof is complete. □

Proof of theorem 5.1. Let x' be a local minimizer of $\mathcal{K}_{k,\mathrm{reg}}$, and assume that $x' \notin P_k$. We first assume that all values x'_j are non-zero, and let α_j be corresponding angles. The set

$T = \{j\colon\ |x_j| \leqslant |\tilde{x}_{k+1}|\}$ clearly satisfies $\#T \geqslant n-k$, so we can use lemma 8.6 on the matrix with columns $\{\mathrm{e}^{\mathrm{i}\alpha_j}a_j\}_{j=1}^n$ to pick two indices p and q such that

$$\|\mathrm{e}^{\mathrm{i}\alpha_p}a_p - \mathrm{e}^{\mathrm{i}\alpha_q}a_q\|_2^2 < 2. \tag{65}$$

By the choice of T, we also have that lemma 8.8 applies. Let $x(t)$ be as in that lemma. It then follows that $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{K}_{k,\mathrm{reg}}(x(t))$ exists at 0 and equals

$$-4 + 2\|\mathrm{e}^{\mathrm{i}\alpha_p}a_p - \mathrm{e}^{\mathrm{i}\alpha_q}a_q\|_2^2 < 0.$$

This contradicts the assumption that x is outside P_k, which hence must be false.

We still need to consider the case when some values x'_j are 0. In this case we pick x_q as in lemma 8.8 and we let p be any index such that $x_p = 0$. The angle α_p can now be chosen such that (65) holds, which leads to a contradiction as before.

It is now established that all local minimizers of $\mathcal{K}_{k,\mathrm{reg}}$ lie in P_k, and clearly they are also local minimizers of $\mathcal{K}_k$ in view of $\mathcal{K}_k \geqslant \mathcal{K}_{k,\mathrm{reg}}$ and the fact that these two coincide on P_k. Next we turn to prove that they exist. Fix $R > R_0$ as in lemma 8.7 and let x' be a global minimizer of $\mathcal{K}_{k,\mathrm{reg}}$ in $(R\mathbb{D})^n$. By the lemma we have $|\tilde{x}'_{k+1}| < R/2$, so any perturbation $x(t)$ as considered in lemma 8.8 stays within $(R\mathbb{D})^n$. With this at hand, we conclude as above that $x' \in P_k$.

However, on P_k both $\mathcal{K}_{k,\mathrm{reg}}(x)$ and $\mathcal{K}_k(x)$ coincide with simply $\|Ax-b\|^2$, the minimum of which is attained by the proof of lemma 8.1. We conclude that $\mathcal{K}_{k,\mathrm{reg}}$ do attain its global minima, and that $\mathcal{K}_{k,\mathrm{reg}}$ and $\mathcal{K}_k$ share global minimizers. □

ORCID iDs

Daniele Gerosa https://orcid.org/0000-0002-0569-8677

References

[1] Adcock B and Hansen A C 2016 Generalized sampling and infinite-dimensional compressed sensing *Found. Comput. Math.* **16** 1263–323
[2] Adcock B, Hansen A C, Poon C and Roman B 2017 Breaking the coherence barrier: a new theory for compressed sensing *Forum of Mathematics, Sigma* vol 5 (Cambridge: Cambridge University Press)
[3] Andersson F, Carlsson M and Olsson C 2017 Convex envelopes for fixed rank approximation *Optim. Lett.* **11** 1–13
[4] Attouch H, Bolte J and Svaiter B F 2013 Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods *Math. Program.* **137** 91–129
[5] Bauschke H H *et al* 2017 *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* vol 011 (Berlin: Springer)
[6] Blanchard J D, Cartis C and Tanner J 2011 Compressed sensing: how sharp is the restricted isometry property? *SIAM Rev.* **53** 105–25
[7] Blumensath T and Davies M E 2008 Iterative thresholding for sparse approximations *J. Fourier Anal. Appl.* **14** 629–54
[8] Blumensath T and Davies M E 2009 Iterative hard thresholding for compressed sensing *Appl. Comput. Harmon. Anal.* **27** 265–74
[9] Bredies K, Lorenz D A and Reiterer S 2015 Minimization of non-smooth, non-convex functionals by iterative thresholding *J. Optim. Theory Appl.* **165** 78–112
[10] Breheny P and Huang J 2011 Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection *Ann. Appl. Stat.* **5** 232

[11] Candès E J 2008 The restricted isometry property and its implications for compressed sensing *C. R. Math.* **346** 589–92
[12] Candès E J, Li X, Ma Y and Wright J 2011 Robust principal component analysis? *J. ACM* **58** 11:1–11:37
[13] Candès E J, Romberg J and Tao T 2006 Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information *IEEE Trans. Inform. Theory* **52** 489–509
[14] Candès E J, Romberg J K and Tao T 2006 Stable signal recovery from incomplete and inaccurate measurements *Commun. Pure Appl. Math.* **59** 1207–23
[15] Candes E J and Tao T 2005 Decoding by linear programming *IEEE Trans. Inform. Theory* **51** 4203–15
[16] Carlsson M 2016 On convexification/optimization of functionals including an l2-misfit term (arXiv:1609.09378)
[17] Carlsson M 2019 On convex envelopes and regularization of non-convex functionals without moving global minima *J. Optim. Theory Appl.* **183** 66–84
[18] Carlsson M, Gerosa D and Olsson C 2019 An un-biased approach to low rank recovery (arXiv:1909.13363)
[19] Carlsson M, Tourneret J-Y and Wendt H 2019 Unbiased group-sparsity sensing using quadratic envelopes *2019 IEEE 8th Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* (IEEE) pp 425–9
[20] Rick C 2007 Exact reconstruction of sparse signals via nonconvex minimization *IEEE Signal Process. Lett.* **14** 707–10
[21] Chen S S, Donoho D L and Saunders M A 2001 Atomic decomposition by basis pursuit *SIAM Rev.* **43** 129–59
[22] Combettes P L and Wajs V R 2005 Signal recovery by proximal forward-backward splitting *Multiscale Model. Simul.* **4** 1168–200
[23] Donoho D L 2006 For most large underdetermined systems of linear equations the minimal ℓ1-norm solution is also the sparsest solution *Commun. Pure Appl. Math.* **59** 797–829
[24] Donoho D L, Elad M and Temlyakov V N 2005 Stable recovery of sparse overcomplete representations in the presence of noise *IEEE Trans. Inform. Theory* **52** 6–18
[25] Elhamifar E and Vidal R 2013 Sparse subspace clustering: algorithm, theory, and applications *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 2765–81
[26] Fan J and Li R 2001 Variable selection via nonconcave penalized likelihood and its oracle properties *J. Am. Stat. Assoc.* **96** 1348–60
[27] Fan J *et al* 2004 Nonconcave penalized likelihood with a diverging number of parameters *Ann. Stat.* **32** 928–61
[28] Fan J, Xue L and Zou H 2014 Strong oracle optimality of folded concave penalized estimation *Ann. Stat.* **42** 819
[29] Feng C, Au W S A, Valaee S and Tan Z 2010 Compressive sensing based positioning using rss of wlan access points *2010 Proc. IEEE INFOCOM* pp 1–9
[30] Foucart S and Rauhut H 2013 An invitation to compressive sensing *A Mathematical Introduction to Compressive Sensing* (Berlin: Springer) pp 1–39
[31] Kong C and Lucey S 2016 Prior-less compressible structure from motion *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 4123–31
[32] Larsson V and Olsson C 2016 Convex low rank approximation *Int. J. Comput. Vis.* **120** 194–214
[33] Loh P-L and Wainwright M J 2013 Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima *Advances in Neural Information Processing Systems* pp 476–84
[34] Loh P-L, Wainwright M J *et al* 2017 Support recovery without incoherence: a case for nonconvex regularization *Ann. Stat.* **45** 2455–82
[35] Mazumder R, Friedman J H and Hastie T 2011 Sparsenet: coordinate descent with nonconvex penalties *J. Am. Stat. Assoc.* **106** 1125–38
[36] Natarajan B K 1995 Sparse approximate solutions to linear systems *SIAM J. Comput.* **24** 227–34
[37] Candes E J, Wakin M B and Boyd S P 2008 Enhancing sparsity by reweighted l^1 minimization *J. Fourier Anal. Appl.* **14** 877–905
[38] Nikolova M 2013 Description of the minimizers of least squares regularized with ℓ_0-norm. Uniqueness of the global minimizer *SIAM J. Imaging Sci.* **6** 904–37
[39] Nikolova M 2016 Relationship between the optimal solutions of least squares regularized with ℓ0-norm and constrained by k-sparsity *Appl. Comput. Harmon. Anal.* **41** 237–65

[40] Olsson C, Eriksson A and Hartley R 2010 Outlier removal using duality *2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition* 1450–7
[41] Pan Z and Zhang C 2015 Relaxed sparse eigenvalue conditions for sparse estimation via non-convex regularized regression *Pattern Recognit.* **48** 231–43
[42] Qaisar S, Bilal R M, Iqbal W, Naureen M and Lee S 2013 Compressive sensing: from theory to applications, a survey *J. Commun. Netw.* **15** 443–56
[43] Shi B, Lian Q and Chen S 2016 Compressed sensing magnetic resonance imaging based on dictionary updating and block-matching and three-dimensional filtering regularisation *IET Image Processing* **10** 68–79
[44] Shi B, Lian Q and Chang H 2020 Deep prior-based sparse representation model for diffraction imaging: a plug-and-play method *Signal Process.* **168** 107350
[45] Simon B 2005 *Trace Ideals and Their Applications* (*Mathematical Surveys and Monographs* vol 120) (Providence, RI: American Mathematical Society)
[46] Soubies E, Blanc-Féraud L and Aubert G 2015 A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem *SIAM J. Imaging Sci.* **8** 1607–39
[47] Wainwright M J 2009 Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso) *IEEE Trans. Inform. Theory* **55** 2183–202
[48] Selesnick I 2017 Sparse regularization via convex analysis *IEEE Trans. Signal Process.* **65** 4481–94
[49] Wang Y, Yin W and Zeng J 2019 Global convergence of admm in nonconvex nonsmooth optimization *J. Sci. Comput.* **78** 29–63
[50] Wang Z, Liu H and Zhang T 2014 Optimal computational and statistical rates of convergence for sparse nonconvex learning problems *Ann. Stat.* **42** 2164
[51] Wright J, Yang A Y, Ganesh A, Sastry S S and Yi Ma Y 2009 Robust face recognition via sparse representation *IEEE Trans. Pattern Anal. Mach. Intell.* **31** 210–27
[52] Zhang C-H *et al* 2010 Nearly unbiased variable selection under minimax concave penalty *Ann. Stat.* **38** 894–942
[53] Zhang C-H and Zhang T 2012 A general theory of concave regularization for high-dimensional sparse estimation problems *Stat. Sci.* **27** 576–93
[54] Zou H and Li R 2008 One-step sparse estimates in nonconcave penalized likelihood models *Ann. Stat.* **36** 1509