An unbiased approach to compressed sensing

Marcus Carlsson∗  Daniele Gerosa∗  Carl Olsson†

Abstract

In compressed sensing a sparse vector is approximately retrieved from an under-determined 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 \( x_0 \) and \( \epsilon \) 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 \( \| \hat{x} - x_0 \|_2 \leq C \| \epsilon \|_2 \) 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/non-smooth optimization

MSC2010: 49J25, 49M20, 65K10, 90C26, 90C27

1 Introduction

1.1 Background

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

\[
\arg\min_{x: \|Ax-b\|_2 < \varepsilon} \text{card}(x),
\]

where \( \varepsilon > 0 \) is some allowed tolerance of the error and \( x_0 \) lies in \( \mathbb{R}^n \) or \( \mathbb{C}^n \). Problem \([1]\) is NP-hard \([27]\) and a popular approach, commonly referred to as “compressed sensing”, is to replace \( \text{card}(x) \) with the convex function \( \|x\|_1 \), i.e.

\[
\arg\min_{x: \|Ax-b\|_2 < \varepsilon} \|x\|_1.
\]

∗Centre for Mathematical Sciences, Lund University, mc.gerosa.calle@maths.lth.se
†Electrical Engineering, Chalmers University of Technology, caols@chalmers.se

1
This method goes back (at least) to the 70’s (see the introduction of [14] for a nice historical overview) but received increasing attention in the late 90’s due to the work by Chen, Donoho and Saunders [18] on what they called basis pursuit, which amounts to solving

$$\text{arg min} \{ \lambda \|x\|_1 + \frac{1}{2} \|Ax - b\|_2^2 \}$$  \hspace{1cm} (3)

for a suitable choice of parameter $\lambda$ (playing the role of $\varepsilon$ in the $\ell^1$-version of (1)). In fact, (3) is the dual problem of (2) in the sense that for each $\varepsilon$ there is a $\lambda$ such that the solution of (2) and (3) coincides. The method received massive attention after the works of Candès and coworkers in the early 2000, and the term compressed sensing was coined. In [12], Candès, Romberg and Tao proved the surprising result that, given a sparse vector $x_0$ and a measurement

$$b = Ax_0 + \epsilon,$$  \hspace{1cm} (4)

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

$$\|\hat{x} - x_0\|_2 < C_K \|\epsilon\|_2,$$  \hspace{1cm} (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 properties of $A$, which in a separate publication (Theorem 1.6, [13]) was shown to hold with “overwhelming probability”. The result is indeed surprising, in Figure 1 we display the one-dimensional counterparts of $\text{card}(x)$ and $\|x\|_1$, demonstrating that $\text{card}(x)$ and $\|x\|_1$ are indeed quite different functionals.

Figure 1: Illustration of penalties.

However, the definition “overwhelming probability” is asymptotic in its nature, and therefore it is not clear if (5) is valid in a moderately sized application (it often is not, see Section 2.3). This continues to be the state of the art, see e.g. [1, 2, 10] which provides asymptotic theorems about when compressed sensing works in concrete setups. Moreover, whereas very strong recovery results were reported e.g. in [11, 17, 20] for the case of exact data $b = Ax_0$, in the presence of noise (c.f. (4)) the method gives a well known bias (see e.g. [21, 26]). The $\ell^1$ 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 towards 0 from the sought point $x_0$. This has led to
a large amount of non-convex suggestions to replace the \( \ell_1 \)-penalty, see e.g. [8, 7, 6, 4, 17, 30, 38, 22, 35, 31, 24, 23, 37, 36, 25, 14, 9, 21, 26]. These methods are tailor-made for the sparsity problem, and upon changing the non-convex penalty card \((x)\) it is not clear what to do. We consider now the general problem of minimizing

\[
 f(x) + \|Ax - b\|^2_2
\]

where \( f \) is some non-convex penalty and \( x \) is a vector in some linear space, not necessarily \( \mathbb{R}^n \). For example, if the desired cardinality \( K \) is known a priori, we can take \( f \) to be the indicator function \( \iota_{P_K} \) of the set \( P_K = \{ x : \text{card}(x) \leq K \} \) in which case \((6)\) reduces to

\[
 \arg \min_{\text{card}(x) \leq K} \|Ax - b\|.
\]

In [16] the “quadratic envelope” \( Q_2(f) \) was introduced, where \( Q_2 \) is the “quadratic biconjugate” (apart from the name, this transform was introduced already in [15], see Figure 2 for an illustration). It has the property that \( Q_2(f)(x) + \|x\|^2 \) is the lower semi-continuous convex envelope of \( f(x) + \|x\|^2 \), and the relationship between

\[
 Q_2(f)(x) + \|Ax - b\|^2_2
\]

and the original functional in \((6)\) was investigated. The main result is that, given \( \|A\| < 1 \), the set of local minimizers to \((8)\) is a subset of the local minimizers of \((6)\), and most importantly that the global minimizers coincide. In the particular case of \( f(x) = \text{card}(x) \) the functional \((8)\) has previously been introduced by Zhang [36] under the name MCP (Minimax Concave Penalty) and independently by Aubert, Blanc-Feraud and Soubies [33] under the name CE\( \ell_0 \). It also shows up in earlier publications, for example (2.4) in [21], but it seems like [36] is the first comprehensive performance study and [33] 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. \( Q_2(\text{card}) \)) is the oracle solution (for certain choice of parameters). On the other hand, \( Q_2(\iota_{P_K}) \) is a new object that has only appeared previously in earlier publications by the authors of the present article. In this article we provide theoretical results of the type \((5)\) for the two concrete functionals \( Q_2(\text{card}) \) and \( Q_2(\iota_{P_K}) \). A more extensive discussion of previous results concerning MCP/\( CE\ell_0 \) is found in Section 2.5, as well as other related results on non-convex optimization.
1.2 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 conditions on \( A \) which imply that any local minima of (\ref{eq:original_problem}) for \( f(x) = \text{card}(x) \) 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 global minima. In the case of \( f = \iota_K \) we take this one step further and give conditions under which (\ref{eq:original_problem}) has a unique local minimizer, which hence must be the global minimizer as well as the solution to the original problem (\ref{eq:original_problem}).

In the case when \( b = Ax_0 + \epsilon \) 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. Secondly, we obtain an estimate corresponding to (\ref{eq:oracle_lasso}) where the involved constants are significantly smaller than \( C_K \). Thirdly (and most importantly), the method seems to work better in practice, at least in the setting when \( A \) has normalized Gaussian random columns. In particular, for reasonable values of noise (e.g. \( SNR \approx 4 \) for the case of a 100x200 matrix \( A \), see Section 2.4) we can find the oracle solution using the Forward Backward Splitting algorithm.

In Section 2 we present highlights from the theory, show some numerical results and compare with the traditional \( \ell^1 \)-method (\ref{eq:lasso}). In 2.5 we give a brief review of the field. The remainder of the paper, Sections 3-5, are devoted to developing the theory.

2 Summary of Main Results and Innovations

2.1 Sparse recovery via \( Q_2(\text{card}) \)

We return to the first problem of minimizing (\ref{eq:original_problem}) for \( f = \text{card}(x) \) i.e.

\[
\mathcal{K}(x) = \mu \text{card}(x) + \|Ax - b\|^2
\]  

(9)

(where we introduce the parameter \( \mu \) to control the tradeoff between sparsity and data-fit) which we regularize with

\[
\mathcal{K}_{reg}(x) = Q_2(\mu \text{card})(x) + \|Ax - b\|^2.
\]  

(10)

The graph of \( Q_2(\text{card}) \) is depicted in Figure 1. Is it true that unique global minimizers exist (recall that they are the same given \( \|A\| < 1 \))? It is easy to see that this is not the case in general, just consider the case of a 2x4—matrix \( A \) such that every pair of columns are linearly independent, and let \( \mu \) be such that the global minimum is attained when \( \|Ax - b\| = 0 \). In this case we have \( \binom{4}{2} = 6 \) choices that all give the global minimum. However, in the above example there exists no “sparse” solution, for 2 equals the row-dimension of the matrix, and by sparse we mean a number much smaller than this. For an \( m \times n \) matrix \( A \), with columns sampled from the unit sphere of \( \mathbb{R}^n \), this is formalized in Lemma
2.1 of [20], where it is shown that (1) with $\varepsilon = 0$ has a unique sparse solution (with probability 1) if $b = Ax_0$ and $\text{supp} \ x_0 < m/2$, which is an upper bound of how much sparsity one needs in order to have a well posed sparse problem.

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

- $x'$ is automatically a global minimizer
- any other stationary point $x''$ of (10) satisfies $\text{card}(x'') \gg \text{card}(x')$.

We remind the reader that $A$ satisfies a Restricted Isometry Property for integer $k$, if any $k$ columns of $A$ behaves approximately as an isometry, in the sense that the resulting matrix can be bounded below by $\sqrt{1-\delta_k Id}$ and bounded above by $\sqrt{1+\delta_k Id}$, where $\delta_k$ is the restricted isometry constant for $k$. Classical results from compressed sensing literature usually require that the numbers $\delta_k$ are small, something which we have found is hard to fulfill in practice. For example, the famous estimate (5) holds under the assumption that

$$\delta_{3K} + 3\delta_{4K} < 2. \quad (11)$$

Our numerical evaluation (see Section 2.3) shows that if $K = 5$ this condition is usually not satisfied for a Gaussian random matrix $A$ (with normalized columns) of size $m \times n$, unless $m$ is (at least) around 500. The statement that RIP holds with overwhelming probability [13] is therefore somewhat misleading, since it is based on an asymptotic estimate. For a small size matrix of the type discussed above it typically never applies (more on this in Section 2.3). That the RIP-conditions are hard to satisfy in practice is well known by the community, and has led to interesting new contributions about efficiency of $\ell^1$ without RIP, given that the problem is sampled in a certain way, see e.g. [11, 2]. However, these results are asymptotic in nature and do not apply in as general situation as the ones we will present here.

We base the theory of this paper on the “Restricted Linear Independence Property” (RLIP), basically constituting the lower estimate of the RIP. More precisely, we define

$$\beta_k = \inf\{\frac{\|Ax\|}{\|x\|} : x \neq 0, \text{ card}(x) \leq k\} \quad (12)$$

for $k = 1 \ldots n$. We say that $A$ satisfies RLIP with respect to the property $P_K = \{x : \text{card}(x) \leq K\}$ if $\beta_K \neq 0$. In other words $A$ is RLIP with respect to this property if and only if any $K$ chosen columns of $A$ are linearly independent. The relationship with RIP is as follows; if $A$ satisfies RIP with constant $\delta_k$ then it satisfies RLIP with $\beta_k \geq \sqrt{1-\delta_k}$, whereas the converse often does not hold.

To give an idea of the type of results proven in this paper, we first present two corollaries of theorems in Section 4.
**Corollary 2.1.** Let $A$ have normalized columns and let $x'$ be a stationary point of $K_{\text{reg}}$ with $\text{card}(x') = K$ and set $z' = (I - A^*A)x' + A^*b$. Assume that

$$|z'_i| \notin \left[ \beta_2^2K\sqrt{\mu}, \frac{1}{\beta_2^2K}\sqrt{\mu} \right], \quad 1 \leq i \leq n. \quad (13)$$

If

$$\|Ax' - b\|_2^2 < \mu, \quad (14)$$

then $x'$ is a unique global minimum of both $K$ and $K_{\text{reg}}$. Moreover, any other stationary point $x''$ has a larger support.

The statement is a combination of Theorem 4.2 and 4.5 (for the choice $N = 2K$). Upon assuming a bit more in (13) and (14), we may also conclude that $x''$ has substantially larger support. As a curious remark, note that $\beta_2^2 > 0$ forces $m \geq 2K$ which is precisely the upper bound given by Lemma 2.1 in [20] mentioned above.

In the case when $b = Ax_0 + \epsilon$, as discussed earlier, we can say more. Below we state Theorem 4.8 for the particular case $N = 2K$.

**Corollary 2.2.** Suppose that $b = Ax_0 + \epsilon$ and set $\text{card}(x_0) = K$. Assume that $\|\epsilon\| < \beta_2^2\sqrt{\mu}$ and

$$|x_{0,j}| > \left( \frac{1}{\beta_2^2K} + 1 \right)\sqrt{\mu}, \quad j \in \text{supp } x_0.$$

Then there exists a unique global minimum $x'$ to $K_{\text{reg}}$ as well as $K$, with the property that $\text{supp } x' = \text{supp } x_0$, that

$$\|x' - x_0\| \leq \|\epsilon\| \frac{1}{\beta_2^2K},$$

and that $\text{card}(x'') > K$ for any other stationary point $x''$ of $K_{\text{reg}}$.

Note that the conditions on $\epsilon$ and $x_0$ 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.

### 2.2 Known “model order”.

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 [6]. 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. In this case we set

$$K_K(x) = \iota_{P_K}(x) + \|Ax - b\|^2 \quad (15)$$

(where the subindex $K$ separates the notation from (9)) which we regularize with

$$K_{K,\text{reg}}(x) = \mathcal{Q}_2(\iota_{P_K})(x) + \|Ax - b\|^2. \quad (16)$$

The result corresponding to Corollary 2.1 reads as follows
Corollary 2.3. Let \( A \) have normalized columns such that no pair is orthogonal, and assume that \( n \geq m + K + 2 \). Any local minimizer \( x' \) of \( K_{K,\text{reg}} \) then lies in \( P_K \). Moreover, set \( z' = (I - A^*A)x' + A^*b \), let \( \tilde{z}' \) contain the elements of \( z' \) sorted after decreasing magnitude, and assume that

\[
|\tilde{z}'_{k+1}| < (2\beta_{2K}^2 - 1)|\tilde{z}'_k|.
\]

Then \( x' \) is the unique global minimum of \( K_K \) and \( K_{K,\text{reg}} \).

The result is a combination of Proposition 5.1 and Theorem 5.2 and 5.5. We remark that in the typical compressed sensing application, \( A \) is a matrix with \( m << n \) and \( K << m \). If the columns of \( A \) are normalized random, then the conditions on \( A \) are satisfied with probability 1. Moreover, any subset of \( 2K \) columns will be close to an isometry as long as \( 2K << m \), so it is not unreasonable to expect that \( \beta_{2K} \approx 1 \). In this case the assumption (17) is quite reasonable since \( |\tilde{z}'_{k+1}| \leq |\tilde{z}'_k| \) by construction and \( 2\beta_{2K}^2 - 1 \approx 1 \). The size of \( \beta_{2K} \) in the above scenario is further discussed in subsection 2.3.

We now consider the case when \( b = Ax_0 + \epsilon \) and we wish to retrieve \( x_0 \), where \( \text{card}(x_0) = K \). By combining Proposition 5.6 and Theorem 5.7 we have (for \( A \) as in the previous corollary);

Corollary 2.4. If \( \beta_{2K} > \frac{1}{\sqrt{2}} \) and \( |x_{0,j}| > \left( \frac{1}{2\beta_{2K}^2 - 1} + \frac{1}{\beta_{2K}} \right) \|\epsilon\| \) for all \( j \in S \) then there exists a unique local minimizer \( x' \) to \( K_{K,\text{reg}} \) with \( \text{supp}(x') = \text{supp}(x_0) \). This is the global minimum of both \( K_K \) and \( K_{K,\text{reg}} \) and moreover it satisfies \( \|Ax' - b\| \leq \|\epsilon\| \), \( \text{supp}(x') = \text{supp}(x_0) \) and

\[
\|x' - x_0\| \leq \frac{\|\epsilon\|}{\beta_{2K}}.
\]

If \( \beta_{2K} \approx 1 \) the condition is \( |x_{0,j}| \geq 2\|\epsilon\| \) and the conclusion \( \|x' - x_0\| \lesssim \|\epsilon\| \). We further remark that \( x' \) in Corollaries 2.2 and 2.4 is 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 \( A_Sx = b \) 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 [12]). 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 \( \|x' - x_0\| \). Indeed, finding the correct support is often used as a measurement of success in

Figure 3: Plot of \( \frac{1}{\beta_{2K}} \) for a 17×25 matrix \( A \) with normalized random columns.
numerical sections on the topic [7, 14, 25]. Apart from [7] which studies the minimization of (9) itself (and performs poor in practice, see Figure 5), we have not been able to locate any results in the literature which claim to find $S$, as Corollaries 2.2 and 2.4.

In [28] and [29], the minimizers of the un-regularized functionals $K$ and $\mathcal{K}_K$ are studied. Corollaries 2.1-2.4 provide new results and extensions of this line of research.

### 2.3 On the size of RIP/RLIP-constants

We are interested in estimating the size of the constants $\beta_K$, $\delta_K$ as well as compare the upper bound $C_K$ in (5) with $1/\beta_K$ in the corresponding estimates of Corollary 2.2 and 2.4. We focus on matrices $A$ of size $m \times n$ where $n > m$ and the columns are generated from a random Gaussian distribution and then normalized so that $\|A\|_{\infty,\text{col}} = 1$, in accordance with [12, 20] as well as the the assumptions in this paper.

Figure 3 shows numerical computations of $1/\beta_K$ for random matrices of size $17 \times 25$. Here we do not plot $C_K$ since the requirement $\delta_{3K} + 3\delta_{4K} < 2$ turned out to almost never be fulfilled when $K > 1$. In contrast, Corollary 2.2 apply whenever $\beta_{2K} > 0$ and Corollary 2.4 when $\beta_{2K} > 1/2$.

In order to compare $1/\beta_K$ with $C_K$ for a moderately sized application, we would like to compute these for a $256 \times 512$-matrix, say. However, due to the combinatorial nature of the constants $\delta_j$ and $\beta_k$, it is not possible to compute them for matrices with more columns than $\approx 30$ (on a standard laptop at least). Nevertheless a $256 \times 25$-matrix can be seen as the first portion of a $256 \times 512$-matrix, and from this perspective the values obtained in the $256 \times 25$-case serve as lower bounds of the true values. It turns out that (11) typically does not hold for $K > 2$, while $\beta_{2K} > 1/2$ holds for all $K$ up to 25.

Finally, considering $512 \times 25$-matrices, we do have that (11) is satisfied in general, and Figure 4 plots a graph of $1/\beta_K$ versus $C_K$. From the right graph...
we also see that values of $\beta_K$ are very decent, around 0.8, for $K$ near 20.

### 2.4 Numerical Recovery Results

In [14] 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 $2K < m$ in the present paper). However, in the presence of noise, performance seems to drop drastically. In Figure 7 (of the same paper) we see an example where $K = 8$, $m = 72$ and $n = 256$. This is in line with the predictions of [24], which use $K = \sqrt{m}$ in their numerical section 4.3.

![Figure 5](image)

**Figure 5:** $\|x' - x_0\|$ (left) and $\|x' - x_S\|$ (right) versus $\|\epsilon\|$ for the 5 methods (3), (9)-(10) and (15)-(16). The methods based on $Q_2(\text{card})$ and $Q_2(\iota_{PK})$ work perfectly down to $\text{SNR} \approx 4$.

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), (10) and (16) for $b = Ax_0 + \epsilon$ for different levels of noise $\|\epsilon\|$ between 0 and 5. The vector $x_0$ has random entries between 2 and 4 in magnitude, and a total magnitude $\|x_0\| = 11$.

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 [15] or Section 6 of [16]). In Section 5 of [4] the convergence of FBS for the unregularized problems (9) and (15) is considered, but with no analysis of performance. This has also been proposed earlier in [6] where it is compared against matching pursuit. For this reason, we also included graphs for the result of minimizing (9) and (15).

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. [19]). For the $\ell^1$-problem (3) we used the
corresponding to the recommendations in Section 5.2 of \[18\]. For (9)-(10) we used $\mu = 1$ and $K$ was set to 10 for (15)-(16).

If the values of $\beta_{2K}$ and $\beta_K$ are $\approx 1$, then the conditions in Corollary 2.2 hold given that $2\sqrt{\mu} < \min\{|x_{0,j}| : |x_{0,j}| \neq 0\}$ which in our case is 2.05 and $\|\epsilon\| \leq \sqrt{\mu}$, whereas the conditions in Corollary 2.4 hold as long as $2\|\epsilon\| < 2.05$. In both cases, the estimate for $\|x' - x_0\|$ reads $\|x' - x_0\| \lesssim \|\epsilon\|$ which is supposed to hold for $\|\epsilon\| \lesssim 1$. As can be seen from the graph in Figure 5 (left) the true bound (for this particular example) seems to be $\|x' - x_0\| \lesssim \frac{1}{\sqrt{3}}\|\epsilon\|$ for both (10) and (16), whereas the true constant for $\ell^1$ is around 1 (despite $C_{10} = \infty$ as argued earlier). The unregularized cardinality problem (9), a.k.a. iterative hard thresholding, seems to perform poorly (with our parameters) whereas (15) (a.k.a. the $K$-sparse algorithm [6]) seems to do a decent job, similar to $\ell^1$ in performance for higher noise levels. In both cases corresponding performance for the regularized versions is significantly better, indicating that the regularization by $S_2$ indeed has a crucial effect. 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 (10) and (16) for moderate noise levels. However, both these methods are designed to find the oracle solution $x_S$, not $x_0$, so to evaluate this performance we include in Figure 5 (right) also the graph of $\|x' - x_S\|$ versus $\|\epsilon\|$. From this we deduce that both work perfectly until $\|\epsilon\| = 3$, but that (10) deteriorates substantially faster beyond this point. In other words, in this example both methods based on $Q_2(\text{card})$ and $Q_2(\text{P}O)$ work as expected down to SNR around 4. Note that, in the best case scenario $\beta_K = \beta_{2K} = 1$, and then a simple computation shows that Corollary 2.2 and 2.4 then applies for SNR down to $2\sqrt{10} \approx 6$, and hence there is almost perfect harmony between theory and numerical results. More precisely, we can allow 50% more noise in practice than predicted by the theory.

In Figure 6 we show the same graphs except that now $A$ has size $50 \times 200$. Clearly this has a significant impact on performance. In particular, although (10) and (16) still do better than traditional $\ell^1$-minimization, there is no longer a significant difference. This could indicate that the convex $\ell^1$-method is more reliable in very difficult scenarios, as opposed to the non-convex methods suggested here, but this would have to be further investigated to be confirmed. A
drawback of \( \ell_1 \)-methods is that one often needs to find a suitable \( \lambda \), which leads to slow evaluation in practice.

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 \( Ax = b \). In our final graph Figure 7 we plot a histogram of the cardinality of \( x' \) over 50 trials with the noise level \( \| \epsilon \| = 25 \). Using \( Q_2(\text{card}) \) and \( x_{LS} \) as starting point. For this noise level and starting point, \( Q_2(\epsilon_{1.0}) \) still works perfectly, which is why its performance is excluded, the histogram hits 50 at \( K = 10 \). It is interesting to note the following dichotomy, either the cardinality is around 10, or substantially larger, in harmony with the results presented in this paper (e.g. Corollary 2.1 and more generally Theorem 4.2).

2.5 Brief review of related results

As previously mentioned, \( Q_2(\text{card}) \) was introduced previously in [36] and [33], and appeared earlier also e.g. in [21], although in this paper they move on to introduce yet another penalty called SCAD which we discuss further below. Needless to say, we are not the first group to address the shortcomings of traditional \( \ell_1 \)-minimization by use of non-convex penalties. In fact, even before the birth of compressed sensing, the shortcomings of \( \ell_1 \)-techniques were debated and non-convex alternatives were suggested, we refer to [21] for an overview of early publications on this issue. Moreover, shortly after publishing the celebrated result (5), Candès, Wakin and Boyd suggested an improvement called “Reweighted \( \ell_1 \)-minimization” [14] 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 \( \epsilon \) is a parameter chosen by the user. Figure 1 shows the functions \( \text{card}(x) \), \( |x| \) and \( \log(0.1 + |x|) - \log(0.1) \) as well as \( Q_2(\text{card}) \). As is clear to see, \( \log(0.1 + |x|) - \log(0.1) \) is closer to \( \text{card}(x) \) than \( |x| \), which may explain the better performance by reweighted \( \ell_1 \)-minimization reported in [14]. The functional \( Q_2(\text{card}) \) is even closer to \( \text{card}(x) \), and while this certainly is one reason behind the superior theoretical results reported in this paper, it is not clear that it is beneficial in practice since it may lead to an increased probability of getting stuck in local minima. Indeed, suppose one has a non-sparse solution to \( Ax' = b \) where all non-zero elements of \( x' \) are in the flat part of \( Q_2(\text{card}) \), then we clearly have an (undesired) local minima of \( K_{\text{reg}}(x) = Q_2(\text{card})(x) + \| Ax - b \|_2^2 \), whose presence for high levels of noise is clearly visible in Figure 7. This is further
studied in [33] where it is shown that, for a particular choice of $7 \times 15$-matrix $A$ and other parameters, the original functional (9) has roughly 16000 (!) local minima, whereas around 5000 of these remain as stationary points for the regularization (10). Several of these stationary points turn out to not be local minima, and hence the authors provide a macro algorithm to avoid non-local minima. In the same vein, Zhang [36] proposes to iteratively update relevant parameters to reach the desired global minima with higher probability.

While these results speak a bit in favor of using another method such as reweighted $\ell^1$ or $Q_2(\underline{t}_K)$, which according to Corollary 2.4 does not suffer from the same drawbacks (under mild assumptions), favorable results for $Q_2(\text{card})$ were reported in [25], which compares the use of $Q_2(\text{card})$/MCP with $\ell^1$ and reweighted $\ell^1$ (called LSP in [25]) as well as SCAD (which has similar performance as $Q_2(\text{card})$). 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. At first glance, this seems to contradict the findings of [33] reported above. However, in their experiments they do not use $b = Ax_0$ for some sparse $x_0$ and they also do not use a matrix $A$ with good RLIP-properties.

The best theoretical justification to support the use of $Q_2(\text{card})$ seems to be Corollary 1 in [25] which, under a number of assumptions, prove that $\mathcal{K}_{\text{reg}}$ does have a unique stationary point with high probability, and provide an estimate of the type (5). The setting of [25] is rather different and we have not been able to verify reasonable values of the involved constants $c_1, c_2, c_3, R, \mu, \lambda, \eta, c_1, c_2, c_3$ and $\gamma$ in order to compare the strength of our Corollary 2.2 with Corollary 1 in [25]. We simply note that they point in the same direction and that Corollary 2.2 holds under simpler conditions. The same remark goes for Theorem 3 in [30] and Theorem 1 in [21], which provide conditions under which a class of non-convex optimization problems give the same estimate as the “oracle estimator”, with a high probability.

The papers [4, 6] considers [4] for the cases $f(x) = \text{card}(x)$ as well as $f(x) = \underline{t}_K(x)$, 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. Incidentally, this article in combination with Section 2.4 of [15] (Section 6 of [10]) shows that FBS also converges for [5] under very soft assumptions. For the case $f(x) = \text{card}(x)$, [4] goes a bit further and actually provide an estimate of the type (5) with the good value $C_K = 5$ (independent of $K$), however under the assumption $\delta_3K < 1/8$ which, in the light of the Section 2.3 is not easy to satisfy.

Many other non-convex penalties have been proposed over the years [31, 8, 17, 30, 38, 22, 35, 24, 23, 37, 33, 25, 14, 9, 21, 26], and we make no attempt to review them here. The introduction of [25] contains a recent overview. A common denominator seems to be that the penalty function has the form $p(x) = \sum_j p_j(x_j)$ where $p_j$ are functions on $\mathbb{R}$ (except the recent contribution [31]). In this sense, $Q_2(\underline{t}_K)$ stands out as an interesting deviation.
3 Uniqueness of minimizers and stationary points with the desired property

We now turn to the heart of the matter, namely uniqueness of sparse minimizers of $K_{\text{reg}}$, more precisely minimizers in a given $K$-sparse set $P_K$. As noted in the introduction, this is not possible without imposing additional conditions on $A$ as well as $b$. In this section we provide such a condition and in the coming ones we show what it entails in practice for the sparsity (and $K$-sparsity) problem.

We first introduce the concept of a stationary point for a non-convex functional $g$, which in practice is easier to find than local minimizers. We recall that the Fréchet subdifferential $\hat{\partial} g(x)$ is the set of vectors $v$ with the property that

$$\liminf_{y \to x, y \neq x} g(y) - g(x) - \langle v, y - x \rangle \geq 0.$$ 

We say that a point $x$ is a stationary point of $g$ if $0 \in \hat{\partial} g(x)$. For the case when $g$ is a sum of a convex function $g_c$ and a differentiable function $g_d$, it is easy to see that $x$ is a stationary point if and only if $-\nabla g_d(x) \in \partial g_c(x)$ where $\partial g_c(x)$ denotes the usual subdifferential. Set

$$G(x) = \frac{1}{2} Q_2(f)(x) + \frac{1}{2} \|x\|_2^2,$$  \hspace{1cm} (18)

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

$$K_{\text{reg}}(x) = 2G(x) - \|x\|_2^2 + \|Ax - b\|_2^2$$  \hspace{1cm} (19)

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

$$(I - A^*A)x' + A^*b \in \partial G(x').$$  \hspace{1cm} (20)

Given any $x$, we therefore associate with it a new point $z$ via

$$z = (I - A^*A)x + A^*b.$$  \hspace{1cm} (21)

The importance of $z$ is due to the following simple observation.

**Proposition 3.1.** Let $x'$ and $x''$ be distinct stationary points of $K_{\text{reg}}$ such that $x'' - x' \in P_K$. Then

$$\Re \langle z'' - z', x'' - x' \rangle \leq (1 - \beta_K^2) \|x'' - x'\|_2^2.$$  \hspace{1cm} (22)

The above proposition will mainly be used backwards, i.e. we will show that (22) does not hold and thereby conclude that $x'' - x' \notin P_K$.

**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

\[
\text{Re} \langle z'' - z', x'' - x' \rangle = \|x'' - x'\|^2 - \|A(x'' - x')\|^2 \leq (1 - \beta_k^2) \|x'' - x'\|^2,
\]

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 \( K_{\text{reg}} \) if and only if it solves the convex problem

\[
x' \in \arg \min_x Q_2(f)(x) + \|x - z'\|^2.
\]

Note the absence of \( A \) in the above formula, which in particular implies that \( Q_2(f)(x) + \|x - z'\|^2 \) is the convex envelope of \( f(x) + \|x - z'\|^2 \).

**Proof.** As noted in (20), \( x' \) is a stationary point of \( K_{\text{reg}} \) if and only if \( z' \in \partial G(x') \). By the same token, \( x' \) is a stationary point of

\[
Q_2(f)(x) + \|x - z'\|^2 = 2G(x) - 2 \langle x, z' \rangle + \|z'\|^2
\]

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

\( \square \)

### 4 The sparsity problem

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

\[
Q_2(\mu \text{card})(x) = \sum_{j=1}^n \mu - \max\{\sqrt{\mu} - |x_j|, 0\}^2, \quad x \in \mathbb{R}^n.
\]

(23)

To recapitulate, we want to minimize (9), i.e.

\[
K(x) = \mu \text{card}(x) + \|Ax - b\|^2
\]

(24)

which we replace by (10), i.e.

\[
K_{\text{reg}}(x) = Q_2(\mu \text{card})(x) + \|Ax - b\|^2.
\]

(25)
4.1 Equality of minimizers for $K$ and $K_{\text{reg}}$

As noted by Aubert, Blanc-Feraud and Soubies (see Theorems 4.5 and 4.8 in [33]), $K_{\text{reg}}$ has the same global minima and potentially fewer local minima than $K$ if

$$
\|A\|_{\infty,\text{col}} = \sup_i \|a_i\|_2 \leq 1,
$$

(26)

where $a_i$ denotes the columns of $A$. Below we (essentially) reproduce their statement in the terminology of this paper, and also include a proof for completeness.

**Theorem 4.1.** If $\|A\|_{\infty,\text{col}} < 1$, then any local minimizer of $K_{\text{reg}}$ is a local minimizer of $K$, and the (nonempty) set of global minimizers coincide. If merely $\|A\|_{\infty,\text{col}} = 1$, then any connected component of global minima of $K_{\text{reg}}$ includes at least two global minima of $K$.

**Proof.** We first establish that $\inf K = \inf K_{\text{reg}}$. Since $K \geq K_{\text{reg}}$ it suffices to show that $\inf K \leq \inf K_{\text{reg}}$. Suppose not and let $x_0$ be a point such that $K_{\text{reg}}(x_0) < \inf K$. Then there must be some index $j$ such that the corresponding value in $Q_2(\mu \card)(x_0)$ is different from $\mu \card(a_{0,j})$, which (see (23)) implies that

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

(27)

But then

$$
\partial^2_{x_{0,j}} K_{\text{reg}}(x_0) = -2 + 2\|a_j\|^2 \leq 0.
$$

(28)

It follows that we can redefine $x_{0,j}$ to equal either 0 or $\sqrt{\mu}$, so that the resulting point $x_1$ satisfies $K_{\text{reg}}(x_1) \leq K(x_0)$. We can now continue like this for another index $j$ such that (27) 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 $K(x_N) = K_{\text{reg}}(x_N) < \inf K$, a contradiction.

It is easy to see that $K$ has global minimizers, and by the above argument these are also global minimizers for $K_{\text{reg}}$. Now let $x_0$ be a global minimizer of $K_{\text{reg}}$ but not for $K$. As before we have that (27) and (28) holds for some index $j$, and we clearly must have equality in the latter. If $\|A\|_{\infty,\text{col}} < 1$ this is a contradiction, and hence we see that the global minimizers of $K$ and $K_{\text{reg}}$ coincides. If merely $\|A\|_{\infty,\text{col}} = 1$, then we can redefine $x_{0,j}$ to equal either 0 or $\sqrt{\mu}$ without changing the value of $K_{\text{reg}}$, and as before this process eventually leads to a point $x_N$ which is also a global minimizer for $K$. By the construction, there are at least two such points.

It remains to prove the statement about local minimizers. If $x_0$ is a local minimizer of $K_{\text{reg}}$ and $\|A\|_{\infty,\text{col}} < 1$, we immediately get a contradiction from (28) unless $K(x_0) = K_{\text{reg}}(x_0)$. In view of $K \geq K_{\text{reg}}$, this establishes the claim.

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 $\beta_N > 0$, 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 $K_{reg}$, let $z'$ be given by (21), and assume that

$$|z'_i| \notin \left[ \beta N \sqrt{\mu}, \frac{1}{\beta N} \sqrt{\mu} \right]$$

for all $i \in \{1, \ldots, n\}$ (where the condition is automatically fulfilled if $\beta N > 1$.) If $x''$ is another stationary point of $K_{reg}$ then $\text{card}(x'' - x') > N$.

Note that we allow $\beta N > 1$ 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 the section. Clearly, we will rely on Proposition 3.1, which requires an investigation of the functional $G$ (18) and in particular its sub-differential. Introducing the function $g$ as

$$g(x) = \begin{cases} \mu + \frac{x^2}{2} & |x| \geq \sqrt{\mu} \\ \sqrt{\mu} |x| & 0 \leq |x| \leq \sqrt{\mu} \end{cases}$$

we get

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

Its sub-differential is given by

$$\partial g(x) = \begin{cases} \{x\} & |x| \geq \sqrt{\mu} \\ \{ \sqrt{\mu} \frac{x}{|x|} \} & 0 < |x| < \sqrt{\mu} \\ \sqrt{\mu} \mathbb{D} & x = 0 \end{cases}$$

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 8 illustrates $g$ and its sub-differential.

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

**Lemma 4.3.** Assume that $z_0 \in \partial g(x_0)$ and $\beta N < 1$. If

$$|z_0| > \frac{\sqrt{\mu}}{\beta N}$$

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

$$\text{Re}(z_1 - z_0)(x_1 - x_0) > (1 - \beta^2 N)|x_1 - x_0|^2.$$
that it is larger than 1
consider the angle as an argument. Recall be real. If \( x_0 \)
Summing up, we have that \( x \)
which then automatically applies to \( x_0 \) and which shows that the quotient is minimized when \( x_0 = x_0 \) and in particular that \( x_0 \in \mathbb{R} \) and \( x_0 > \frac{\sqrt{\mu}}{\beta^2_N} \).

To prove the result we now minimize the quotient \( \frac{\text{Re}(z_1 - z_0)(x_1 - x_0)}{|x_1 - x_0|^2} \) and show that it is larger than \( 1 - \beta^2_N \). There are three cases to consider; \( |x_1| = 0 \), \( 0 < |x_1| < \sqrt{\mu} \) and \( |x_1| \geq \sqrt{\mu} \). The latter case is easy since then \( z_1 - z_0 = x_1 - x_0 \) and \( 1 - \beta^2_N < 1 \), which yields the desired conclusion.

For the two other cases we first show that \( z_1 \) and \( x_0 \) can be assumed to be real. If \( x_1 = 0 \) then the optimization of the above quotient is equivalent to minimization of \( -\text{Re}(z_1) \) over \( z_1 \in \sqrt{\mu} \mathbb{D} \), since \( x_0 = z_0 \) is real and positive, which is clearly minimized in \( z_1 = \sqrt{\mu} \). For the middle case, \( z_1 \) and \( x_1 \) have the same angle with \( \mathbb{R} \), and \( |x_1| < |z_1| < \infty \). 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

\[
\frac{\text{Re}(z_1 - z_0)(x_1 - x_0)}{|x_1 - x_0|^2} = \frac{R|x_1|^2 - 2R\text{Re}x_1x_0 + |x_0|^2}{|x_1|^2 - 2R\text{Re}x_1x_0 + |x_0|^2} = \frac{R^2 - R^2 - 1}{R^2 - R^2 - 1} = \frac{|z_1 - z_0|}{|x_1 - x_0|} (36)
\]

which shows that the quotient is minimized when \( x_1 \) is real and \( 0 < x_1 < \sqrt{\mu} \) (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 non-negative and \( 0 \leq x_1 < \sqrt{\mu} \). This simplifies the quotient (36) to \( \frac{z_1 - z_0}{x_1 - x_0} \). We now hold \( x_1 \), \( z_1 \) fixed and consider \( x_0 \) as the variable. Since \( z_1 \geq x_1 \), it is easy to see that this is minimized for \( x_0 \) as small as possible, i.e. \( x_0 = \frac{\sqrt{\mu}}{\beta^2_N} \). With this at hand, the minimum of \( \frac{|z_1 - z_0|}{|x_1 - x_0|} \) is clearly attained at \( x_1 = 0 \) and \( z_1 = \sqrt{\mu} \).

Summing up, we have that

\[
\frac{\text{Re}(z_1 - z_0)(x_1 - x_0)}{|x_1 - x_0|^2} > \frac{|z_1 - z_0|}{|x_1 - x_0|} > \frac{|\sqrt{\mu} - \frac{\sqrt{\mu}}{\beta^2_N}|}{|\frac{\sqrt{\mu}}{\beta^2_N}|} = 1 - \beta^2_N.
\]

Figure 8: The function \( g(x) \) (left) and its sub-differential \( \partial g(x) \) (right). Note that the sub-differential contains a unique element everywhere except at \( x = 0 \).

Proof. By rotational symmetry (i.e. \( \partial g(\exp(i\phi)x) = \exp(i\phi)\partial g(x) \)), it is no restriction to assume that \( z_0 > 0 \). By \( \frac{1}{\beta^2_N} > 1 \) and (33), we see that \( z_0 > \sqrt{\mu} \), and hence the identity \( z_0 \in \partial g(x_0) \) and (32) together imply that \( z_0 = x_0 \) and in particular that \( x_0 \in \mathbb{R} \) and

\[
x_0 > \frac{\sqrt{\mu}}{\beta^2_N}.
\]

(35)
Lemma 4.4. Assume that $z_0 \in \partial g(x_0)$ and $\beta_N < 1$. If

$$|z_0| < \beta_N^2 \sqrt{\mu}$$

(37)

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

$$\Re(z_1 - z_0)(x_1 - x_0) > (1 - \beta_N^2)|x_1 - x_0|^2.$$  

(38)

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 $(0, \sqrt{\mu}]$. For a fixed radius $R = \frac{|z_1|}{|x_1|}$ the quotient 

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

is smallest when both $x_1$ and $z_1$ are real valued and positive. Since $x_1 \neq 0$ we have $z_1 = \max(x_1, \sqrt{\mu})$. It is also easy to see that $z_0 = \beta_N^2 \sqrt{\mu}$ gives a minimum value for any positive choice of $x_1$, $z_1$. This reduces the problem to finding the minimum of

$$x_1 \mapsto \max(x_1, \sqrt{\mu}) - \beta_N^2 \sqrt{\mu}$$

which by basic calculus equals $(1 - \beta_N^2)$, as desired.

We are now ready to prove Theorem 4.2.

Proof of Theorem 4.2. By Proposition 3.1 it suffices to verify

$$\Re(z'' - z', x'' - x') > (1 - \beta_N^2)|x'' - x'|^2, \quad x'' \neq x'.$$  

(39)

Suppose first that $\beta_N < 1$. Since $\partial G(x) = \sum_{j=1}^n \partial g(x_j)$, Lemmas 4.3 and 4.4 imply that

$$\Re(z'' - z')(x'' - x') > (1 - \beta_N^2)|x'' - x'|^2,$$

for all $i$ with $x''_i - x'_i \neq 0$. Since $x''_i - x'_i = 0$ gives $(z''_i - z'_i)(x''_i - x'_i) = 0$ summing over $i$ gives the result.

Suppose now that $\beta_N > 1$. By (39) it suffices to prove that $\Re(z'' - z', x'' - x') \geq 0$ for all $x'' \neq x'$. Fix $i$ in $\{1, \ldots, n\}$. By rotational symmetry it is easy to see that we can assume that $x'_i, z'_i \geq 0$. Moreover, for fixed values of $|z''_i|$ and $|x''_i|$ (but variable complex phase) it is easy to see that $\Re(z''_i - z')(x''_i - x'_i)$ achieves min when these are also real, i.e. we can assume that $x''_i, z''_i \in \mathbb{R}$. Since the graph of $\partial g$ is non-decreasing it follows that $(z''_i - z')(x''_i - x'_i) \geq 0$ for all $i$, as desired.

It remains to consider the case when $\beta_N = 1$, and as above we reach a contradiction if we prove that $\Re(z'' - z', x'' - x') > 0$. Again we can assume that $x'_i, z'_i \geq 0$ and that $x''_i, z''_i \in \mathbb{R}$. Then (29) implies that $z''_i \neq \sqrt{\mu}$ for all $1 \leq i \leq n$, which via $z''_i \in \partial g(x'_i)$ also implies that $x''_i \notin (0, \sqrt{\mu}]$. If $x''_i \neq x'_i$ we must have $x''_i = x_i$ for some $i$. Using that $z''_i \in \partial g$, examination of (32) yields that also $z''_i \neq z'_i$. With this at hand we see that the left hand side of (39) is strictly positive, whereas the right equals 0, which again is a contradiction.

\qed
4.3 Conditions on global minimality

Theorem 4.5. Let $A$ satisfy $\|A\|_{\infty,col} \leq 1$, let $x'$ be a stationary point of $K_{reg}$ and let $z'$ be given by (21). Assume that

$$|z'_i| \notin \left[ \frac{\beta_N^2}{\sqrt{\mu}}, \frac{1}{\beta_N^2} \right], \quad 1 \leq i \leq n. \quad (40)$$

If

$$2 \mu \text{card}(x') + \|Ax' - b\|^2_2 < \mu N + \mu, \quad (41)$$

then $x'$ is the unique global minimum of $K$ and $K_{reg}$.

Proof. Set $K = \text{card}(x')$ and assume that $x'$ is not the unique global minimizer of $K_{reg}$. By Theorem 4.1, there exists a global minima $x'' \neq x'$ for both $K_{reg}$ and $K$, which hence is a stationary point of $K_{reg}$. Theorem 4.2 then shows that $\text{card}(x'') \geq N - K + 1$. It follows that

$$K_{reg}(x'') - K_{reg}(x') \geq K(x'') - K(x') \geq \mu(N - K + 1) - (\mu K + \|Ax' - b\|^2_2) > 0$$

by (40), $K(x'') = K_{reg}(x'')$ and $K(x') \geq K_{reg}(x')$. This is a contradiction, and hence $x'$ must be the unique global minimizer of $K_{reg}$. By Theorem 4.1 it then follows that $x'$ is also unique minimizer of $K$. \hfill \Box

4.4 Noisy data.

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 $\epsilon$ is noise and $x_0$ is sparse. More precisely we set $S = \text{supp } x_0$ where we assume that $\#S = K$ is much smaller than $m$ — the amount of rows in $A$ (i.e. number of measurements). 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 $x_S$ denote the least squares solution to $A_Sx_S = b$. Note that this is the so called “oracle solution” discussed in the introduction, which can also be written $x_S = (A_S^* A_S)^{\dagger} A_S^* b$ where $(A_S^* A_S)^{\dagger}$ denotes the Moore-Penrose inverse. The below proposition shows that the oracle solution is under mild assumptions a local minimizer of $K_{reg}$, which we denote by $x'$ for notational consistency.

Proposition 4.6. Let $A$ satisfy $\|A\|_{\infty,col} \leq 1$. If $\|\epsilon\| < \sqrt{\mu}$ and

$$|x_{0,j}| > \sqrt{\mu} + \frac{\|\epsilon\|}{\beta_K}$$

for all $j \in S$ then the oracle solution $x' = x_S$ is a strict local minimum to $K_{reg}$ with $\text{supp } (x') = \text{supp } (x_0)$. We also have $|x'_j| > \sqrt{\mu}$, $j \in S$, $\|Ax' - b\| \leq \|\epsilon\|$, and

$$\|x' - x_0\| \leq \frac{\|\epsilon\|}{\beta_K}.$$

19
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 $(A_S^* A_S)^{-1} A_S^* \epsilon$ which gives the solution

$$x' = x_0 + (A_S^* A_S)^{-1} A_S^* \epsilon = x_0 + \delta,$$

where we set $(A_S^* A_S)^{-1} A_S^* \epsilon = \delta$. By construction of the Moore-Penrose inverse, $\text{supp } \delta \subseteq S$, and hence

$$A\delta = A_S \delta = P_{\text{Ran} A_S} \epsilon,$$

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

$$\|\delta\| \leq \frac{\| A_S \delta \|}{\beta_K} = \frac{\| P_{\text{Ran} A_S} \epsilon \|}{\beta_K} \leq \|\epsilon\|/\beta_K,$$

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

$$|x'_j| \geq |x_{0,j} - |\delta_j| > \sqrt{\mu} + \|\epsilon\|/\beta_K - \|\epsilon\|/\beta_K = \sqrt{\mu}, \quad j \in S. \quad (42)$$

This also gives $\text{supp } x' = \text{supp } x_0$ since we already have shown $\text{supp } x' \subseteq \text{supp } x_0 \cup \text{supp } \delta \subseteq S$.

We now consider $Ax' - b$, which equals

$$Ax' - b = A_S x' - b = A_S x_0 + A_S (A_S^* A_S)^{-1} A_S^* \epsilon - (A_S x_0 + \epsilon) = (P_{\text{Ran} A_S} - I) \epsilon = -P_{\text{Ran} A_S} \epsilon$$

and hence $\|Ax' - b\| \leq \|\epsilon\|$.

It remains to prove that $x'$ is a local minimum of $K_{\text{reg}} = Q_2(\mu \text{card}) + \|Ax - b\|_2^2$. To this end, consider $K_{\text{reg}}(x' + v)$. Since $|x'_j| > \sqrt{\mu}$ for $j \in S$, the term $Q_2(\mu \text{card})$ is flat for the corresponding indices of $v$. We get

$$K(x' + v) = \sum_{j \in S^c} (2\sqrt{\mu}|v_j| - v_j^2) + 2 \langle v, A^*(Ax' - b) \rangle + \|Av\|^2 + K_{\text{reg}}(x'). \quad (44)$$

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 [44] then simplifies to

$$2 \left( \sum_{j \in S^c} \sqrt{\mu}|v_j| + v_j \langle a_j, Ax' - b \rangle \right) - \sum_{j \in S^c} v_j^2 + \|Av\|^2 + K_{\text{reg}}(x'). \quad (45)$$

By the Cauchy-Schwartz inequality and [43] we have

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

It follows that the term $\sum_{j \in S^c} \sqrt{\mu}|v_j| + v_j \langle a_j, Ax' - b \rangle$ in [45] can be estimated from below by $\rho \sqrt{\sum_{j \in S^c} v_j^2}$ for some $\rho > 0$, and hence that

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

20
for $v$ in a neighborhood of 0, as long as $\sum_{j \in S} v_j^2 \neq 0$. To have $K_{\text{reg}}(x' + v) \leq K_{\text{reg}}(x')$ we thus need $\text{supp} \ v \subset S$, as seen from (45). But then (45) reduces to $\|Av\|^2 + K_{\text{reg}}(x')$, and since $\beta_K > 0$ it follows that $\|Av\|^2 > 0$ 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 (21). We need conditions such that (40) holds for $z'$, i.e. $|z'_j| \notin \left[ \beta_N^2 \sqrt{\mu}, \frac{1}{\beta_N^2 \sqrt{\mu}} \right]$. (46)

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 4.7.** Let $A$ satisfy $\|A\|_{\infty,\text{col}} \leq 1$. If $\|\epsilon\| < \beta_N^2 \sqrt{\mu}$ and

$$|x_{0,j}| > \frac{\sqrt{\mu}}{\beta_N^2} + \frac{\beta^2 N \sqrt{\mu}}{\beta_K}, \quad j \in \text{supp} \ x_0,$$

then (46) holds.

**Proof.** Using (43) we get

$$z' = (I - A^*A)x' + A^*b = x' - A^*(Ax' - b) = x' + A^*P_{(\text{Ran} A_S)^\perp} \epsilon. \quad (48)$$

Since $A^*P_{(\text{Ran} A_S)^\perp}$ is 0 on rows with index $j \in S$ (being a scalar product of a vector in $\text{Ran} A_S$ 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

$$|z'_j| > \frac{1}{\beta_N^2 \sqrt{\mu}}, \quad j \in S$$

holds whenever

$$|x_{0,j}| > \frac{\sqrt{\mu}}{\beta_N^2} + \frac{\|\epsilon\|}{\beta_K}, \quad j \in S,$$

which is true by the assumptions. For the remaining $z'_j$, (i.e. $j \in S^c$), we have $x'_j = 0$ so

$$|z'_j| = |(A^*P_{(\text{Ran} A_S)^\perp}) \epsilon| = |\langle P_{(\text{Ran} A_S)^\perp} \epsilon, a_j \rangle| \leq \|A\|_{\infty,\text{col}} \|\epsilon\| \leq \beta_N^2 \sqrt{\mu}, \quad (50)$$

which establishes (46).

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

**Theorem 4.8.** Suppose that $b = Ax_0 + \epsilon$ where $A$ is an $m \times n$-matrix with $\|A\|_{\infty,\text{col}} \leq 1$ and set card$(x_0) = K$. Let $N \geq 2K$ and assume that $\|\epsilon\| < \beta_N^2 \sqrt{\mu}$ and

$$|x_{0,j}| > \left( \frac{1}{\beta_N^2} + 1 \right) \sqrt{\mu}, \quad j \in \text{supp} \ x_0.$$
Then the oracle solution \( x' = x_S \) is a unique global minimum to \( K_{\text{reg}} \) as well as \( K \), with the property that \( \text{supp} x' = \text{supp} x_0 \), that 
\[
\| x' - x_0 \| \leq \frac{\| \epsilon \|}{\beta_K},
\]
and that \( \text{card}(x'') > N - K \) for any other stationary point \( x'' \) of \( K_{\text{reg}} \).

**Proof.** All the statements follow by Theorem 4.2, Theorem 4.5 and Proposition 4.6, so we just need to check that these apply. Note that 
\[
\beta_N \leq \beta_K \leq \| A \|_{\infty,\text{col}} \leq 1
\]
which will be used repeatedly.

We begin to verify that Proposition 4.6 applies, which is easy by noting that 
\[
\sqrt{\mu} \leq \frac{\| \epsilon \|}{\beta_K} \leq \frac{\sqrt{\mu}}{\beta_N} \leq \frac{\sqrt{\mu}}{\beta_K} + \sqrt{\mu} < |x_{0,j}|.
\]

Now, to verify that the two theorems apply we need to check the conditions (46), which follow if we show that Proposition 4.7 applies. The estimate on \( \| \epsilon \| \) is satisfied by assumption and the other follows by noting that 
\[
\beta_K^2 \beta_N \| x' - x_0 \| \leq \| \epsilon \| \beta_K^2 \beta_N \leq \sqrt{\mu} < \| x_0 \|.
\]

As usual, a simpler statement is found by setting \( N = 2K \), which gives the loosest conditions to verify, (see Corollary 2.2). The only difference in the conclusion concerns the cardinality of other local minima, since the estimate on \( \| x' - x_0 \| \) only depends on \( \beta_K \).

5 *Known model order; the \( K \)-sparsity problem*

Let \( P_K = \{ x : \text{card}(x) \leq K \} \) 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 
\[
\arg \min_{\text{card}(x) \leq K} \| Ax - b \|
\]
is equivalent to finding the minimum of 
\[
K_{\mathcal{K}}(x) = \iota_{P_K}(x) + \| Ax - b \|^2,
\]
(52)
(where we put a subindex \( K \) to distinguish from \( K \) in the previous section). Again, we will approach this problem by using 
\[
K_{K,\text{reg}}(x) = Q_2(\iota_{P_K})(x) + \| Ax - b \|^2.
\]
This is in some ways much simpler than the situation in the previous sections, for example all local minimizers of \( K_{\mathcal{K}} \) are clearly in \( P_K \). On the other hand,
\( Q_2(\iota P K) \) 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

\[
Q_2(\iota P K)(x) = \frac{1}{k^*} \left( \sum_{j>K-k^*} |\tilde{x}_j| \right)^2 - \sum_{j>K-k^*} |\tilde{x}_j|^2
\]  

(53)

where \( k^* \) is the largest value in \( 1...K \) for which the non-increasing sequence

\[
s(k) = \left( \sum_{j>K-k} |\tilde{x}_j| \right) - k |\tilde{x}_{K+1-k}|
\]

(54)

is non-negative (note that it clearly is non-negative for \( k = 1 \)). Although it is not very clear from the above expression, \( Q_2(\iota P K) \) is known to be continuous (see e.g. Proposition 3.2 in [10]), and this will be used without comment below.

We first show that the global minima of \( K_{K,\text{reg}} \) and \( K_K \) coincide.

5.1 Equality of minimizers and \( K \)-feasibility

In order to provide a theorem similar to 4.1, we need a technical condition on the columns \( a_1, \ldots, a_n \) of \( A \). We say that \( A \) is \( K \)-feasible if \( \|A\|_{\infty,\text{col}} \leq 1 \) and for any subset of \( n-K \) columns, we can pick two such that \( \|a_i - a_j\|_2 \leq 2 \). We say that \( A \) is strictly \( K \)-feasible if the inequality is strict. This is very easy to satisfy, the following proposition lists conditions that imply \( K \)-feasibility in \( \mathbb{R} \) and \( \mathbb{C} \) respectively.

**Proposition 5.1.** If we work over \( \mathbb{R} \), any \( A \) with \( \|A\|_{\infty,\text{col}} \leq 1 \) and \( n \geq m + K + 2 \) is \( K \)-feasible. If we add the condition that \( \langle a_i, a_j \rangle \neq 0 \) for all pairs, or that \( \|A\|_{\infty,\text{col}} < 1 \), then strict \( K \)-feasibility follows. The same follows over \( \mathbb{C} \) if \( n \geq 2m + K + 2 \). Another condition ensuring strict \( K \)-feasibility, which works in both \( \mathbb{R} \) and \( \mathbb{C} \), is that \( \|A\|_{\infty,\text{col}} \leq 1 \) and at least \( nK \) of the values \( \{\text{Re} \langle a_i, a_j \rangle\}_{i>j} \) are positive (repetitions allowed).

We remark that it is possible to choose \( 2m + 1 \) vectors in \( \mathbb{C}^m \) such that \( \|a_i - a_j\|_2 > 2 \), just consider a simplex in \( \mathbb{R}^{2m} \) with equal sidelengths and all corners on the unit sphere. The condition that \( n \geq 2m + K + 2 \) is a bit unfortunate, since it rules out the common situation \( n = 2m \). This is why we added the final part of the proposition.

**Proof.** If the chosen subset contain a zero vector, the conclusion is immediate, so we can assume that this is not the case. For any \( m+2 \) vectors \( a_1, \ldots, a_{m+2} \) in \( \mathbb{R}^m \), we can always pick two such that \( \langle a_i, a_j \rangle \geq 0 \), which follows from a simple induction argument [33]. The first two conclusions regarding \( \mathbb{R}^m \) follow immediately by this observation. Since \( \mathbb{C}^m \) is isomorphic with \( \mathbb{R}^{2m} \), the corresponding result for \( \mathbb{C} \) follows. Finally, the elements above the diagonal in \( (\langle a_i, a_j \rangle)_{i,j=1}^n \)
are $n(n - 1)/2$ in number. When we consider a subset of columns of cardinality $n - K$, we remove a total of $(n - K)K + (K - 1)K/2$ of these. Since $nK$ is a bigger number, we are certain that at least one positive value remain.

To illustrate a concrete example, which sometimes appears in applications, we consider the concatenation of a discrete Fourier transform matrix and an identity matrix. To see what the above proposition entails in this case, suppose that $m = 4k$ or $m = 4k + 1$. Each column of the Fourier matrix gives rise to at least $k$ positive values in its scalar products with the canonical basis coming from the identity matrix, so in total we have at least $mk$ positive off diagonal elements in the Fourier matrix. This gives the condition $2mK \leq mk$, i.e. $K \leq m/8$, which is acceptable for relevant applications. We now develop the theory for $K$-feasible matrices, starting with the analog of Theorem 4.1.

**Theorem 5.2.** Let $A$ be $K$-feasible. Then the global minimizers of $K_{K,reg}$ and $K_K$ coincide, and all lie in $P_K$. If $A$ is strictly $K$ feasible, then any local minimizer of $K_{K,reg}$ lies in $P_K$ and is a minimizer of $K_K$.

**Proof.** We first treat global minimizers. Clearly all minimizers of $K_K$ are in $P_K$. Since $K_K \geq K_{K,reg}$ and they coincide on $P_K$, it suffices to show that a given global minimizer of $K_{K,reg}$ is in $P_K$. This is annoyingly difficult, it is even a problem to show that $K_{K,reg}$ has global minima. We begin by showing that large values of $x$ are not candidates for global minima if they are in the set

$$U = \left\{ x \neq 0 : \frac{\sum_{j=K+1}^{n} |\tilde{x}_j|}{\sum_{j=1}^{K} |\tilde{x}_j|} \geq \frac{1}{2K} \right\}. $$

We have $Q_2(tP_K)(x) > 0$ for all $x \in U$. One way to see this is by Corollary 4.4 in [16] (or Theorem 2.20 in [13]), which states that there exists a direction $v$ such that $t \rightarrow Q_2(tP_K)(x + tv) > 0$ has negative second derivative (since $K_{K,reg}(x) < \infty = K_K(x)$), and this would imply that $Q_2(tP_K)$ could take negative values, contradicting Proposition 3.2 in [16] (or Proposition 2.1 in [15]). This can also be deduced by a more careful analysis of (53), which we will perform below. Put

$$\epsilon = \inf \{ Q_2(tP_K)(x) : x \in U, \|x\|_2 = 1 \}. \quad (55)$$

Since we are minimizing a continuous non-zero function over a compact set, $\epsilon > 0$. We now show that large values of $x \in U$ yield large values of $Q_2(tP_K)(x)$. Let us write $s = s_x$ for (53), when there is a need to make the dependence on $x$ clear. The function $s$ is radially dependent, i.e. $s_{tx} = ts_s$ for $t \in \mathbb{R}$, and hence $k_s$ is radially independent. Looking at the expression for $Q_2(tP_K)$ we see that

$$Q_2(tP_K)(tx) = t^2 Q_2(tP_K)(x) \quad t \in \mathbb{R}. $$

Note that $K_K(0) = K_{K,reg}(0) = \frac{1}{2} \|b\|^2$ so the global minimum is less than or equal to this. If $R$ is such that $\epsilon R^2 > \frac{1}{2} \|b\|^2$, it finally follows that no point $x \in U$ with $\|x\| > R$ can be less than $\frac{1}{2} \|b\|^2$. 

24
We remark in passing that \( x = 0 \) cannot be a global minimizer of \( K_{K,\text{reg}} \) unless \( b \) is perpendicular to the range of \( A \), in which case the theorem holds trivially. To see this, note that \( Q_2(t P_K)(0) = 0 \) and \( Q_2(t P_K) \geq 0 \), so if \( x = 0 \) is the global minimum of \( K_{K,\text{reg}} \) the gradient of \( \frac{1}{2} \| A x - b \|^2 \) must necessarily be zero at \( x = 0 \), i.e. \( A^*(-b) = 0 \).

Now let \( G \) be the set of global minimizers for \( K_{K,\text{reg}} \) restricted to \([ -R, R ]^n\). This set is clearly closed, non-empty and bounded. Now let \( G_n \subset G \) be the subset where \( |\tilde{x}_n| \) attains its minimum over \( G \), let \( G_{n-1} \subset G_n \) be the subset where \( |\tilde{x}_{n-1}| \) is minimized, and so on until we reach \( G_{K+1} \), which still is closed, non-empty and bounded. Suppose that \( G_{K+1} \not\subset P_K \) and pick \( x \in G_{K+1} \setminus P_K \). We now show that this is impossible. First of all, note that any two columns of \( A \) and corresponding two elements in \( x \) may (simultaneously) switch positions and sign, without affecting the problem, so it is no restriction to assume that \( |\tilde{x}| = x \), which we now do.

If \( x_{K+1} = R \) then this must also be the case for \( x_1, \ldots, x_K \), and hence \( \|x\| > R \) and \( x \in U \), which is impossible by the earlier conclusions. Thus \( x_{K+1} < R \).

Use \( K \)-feasibility of \( A \) to pick two indices \( j > i > K \) such that \( \|a_i - a_j\|^2 \leq 2 \). Consider the function
\[
x(t) = x + te_i - te_j.
\]
We shall show that this function stays in \( G \) for small values of \( t > 0 \), which contradicts the construction of \( G_{K+1} \). Note that it stays in \([ -R, R ]^n \) for sure, due to \( x_{K+1} < R \).

A complicating factor is the fact that we may fail to have \( |\tilde{x}(t)| = x(t) \) for \( t > 0 \), i.e. this vector is not necessarily non-increasing as a function of its index. As long as \( x_K > x_i \), we can pick both \( i, j > K \) such that \( x(t) \) is non-increasing for small values of \( t > 0 \). We consider this case first. All values of \( s_x(t)(k) \) in \( [54] \) are then unaltered by \( t \), and hence small perturbations do not affect \( k^* \). With this at hand, it follows that the first term in the expression \( [53] \) for \( Q_2(i P_K)(x(t)) \) is unaffected by small changes in \( t \), whereas the latter term is a quadratic polynomial starting with \(-2t^2\). The quadratic term in the expression \( \| Ax(t) - b \|^2 \) on the other hand is \( \|a_i - a_j\|^2 t^2 \). Using \( \|A\|_{\infty,\text{col}} \leq 1 \), it follows that
\[
\frac{d^2}{dt^2} K_{K,\text{reg}}(x(t)) = -4 + 2\|a_i - a_j\|^2 \leq 0.
\]
Thus \( K_{K,\text{reg}}(x(t)) \) is linear in a neighborhood of \( 0 \), and since \( x(0) \in G \) it must actually be a constant. It follows that \( x(t) \in G \) for small values of \( t > 0 \), which is a contradiction as we noted before.

It remains to consider the case when \( x_K = x_i \). In this case we can make \( x(t) \) non-increasing (of its index) for small (fixed) values of \( t > 0 \), upon changing \( i \) so that \( x_i \) is the first value to equal \( x_K \), but now \( i \leq K \) and the independence of \( k_i \) is less clear. Let \( k_i \) be such that \( i = K+1 - k_i \). We will need the following observations about \( s(k) \); if \( x_{K+1-k} = x_{K+1-(k+1)} \) then \( s(k) = s(k+1) \) so we always have
\[
x_{K+1-k} < x_{K+1-(k+1)} = x_{K-k}.
\]
Since \( x \notin P_K \), we also have \( s_x(1) > 0 \). Combined this gives \( s_x(k_i) = s_x(1) > 0 \) and hence \( k_* \geq k_i \). If we now consider \( s_x(t)(k) \) as functions of \( t \), the inequality \( s_x(t)(k_i) > 0 \) is stable with respect to small perturbations. On the other hand, the values of \( s_x(t)(k) \) for \( k > k_i \) are unaffected by small values of \( t \) (they cancel out in the sum of \((\ref{eq:sum})\)), and so we conclude that \( k_* \) is unaffected by small values of \( t \). It follows that both terms depending on \( t \) in \( x(t) \) have indices beyond \( K - k_* \), which precisely as before yields that \((\ref{eq:47})\) holds, a contradiction.

To sum up, we have so far shown that any element \( x \in G_{K+1} \setminus P_K \) necessarily satisfies \( |\tilde{x}_{K+1}| < R \) and \( |\tilde{x}_n| = 0 \). As grande finale we will now show that this is impossible. Let \( j \) be such that \( x_j = 0 \) and pick \( i \) such that \( |x_i| \leq |x_K| \) and \( \langle a_i, a_j \rangle \neq 0 \). That this can be done follows from the basic fact that \( m + 1 \) non-zero vectors in \( \mathbb{R}^m \) can not be mutually orthogonal (suppose for the moment that no \( a_i \) is identically 0). Since \( x_j = 0 \), it is no restriction to assume that \( \langle a_i, a_j \rangle > 0 \), which we do for simplicity. Consider again \( x(t) = x + te_i - te_j \). The previous analysis can now be repeated without modification, to conclude that \((\ref{eq:47})\) holds whether \( i > K \) or not. But in this case \((\ref{eq:47})\) has a strictly negative value, which contradicts that \( x \) is in \( G \) to begin with. In the case when one \( a_i = 0 \), we have \( \|a_i - a_j\| \leq 1 \) for any other \( a_j \), which also yields a contradiction in \((\ref{eq:47})\).

We can finally conclude that any minimizer of \( K_{K,reg} \) in \([-R, R^n]\) is in \( P_K \) and thus also a minimizer of \( K_K \). Since \( R \) could be arbitrarily large, the conclusion holds also in \( \mathbb{R}^n \). The set of global minimizers of \( K_{K,reg} \) is thus closed and non-empty. However, the remaining argument becomes easier if we keep \( R \) fixed where it is for a while longer.

Now consider a path-connected component \( H \) of the set of global minimizers in \([-R, R^n]\). Repeating the entire above argument, we see that \( P_K \cap H \neq \emptyset \). Assume now that it contains points that are not in \( P_K \). Since \( \sum_{j=K+1}^n |\tilde{x}_j| \) is a continuous function on \( H \) which is 0 at \( P_K \), there are points in \( H \) with arbitrarily small quotient \( (\sum_{j=K+1}^n |\tilde{x}_j|)/|\tilde{x}_1| \). Recall that we have ruled out the case \( 0 \in H \) early on in the proof, so the quotient is a continuous function on \( H \). Let \( I \) be a level set of this function. We repeat the construction of sets \( I_n \supset I_{n-1} \supset \ldots \supset I_{K+1} \) precisely as we did with \( G \). Pick a concrete \( x \in I_{K+1} \), and as before it is no restriction to assume that \( x = |\tilde{x}| \). As before we see that it is impossible to have \( x_n = 0 \), and as before we can pick \( i, j \) such that \( \langle a_i, a_j \rangle \geq 0 \) where \( i, j > K \). We define \( x(t) \) via \((\ref{eq:40})\), and establish as before that the \((\ref{eq:47})\) must hold, whereby we get that \( x(t) \in I_{K+1} \) for small values of \( t \), and this contradicts how the sets \( I_1, \ldots, I_K \) were chosen.

By this we get that all global minimizers of \( K_{K,reg} \) in \([-R, R^n]\) must lie in \( P_K \), and since \( R \) can be arbitrarily large, the proof about global minimizers is complete.

Finally if we assume that \( A \) is strictly \( K \)-feasible and that \( x \) is a local minimizer, then we can always find \( i, j \) such that \((\ref{eq:47})\) holds, which is a contradiction since \(-4 + 2\|a_i - a_j\|^2 < 0 \) in this case.
\[\Box\]
5.2 On the uniqueness of sparse stationary points

We now give a condition, similar to (29) 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 5.3.** Let \( x' \) be a stationary point of \( K_{K,\text{reg}} \) with cardinality \( K \), let \( z' \) be given by (21), and assume that
\[
|\tilde{z}_{K+1}'| < (2\beta^2_{2K} - 1)|\tilde{z}_K|.
\]
If \( x'' \) is another stationary point of \( K_{K,\text{reg}} \) then \( \text{card}(x'') > K \).

Again, we allow \( \beta_{2K} > 1 \) in the above theorem, in which case the condition on \( z \) is automatically satisfied. We begin with a lemma. Recall \( G \) given by (18), i.e. \( \frac{1}{2}Q_2(t_{P_K})(x) + \frac{1}{2}||x||_2^2 \) in the present case. We need an expression for \( \partial G(x) \) for \( x \in P_K \).

**Lemma 5.4.** If \( x \in P_K \) then \( z \in \partial G(x) \) if and only if \( \tilde{z}_j = \hat{x}_j \) for \( j = 1, \ldots, K \) and \( \tilde{z}_j \in |\hat{x}_k|D \) for \( j > K \).

**Proof.** Since \( Q_2(t_{P_K}) + ||x||_2^2 \) is the l.s.c. convex envelope of \( t_{P_K} + ||x||_2^2 \), we have that \( G(x) = \frac{1}{2}Q_2(t_{P_K}) + \frac{1}{2}||x||_2^2 \) is the double Fenchel conjugate of \( \frac{1}{2}t_{P_K} + \frac{1}{2}||x||_2^2 \). The Fenchel conjugate of the latter is easily computed to
\[
G^*(y) = \frac{1}{2} \sum_{j=1}^{K} |\tilde{y}_j|^2.
\]
By the well-known identity \( z \in \partial G(x) \Leftrightarrow x \in \partial G^*(z) \) (see e.g. Proposition 16.9 in [32]) we have \( z \in \partial G(x) \) if and only if
\[
z = \arg \max_{z} \Re \langle x, z \rangle - \frac{1}{2} \sum_{j=1}^{K} |\tilde{z}_j|^2.
\]
By standard results on reordering of sequences (see e.g. Ch. 1 in [32]), this implies that the reordering of \( z \) to \( \tilde{z} \) can be chosen identical to that of \( x \), which implies that \( \langle x, z \rangle = \langle \tilde{x}, \tilde{z} \rangle \). Combined with \( \hat{x}_j = 0 \) for \( j > K \), we see that (60) turns into
\[
z = \frac{1}{2} \arg \max_{z} - \sum_{j=1}^{K} |\hat{x}_j - \tilde{z}_j|^2.
\]
The lemma now easily follows.

**Proof of Theorem 5.3.** If \( \text{card}(x'') \leq K \) we clearly have \( x'' - x' \in P_{2K} \) and both \( z' \) and \( z'' \) have the structure stipulated in Lemma 5.4. Let \( I' = \text{supp } x' \) and \( I'' = \text{supp } x'' \). Then \( \Re \langle z'' - z', x'' - x' \rangle \) can be written
\[
\Re \left( \sum_{i \in I'} |x''_i - x'_i|^2 + \sum_{i \notin I'} (x''_i - z''_i)x''_i + \sum_{i \notin I'} (x''_i - z'_i)x''_i \right).
\]
As before we want to reach a contradiction to Proposition 3.1 i.e. we want to prove \( \Re \langle z'' - z', x'' - x' \rangle > (1 - \beta_{2K}^2)||x'' - x'||_2^2 \). Note that
\[
||x'' - x'||_2^2 = \sum_{i \in I'} |x''_i - x'_i|^2 + \sum_{i \notin I'} |x''_i|^2 + \sum_{i \in I''} |x''_i|^2, \tag{63}
\]
that the first term in (62) and (63) are the same, and that \( \beta_{2K}^2 > 0 \). Since the second and third sums have the same number of terms it suffices to show that
\[
\Re(x'_i - z''_i)x''_i + (x''_i - z'_j)x''_j > (1 - \beta_{2K}^2)(|x'_i|^2 + |x''_j|^2), \tag{64}
\]
for any pair \( i \in I', j \notin I'' \) and \( j \notin I', j \in I'' \). This in turn will follow upon showing that
\[
|z''_i||x'_i| + |z''_j||x''_j| < \beta_{2K}^2(|x'_i|^2 + |x''_j|^2).
\]
By Lemma 5.4 it is easy to see that \( |z''_i| < |z''_j| = |x''_j| \) and by assumption we also have \( |z_j| < (2\beta_{2K}^2 - 1)|z'_i| = (2\beta_{2K}^2 - 1)|x'_i| \). Thus
\[
|z''_i||x'_i| + |z''_j||x''_j| < |z''_j||x'_i| + (2\beta_{2K}^2 - 1)|x'_i||x''_j| = 2\beta_{2K}^2|x'_i||x''_j| \leq \beta_{2K}^2(|x'_i|^2 + |x''_j|^2),
\]
as desired. \( \square \)

### 5.3 Conditions on global minimality.

The statements in this section are actually a bit stronger than the corresponding ones in Section 4.3.

**Theorem 5.5.** Let \( A \) be \( K \)-feasible and let \( x' \in P_K \) be a stationary point of \( K_{K, reg} \). Let \( z' \) be given by (21) and assume that (59) applies. Then \( x' \) is a unique global minimizer of \( K_K \) and \( K_{K, reg} \). If \( A \) is strictly \( K \)-feasible, then there are no other local minimizers either.

**Proof.** By Theorem 5.2 there exists \( x'' \in P_K \) which is a global minimizer for both \( K_K \) and \( K_{K, reg} \). If \( x' \neq x'' \) this would contradict Theorem 5.3. If in addition \( A \) is strictly \( K \)-feasible, then Theorem 5.2 says that any local minimizer is a stationary point in \( P_K \), which is impossible by Theorem 5.3. \( \square \)

### 5.4 Noisy data.

We now assume that \( b \) is of the form \( Ax_0 + \epsilon \) where \( \epsilon \) is noise and \( x_0 \) is sparse. More precisely we set \( S = \text{supp} \ x_0 \) where we assume that \#\( S = K \). Let \( A_S \) denote the matrix obtained from \( A \) by setting columns outside of \( S \) to 0. By minor modifications of the proof of Proposition 4.6 we obtain.

**Proposition 5.6.** Let \( A \) satisfy \( \|A\|_{\infty, \text{col}} \leq 1 \). If \( |x_{0,j}| > \left( 1 + \frac{1}{\beta_K} \right) \|\epsilon\| \) for all \( j \in S \) then the oracle solution \( x' = x_S \) is a strict local minimizer of \( K_{K, reg} \) with \( \text{supp} \ (x') = \text{supp} \ (x_0) \). This also satisfies \( |x'_j| > \|\epsilon\|, j \in S \), \( \|Ax' - b\| \leq \|\epsilon\| \) and
\[
\|x' - x_0\| \leq \frac{\|\epsilon\|}{\beta_K}.
\]
Proof. We assume for simplicity that we work over \( \mathbb{R} \), and as in Proposition 4.6 we let \( x' \) be the oracle solution. All estimates of Proposition 4.6 go through with minor modifications, for example (42) is replaced by

\[
|x'_j| \geq |x_{0,j}| - |\delta_j| > |x_{0,j}| - \frac{\|\epsilon\|_K}{\beta K}, \quad j \in S
\]

which shows that \( \text{supp}(x') = \text{supp}(x_0) \) as well as \( |x'_j| > \|\epsilon\| \) for \( j \in S \). The only real difference is the proof that \( x' \) is a local minimizer of \( K_{K,\text{reg}} \), so we consider \( K_{K,\text{reg}}(x' + v) \) where as usual we can assume that \( x' = |\tilde{x}'| \). Since \( x' \) solves the least squares problem posed initially, the vector \( A^*_S(Ax' - b) = A^*_S(A_Sx' - b) \) must be 0, and so

\[
K_{K,\text{reg}}(x' + v) = Q_2 (\iota_{P_K})(x' + v) + 2 \sum_{j=K+1}^n v_j \langle a_j, Ax' - b \rangle + \|Av\|^2 + K_{K,\text{reg}}(x').
\]

Let \( k_1 \) be such that \( x_{K+1-k_1} = x_K \) but \( x_{K-k_1} > x_K \), and note that \( k_2 \) will depend on \( v \) but will always satisfy \( k_2 \leq k_1 \). With this in mind \( Q_2 (\iota_{P_K})(x' + v) \) becomes

\[
\frac{1}{k_2} \left( k_2 x_K + \sum_{j=K-k_2+1}^K v_j + \sum_{j=K+1}^n |v_j| \right)^2 - \sum_{j=K-k_2+1}^K (x_K + v_j)^2 - \sum_{j=K+1}^n v_j^2.
\]

Upon inspection there is a lot of cancelation and the expression reduces to \( 2x_K \sum_{j=K+1}^n |v_j| \) plus quadratic terms in \( v \). Returning to the expression for \( K_{K,\text{reg}}(x' + v) \) and collecting all quadratic contributions from \( v \) in a term \( q(v) \), we see that \( K_{K,\text{reg}}(x' + v) - K_K(x') \) equals

\[
2x_K \left( \sum_{j=K+1}^n |v_j| \right) + 2 \left( \sum_{j=K+1}^n v_j \langle a_j, Ax' - b \rangle \right) + q(v).
\]

Since \( |\langle a_j, Ax' - b \rangle| \leq \|\epsilon\| \) and \( x_K > \|\epsilon\| \), it is easy to see that there exist some constant \( \rho > 0 \) such that

\[
2x_K \left( \sum_{j=K+1}^n |v_j| \right) + 2 \left( \sum_{j=K+1}^n v_j \langle a_j, Ax' - b \rangle \right) \geq \rho \sqrt{\sum_{j=K+1}^n v_j^2}
\]

near 0. As in the proof of Proposition 4.6 we conclude that we must have \( \sum_{j=K+1}^n v_j^2 = 0 \) in order for \( K_{K,\text{reg}}(x' + v) \leq K_{K,\text{reg}}(x') \) to be possible. However, for \( v \) with \( \text{supp}(v) \subset S, Q_2 (\iota_{P_K})(x' + v) = 0 \) and so

\[
K_{K,\text{reg}}(x' + v) = \|A(x' + v) - b\|^2 \geq \|Ax' - b\|^2 = K_{K,\text{reg}}(x'),
\]

and the proof is complete.

\[\square\]
To go from saying that $x'$ is a strict local minimizer to saying unique global minimizer is now a short step.

**Theorem 5.7.** Let $A$ be $K$-feasible and $\beta_{2K} > 1/\sqrt{2}$. If

$$|x_{0,j}| > \left(\frac{1}{2\beta_{2K}^2} - 1 + \frac{1}{\beta_K}\right)\|\epsilon\|, \quad j \in S$$

then the oracle solution $x'$ in the above proposition is a global minimum of $\mathcal{K}_K$ and $\mathcal{K}_{K,reg}$. Moreover, if $A$ is strictly $K$-feasible then $\mathcal{K}_{K,reg}$ has no other local minimizers.

**Proof.** Proposition 5.6 clearly applies and ensures that $x'$ is a local minimizer. We now check that (59) applies for $z'$ given by (21), i.e. we want to check that $|\tilde{z}_{K+1}'| < (2\beta_{2K}^2 - 1)|\tilde{z}_K'|$. Note that $|\tilde{z}_{K+1}'| \leq \|\epsilon\|$ by the same estimate as (50). Moreover, since $z' \in \partial G(x')$, Lemma 5.4 implies that it suffices to show that $\|\epsilon\| < (2\beta_{2K}^2 - 1)|\tilde{x}_K'|$, which by (65) follows if

$$\|\epsilon\| < (2\beta_{2K}^2 - 1)\left(|\tilde{x}_{0,K}'| - \|\epsilon\|/\beta_K\right)$$

which easily is seen to be equivalent with the condition in the statement. The desired conclusions now follow by Theorem 5.5.

**References**

[1] Ben Adcock and Anders C Hansen. Generalized sampling and infinite-dimensional compressed sensing. *Foundations of Computational Mathematics*, 16(5):1263–1323, 2016.

[2] Ben Adcock, Anders C Hansen, Clarice Poon, and Bogdan Roman. Breaking the coherence barrier: A new theory for compressed sensing. In *Forum of Mathematics, Sigma*, volume 5. Cambridge University Press, 2017.

[3] Fredrik Andersson, Marcus Carlsson, and Carl Olsson. Convex envelopes for fixed rank approximation. *Optimization Letters*, pages 1–13, 2017.

[4] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[5] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 2011. Springer, 2017.

[6] Thomas Blumensath and Mike E Davies. Iterative thresholding for sparse approximations. *Journal of Fourier analysis and Applications*, 14(5-6):629–654, 2008.
[7] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

[8] Kristian Bredies, Dirk A Lorenz, and Stefan Reiterer. Minimization of non-smooth, non-convex functionals by iterative thresholding. *Journal of Optimization Theory and Applications*, 165(1):78–112, 2015.

[9] Patrick Breheny and Jian Huang. Coordinate descent algorithms for non-convex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232, 2011.

[10] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011.

[11] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[12] Emmanuel J Candès, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[13] Emmanuel J Candés and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[14] Emmanuel J Candès, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted $l^1$ minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[15] Marcus Carlsson. On convexification/optimization of functionals including an $l_2$-misfit term. *arXiv preprint arXiv:1609.09378*, 2016.

[16] Marcus Carlsson. On convex envelopes and regularization of non-convex functionals without moving global minima. *Journal of Optimization Theory and Applications, to appear*, 2019.

[17] Rick Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

[18] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[19] Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[20] David L Donoho. For most large underdetermined systems of linear equations the minimal $\ell^1$-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.
[21] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[22] Jianqing Fan, Heng Peng, et al. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

[23] Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, 42(3):819, 2014.

[24] Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with non-convexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[25] Po-Ling Loh, Martin J Wainwright, et al. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

[26] Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

[27] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[28] Mila Nikolova. Description of the minimizers of least squares regularized with \ell_0-norm. uniqueness of the global minimizer. *SIAM Journal on Imaging Sciences*, 6(2):904–937, 2013.

[29] Mila Nikolova. Relationship between the optimal solutions of least squares regularized with l0-norm and constrained by k-sparsity. *Applied and Computational Harmonic Analysis*, 41(1):237–265, 2016.

[30] Zheng Pan and Changshui Zhang. Relaxed sparse eigenvalue conditions for sparse estimation via non-convex regularized regression. *Pattern Recognition*, 48(1):231–243, 2015.

[31] Ivan Selesnick. Sparse regularization via convex analysis. *Submitted, Feb*, 2017.

[32] Barry Simon. Trace ideals and their applications, volume 120 of mathematical surveys and monographs. *American Mathematical Society, Providence, RI,*, 2005.

[33] Emmanuel Soubies, Laure Blanc-Féraud, and Gilles Aubert. A continuous exact l0 penalty (cel0) for least squares regularized problem. *SIAM Journal on Imaging Sciences*, 8(3):1607–1639, 2015.

[34] Hagen von Eitzen. https://math.stackexchange.com/questions/1443226/maximum-vector-in-bbb-rn-which-any-two-of-them-has-dot-product-less-than-0.
[35] Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of Statistics*, 42(6):2164, 2014.

[36] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[37] Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, pages 576–593, 2012.

[38] Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509, 2008.