New Insights on the Optimality Conditions of the l2-l0 Minimization Problem

Emmanuel Soubies, Laure Blanc-Féraud, Gilles Aubert

JMIV manuscript No.
(will be inserted by the editor)

New Insights on the Optimality Conditions of the ℓ_2-ℓ_0 Minimization Problem

Emmanuel Soubies · **Laure Blanc-Féraud** · **Gilles Aubert**

Received: date / Accepted: date

Abstract This paper is devoted to the analysis of necessary (not sufficient) optimality conditions for the ℓ_0-regularized least-squares minimization problem. Such conditions are the roots of the plethora of algorithms that have been designed to cope with this NP-hard problem. Indeed, as global optimality is, in general, intractable, these algorithms only ensure the convergence to suboptimal points that verify some necessary optimality conditions. The degree of restrictiveness of these conditions is thus directly related to the performance of the algorithms. Within this context, our first goal is to provide a comprehensive review of commonly used necessary optimality conditions as well as known relationships between them. Then, we complete this hierarchy of conditions by proving new inclusion properties between the sets of candidate solutions associated to them. Moreover, we go one step further by providing a quantitative analysis of these sets. Finally, we report the results of a numerical experiment dedicated to the comparison of several algorithms with different optimality guaranties. In particular, this illustrates the fact that the performance of an algorithm is related to the restrictiveness of the optimality condition verified by the point it converges to.

Keywords ℓ_0-regularized least-squares · CEL0 · Exact relaxation · Minimizers · Optimality conditions

The authors would like to dedicate this work to the memory of Mila Nikolova who passed away in June 2018. Mila made significant contributions to the understanding of the properties of the minimizers of nonconvex regularized least-squares [16, 17, 29–34]. Among them, she published a couple of very instructive papers [31, 32] that provide an in-depth analysis of the minimizers of Problem (1). *These works, as well as exiting discussions with Mila herself, have greatly inspired and contributed to the analysis reported in the present paper.*

Emmanuel Soubies
IRIT, Université de Toulouse, CNRS, France.
E-mail: emmanuel.soubies@irit.fr

Laure Blanc-Féraud
Université Côte d'Azur, CNRS, INRIA, I3S, France.
E-mail: laure.blancferaud@cnrs.fr

Gilles Aubert
Université Côte d'Azur, UNS, LJAD, France.
E-mail: gilles.aubert@unice.fr

1 Introduction

Sparse models are widely used in machine learning, statistics, and signal/image processing applications (*e.g.*, coding, inverse problems, variable selection, or image decomposition). They usually lead to optimization problems that are known to be very hard, which makes them even more interesting from a scientific point of view. For instance, given a linear operator $\mathbf{A} \in \mathbb{R}^{M \times N}$ (with generally $M \ll N$) and a sparse signal $\mathbf{x} \in \mathbb{R}^N$, a standard problem aims at recovering $\mathbf{x}$ from the noisy measurement $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ represents a vector of noise. When the noise is Gaussian, one would ideally solve

$$\hat{\mathbf{x}} \in \left\{ \arg \min_{\mathbf{x} \in \mathbb{R}^N} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \lambda \|\mathbf{x}\|_0 \right\}, \tag{1}$$

or its constrained counterparts, where $\|\cdot\|_0$ denotes the ℓ_0 pseudo-norm that counts the number of non-zero entries of $\mathbf{x}$. Within this context, the present paper is devoted to the analysis of necessary optimality conditions for the challenging optimization problem (1), whose objective function is denoted from now on by F_0.

1.1 Solving (1): A Brief Literature Review

The combinatorial nature of the ℓ_0 pseudo-norm makes Problem (1) belonging to the NP-hard class of complexity [28]. Yet, when the size of $\mathbf{x}$ does not exceed hundreds of variables, it can be solved exactly through mixed integer programming together with branch-and-bound algorithms [6].

The picture is radically different for large-scale inverse problems (*e.g.,* in imaging) as the aforementioned approaches become computationally prohibitive. A popular alternative leverages efficient convex optimization tools through the ℓ_1-relaxation of Problem (1). Needless to say that this convex relaxation enjoys optimality guaranties—under some (restrictive) hypothesis on $\mathbf{A}$—which are the roots and have made the success of compressed sensing [8,15]. Other convex relaxations can be obtained through the convex non-convex strategy introduced in [40,39].

Besides, direct approaches to find approximate solution of (1) have also been extensively studied. These include greedy methods [2,4,12,23–25,36,43–46] as well as iterative thresholding algorithms [5,14,19,27].

Finally, there exist alternatives approaches that focus on continuous, yet nonconvex, relaxation of (1). Within this framework, the standard practice aims at replacing the ℓ_0 term in (1) by a continuous sparsity promoting penalty $\Phi : \mathbb{R}^N \to \mathbb{R}$. Problem (1) is then relaxed as

$$\hat{\mathbf{x}} \in \left\{ \arg\min_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2 + \Phi(\mathbf{x}) \right\}. \tag{2}$$

Numerous penalties have been proposed and analyzed in the literature [7,9,13,18,20,37,38,48,49,51,52]. Moreover, some of them lead to exact relaxations of F_0 in the sense that their minimizers coincide [11,41,42].

1.2 Necessary Optimality Conditions

Because Problem (1) is NP-hard, not only one cannot expect, in general, to attain an optimal point, but verifying the optimality of a point $\hat{\mathbf{x}}$ is also, in general, intractable. Hence, there has been an increasing interest in studying tractable necessary (not sufficient) optimality conditions for Problem (1) [4,32,47], its constrained counterpart [2,3], as well as links between these different formulations [33,50]. Such conditions are important tools as they constitute a criteria to compare the aforementioned suboptimal methods. The stronger (*i.e.,* the more restrictive) the optimality condition verified by points attained by a given algorithm, the "better" the algorithm. Moreover, these conditions can also give rise to new iterative algorithms [2,4,44]. Although a variety of necessary optimality conditions with different degree of sophistication have been defined, analyzed, and hierarchized (see Section 2 and Figure 1), there is a lack of connection between some of them. In particular, the relation between conditions that are based on the support and those that derive from exact continuous relaxations [11,41,42] have not been studied yet.

1.3 Contributions and Roadmap

To complete the picture of hierarchy between existing necessary optimality condition for (1) (see Section 2), we consider the continuous exact ℓ_0 (CEL0) penalty [41]

$$\Phi(\mathbf{x}) = \lambda N - \sum_{i=1}^{N} \frac{\|\mathbf{a}_i\|^2}{2}\left(|x_i| - \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|}\right)^2 \mathbb{1}_{\left\{|x_i|\leq \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|}\right\}}, \tag{3}$$

where $\mathbf{a}_i \in \mathbb{R}^M$ denotes the ith column of $\mathbf{A}$. In the sequel, we refer to the associated objective function in (2) as $\tilde{F}$. Then, our contributions are as follows:

- We derive necessary and sufficient conditions ensuring that a critical point of $\tilde{F}$ is a strict local minimizer of $\tilde{F}$ (Section 3),
- We prove new relationships between necessary optimality conditions as summarized in Figure 1, building upon the prior work [4] (Section 4),
- We provide a quantitative analysis of optimal points with respect to the penalty parameter λ (Section 5),
- We compare the performance of several algorithms which are proven to converge to points verifying different necessary optimality conditions (Section 6).

1.4 Notations

Scalars and functions are denoted by italic letters. Vectors are denoted by bold lowercase letters and matrices by bold uppercase letters. We define the set $\mathbb{I}_N = \{1,\dots,N\}$. The zero vector of $\mathbb{R}^N$ is denoted $\mathbf{0}_{\mathbb{R}^N}$ and we write $\mathbf{e}_i$ the ith vector of the natural basis of $\mathbb{R}^N$. Given a column vector $\mathbf{x} = [x_1 \cdots x_N]^T$, its p-norm is defined as $\|\mathbf{x}\|_p = (\sum_{i=1}^N |x_i|^p)^{\frac{1}{p}}$. When not specified, $\|\cdot\| = \|\cdot\|_2$. The open ℓ_p-ball, $p \in [0,\infty]$, of radius $\eta > 0$ centered at $\mathbf{x} \in \mathbb{R}^N$ is denoted $\mathcal{B}_p(\mathbf{x},\eta) = \{\mathbf{u} \in \mathbb{R}^N : \|\mathbf{u}-\mathbf{x}\|_p < \eta\}$. Then, we write its closure as $\bar{\mathcal{B}}_p(\mathbf{x},\eta)$. We denote the support of $\mathbf{x} \in \mathbb{R}^N$ as $\sigma_{\mathbf{x}} = \{i \in \mathbb{I}_N : x_i \neq 0\}$. Hence, we have $\|\mathbf{x}\|_0 = \#\sigma_{\mathbf{x}}$ where the prefix $\#$ stands for cardinality. Let $\omega \subseteq \mathbb{I}_N$, then we define $\mathbf{x}_\omega \in \mathbb{R}^{\#\omega}$ ($\mathbf{A}_\omega \in \mathbb{R}^{M\times\#\omega}$, respectively) as the restriction of $\mathbf{x} \in \mathbb{R}^N$ ($\mathbf{A} \in \mathbb{R}^{M\times N}$, respectively) to the elements (the columns, respectively) indexed by ω. Moreover, for the sake of simplicity, $\mathbf{a}_i = \mathbf{A}_{\{i\}}$ denotes the ith column of $\mathbf{A}$. Given $\mathbf{x} \in \mathbb{R}^N$ and the index $i \in \mathbb{I}_N$, we define the vector $\mathbf{x}^{(i)} = \mathbf{x} - x_i\mathbf{e}_i = [x_1 \cdots x_{i-1}\, 0\, x_{i+1} \cdots x_N]^T$.

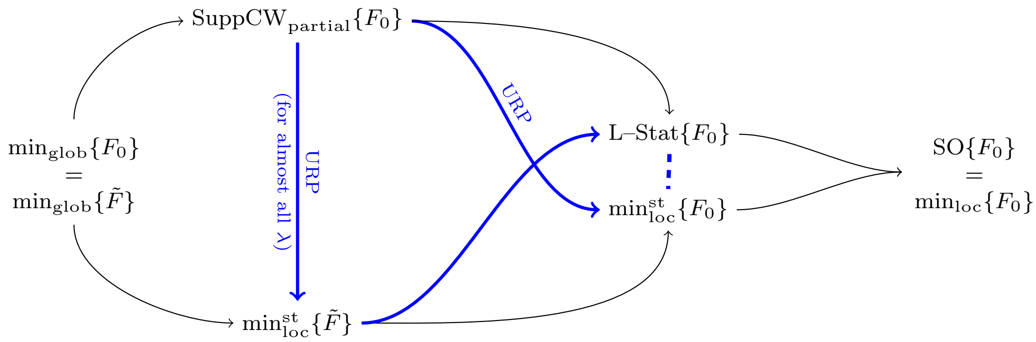

Fig. 1: Hierarchy between optimality conditions. Arrows stand for inclusion (*i.e.,* $A \to B$ means $A \subseteq B$) while the dotted line indicates that there is no inclusion property between the two sets. New results from the present paper are highlighted with thick blue lines. The employed terminology is as follows: $\min_{\text{glob}}\{\cdot\}$ for global minimizers, $\min_{\text{loc}}\{\cdot\}$ for local minimizers, $\min^{\text{st}}_{\text{loc}}\{\cdot\}$ for stirct local minimizers, SuppCW$_{\text{partial}}\{\cdot\}$ for partial support coordinate-wise points, L–Stat$\{\cdot\}$ for L-stationary points (with $L \geq \|\mathbf{A}\|^2$), and SO$\{\cdot\}$ for support optimal points. URP means that the inclusion property is valid under the unique representation property (see Theorem 5). Finally, the references where these relationships (including redundant ones) were established are summarized in Table 1.

2 Necessary Optimality Conditions: Definitions and Known Relationships

In this section, we review necessary optimality conditions for Problem (1) that have been studied over the past. The upcoming definitions are mainly related to the two papers [4,32].

Definition 1 (Support optimality [4]) A point $\mathbf{x} \in \mathbb{R}^N$ is said to be support optimal (SO) for (1) if

$$\mathbf{x} \in \left\{ \arg\min_{\mathbf{u}\in\mathbb{R}^N} \|\mathbf{A}\mathbf{u} - \mathbf{y}\|^2 \ \text{ s.t. } \ \sigma_{\mathbf{u}} \subseteq \sigma_{\mathbf{x}} \right\}, \tag{4}$$

or, equivalently, if $\mathbf{x}$ is such that

$$\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle = 0 \quad \forall i \in \sigma_{\mathbf{x}} \tag{5}$$

This support optimality condition, introduced by the authors in [4], should not be confused with the optimality of the support (*i.e.,* when $\mathbf{x}$ shares its support with a global minimizer $\hat{\mathbf{x}}$ of F_0). Instead, Definition 1 characterizes points that are optimal in the least-squares sense for a given support. Actually, from [32, Corollary 2.5], SO points are local minimizers of F_0, and conversely.

Among these local minimizers (or SO points) of F_0, an important subset contains the *strict* local minimziers, *i.e.,* points $\mathbf{x} \in \mathbb{R}^N$ such that

$$\exists \varepsilon > 0,\ \forall \mathbf{u} \in \mathcal{B}_2(\mathbf{x}, \varepsilon),\ F_0(\mathbf{x}) < F_0(\mathbf{u}). \tag{6}$$

The relevance of these points comes from [32, Theorem 4.4] which states that global minimizers of F_0 are strict. Hence, strict local optimality—characterized by Theorem 1—is a necessary optimality condition for F_0 that is stronger than SO.

Table 1: References establishing the hierarchy between optimality conditions. The top part of the table corresponds to the links depicted in Figure 1 while the bottom part contains additional (redundant) results for completeness.

Property	Reference
$\min_{\text{glob}}\{F_0\} = \min_{\text{glob}}\{\tilde{F}\}$	**Corollary 1**
$\min_{\text{glob}}\{F_0\} \subseteq \text{SuppCW}_{\text{partial}}\{F_0\}$	[4, Theorem 4.14]
$\min_{\text{glob}}\{\tilde{F}\} \subseteq \min^{\text{st}}_{\text{loc}}\{\tilde{F}\}$	**Corollary 1**
$\min^{\text{st}}_{\text{loc}}\{\tilde{F}\} \subseteq \text{L–Stat}\{F_0\}$	**Theorem 4**
$\min^{\text{st}}_{\text{loc}}\{\tilde{F}\} \subseteq \min^{\text{st}}_{\text{loc}}\{F_0\}$	[41, Corollary 4.9]
$\text{SuppCW}_{\text{partial}}\{F_0\} \subseteq \text{L–Stat}\{F_0\}$	[4, Theorem 4.17]
$\text{SuppCW}_{\text{partial}}\{F_0\} \subseteq \min^{\text{st}}_{\text{loc}}\{F_0\}$	**Theorem 5**
$\text{L–Stat}\{F_0\} \subseteq \text{SO}\{F_0\}$	[4, Theorem 4.11]
$\min^{\text{st}}_{\text{loc}}\{F_0\} \subseteq \min_{\text{loc}}\{F_0\}$	By definition
$\text{SO}\{F_0\} = \min_{\text{loc}}\{F_0\}$	[32, Corollary 2.5]
$\text{SuppCW}_{\text{partial}}\{F_0\} \subseteq \min^{\text{st}}_{\text{loc}}\{\tilde{F}\}$	**Theorem 6**
$\text{L–Stat}\{F_0\}$ - - - $\min^{\text{st}}_{\text{loc}}\{F_0\}$	**Section 4.2**
$\min_{\text{glob}}\{F_0\} \subseteq \min^{\text{st}}_{\text{loc}}\{F_0\}$	[32, Theorem 4.4]
$\min_{\text{glob}}\{F_0\} \subseteq \text{L–Stat}\{F_0\}$	[4, Theorem 4.10]

Theorem 1 (Strict local optimality for F_0 [32]) *A point $\mathbf{x} \in \mathbb{R}^N$ is a strict local minimizer of F_0 if and only if it is SO and* $\text{rank}(\mathbf{A}_{\sigma_{\mathbf{x}}}) = \#\sigma_{\mathbf{x}}$.

From Theorem 1, one can easily obtain a strict local minimizer of F_0 by choosing a support $\omega \in \mathbb{I}_N$ such that $\text{rank}(\mathbf{A}_\omega) = \#\omega$ and solving the restricted normal equations $(\mathbf{A}_\omega)^T\mathbf{A}_\omega\mathbf{x}_\omega = (\mathbf{A}_\omega)^T\mathbf{y}$. As the number of such supports is upper-bounded by $\sum_{k=0}^{\min(M,N)}\binom{N}{k}$, strict local optimality, although stronger than SO, is not really discriminant.

Another consequence of Theorem 1 is that the main difficulty in solving Problem (1) lies in the determination of the support of the solution. This naturally led researchers to investigate necessary optimality conditions that are based on the support. Most of them exploit the fact that the function value should not increase if one performs a small change in the support between SO points. For instance, a simple condition—which is at the core of greedy algorithms such as OLS [12] and SBR [43]—is to test if any insertion and/or removal in the support does not increase the objective function. Allowing a higher degree of complexity, these insertion and removal tests can be completed by any swap between support and non-support elements. However, the verification of such conditions requires the computation of at least N SO points (*i.e.,* solving at least N times (4)) which may lead to computational limitations for large-scale problems. Therefore, for the sake of tractability, we consider partial versions of these conditions as proposed in [4].

Definition 2 (Partial support coordinate-wise optimality [4]) A point $\mathbf{x} \in \mathbb{R}^N$ is said to be partial support coordinate-wise (CW) optimal for (1) if it is SO and verifies $\|\mathbf{x}\|_0 \leq \min\{M, N\}$ as well as

$$F_0(\mathbf{x}) \leq \min\{F_0(\mathbf{u}) : \mathbf{u} \in \mathcal{U}\}, \tag{7}$$

with

$$\mathcal{U} = \begin{cases} \{\mathbf{u}_\mathbf{x}^+\} & \text{if } \|\mathbf{x}\|_0 = 0, \\ \{\mathbf{u}_\mathbf{x}^-, \mathbf{u}_\mathbf{x}^{\text{swap}}, \mathbf{u}_\mathbf{x}^+\} & \text{if } \|\mathbf{x}\|_0 \in (0, \min\{M, N\}), \\ \{\mathbf{u}_\mathbf{x}^-\} & \text{if } \|\mathbf{x}\|_0 = \min\{M, N\}, \end{cases} \tag{8}$$

where $\mathbf{u}_\mathbf{x}^-$, $\mathbf{u}_\mathbf{x}^{\text{swap}}$, and $\mathbf{u}_\mathbf{x}^+$ are SO points defined by

$$\mathbf{u}_\mathbf{x}^- \in \left\{\arg\min_{\mathbf{u} \in \mathbb{R}^N} \|\mathbf{A}\mathbf{u} - \mathbf{y}\|^2 : \sigma_\mathbf{u} \subseteq \sigma_\mathbf{x} \backslash \{i_\mathbf{x}\}\right\}$$

$$\mathbf{u}_\mathbf{x}^{\text{swap}} \in \left\{\arg\min_{\mathbf{u} \in \mathbb{R}^N} \|\mathbf{A}\mathbf{u} - \mathbf{y}\|^2 : \sigma_\mathbf{u} \subseteq \sigma_\mathbf{x} \backslash \{i_\mathbf{x}\} \cup \{j_\mathbf{x}\}\right\}$$

$$\mathbf{u}_\mathbf{x}^+ \in \left\{\arg\min_{\mathbf{u} \in \mathbb{R}^N} \|\mathbf{A}\mathbf{u} - \mathbf{y}\|^2 : \sigma_\mathbf{u} \subseteq \sigma_\mathbf{x} \cup \{j_\mathbf{x}\}\right\}$$

for a chosen $i_\mathbf{x} \in \sigma_\mathbf{x}$ and $j_\mathbf{x} \in \mathbb{I}_N \backslash \sigma_\mathbf{x}$.

In Definition 2, there are different ways to choose $i_\mathbf{x} \in \sigma_\mathbf{x}$ and $j_\mathbf{x} \in \mathbb{I}_N \backslash \sigma_\mathbf{x}$. In this work, we follow the selection rule introduced by the authors in [4], which reads as follows,

$$i_\mathbf{x} \in \left\{\arg\min_{k \in \sigma_\mathbf{x}} |x_k|\right\}, \tag{9}$$

$$j_\mathbf{x} \in \left\{\arg\max_{k \in \mathbb{I}_N \backslash \sigma_\mathbf{x}} |\langle \mathbf{a}_k, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle|\right\}. \tag{10}$$

In other words, $i_\mathbf{x}$ corresponds to the element of the support with the smallest amplitude while $j_\mathbf{x}$ corresponds to the zero element for which the gradient of the data-fidelity term is maximal. Moreover it turns out that this choice is important to prove Theorem 6.

Remark 1 As opposed to [4], we refer to points verifying (7) as "partial *support* CW". We added the term "support" to prevent any confusion with the traditional definition of CW minimizers [2,21]. However, to simplify the presentation, we discard the analysis of traditional CW minimizers as they do not bring additional information to Figure 1. In addition, we slightly modified the definition of these partial support CW points, compared to the definition proposed in [4], with the introduction of equation (8). First, it allows to properly deal with the fact that $i_\mathbf{x}$ cannot be defined when $\mathbf{x} = \mathbf{0}_{\mathbb{R}^N}$ (let us recall that $\mathbf{0}_{\mathbb{R}^N}$ is always an SO point). Second, it avoids comparing $F_0(\mathbf{x})$ with $F_0(\mathbf{u}_\mathbf{x}^+)$ when $\|\mathbf{x}\|_0 = \min\{M, N\}$ as in that case $\mathbf{u}_\mathbf{x}^+$ is never uniquely defined.

We also consider another necessary optimality condition, namely L-stationarity, which became popular due to its relation with the iterative hard thresholding (IHT) algorithm [5]. More precisely, L-stationarity points for $L \geq \|\mathbf{A}\|^2$ are fixed points of the IHT algorithm [1,5].

Definition 3 (L-stationarity [47,4]) A point $\mathbf{x} \in \mathbb{R}^N$ is said to be L-stationary for (1) ($L > 0$), if

$$\mathbf{x} \in \left\{\arg\min_{\mathbf{u} \in \mathbb{R}^N} \frac{1}{2}\|T_L(\mathbf{x}) - \mathbf{u}\|^2 + \frac{\lambda}{L}\|\mathbf{u}\|_0\right\}, \tag{11}$$

where $T_L(\mathbf{x}) = \mathbf{x} - L^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{y})$.

Finally, although barely exploited so far, exact continuous relaxations (2) such as CEL0 open the door to new necessary optimality conditions for the initial Problem (1). Indeed, there is strong links between minimizers of F_0 and those of the CEL0 relaxation $\tilde{F}$. We recall these results in Theorem 2 and Corollary 1.

Theorem 2 (Links between (1) **and** (2)-(3) **[41, 11])** *Let $\mathcal{L}_0$ ($\tilde{\mathcal{L}}$, respectively) be the set of local minimizers of F_0 ($\tilde{F}$, respectively). Let $\mathcal{G}_0 \subseteq \mathcal{L}_0$ ($\tilde{\mathcal{G}} \subseteq \tilde{\mathcal{L}}$, respectively) be the corresponding subset of global minimizers. Then,*

1. *there exists a simple thresholding rule $\mathcal{T} : \mathbb{R}^N \to \mathbb{R}^N$, defined by*

$$[\mathcal{T}(\mathbf{x})]_i = \begin{cases} x_i & \text{if } |x_i| \geq \sqrt{2\lambda}/\|\mathbf{a}_i\|, \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

such that for any $\mathbf{x} \in \tilde{\mathcal{L}}$, $\mathcal{T}(\mathbf{x}) \in \mathcal{L}_0$

2. $\mathcal{G}_0 \subseteq \tilde{\mathcal{G}}$.

In other words, global minimizers of F_0 are preserved by $\tilde{F}$ and from each local minimizer of $\tilde{F}$ one can easily extract a local minimizer of F_0. Moreover, there is no converse property for the point 1 of Theorem 2. Hence, some local (not global) minimizers of F_0 can be removed by $\tilde{F}$.

To simplify the presentation and discard the need of the thresholding rule $\mathcal{T}$ in Theorem 2, we work under Assumption 1.

Assumption 1. *When F_0 does not admit a unique global minimizer, every pair $(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2)$ of global minimizers ($\hat{\mathbf{x}}_1 \neq \hat{\mathbf{x}}_2$) verify $\|\hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2\|_0 > 1$.*

Assumption 1 implies the following corollary of Theorem 2 whose proof is detailled in Appendix C.

Corollary 1 *Under Assumption 1, global minimizers of F_0 and $\tilde{F}$ coincide (i.e., $\mathcal{G}_0 = \tilde{\mathcal{G}}$). Moreover, they are stricts for both F_0 and $\tilde{F}$.*

A consequence of Corollary 1 is that, under Assumption 1, strict local optimality for $\tilde{F}$ is a necessary optimality condition for F_0.

Remark 2 Assumption 1 is always fulfilled when F_0 admits a unique global minimizer. Moreover, when F_0 admits multiple global minimizers, Assumption 1 breaks only for a finitely number of λ values (see Appendix B).

To conclude this section, known relationships between these necessary optimality conditions are illustrated in Figure 1 (black arrows). The associated references are provided in Table 1.

3 Description of the Strict Minimizers of $\tilde{F}$

In [41], only the relation between local (not global) minimizer of $\tilde{F}$ and minimizers of F_0 is studied. The question to know how to recognize critical points of $\tilde{F}$ which are local minimizers is not addressed. In this section, we derive necessary and sufficient conditions which ensure that a critical point of $\tilde{F}$ is a strict local minimizer.

We first recall the characterization of the critical points of CEL0.

Proposition 1 (Critical points of CEL0 [41]) *Let $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{z} \in \mathbb{R}^N$ be such that $z_i = \langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \rangle$. Then $\mathbf{x}$ is a critical point of the CEL0 relaxation $\tilde{F}$ if and only if $\forall i \in \mathbb{I}_N$*

$$x_i \in \begin{cases} \{0\} & \text{if } |z_i| < \sqrt{2\lambda}\|\mathbf{a}_i\| \\ -\mathrm{sign}(z_i) \times \left[0, \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|}\right] & \text{if } |z_i| = \sqrt{2\lambda}\|\mathbf{a}_i\| \\ \{-z_i/\|\mathbf{a}_i\|^2\} & \text{if } |z_i| > \sqrt{2\lambda}\|\mathbf{a}_i\| \end{cases} \quad (13)$$

Moreover, given a critical point $\mathbf{x} \in \mathbb{R}^N$ of CEL0, we introduce the two following sets,

$$\sigma_{\mathbf{x}}^- = \left\{ i \in \mathbb{I}_N : 0 < |x_i| < \tfrac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|} \right\}, \quad (14)$$

$$\sigma_{\mathbf{x}}^+ = \left\{ i \in \mathbb{I}_N : |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \rangle| = \sqrt{2\lambda}\|\mathbf{a}_i\| \right\}. \quad (15)$$

Clearly, from Proposition 1, we have $\sigma_{\mathbf{x}}^- \subseteq \sigma_{\mathbf{x}}^+$.

We now derive in Theorem 3 a necessary and sufficient condition to recognize critical points that are strict local minimizers of $\tilde{F}$. The proof is provided in Appendix D.

Theorem 3 (Strict local optimality for $\tilde{F}$) *A critical point $\mathbf{x} \in \mathbb{R}^N$ of $\tilde{F}$ is a strict local minimizer of $\tilde{F}$ if and only if $\sigma_{\mathbf{x}}^+ = \emptyset$ and $\mathrm{rank}(\mathbf{A}_{\sigma_\mathbf{x}}) = \#\sigma_\mathbf{x}$.*

4 New Relationships Between Optimality Conditions

In this section, we provide new relationships between the necessary optimality conditions depicted in Figure 1. We distinguish inclusion properties (blue arrows in Figure 1) from partial inclusion properties (dotted lines in Figure 1).

4.1 Inclusion Properties

In Theorems 4 and 5, we show that strict local minimizers of $\tilde{F}$ are L-stationary points and that partial support CW points are strict local minimizers of F_0. Moreover, we prove in Theorem 6 that partial support CW points are also strict local minimizers of $\tilde{F}$. These results constitute three new inclusion properties between the necessary optimality conditions of Figure 1. The proofs of Theorems 4 and 6 are provided in Appendices E and F, respectively.

Theorem 4 ($\mathbf{min}_{\mathbf{loc}}^{\mathbf{st}}\{\tilde{F}\} \Rightarrow$ L-stationary) *Let $\mathbf{x} \in \mathbb{R}^N$ be a strict local minimizer of $\tilde{F}$. Then it is a L-stationary point of (1) for any $L \geq \max_{i \in \mathbb{I}_N} \|\mathbf{a}_i\|^2$.*

Note that $\|\mathbf{A}\|^2 = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|^2 \geq \max_{i \in \mathbb{I}_N} \|\mathbf{a}_i\|^2$ and thus Theorem 4 includes $L \geq \|\mathbf{A}\|^2$.

Theorem 5 ($\mathbf{SuppCW_{partial}}\{F_0\} \Rightarrow \mathbf{min}_{\mathbf{loc}}^{\mathbf{st}}\{F_0\}$) *Let $\mathbf{A}$ satisfy the unique representation property (URP)[1]. Let $\mathbf{x} \in \mathbb{R}^N$ be a partial support CW point of (1) as specified in Definition 2. Then it is a strict local minimizer of F_0.*

[1] A matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ satisfies the URP [22] if any $\min\{M, N\}$ columns of $\mathbf{A}$ are linearly independent.

Proof. Let $\mathbf{x} \in \mathbb{R}^N$ be a partial support CW point. Then, from Definition 2, it is an SO point (*i.e.,* a local minimizer of F_0) such that $\|\mathbf{x}\|_0 \leq \min\{M, N\}$. It follows that $\mathrm{rank}(\mathbf{A}_{\sigma_\mathbf{x}}) = \#\sigma_\mathbf{x}$ because $\mathbf{A}$ satisfies the URP. Finally, Theorem 1 allows to conclude that $\mathbf{x}$ is a strict local minimizer of F_0. □

Theorem 6 ($\mathbf{SuppCW_{partial}}\{\boldsymbol{F_0}\} \Rightarrow \mathbf{min^{st}_{loc}}\{\tilde{\boldsymbol{F}}\}$) *Let* $\mathbf{A}$ *satisfy the URP and have unit norm columns. Then, for all* $\lambda \in \mathbb{R}_{>0} \backslash \Lambda$ *(where* Λ *is a subset of* $\mathbb{R}_{>0}$ *whose Lebesgue measure is zero), each partial support CW point of* (1) *is a strict local minimizer of* $\tilde{F}$*.*

Remark 3 The unit norm assumption for the columns of $\mathbf{A}$ in Theorem 6 can be easily relaxed by modifying the definition of $i_\mathbf{x}$ and $j_\mathbf{x}$ in (9) and (10) as

$$i_\mathbf{x} \in \left\{ \arg \min_{k \in \sigma_\mathbf{x}} \|\mathbf{a}_k\| \, |x_k| \right\}, \tag{16}$$

$$j_\mathbf{x} \in \left\{ \arg \max_{k \in \mathbb{I}_N \backslash \sigma_\mathbf{x}} |\langle \mathbf{a}_k, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle| / \|\mathbf{a}_k\| \right\}. \tag{17}$$

Remark 4 Inquisitive minds will have observed that the proof of Theorem 6 does not make use of $\mathbf{u}_\mathbf{x}^{\mathrm{swap}}$ in Definition 2. Hence, the result of Theorem 6 is also valid for a weaker partial support CW optimality condition that involves only insertions and deletions (no swap) on the support. Note that the connections between this alternative optimality condition and the L-stationarity as well as the strict local optimality of for F_0 are then trivial by chaining.

4.2 A Partial Inclusion Property

We provide here some evidence that the relationship between L-stationarity and strict local optimality for F_0 does not constitute an inclusion property.

Given a strict local minimizer $\mathbf{x}$ of F_0 and $L > 0$, one can easily get a value of λ (from Lemma 2 in Appendix A) for which $\mathbf{x}$ is not a L-stationary point. This fact is illustrated in Figure 2 (see Section 5) which reveals that some strict local minimizers of F_0 are not L-stationary points. Conversely, L-stationary points are not necessarily strict local minimizers of F_0 as their characterization in Lemma 2 does not impose $\mathbf{A}_{\sigma_\mathbf{x}}$ to be full rank. We refer the reader to [4, Example 4.18] for an example of a problem admitting an infinite number of L-stationary points which are thus nonstrict local minimizers of F_0.

5 Quantifying Optimal Points

The relationships between the necessary optimality conditions that are summarized in Figure 1 and Table 1 are qualitative and not quantitative. The "restrictiveness" of these conditions is not reflected by these theoretical results. However, this notion is of a fundamental importance as one seeks necessary optimality conditions with a strong power of discrimination among candidate solutions (*i.e.,* local minimizers of F_0).

From now on, we denote by $\mathcal{S}_0$ the set of strict local minimizers of F_0. According to Theorem 1, $\mathcal{S}_0$ contains a finite number of points. Then, we define three subsets of $\mathcal{S}_0$, namely

- $\tilde{\mathcal{S}} = \{\mathbf{x} \in \mathcal{S}_0 : \mathbf{x} \text{ strict local minimizer of } \tilde{F}\}$
- $\mathcal{S}_{\mathrm{CW}} = \{\mathbf{x} \in \mathcal{S}_0 : \mathbf{x} \text{ partial support CW point}\}$
- $\mathcal{S}_{\mathrm{L}} = \{\mathbf{x} \in \mathcal{S}_0 : \mathbf{x} \text{ L-stationary point}\}$

From the diagram in Figure 1, these sets are non-empty as they contain at least the global minimizer(s) of Problem (1) (see [32] for a discussion about the existence of global minimizers). Moreover, because these sets are defined as subsets of $\mathcal{S}_0$, one may wonder if all strict local minimizers of $\tilde{F}$ are contained in $\tilde{\mathcal{S}}$ (and similarly for $\mathcal{S}_{\mathrm{CW}}$ and $\mathcal{S}_{\mathrm{L}}$). Actually, this is not the case for $\mathcal{S}_{\mathrm{L}}$ (see Section 4.2). However, from [41, Corollary 4.9] (Theorem 5, respectively) we have that $\tilde{\mathcal{S}}$ ($\mathcal{S}_{\mathrm{CW}}$, respectively) contains all the strict local minimizers of $\tilde{F}$ (all the partial support CW points, under the URP of $\mathbf{A}$, respectively).

5.1 Numerical Experiment

Our goal is to quantify the number of strict local minimizers[2] of F_0 that satisfy a given necessary optimality condition (*e.g.,* L-stationarity, strict local optimality for $\tilde{F}$, partial support-CW optimality). Hence, we first need to compute all strict local minimizers of F_0. Following Theorem 1, this can be achieved by solving

$$(\mathbf{A}_\omega)^T \mathbf{A}_\omega \mathbf{x}_\omega = (\mathbf{A}_\omega)^T \mathbf{y} \tag{18}$$

for any support $\omega \in \bar{\Omega}$ where

$$\bar{\Omega} = \bigcup_{r=0}^{M} \Omega_r, \; \Omega_r = \{\omega \in \mathbb{I}_N : \#\omega = r = \mathrm{rank}(\mathbf{A}_\omega)\}. \tag{19}$$

Because the cardinality of $\bar{\Omega}$ explodes quickly with the dimension N (and M), we restrict this experiment to small-size problems ($M = 5$ and $N = 10$). Then, it is noteworthy to mention that the computation of strict local minimizers of F_0 does not depend on λ [32, Remark 5]. Hence, F_0 admits the same set of strict local minimizers for any value of λ (*i.e.,* $\mathcal{S}_0$ is independent of λ), which is not the case for the others optimality conditions that we study. For a set of values of λ, we

[2] Nonstrict local minimizers are uncountable by definition.

thus determine the amount of strict local minimizers of F_0 that verify a given necessary optimality condition.

To summarize, given $\mathbf{A} \in \mathbb{R}^{5\times 10}$ and $\mathbf{y} \in \mathbb{R}^5$, we proceed as follows:

1. Compute all strict local minimizers of $F_0 \to \mathcal{S}_0$,
2. For each $\lambda \in \{\lambda_1, \dots, \lambda_P\}$, determine the subset of $\mathcal{S}_0$ that contains points verifying a given necessary optimality condition (*i.e.*, $\tilde{\mathcal{S}}$, $\mathcal{S}_{\mathrm{CW}}$, $\mathcal{S}_{\mathrm{L}}$),
3. Repeat 1-2 for different $\mathbf{A} \in \mathbb{R}^{5\times 10}$ and $\mathbf{y} \in \mathbb{R}^5$, and draw the average evolution of $\#\tilde{\mathcal{S}}$, $\#\mathcal{S}_{\mathrm{CW}}$, and $\#\mathcal{S}_{\mathrm{L}}$, with respect to λ.

We consider three ways of generating $\mathbf{A} \in \mathbb{R}^{5\times 10}$ and $\mathbf{y} \in \mathbb{R}^5$,

- The entries of $\mathbf{A}$ and $\mathbf{y}$ are drawn from a zero mean unit variance normal distribution,
- The entries of $\mathbf{A}$ and $\mathbf{y}$ are drawn from a uniform distribution on $[0, 1]$,
- $\mathbf{A}$ is a "sampled Toeplitz" matrix built from a Gaussian kernel ($A_{ij} = \exp(-(x_i - 0.5(j-1)/M)^2/(2\sigma^2))$ where $\{x_i\}_{i=1}^M$ are M uniform sampling points of $[0, 1]$ and $\sigma^2 = 0.04$). The entries of $\mathbf{y}$ are drawn from a zero mean unit variance normal distribution.

Remark 5 Note that we have $\#\mathcal{S}_0 \le \#\bar{\Omega}$, *i.e.*, two supports $(\omega, \omega') \in \bar{\Omega}^2$ can lead to the same strict local minimizer $\mathbf{x}$. The equality or the strict inequality depends on $\mathbf{y}$. However, the equality is verified for all $\mathbf{y} \in \mathbb{R}^M \backslash \mathcal{Y}$, where

$$\mathcal{Y} = \bigcup_{\omega \in \bar{\Omega}} \bigcup_{i=1}^{\#\omega} \{\mathbf{y} \in \mathbb{R}^M : \langle \mathbf{e}_i, (\mathbf{A}_\omega^T \mathbf{A}_\omega)^{-1} \mathbf{A}_\omega^T \mathbf{y} \rangle = 0\}. \quad (20)$$

Moreover, $\mathbb{R}^M \backslash \mathcal{Y}$ contains a dense open subset in $\mathbb{R}^M$ [32, Lemma 3.7]. Hence, in our experiments, we control that $\mathbf{y} \in \mathbb{R}^M \backslash \mathcal{Y}$. This ensures that all the strict local minimizers computed with (18)-(19) are distinct.

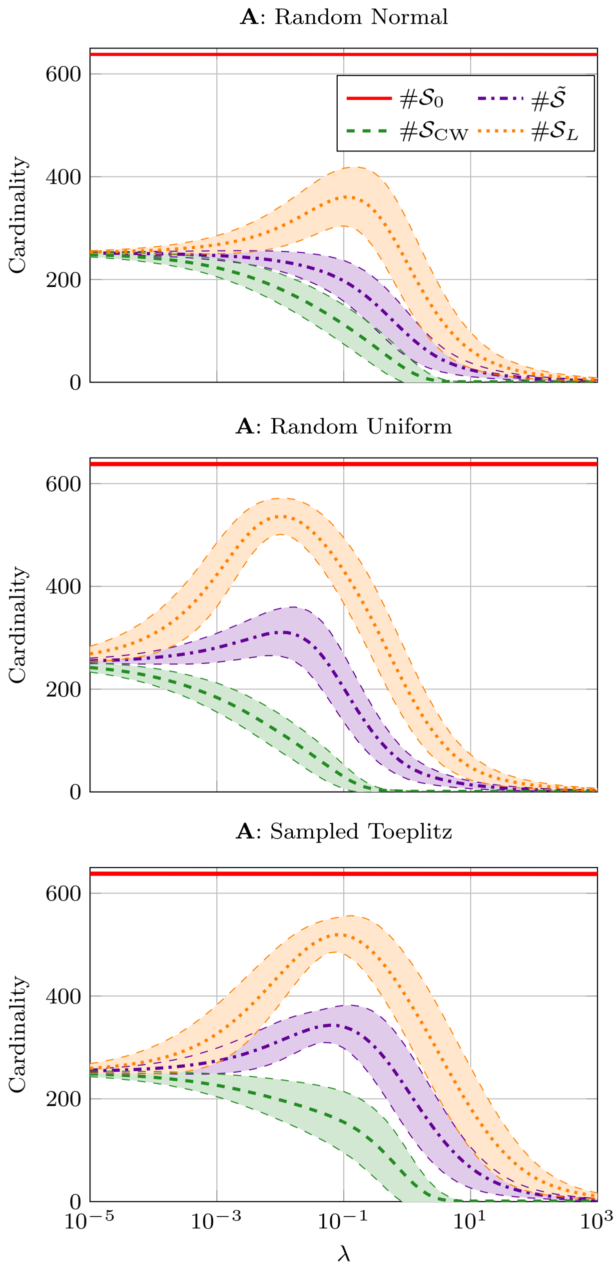

Fig. 2: Cardinality of $\tilde{\mathcal{S}}$, $\mathcal{S}_{\mathrm{CW}}$, and $\mathcal{S}_{\mathrm{L}}$, with respect to λ. The curves correspond to an average value (with standard deviation) over 1000 generations of $\mathbf{A} \in \mathbb{R}^{5\times 10}$ and $\mathbf{y} \in \mathbb{R}^5$. As a reference, we plot the value of $\#\mathcal{S}_0$ that does not depend on λ [32, Remark 5].

The evolution of $\#\tilde{\mathcal{S}}$, $\#\mathcal{S}_{\mathrm{CW}}$, and $\#\mathcal{S}_{\mathrm{L}}$ with respect to λ is depicted in Figure 2. First, one can see that both strict local optimality for $\tilde{F}$ and partial support CW optimality are stronger conditions than L-stationarity. This illustrates and completes the results provided by Theorem 4 and [4, Theorem 4.17]. Second, in light of Theorem 6, the results reported in Figure 2 reveal that, in general, there is less partial support CW points that strict local minimizers of $\tilde{F}$.

Then, a remarkable observation is that, for large and small values of λ, these three necessary optimality conditions are equivalent. Moreover, this is true for any operator $\mathbf{A}$ and we shall provide a theoretical justification of this behaviour in Section 5.2. Finally, for intermediate values of λ, we observe different behaviours depending on the choice of $\mathbf{A}$. In particular, the situation is more favorable when the entries of $\mathbf{A}$ are generated from a normal distribution rather than a uniform distribution or a sampled Toeplitz matrix. Given that the former (*i.e.*, i.i.d. normal entries) leads to matrices $\mathbf{A}$ with good restricted isometry property (RIP), this observation is supported by the recent work [10].

5.2 Theoretical Explanation

We provide in Theorem 7 a theoretical justification of the observations made with the experiment presented in Section 5.1. The proof is detailled in Appendix G.

Theorem 7 *Let $\mathcal{S}_0$ be the set of strict local minimizers of F_0. Let $\tilde{\mathcal{S}}$, $\mathcal{S}_{\mathrm{CW}}$, and $\mathcal{S}_{\mathrm{L}}$ be the subsets of $\mathcal{S}_0$ containing the strict local minimizers of $\tilde{F}$, the partial support CW points, and the L-stationary points, respectively. Finally, define*

$$\mathcal{X}_{\mathrm{LS}} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2, \tag{21}$$

the solution set of the un-penalized least-squares problem. Then, for all $\mathcal{S} \in \{\tilde{\mathcal{S}}, \mathcal{S}_{\mathrm{L}}, \mathcal{S}_{\mathrm{CW}}\}$, there exists (under the URP of $\mathbf{A}$ for $\mathcal{S}_{\mathrm{CW}}$) $\lambda_0 > 0$ and $\lambda_\infty > 0$ such that

1. $\forall\lambda \in (0, \lambda_0)$, $\mathcal{S} = (\mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}})$,
2. $\forall\lambda \in (\lambda_\infty, +\infty)$, $\mathcal{S} = \{\mathbf{0}_{\mathbb{R}^N}\}$.

Hence, when λ is sufficiently small, the three studied necessary optimality conditions are equivalent in the sense that, among the strict local minimizers of F_0, they preserve only those which are solution of the un-penalized least squares problem. For the experiment reported in Figure 2, and following Remark 5, there is exactly $\binom{10}{5} = 252$ 5-sparse strict minimizers of F_0 (under the URP of $\mathbf{A}$ and $\mathbf{y} \notin \mathbb{R}^M \backslash \mathcal{Y}$ defined in Remark 5) that are also solutions of the un-penalized least-squares problem. This value corresponds to the limiting value for small λ that we observe in Figure 2. Then, the second point of Theorem 7 states that these necessary optimality conditions are also equivalent when λ is large as they allow to discard all the non-zero strict local minimizers of F_0. Note that, for such large values of λ, $\mathbf{0}_{\mathbb{R}^N}$ is the global minimizer of F_0.

Although $\tilde{\mathcal{S}}$, $\mathcal{S}_{\mathrm{CW}}$, and $\mathcal{S}_{\mathrm{L}}$ are completely characterized by Theorem 7 for extreme values of λ, a theoretical analysis of theses sets for intermediates values of λ remains an interesting open question. Clearly, from the experiments of Figure 2, such an analysis will depend on the operator $\mathbf{A}$.

6 Algorithms

Let $\mathcal{A}$ denotes an algorithm that is proven to converge to a point that satisfy one of the necessary optimality conditions studied in the present paper. Then, according to the inclusion properties of Table 1 as well as the analysis conducted in Section 5, one can expect that the efficiency of $\mathcal{A}$ to minimize F_0 depends on the necessary optimality condition it guarantees to converge to. In this section, we propose a numerical illustration of this claim. We consider four algorithms:

- CowS: the CW support optimatity (CowS) algorithm. It is a greedy method that converges to a support CW point [4].
- IHT: the iterative hard thresholding (IHT) algorithm that ensures the convergence to an L-stationary point [1,4,5].
- FBS-CEL0: the forward-backward splitting (FBS) algorithm applied to the CEL0 exact relaxation $\tilde{F}$ defined by equations (2)-(3). FBS ensures the convergence to a stationary point of $\tilde{F}$ [1].
- IRL1-CEL0: the iterative reweighted-ℓ_1 (IRL1) algorithm [35] also used to obtain a stationary point of the CEL0 exact relaxation $\tilde{F}$.

Note that FBS-CEL0 and IRL1-CEL0 do not ensure the convergence to a strict local minimizer of $\tilde{F}$, but only the convergence to a stationary point of $\tilde{F}$. Hence, the results reported hereafter for the use of the CEL0 relaxation could be improved by the design of an algorithm that ensures the convergence to a strict local minimizer of $\tilde{F}$.

6.1 Description of the Experiment

For $K = 50$ instances of Problem (1) (*i.e.,* instances of $\mathbf{A}$ and $\mathbf{y}$) and two values of the penalty parameter $\lambda \in \{10^{-8}, 10^{-3}\}$, we execute the four algorithms with the initialization $\mathbf{x}^0 = \mathbf{0}_{\mathbb{R}^N}$. Following the experiment conducted in Section 5.1, we consider three ways of generating the matrix $\mathbf{A}$ of size $M = 100$ and $N = 256$: (i) i.i.d. entries drawn from a normal distribution, (ii) i.i.d. entries drawn from a uniform distribution, (iii) "sampled Toeplitz" matrix with a Gaussian kernel ($A_{ij} = \exp(-(x_i - 0.4(j-1)/M)^2/(2\sigma^2))$ where $\{x_i\}_{i=1}^M$ are M uniform sampling points of $[0,1]$ and $\sigma^2 = 10^{-4}$). Then, the columns of $\mathbf{A}$ are normalized and the measurements $\mathbf{y} \in \mathbb{R}^M$ are generated according to

$$\mathbf{y} = \mathbf{A}\mathbf{x}^\star + \mathbf{n}, \tag{22}$$

where $\mathbf{x}^\star$ is a 30-sparse vector (*i.e.,* $\|\mathbf{x}^\star\|_0 = 30$) whose non-zero entries are drawn from a normal distribution, and $\mathbf{n}$ is a vector of Gaussian noise with standard deviation 10^{-2}.

6.2 Results

For each algorithm and each instance of Problem (1), Figure 3 reports the value of the ratio $F_0(\hat{\mathbf{x}})/F_0(\mathbf{x}^0)$, where $\hat{\mathbf{x}} \in \mathbb{R}^N$ is the output of the algorithm.

For $\lambda = 10^{-8}$, we observe the same behaviour independently of the way $\mathbf{A}$ is generated. Indeed, this situation corresponds to the small λ regime described by

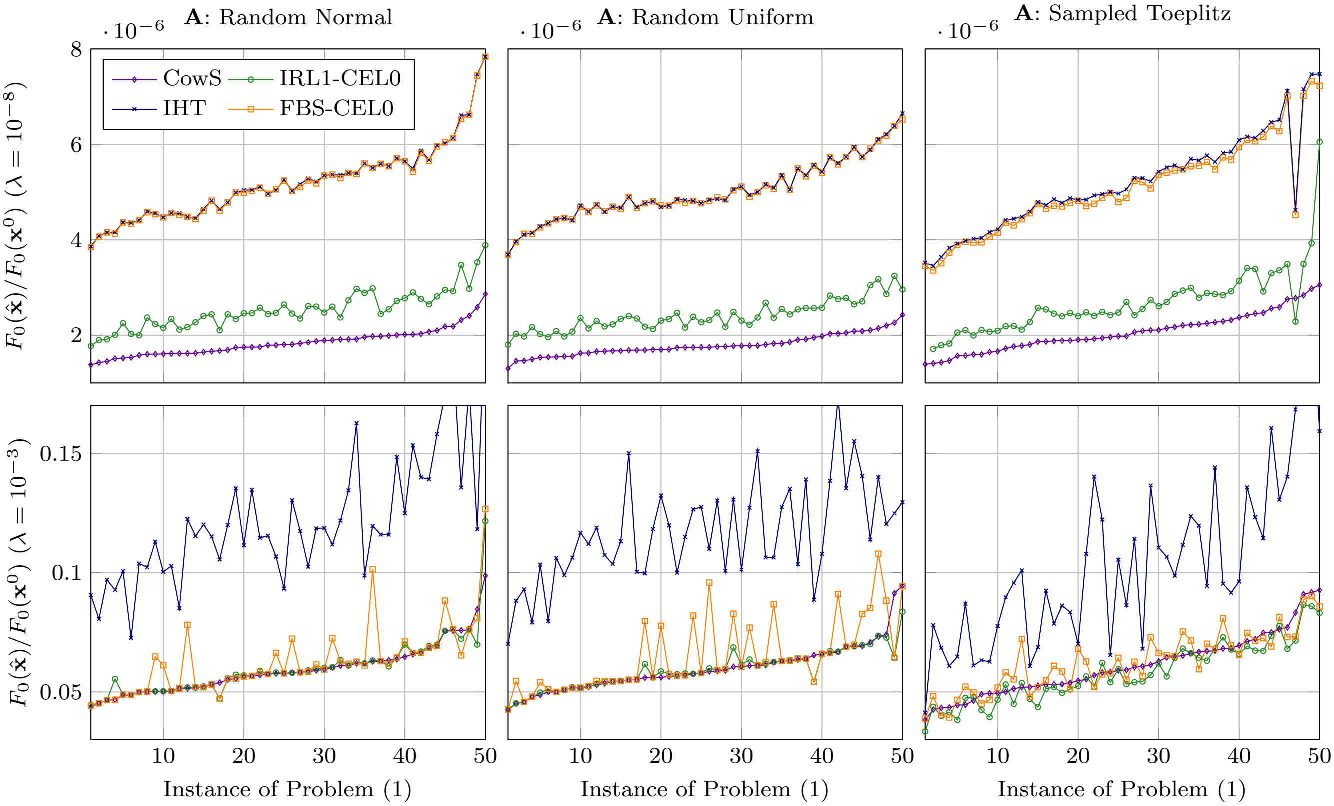

Fig. 3: Value of the ratio $F_0(\hat{\mathbf{x}})/F_0(\mathbf{x}^0)$ where $\hat{\mathbf{x}}$ is the output of the minimization algorithm and $\mathbf{x}^0 = \mathbf{0}_{\mathbb{R}^N}$ is the initial point. For each type of matrix $\mathbf{A}$, results are reported for $K = 50$ instances of Problem (1) with two values of λ, and for the four algorithms CowS, IHT, FBS-CEL0, and IRL1-CEL0. The K instances of Problem (1) are sorted according to the value of the ratio $F_0(\hat{\mathbf{x}})/F_0(\mathbf{x}^0)$ obtained with the CowS method.

Theorem 7 and illustrated by the experiment reported in Section 5.1. The differences that one can observe between the algorithms come from the fact that IHT, FBS-CEL0, and IRL1-CEL0, converge to nonstrict local minimizers of F_0 such that $\|\hat{\mathbf{x}}\|_0 > M$ (such points are not studied in Section 5 as they are uncountable). More precisely, on average we have $\|\hat{\mathbf{x}}_{\text{FBS}-\text{CEL0}}\|_0 = \|\hat{\mathbf{x}}_{\text{IHT}}\|_0 = N$ and $M < \|\hat{\mathbf{x}}_{\text{IRL1}-\text{CEL0}}\|_0 < N$, which explain the observed differences between the corresponding curves. As opposed to these three methods, the CowS algorithm always converges to a point $\hat{\mathbf{x}}$ with $\|\hat{\mathbf{x}}\|_0 \leq M$. In particular, this is due to its greedy behaviour. Note that we propose this small λ situation as an additional illustration of the results developed in Section 5. Indeed, this regime (*i.e.,* $\lambda < \lambda_0$ of Theorem 7) has no practical interest as one can easily obtain a global minimizer of F_0 by selecting a support $\omega \subseteq \mathbb{I}_N$ such that $\text{rank}(\mathbf{A}_\omega) = \text{rank}(\mathbf{A})$. For similar reasons, we do not present the case for a large value of λ.

The interesting regime corresponds to "intermediate" values of λ which are represented by $\lambda = 10^{-3}$ in Figure 3. One can clearly see that IHT—that ensures the convergence to an L-stationary point—presents the worst performance. This observation is in agreement with the results of Section 5 as Cows, FBS-CEL0, and IRL1-CEL0 ensure a convergence to a point verifying a stronger optimality condition than L-stationarity. Then, when the entries of $\mathbf{A}$ are generated according to normal and uniform distributions, the CowS method tends to provide slightly better results than the minimization of the CEL0 relaxation $\tilde{F}$ with either FBS or IRL1. However, for a sampled Toeplitz matrix $\mathbf{A}$, the results are more mitigated, showing that both the reached necessary optimality condition and the algorithm are driving the quality of the minimization.

7 Discussion

Support-Based Optimality Conditions One can easily get alternative support-based necessary optimality conditions by i) modifying the selection rule in Definition 2, or ii) modifying the set $\mathcal{U}$ in (8) (*i.e.,* the neighbours supports). By increasing $\mathcal{U}$, one would likely define a stronger necessary optimality condition. The price to pay, however, is a larger computational cost to verify this condition for a given point $\mathbf{x} \in \mathbb{R}^N$. Furthermore, this cost is directly related to the one of the greedy method that can be derived from such a support-based condition.

Hence, there is a trade-off to find between the restrictiveness of the condition and the computational burden it generates. For instance, in light of Remark 4, $\mathbf{u}_{\mathbf{x}}^{\text{swap}}$ can be removed from $\mathcal{U}$ in Definition 2 without any change in the hierarchy presented in Figure 1. This leads to a (probably) weaker condition but faster to verify as $\mathbf{u}_{\mathbf{x}}^{\text{swap}}$ do not need to be computed. One can also define a heavier but stronger condition that compares the current value of $F_0(\mathbf{x})$ to the one obtained by any insertion or deletion in the support of $\mathbf{x}$. Actually, this condition is at the core of the single best replacement (SBR) algorithm [43]. Hence, as a by product of the results developed in the present paper, we get that the SBR algorithm is ensured to converge to a strict local minimizer of the CEL0 relaxation $\tilde{F}$. Finally, note that adding to the optimality condition used in SBR the test of any swap in the support of $\mathbf{x}$ is computationally prohibitive as one would have to solve $\#\sigma_{\mathbf{x}} \times (N - \#\sigma_{\mathbf{x}})$ additional linear systems at each iteration of the greedy method.

Greedy versus Variational Approaches While support-based conditions naturally lead to greedy approaches, exact continuous relaxations of F_0 [11,42], which include CEL0 [41], open the door to a variety of variational approaches dedicated to nonsmooth nonconvex optimization. It is noteworthy to mention that these methods could not (in general) be used directly on the initial problem (due to the discontinuity of F_0). In this work, we restricted our analysis to optimality conditions that derive from the CEL0 functional itself, rather than the algorithm used to minimize it (as opposed to the L-stationarity for F_0). Hence, our results are independent of any algorithm. Yet, given an algorithm, a specific analysis of its fixed points would also be of interest as some strict local minimizers (or critical points) of $\tilde{F}$ can potentially not be fixed points of the algorithm.

The inclusion property derived in Theorem 6 as well as the numerical experiment reported in Section 5 play in favor of greedy-based necessary conditions. However, the results presented in Section 6 reveal that, in practice, the associated algorithms are comparable in terms of their ability to minimize F_0. This observation is in line with the fact that the algorithm itself can escape from non-(globally)-optimal points of $\tilde{F}$. Moreover, for large-scale problems such as in imaging sciences, the computational burden of greedy methods may make the use of exact relaxations together with variational approaches preferable. Finally, we would like to stress out that, for moderate-size problems, the exact continuous relaxation $\tilde{F}$ can be globally minimized using Lasserre's hierarchies [26].

Appendices

A Preliminary Lemmas

In this section we provide two technical lemmas that are used in some of the proofs detailed in the next appendices. The following developments make use of the notations $\sigma_{\mathbf{x}}^-$ and $\sigma_{\mathbf{x}}^+$ that are defined in (14) and (15), respectively. Other notations can be found in Section 1.4.

Lemma 1 *Let $\mathbf{x} \in \mathbb{R}^N$ be a local minimizer of $\tilde{F}$ and set $s_i = \operatorname{sign}(\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \rangle)$. Then,*

1. *$\forall i \in \sigma_{\mathbf{x}}^+$, $\exists \mathcal{T}_i \subseteq [0, \sqrt{2\lambda}/\|a_i\|]$, a non-degenerate interval of $\mathbb{R}$, such that $|x_i| \in \mathcal{T}_i$ and $\forall t \in \mathcal{T}_i$,*
$$\bar{\mathbf{x}} = \mathbf{x}^{(i)} - s_i t \mathbf{e}_i$$
is a local minimizer of $\tilde{F}$.
2. *if $\mathbf{x}$ is a global minimizer, then $\forall i \in \sigma_{\mathbf{x}}^+$, $\mathcal{T}_i = [0, \sqrt{2\lambda}/\|a_i\|]$ and $\bar{\mathbf{x}}$ is a global minimizer.*

Proof. Let $i \in \sigma_{\mathbf{x}}^+$ and $f : [0, \sqrt{2\lambda}/\|\mathbf{a}_i\|] \to \mathbb{R}$ be the restriction of $\tilde{F}$ defined by

$$f(t) = \tilde{F}(\mathbf{x}^{(i)} - s_i t \mathbf{e}_i), \ \forall t \in \left[0, \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|}\right]. \tag{23}$$

Denoting, $\phi_i(x) = \lambda - \frac{\|\mathbf{a}_i\|^2}{2}\left(|x| - \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|}\right)^2 \mathbb{1}_{\left\{|x| \le \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|}\right\}}$, we can rewrite f as

$$\begin{aligned} f(t) &= \frac{1}{2}\|\mathbf{A}\mathbf{x}^{(i)} - s_i t \mathbf{a}_i - \mathbf{y}\|^2 + \sum_{j \neq i} \phi_j(x_j) + \phi_i(-s_i t) \\ &= C - s_i t \langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \rangle + \frac{\|\mathbf{a}_i\|^2}{2} t^2 + \lambda \\ &\qquad - \frac{\|\mathbf{a}_i\|^2}{2}\left(|s_i t| - \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|}\right)^2 \\ &= C - s_i t \langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \rangle + t\sqrt{2\lambda}\|\mathbf{a}_i\| \\ &= C + t(\sqrt{2\lambda}\|\mathbf{a}_i\| - |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \rangle|) \\ &= C \in \mathbb{R}, \end{aligned} \tag{24}$$

where $C = \frac{1}{2}\|\mathbf{A}\mathbf{x}^{(i)} - \mathbf{y}\|^2 + \sum_{j \neq i} \phi_j(x_j)$ is a constant independent of t. The last equality comes from the fact that, by definition (eq. (15)), $i \in \sigma_{\mathbf{x}}^+ \Rightarrow |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \rangle| = \sqrt{2\lambda}\|\mathbf{a}_i\|$.

We now show the two assersions of Lemma 1.

1. Because $\mathbf{x}$ is a local minimizer of $\tilde{F}$, there exists $\eta > 0$ such that,
$$\forall \mathbf{u} \in \mathcal{B}_2(\mathbf{x}, \eta),\ \tilde{F}(\mathbf{x}) \leq \tilde{F}(\mathbf{u}). \tag{25}$$
Then, $\forall i \in \sigma^+_{\mathbf{x}}$ let
$$\mathcal{T}_i = \left\{ t \in \left[0, \frac{\sqrt{2\lambda}}{\|a_i\|}\right], \mathbf{u} = \mathbf{x}^{(i)} - s_i t \mathbf{e}_i \in \mathcal{B}_2(\mathbf{x}, \eta) \right\}.$$
Clearly, because $\eta > 0$, $\mathcal{T}_i$ is a non-degenerate interval of $\mathbb{R}$. Then,
$$\forall t \in \mathcal{T}_i,\ \exists \eta' \in (0, \eta),\ \text{ s.t. } \mathcal{B}_2(\bar{\mathbf{x}}, \eta') \subset \mathcal{B}_2(\mathbf{x}, \eta), \tag{26}$$
where $\bar{\mathbf{x}} = \mathbf{x}^{(i)} - s_i t \mathbf{e}_i$, and we get
$$\forall \mathbf{u} \in \mathcal{B}_2(\bar{\mathbf{x}}, \eta'),\ \tilde{F}(\bar{\mathbf{x}}) \underset{(24)}{=} \tilde{F}(\mathbf{x}) \underset{(25)\&(26)}{\leq} \tilde{F}(\mathbf{u}), \tag{27}$$
which completes the proof of the first assertion of Lemma 1.
2. Using the fact that $\mathbf{x}$ is a global minimizer of $\tilde{F}$, (24) completes the proof.

□

Lemma 2 *A point $\mathbf{x} \in \mathbb{R}^N$ is L-stationary for $L > 0$ if and only if it is SO and*
$$\begin{cases} |\left\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \right\rangle| \leq \sqrt{2\lambda L} & \text{if } i \in \mathbb{I}_N \backslash \sigma_{\mathbf{x}}, \\ |\left\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \right\rangle| \geq \sqrt{2\lambda / L}\|\mathbf{a}_i\|^2 & \text{if } i \in \sigma_{\mathbf{x}}. \end{cases} \tag{28}$$

Proof. Let $\mathbf{x} \in \mathbb{R}^N$ be a L-stationary point for $L > 0$. Then, from Definition 3, $\mathbf{x}$ verifies (11) which is equivalent, by separability, to: $\forall i \in \mathbb{I}_N$,
$$x_i \in \left\{ \arg\min_{u \in \mathbb{R}} \frac{1}{2}([T_L(\mathbf{x})]_i - u)^2 + \frac{\lambda}{L}|u|_0 \right\}, \tag{29}$$
$$\iff x_i \in \begin{cases} \{0\} & \text{if } |[T_L(\mathbf{x})]_i| < \sqrt{2\lambda/L}, \\ \{0, [T_L(\mathbf{x})]_i\} & \text{if } |[T_L(\mathbf{x})]_i| = \sqrt{2\lambda/L}, \\ \{[T_L(\mathbf{x})]_i\} & \text{if } |[T_L(\mathbf{x})]_i| > \sqrt{2\lambda/L}. \end{cases} \tag{30}$$
Hence, we now shall show that (30) is equivalent to $\mathbf{x}$ SO and (28). We proceed by proving both implications.

$\Longrightarrow$ Let $\mathbf{x}$ be a L-stationary point, then it is SO from [4, Theorem 4.11]. Hence, it follows from Definition 1 that
$$\forall i \in \sigma_{\mathbf{x}},\ 0 = \langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle, \tag{31}$$
$$\iff \forall i \in \sigma_{\mathbf{x}},\ x_i = -\left\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \right\rangle / \|\mathbf{a}_i\|^2. \tag{32}$$
Combining that fact with the expression of $[T_L(\mathbf{x})]_i = x_i - L^{-1}\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle$ we obtain
$$\begin{cases} [T_L(\mathbf{x})]_i = -L^{-1}\left\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \right\rangle & \text{if } i \in \mathbb{I}_N \backslash \sigma_{\mathbf{x}}, \\ [T_L(\mathbf{x})]_i = -\left\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y} \right\rangle / \|\mathbf{a}_i\|^2 & \text{if } i \in \sigma_{\mathbf{x}}. \end{cases} \tag{33}$$
Finally, by injecting (33) into (30) we get (28).

$\Longleftarrow$ Let $\mathbf{x}$ be a SO point such that (28) is verified. As previously, the SO property implies (31)-(32), and thus (33). Finally, injecting (33) into (28) completes the proof. □

B Breaking Assumption 1

For $\lambda > 0$, let $\hat{\mathbf{x}}_1 \in \mathbb{R}^N$ and $\hat{\mathbf{x}}_2 \in \mathbb{R}^N$ be two global minimizers of F_0 such that $\|\hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2\|_0 = 1$. (Note that $\|\hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2\|_0 = 0$ would imply that $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2$.) Then, $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ differ from only one component. Moreover, because global minimizers of F_0 are stricts [32, Theorem 4.4], we necessarily have $\|\hat{\mathbf{x}}_2\|_0 = \|\hat{\mathbf{x}}_1\|_0 - 1$ (or $\|\hat{\mathbf{x}}_1\|_0 = \|\hat{\mathbf{x}}_2\|_0 - 1$ by reversing the role of $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$). It then follows that
$$F_0(\hat{\mathbf{x}}_1) = F_0(\hat{\mathbf{x}}_2) \tag{34}$$
$$\iff \frac{1}{2}\|\mathbf{A}\hat{\mathbf{x}}_1 - \mathbf{y}\|^2 + \lambda\|\hat{\mathbf{x}}_1\|_0 = \frac{1}{2}\|\mathbf{A}\hat{\mathbf{x}}_2 - \mathbf{y}\|^2 + \lambda(\|\hat{\mathbf{x}}_1\|_0 - 1) \tag{35}$$
$$\iff \lambda = \frac{1}{2}\left(\|\mathbf{A}\hat{\mathbf{x}}_2 - \mathbf{y}\|^2 - \|\mathbf{A}\hat{\mathbf{x}}_1 - \mathbf{y}\|^2\right) \tag{36}$$
Hence two such points $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ can be both global minimizers of F_0 for only one value of the regularization parameter λ. This shows that, when F_0 admits multiple global minimizers, Assumption 1 eventually breaks only for a finitely number of λ values.

C Proof of Corollary 1

Let $\mathcal{G}_0$ and $\tilde{\mathcal{G}}$ be the sets of global minimizers of F_0 and $\tilde{F}$ respectively. Then, from Theorem 2, we have $\mathcal{G}_0 \subseteq \tilde{\mathcal{G}}$. Now assume that, under Assumption 1, there exists $\hat{\mathbf{x}} \in \tilde{\mathcal{G}}$ such that $\hat{\mathbf{x}} \notin \mathcal{G}_0$. This implies from Theorem 2 that $\sigma^-_{\hat{\mathbf{x}}} \neq \emptyset$ (*i.e.,* given the definition of $\sigma^-_{\hat{\mathbf{x}}}$ in (14), that there exists $i \in \{1, \ldots, N\}$ such that $|\hat{x}_i| \in (0, \sqrt{2\lambda}/\|\mathbf{a}_i\|)$). From the second point of Lemma 1 (see Appendix A), we can then build a sequence of global minimizers of $\tilde{F}$, denoted $\{\mathbf{x}_k\}_{k=1}^K$, such that $\mathbf{x}_1 = \hat{\mathbf{x}}$ and $\mathbf{x}_{k+1} = \mathbf{x}_k^{(j_k)}$ where $\{j_1, \ldots, j_K\} = \sigma^-_{\hat{\mathbf{x}}}$ and $K = \#\sigma^-_{\hat{\mathbf{x}}}$. In other words, we set one by one the components of $\hat{\mathbf{x}}$ indexed by the elements of $\sigma^-_{\hat{\mathbf{x}}}$ to zero. Note that $\mathbf{x}_K = \mathcal{T}(\hat{\mathbf{x}})$ where $\mathcal{T}$ is the thresholding rule defined in (12).

Considering $\mathbf{x}_{K-1}$, we can either set its j_Kth component to zero and get $\mathbf{x}_K$, or set this component to $-s_{j_K}\sqrt{2\lambda}/\|\mathbf{a}_{j_K}\|$ to obtain another global minimizer of $\tilde{F}$ (see Lemma 1) which we denote by $\tilde{\mathbf{x}}_K$. Moreover, we have by definition that $\sigma^-_{\mathbf{x}_K} = \sigma^-_{\tilde{\mathbf{x}}_K} = \emptyset$, and thus both $\mathbf{x}_K$ and $\tilde{\mathbf{x}}_K$ are global minimizers of F_0 from Theorem 2. However, by construction,
$$\|\mathbf{x}_K - \tilde{\mathbf{x}}_K\|_0 = 1, \tag{37}$$
which contradicts Assumption 1. This proves that $\hat{\mathbf{x}} \in \mathcal{G}_0 \cap \tilde{\mathcal{G}}$ and that $\mathcal{G}_0 = \tilde{\mathcal{G}}$.

Finally, we know that these global minimizers are strict for F_0 from [32, Theorem 4.4]. Hence, according to the fact that $\mathcal{G}_0 = \tilde{\mathcal{G}}$, they are also strict for $\tilde{F}$.

D Proof of Theorem 3

We proceed by proving both implications.

D.1 Proof of $\Longrightarrow$

Let $\mathbf{x} \in \mathbb{R}^N$ be a strict local minimizer of $\tilde{F}$ and assume that $\sigma_{\mathbf{x}}^+ \neq \emptyset$. Then from Lemma 1, for all $i \in \sigma_{\mathbf{x}}^+$, there exists a non-degenerate interval $\mathcal{T}_i \subseteq [0, \sqrt{2\lambda}/\|\mathbf{a}_i\|]$ containing $|x_i|$ such that, $\forall t \in \mathcal{T}_i$, $\bar{\mathbf{x}} = \mathbf{x}^{(i)} - s_i t \mathbf{e}_i$ is another local minimizer of $\tilde{F}$. This contradicts the fact that $\mathbf{x}$ is a strict local minimizer of $\tilde{F}$. Hence $\sigma_{\mathbf{x}}^+ = \emptyset$. Then, the fact that $\mathrm{rank}(\mathbf{A}_{\sigma_{\mathbf{x}}}) = \#\sigma_{\mathbf{x}}$ comes from [41, Corollary 4.9]. The idea is that a strict minimizer of $\tilde{F}$ is necessarily a strict minimizer of F_0 (because $\tilde{F}$ is always lower than F_0). Then, we know from [32, Theorem 3.2] that a strict minimizer of F_0 verifies $\mathrm{rank}(\mathbf{A}_{\sigma_{\mathbf{x}}}) = \#\sigma_{\mathbf{x}}$.

D.2 Proof of $\Longleftarrow$

Let $\mathbf{x} \in \mathbb{R}^N$ be a critical point of $\tilde{F}$ such that $\sigma_{\mathbf{x}}^+ = \emptyset$ and $\mathrm{rank}(\mathbf{A}_{\sigma_{\mathbf{x}}}) = \#\sigma_{\mathbf{x}}$. To prove the result, we will show that there exists $\eta > 0$ such that

$$\forall\, \boldsymbol{\varepsilon} \in \mathcal{B}_2(\mathbf{0}_{\mathbb{R}^N}, \eta),\ \tilde{F}(\mathbf{x} + \boldsymbol{\varepsilon}) > \tilde{F}(\mathbf{x}). \tag{38}$$

D.2.1 Determination of η

Let η be defined as

$$\eta = \min_{i \in \{1,2,3\}} \eta_i, \tag{39}$$

where

$$\eta_1 = \min_{i \in \sigma_{\mathbf{x}}^c} \left(\frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|} \right) \tag{40}$$

$$\eta_2 = \min_{i \in \sigma_{\mathbf{x}}^c} \left(\frac{2(\sqrt{2\lambda}\|\mathbf{a}_i\| - |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle|)}{\|\mathbf{a}_i\|^2} \right) \tag{41}$$

$$\eta_3 = \min_{i \in \sigma_{\mathbf{x}}} \left(|x_i| - \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|} \right), \tag{42}$$

and $\sigma_{\mathbf{x}}^c = \mathbb{I}_N \backslash \sigma_{\mathbf{x}}$. Clearly $\eta > 0$ as $\sigma_{\mathbf{x}}^+ = \emptyset$ implies

- $\forall i \in \sigma_{\mathbf{x}}^c$, $|\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle| < \sqrt{2\lambda}\|\mathbf{a}_i\| \Longrightarrow \eta_2 > 0$,
- $\forall i \in \sigma_{\mathbf{x}}$, $|x_i| > \sqrt{2\lambda}/\|\mathbf{a}_i\| \Longrightarrow \eta_3 > 0$.

D.2.2 Proof of (38)

Let $\boldsymbol{\varepsilon} \in \mathcal{B}_2(\mathbf{0}_{\mathbb{R}^N}, \eta) \setminus \{\mathbf{0}_{\mathbb{R}^N}\}$. Then by definition of η in (39) we have,

$$\forall i \in \sigma_{\mathbf{x}}^c,\ |x_i + \varepsilon_i| = |\varepsilon_i| < \eta \leq \sqrt{2\lambda}/\|\mathbf{a}_i\|, \tag{43}$$

$$\forall i \in \sigma_{\mathbf{x}},\ |x_i + \varepsilon_i| > |x_i| - \eta \geq \sqrt{2\lambda}/\|\mathbf{a}_i\|, \tag{44}$$

By combining the inequalities (43) and (44) with the definition of the CEL0 penalty in (3), we obtain

$$\Phi(\mathbf{x} + \boldsymbol{\varepsilon}) = \sum_{i \in \sigma_{\mathbf{x}}^c} \phi_i(\varepsilon_i) + \sum_{i \in \sigma_{\mathbf{x}}} \phi_i(x_i), \tag{45}$$

where, $\phi_i(x) = \lambda - \frac{\|\mathbf{a}_i\|^2}{2} \left(|x| - \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|} \right)^2 \mathbb{1}_{\left\{|x| \leq \frac{\sqrt{2\lambda}}{\|\mathbf{a}_i\|}\right\}}$.

On the other hand, we have

$$\frac{1}{2}\|\mathbf{A}(\mathbf{x} + \boldsymbol{\varepsilon}) - \mathbf{y}\|^2 = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{2}\|\mathbf{A}\boldsymbol{\varepsilon}\|^2 + \sum_{i \in \mathbb{I}_N} \varepsilon_i \langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle . \tag{46}$$

Using the description of the critical points in Proposition 1 and the fact that $\sigma_{\mathbf{x}}^+ = \emptyset$, we get that $\forall i \in \sigma_{\mathbf{x}}$, $\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle = 0$. Hence the last term in (46) can be simplified as

$$\sum_{i \in \mathbb{I}_N} \varepsilon_i \langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle = \sum_{i \in \sigma_{\mathbf{x}}^c} \varepsilon_i \langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle . \tag{47}$$

Combining equations (45) to (47), we obtain

$$\begin{aligned} \tilde{F}(\mathbf{x} + \boldsymbol{\varepsilon}) = {} & \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{2}\|\mathbf{A}\boldsymbol{\varepsilon}\|^2 + \sum_{i \in \sigma_{\mathbf{x}}} \phi_i(x_i) \\ & + \sum_{i \in \sigma_{\mathbf{x}}^c} \varepsilon_i \langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle + \phi_i(\varepsilon_i) \qquad (48) \\ \geq {} & \tilde{F}(\mathbf{x}) + \frac{1}{2}\|\mathbf{A}\boldsymbol{\varepsilon}\|^2 \\ & + \sum_{i \in \sigma_{\mathbf{x}}^c} \phi_i(\varepsilon_i) - |\varepsilon_i| \, |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle| \qquad (49) \end{aligned}$$

From the expression of ϕ_i and the fact that $\forall \boldsymbol{\varepsilon} \in \mathcal{B}_2(\mathbf{0}_{\mathbb{R}^N}, \eta) \backslash \{\mathbf{0}_{\mathbb{R}^N}\}$, $|\varepsilon_i| < \sqrt{2\lambda}/\|\mathbf{a}_i\|$ $\forall i \in \sigma_{\mathbf{x}}^c$, we have,

$$\begin{aligned} & \phi_i(\varepsilon_i) - |\varepsilon_i| \, |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle| \qquad (50) \\ & = |\varepsilon_i| \left(\sqrt{2\lambda}\|\mathbf{a}_i\| - \frac{\|\mathbf{a}_i\|^2}{2}|\varepsilon_i| - |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle| \right) \qquad (51) \end{aligned}$$

Moreover, $\forall i \in \sigma_{\mathbf{x}}^c$

$$\sqrt{2\lambda}\|\mathbf{a}_i\| - \frac{\|\mathbf{a}_i\|^2}{2}|\varepsilon_i| - |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle| > 0, \tag{52}$$

$$\Longleftrightarrow |\varepsilon_i| < \frac{2}{\|\mathbf{a}_i\|^2} \left(\sqrt{2\lambda}\|\mathbf{a}_i\| - |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle| \right), \tag{53}$$

which is true by definition of η (see (41)). Hence, we can write (49) as

$$\tilde{F}(\mathbf{x}+\boldsymbol{\varepsilon}) \geq \tilde{F}(\mathbf{x}) + \frac{1}{2}\|\mathbf{A}\boldsymbol{\varepsilon}\|^2 + \sum_{i\in\sigma_{\mathbf{x}}^c} \alpha|\varepsilon_i|, \tag{54}$$

where $\alpha > 0$. Finally, because $\mathrm{rank}(\mathbf{A}_{\sigma_{\mathbf{x}}}) = \#\sigma_{\mathbf{x}}$ and $\boldsymbol{\varepsilon} \neq \mathbf{0}_{\mathbb{R}^N}$, at least one of the two following assertions holds true

- $\exists i \in \sigma_{\mathbf{x}}$ such that $|\varepsilon_i| > 0$ and thus $\|\mathbf{A}\boldsymbol{\varepsilon}\|^2 > 0$,
- $\exists i \in \sigma_{\mathbf{x}}^c$ such that $|\varepsilon_i| > 0$ and thus $\alpha|\varepsilon_i| > 0$.

Hence, we have

$$\tilde{F}(\mathbf{x}+\boldsymbol{\varepsilon}) > \tilde{F}(\mathbf{x}), \tag{55}$$

which shows that $\mathbf{x}$ is a strict local minimizer of $\tilde{F}$ and completes the proof.

E Proof of Theorem 4

Let $\mathbf{x} \in \mathbb{R}^N$ be a strict local minimizer of $\tilde{F}$. Hence, it is a critical point of $\tilde{F}$ such that $\sigma_{\mathbf{x}}^+ = \emptyset$ (Theorem 3). From Proposition 1, $\mathbf{x}$ is SO and verifies

$$\begin{cases} |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y}\rangle| < \sqrt{2\lambda}\|\mathbf{a}_i\| \text{ if } i \in \mathbb{I}_N\backslash\sigma_{\mathbf{x}}, \\ |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y}\rangle| > \sqrt{2\lambda}\|\mathbf{a}_i\| \text{ if } i \in \sigma_{\mathbf{x}}. \end{cases} \tag{56}$$

It follows from Lemma 2 that $\mathbf{x}$ is L-stationary for

$$L \geq \max_{i\in\mathbb{I}_N} \|\mathbf{a}_i\|^2 \Rightarrow \forall i, \begin{cases} \sqrt{2\lambda}\|\mathbf{a}_i\| \leq \sqrt{2\lambda L}, \\ \sqrt{2\lambda}\|\mathbf{a}_i\| \geq \sqrt{2\lambda/L}\|\mathbf{a}_i\|^2, \end{cases} \tag{57}$$

which completes the proof.

F Proof of Theorem 6

Let $\mathbf{x} \in \mathbb{R}^N$ be a partial support CW point of Problem (1) for $\lambda \in \mathbb{R}_{>0}\backslash\Lambda$, where

$$\Lambda = \Big\{ \lambda = \left(\left\langle \mathbf{a}_k, \mathbf{A}\mathbf{x}^{(k)} - \mathbf{y}\right\rangle\right)^2/2 \\ \text{for } k \in \{i_{\mathbf{x}}, j_{\mathbf{x}}\} \text{ and } \mathbf{x} \in \min_{\mathrm{loc}}^{\mathrm{st}}\{F_0\}\Big\}, \tag{58}$$

with $i_{\mathbf{x}}$ and $j_{\mathbf{x}}$ defined in (9) and (10), respectively. Clearly, because $\min_{\mathrm{loc}}^{\mathrm{st}}\{F_0\}$ contains a finite number of points [32], Λ has a zero Lebesgue measure.

Under the URP of $\mathbf{A}$, Theorem 5 states that $\mathbf{x}$ is a strict local minimizer of F_0 and it follows from Theorem 1 that $\mathrm{rank}(\mathbf{A}_{\sigma_{\mathbf{x}}}) = \#\sigma_{\mathbf{x}}$. Then, from Theorem 3, we get that $\mathbf{x}$ is a strict local minimizer of $\tilde{F}$ if and only if $\mathbf{x}$ is a critical point of $\tilde{F}$ and $\sigma_{\mathbf{x}}^+ = \emptyset$. According to Proposition 1 together with the definition of $\sigma_{\mathbf{x}}^+$ in (15), these two conditions are equivalent to

$$\begin{cases} |\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle| < \sqrt{2\lambda} & \forall i \in \mathbb{I}_N\backslash\sigma_{\mathbf{x}} \\ |x_i| > \sqrt{2\lambda} & \forall i \in \sigma_{\mathbf{x}}. \end{cases} \tag{59}$$

(We recall that $\mathbf{A}$ is assumed to have unit norm columns in the statement of Theorem 6 and that $\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y}\rangle = \langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle$ for $i \in \mathbb{I}_N\backslash\sigma_{\mathbf{x}}$ and $\langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y}\rangle = -x_i$ for $i \in \sigma_{\mathbf{x}}$.)

Now assume that $\mathbf{x}$ is not a strict local minimizer of $\tilde{F}$. We distinguish two cases from (59)

- $\exists j \in \mathbb{I}_N\backslash\sigma_{\mathbf{x}}$ such that $|\langle \mathbf{a}_j, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle| \geq \sqrt{2\lambda}$. By definition of $j_{\mathbf{x}}$ in (10), we have

  $$|\langle \mathbf{a}_{j_{\mathbf{x}}}, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle| \geq |\langle \mathbf{a}_j, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle| \geq \sqrt{2\lambda} \tag{60}$$

  Hence, the first line of (59) is also violated for $j_{\mathbf{x}}$. Now let $t \in \mathbb{R}$ be such that

  $$\begin{aligned} t &= \arg\min_{v\in\mathbb{R}} \frac{1}{2}\|\mathbf{A}(\mathbf{x} + \mathbf{e}_{j_{\mathbf{x}}}v) - \mathbf{y}\|^2 \\ &= -\langle \mathbf{a}_{j_{\mathbf{x}}}, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle, \end{aligned} \tag{61}$$

  and define $\bar{\mathbf{x}} = \mathbf{x} + \mathbf{e}_{j_{\mathbf{x}}}t$. Given that $\lambda \in \mathbb{R}_{>0}\backslash\Lambda$, we have $t > \sqrt{2\lambda}$. Then,

  $$\begin{aligned} \frac{1}{2}\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\|^2 &= \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \frac{t^2}{2} \\ &\quad + t\langle \mathbf{a}_{j_{\mathbf{x}}}, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle \\ &\underset{(61)}{=} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 - \frac{t^2}{2} \\ &< \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 - \lambda. \end{aligned} \tag{62}$$

  Moreover, by definition of $\mathbf{u}_{\mathbf{x}}^+$, we get

  $$\frac{1}{2}\|\mathbf{A}\mathbf{u}_{\mathbf{x}}^+ - \mathbf{y}\|^2 \leq \frac{1}{2}\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\|^2, \tag{63}$$

  and that $\sigma_{\mathbf{u}_{\mathbf{x}}^+} \subseteq \sigma_{\bar{\mathbf{x}}} = \sigma_{\mathbf{x}} \cup \{j_{\mathbf{x}}\}$. Hence, $\|\mathbf{u}_{\mathbf{x}}^+\|_0 \leq \|\mathbf{x}\|_0 + 1$ and, with (62) and (63), we obtain that $F_0(\mathbf{x}) > F_0(\mathbf{u}_{\mathbf{x}}^+)$. This is in contradiction with the fact that $\mathbf{x}$ is a partial support CW point.
- $\exists i \in \sigma_{\mathbf{x}}$ such that $|x_i| \leq \sqrt{2\lambda}$. Again, from the definition of $i_{\mathbf{x}}$ in (9), we get

  $$|x_{i_{\mathbf{x}}}| \leq |x_i| \leq \sqrt{2\lambda}, \tag{64}$$

  which shows that the second line of (59) is also violated for $i_{\mathbf{x}}$. Moreover, because $\lambda \in \mathbb{R}_{>0} \setminus \Lambda$, we have $|x_{i_{\mathbf{x}}}| < \sqrt{2\lambda}$ and

  $$\begin{aligned} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 &= \frac{1}{2}\|\mathbf{A}\mathbf{x}^{(i_{\mathbf{x}})} - \mathbf{y}\|^2 + \frac{x_{i_{\mathbf{x}}}^2}{2} \\ &\quad + x_{i_{\mathbf{x}}}\left\langle \mathbf{a}_{i_{\mathbf{x}}}, \mathbf{A}\mathbf{x}^{(i_{\mathbf{x}})} - \mathbf{y}\right\rangle \\ &= \frac{1}{2}\|\mathbf{A}\mathbf{x}^{(i_{\mathbf{x}})} - \mathbf{y}\|^2 - \frac{x_{i_{\mathbf{x}}}^2}{2} \\ &> \frac{1}{2}\|\mathbf{A}\mathbf{x}^{(i_{\mathbf{x}})} - \mathbf{y}\|^2 - \lambda. \end{aligned} \tag{65}$$

Moreover, by definition of $\mathbf{u}_{\mathbf{x}}^-$, we have

$$\frac{1}{2}\|\mathbf{A}\mathbf{x}^{(i_{\mathbf{x}})} - \mathbf{y}\|^2 \geq \frac{1}{2}\|\mathbf{A}\mathbf{u}_{\mathbf{x}}^- - \mathbf{y}\|^2. \tag{66}$$

Combining these two last inequalities with the fact that $\|\mathbf{u}_{\mathbf{x}}^-\|_0 \leq \|\mathbf{x}\|_0 - 1$, we obtain $F_0(\mathbf{x}) > F_0(\mathbf{u}_{\mathbf{x}}^-)$. This contradicts the fact that $\mathbf{x}$ is a partial support CW point and completes the proof.

G Proof of Theorem 7

We provide three independent proofs for $\tilde{\mathcal{S}}$, $\mathcal{S}_{\mathrm{CW}}$, and $\mathcal{S}_{\mathrm{L}}$. Before to enter into the details of the proofs, let us recall the reader that $\mathcal{S}_0$ contains a finite number of points that do not depend on λ. Moreover, $\forall \mathbf{x} \in \mathcal{S}_0$

$$\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle = 0 \quad \forall i \in \sigma_{\mathbf{x}}, \tag{67}$$

and such points belong to $\mathcal{X}_{\mathrm{LS}}$ if and only if

$$\begin{aligned} &\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{y} \\ \underset{\text{under (67)}}{\Longleftrightarrow} &\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle = 0,\ \forall i \in \mathbb{I}_N \backslash \sigma_{\mathbf{x}}. \end{aligned} \tag{68}$$

Also let us recall that for $\mathbf{x} \in \mathcal{S}_0$ and $i \in \mathbb{I}_N\backslash\sigma_{\mathbf{x}}$, $\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle = \langle \mathbf{a}_i, \mathbf{A}\mathbf{x}^{(i)} - \mathbf{y}\rangle$.

Finally, for the first statement of Theorem 7, we need to verify that $\mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}}$ is non-empty. This is always true as for any support $\omega \subseteq \mathbb{I}_N$ such that $\mathrm{rank}(\mathbf{A}_\omega) = \#\omega = \mathrm{rank}(\mathbf{A})$, the associated SO point $\mathbf{x}$ (which is unique) belongs to both $\mathcal{S}_0$ and $\mathcal{X}_{\mathrm{LS}}$.

G.1 Proof for $\tilde{\mathcal{S}}$

1. Define

$$\tilde{\lambda} = \min_{\substack{\mathbf{x} \in \mathcal{S}_0\backslash\mathcal{X}_{\mathrm{LS}} \\ \|\mathbf{x}\|_0 < N}} \left\{ \max_{i \in \mathbb{I}_N\backslash\sigma_{\mathbf{x}}} \frac{\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle^2}{2\|\mathbf{a}_i\|^2} \right\}. \tag{69}$$

Clearly, from (68), we have $\tilde{\lambda} > 0$. Then, for all $\lambda < \tilde{\lambda}$ and all $\mathbf{x} \in \mathcal{S}_0\backslash\mathcal{X}_{\mathrm{LS}}$ such that $\|\mathbf{x}\|_0 < N$, we have

$$\lambda < \max_{i \in \mathbb{I}_N\backslash\sigma_{\mathbf{x}}} \frac{\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle^2}{2\|\mathbf{a}_i\|^2} \tag{70}$$

This implies that, for such λ and $\mathbf{x}$, there exists $i \in \mathbb{I}_N \setminus \sigma_{\mathbf{x}}$ such that $|\langle \mathbf{a}_i, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle| > \sqrt{2\lambda}\|\mathbf{a}_i\|$. Then, from Proposition 1, it follows that $\mathbf{x}$ cannot be a critical point of $\tilde{F}$, and thus $\mathbf{x} \notin \tilde{\mathcal{S}}$. As a result, for $\lambda < \tilde{\lambda}$, we have that $\tilde{\mathcal{S}} \subseteq \mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}}$ (recalling that $\tilde{\mathcal{S}} \subseteq \mathcal{S}_0$).
To show the reciprocal inclusion, we define $\lambda_0 = \min\{\tilde{\lambda}, \lambda'\}$, where

$$\lambda' = \min_{\mathbf{x} \in (\mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}})\backslash\{\mathbf{0}_{\mathbb{R}^N}\}} \left\{ \min_{i \in \sigma_{\mathbf{x}}} \frac{(x_i\|\mathbf{a}_i\|)^2}{2} \right\}. \tag{71}$$

It is also easy to see that $\lambda_0 > 0$. Then, for all $\lambda < \lambda_0$ and $\mathbf{x} \in (\mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}})\backslash\{\mathbf{0}_{\mathbb{R}^N}\}$, we get from (71) that for all $i \in \sigma_{\mathbf{x}}$, $|x_i| > \sqrt{2\lambda}/\|\mathbf{a}_i\|$. Hence $\sigma_{\mathbf{x}}^- = \emptyset$ and moreover, with (68), we have $\sigma_{\mathbf{x}}^+ = \emptyset$. The fact that $\mathrm{rank}(\mathbf{A}_{\sigma_{\mathbf{x}}}) = \#\sigma_{\mathbf{x}}$ follows from the fact that $\mathbf{x}$ is a strict local minimizer of F_0 ($\mathbf{x} \in \mathcal{S}_0$). Finally, with (67) and Proposition 1 we get that $\mathbf{x}$ is a critical point of $\tilde{F}$. Hence, for all $\lambda < \lambda_0$, all $\mathbf{x} \in \mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}}$ fulfill the requirement of Theorem 3 and are thus strict local minimizers of $\tilde{F}$ (*i.e.,* $\mathbf{x} \in \tilde{\mathcal{S}}$).

2. Define

$$\lambda_\infty = \max_{\mathbf{x} \in \mathcal{S}_0\backslash\{\mathbf{0}_{\mathbb{R}^N}\}} \left\{ \min_{i \in \sigma_{\mathbf{x}}} \frac{(x_i\|\mathbf{a}_i\|)^2}{2} \right\}. \tag{72}$$

Then, for all $\lambda > \lambda_\infty$ and all $\mathbf{x} \in \mathcal{S}_0\backslash\{\mathbf{0}_{\mathbb{R}^N}\}$, we have

$$\lambda > \min_{i \in \sigma_{\mathbf{x}}} \frac{(x_i\|\mathbf{a}_i\|)^2}{2}. \tag{73}$$

This implies that, for such λ and $\mathbf{x}$, there exists $i \in \sigma_{\mathbf{x}}$ such that $|x_i| < \sqrt{2\lambda}/\|\mathbf{a}_i\|$. Hence, $\sigma_{\mathbf{x}}^- \neq \emptyset$ and thus $\sigma_{\mathbf{x}}^+ \neq \emptyset$ as $\sigma_{\mathbf{x}}^- \subseteq \sigma_{\mathbf{x}}^+$. Then, it follows from Theorem 3 that $\mathbf{x} \notin \tilde{\mathcal{S}}$. As a result, for $\lambda > \lambda_\infty$, we have $\tilde{\mathcal{S}} \subseteq \{\mathbf{0}_{\mathbb{R}^N}\}$. Finally the equality comes from the fact that $\tilde{\mathcal{S}}$ includes the set of global minimizers of $\tilde{F}$ and F_0 (from Corollary 1) which is non-empty [32, Theorem 4.4 (i)].

G.2 Proof for $\mathcal{S}_{\mathrm{L}}$

Using Lemma 2, the proof follows the line of the one for $\tilde{\mathcal{S}}$ (see Section G.1). Hence, we let it to the reader.

G.3 Proof for $\mathcal{S}_{\mathrm{CW}}$

1. Let $\mathbf{x} \in \mathcal{S}_0\backslash\mathcal{X}_{\mathrm{LS}}$ (if applicable, *i.e.,* non-empty, otherwise go to the paragraph before equation (82)). Then we have $\|\mathbf{x}\|_0 < \min\{M, N\}$. Indeed, $\mathbf{x} \in \mathcal{S}_0$ and $\|\mathbf{x}\|_0 = \min\{M, N\}$ would imply that

$$\mathrm{rank}(\mathbf{A}_{\sigma_{\mathbf{x}}}) = \min\{M, N\} = \mathrm{rank}(\mathbf{A}), \tag{74}$$

and thus $\mathbf{x} \in \mathcal{X}_{\mathrm{LS}}$.
Then, by definition of $\mathbf{u}_{\mathbf{x}}^+$ in Definition 2, we have

$$\beta(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 - \frac{1}{2}\|\mathbf{A}\mathbf{u}_{\mathbf{x}}^+ - \mathbf{y}\|^2 > 0, \tag{75}$$

The fact that $\beta(\mathbf{x}) > 0$ comes from the URP of $\mathbf{A}$. Indeed, let us first show that we necessarily have $\mathbf{u}_{\mathbf{x}}^+ \neq \mathbf{x}$ (which is not trivial as by definition $\sigma_{\mathbf{u}_{\mathbf{x}}^+} \subseteq \sigma_{\mathbf{x}} \cup \{j_{\mathbf{x}}\}$, where $j_{\mathbf{x}}$ is defined in (10)). To that end, assume that $\mathbf{u}_{\mathbf{x}}^+ = \mathbf{x}$. Hence, by construction of $\mathbf{u}_{\mathbf{x}}^+$ we get that (normal equations)

$$\langle \mathbf{a}_{j_{\mathbf{x}}}, \mathbf{A}\mathbf{x} - \mathbf{y}\rangle = 0. \tag{76}$$

It follows, by definition of $j_{\mathbf{x}}$ in (10), that

$$\langle \mathbf{a}_j, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle = 0, \ \forall j \in \mathbb{I}_N \backslash \sigma_{\mathbf{x}}. \tag{77}$$

Moreover, from (67), (77) is also true for all $j \in \sigma_{\mathbf{x}}$. This implies that $\mathbf{x} \in \mathcal{X}_{\mathrm{LS}}$ (with (68)) which leads to a contradiction. As a result, we get that $\mathbf{u}_{\mathbf{x}}^+ \neq \mathbf{x}$. Now assume that $\beta(\mathbf{x}) = 0$. Then, again by construction of $\mathbf{u}_{\mathbf{x}}^+$, $\beta(\mathbf{x}) = 0$ implies that for $\omega = \sigma_{\mathbf{x}} \cup \{j_{\mathbf{x}}\}$, both $(\mathbf{u}_{\mathbf{x}}^+)_\omega$ and $\mathbf{x}_\omega$ (which are different) are minimizers of $\mathbf{v} \mapsto \frac{1}{2}\|\mathbf{A}_\omega \mathbf{v} - \mathbf{y}\|^2$. This is in contradiction with the fact that $\mathbf{A}_\omega$ is full rank ($\#\omega \leq \min\{M, N\}$ and URP of $\mathbf{A}$). Hence $\beta(\mathbf{x}) > 0$ and we can define

$$\tilde{\lambda} = \min_{\mathbf{x} \in \mathcal{S}_0 \backslash \mathcal{X}_{\mathrm{LS}}} \beta(\mathbf{x}). \tag{78}$$

It follows that, for all $\lambda < \tilde{\lambda}$ and all $\mathbf{x} \in \mathcal{S}_0 \backslash \mathcal{X}_{\mathrm{LS}}$,

$$\begin{aligned} F_0(\mathbf{x}) - F_0(\mathbf{u}_{\mathbf{x}}^+) &= \beta(\mathbf{x}) + \lambda(\|\mathbf{x}\|_0 - \|\mathbf{u}_{\mathbf{x}}^+\|_0) && (79)\\ &\geq \beta(\mathbf{x}) - \lambda && (80)\\ &> \beta(\mathbf{x}) - \tilde{\lambda} \underset{(78)}{\geq} 0, && (81) \end{aligned}$$

using the fact (for the second line) that $\|\mathbf{u}_{\mathbf{x}}^+\|_0 \leq \|\mathbf{x}\|_0 + 1$. Hence, $F_0(\mathbf{x}) > F_0(\mathbf{u}_{\mathbf{x}}^+)$, which prevents $\mathbf{x}$ from being a partial support CW point. Thus, we have $\mathcal{S}_{\mathrm{CW}} \subseteq \mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}}$ (recalling that $\mathcal{S}_{\mathrm{CW}} \subseteq \mathcal{S}_0$ by definition).

Now, let $\mathbf{x} \in \mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}}$.

We will show that for all $\mathbf{u} \in \mathcal{U}$ defined in (8), we have $F_0(\mathbf{x}) \leq F_0(\mathbf{u})$. This will ensure that $\mathbf{x} \in \mathcal{S}_{\mathrm{CW}}$.

– Case $\mathbf{u}_{\mathbf{x}}^-$: By definition of $\mathcal{U}$ in (8), this case is relevant only for $\mathbf{x} \neq \mathbf{0}_{\mathbb{R}^N}$. Let

$$\beta(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{u}_{\mathbf{x}}^- - \mathbf{y}\|^2 - \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 > 0. \tag{82}$$

The fact that $\beta(\mathbf{x}) > 0$ comes from similar arguments as those used previously. Indeed, here we have $\mathbf{u}_{\mathbf{x}}^- \neq \mathbf{x}$ by definition ($\sigma_{\mathbf{u}_{\mathbf{x}}^-} \subseteq \sigma_{\mathbf{x}} \backslash \{i_{\mathbf{x}}\}$). Then $\#\sigma_{\mathbf{x}} \leq \min\{M, N\}$ (because $\mathbf{x} \in \mathcal{S}_0$) together with the URP of $\mathbf{A}$ allows to conclude. Now, define

$$\tilde{\lambda}^- = \min_{\mathbf{x} \in \mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}}} \beta(\mathbf{x}) / (\|\mathbf{x}\|_0 - \|\mathbf{u}_{\mathbf{x}}^-\|_0). \tag{83}$$

By construction we have $\|\mathbf{x}\|_0 > \|\mathbf{u}_{\mathbf{x}}^-\|_0$ and hence $\tilde{\lambda}^- > 0$. It follows that, for all $\lambda < \tilde{\lambda}^-$ and all $\mathbf{x} \in \mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}}$,

$$\begin{aligned} F_0(\mathbf{u}_{\mathbf{x}}^-) - F_0(\mathbf{x}) &= \beta(\mathbf{x}) - \lambda(\|\mathbf{x}\|_0 - \|\mathbf{u}_{\mathbf{x}}^-\|_0) \\ &\underset{(83)}{>} 0, && (84) \end{aligned}$$

– Case $\mathbf{u}_{\mathbf{x}}^+$: By definition of $\mathcal{U}$ in (8), this case is relevant only when $\|\mathbf{x}\|_0 < \min\{M, N\}$. Then, because $\mathbf{x} \in \mathcal{X}_{\mathrm{LS}}$ we have

$$\langle \mathbf{a}_j, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle = 0, \ \forall j \in \mathbb{I}_N \backslash \sigma_{\mathbf{x}}. \tag{85}$$

Hence, one can choose $j_{\mathbf{x}}$ to be any element of $\mathbb{I}_N \backslash \sigma_{\mathbf{x}}$ according to (10). Moreover, for any of these choices, one will get under the URP of $\mathbf{A}$ that $\mathbf{u}_{\mathbf{x}}^+ = \mathbf{x}$ and thus $F_0(\mathbf{x}) = F_0(\mathbf{u}_{\mathbf{x}}^+)$.

– Case $\mathbf{u}_{\mathbf{x}}^{\mathrm{swap}}$: Again, by definition of $\mathcal{U}$ in (8), this case is relevant only when $\|\mathbf{x}\|_0 \in (0, \min\{M, N\})$. Let

$$\beta(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{u}_{\mathbf{x}}^{\mathrm{swap}} - \mathbf{y}\|^2 - \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 > 0. \tag{86}$$

Once again the fact that $\beta(\mathbf{x}) > 0$ comes from the URP of $\mathbf{A}$ together with the facts that $\mathbf{u}_{\mathbf{x}}^+ = \mathbf{x}$ (see previous point) and that $\sigma_{\mathbf{u}_{\mathbf{x}}^{\mathrm{swap}}} \subseteq \sigma_{\mathbf{x}} \cup \{j_{\mathbf{x}}\}$. Now define

$$\tilde{\lambda}^{\mathrm{swap}} = \min_{\mathbf{x} \in \mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}}} \frac{\beta(\mathbf{x})}{\max\{1, (\|\mathbf{x}\|_0 - \|\mathbf{u}_{\mathbf{x}}^{\mathrm{swap}}\|_0)\}}. \tag{87}$$

It follows that, for all $\lambda < \tilde{\lambda}^{\mathrm{swap}}$ and all $\mathbf{x} \in \mathcal{S}_0 \cap \mathcal{X}_{\mathrm{LS}}$,

$$\begin{aligned} F_0(\mathbf{u}_{\mathbf{x}}^{\mathrm{swap}}) - F_0(\mathbf{x}) &= \beta(\mathbf{x}) - \lambda(\|\mathbf{x}\|_0 - \|\mathbf{u}_{\mathbf{x}}^{\mathrm{swap}}\|_0) \\ &\underset{(87)}{>} 0, && (88) \end{aligned}$$

using the fact that $\|\mathbf{u}_{\mathbf{x}}^{\mathrm{swap}}\|_0 \leq \|\mathbf{x}\|_0$.

Finally taking $\lambda_0 = \min\{\tilde{\lambda}, \tilde{\lambda}^-, \tilde{\lambda}^{\mathrm{swap}}\}$ completes the proof.

2. For $\mathbf{x} \in \mathcal{S}_0 \backslash \{\mathbf{0}_{\mathbb{R}^N}\}$, we define

$$\beta(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{u}_{\mathbf{x}}^- - \mathbf{y}\|^2 - \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 > 0, \tag{89}$$

where $\mathbf{u}_{\mathbf{x}}^-$ is an SO point such that $\|\mathbf{u}_{\mathbf{x}}^-\|_0 \leq \|\mathbf{x}\|_0 - 1$ (see Definition 2). The fact that $\beta(\mathbf{x}) > 0$ follows the same arguments as for (82). Now, let

$$\lambda_\infty = \max_{\mathbf{x} \in \mathcal{S}_0 \backslash \{\mathbf{0}_{\mathbb{R}^N}\}} \beta(\mathbf{x}). \tag{90}$$

It follows that, for all $\lambda > \lambda_\infty$ and all $\mathbf{x} \in \mathcal{S}_0 \backslash \{\mathbf{0}_{\mathbb{R}^N}\}$,

$$\begin{aligned} F_0(\mathbf{u}_{\mathbf{x}}^-) - F_0(\mathbf{x}) &= \beta(\mathbf{x}) + \lambda(\|\mathbf{u}_{\mathbf{x}}^-\|_0 - \|\mathbf{x}\|_0) && (91)\\ &\leq \beta(\mathbf{x}) - \lambda && (92)\\ &< \beta(\mathbf{x}) - \lambda_\infty \underset{(90)}{\leq} 0. && (93) \end{aligned}$$

Hence, $F_0(\mathbf{x}) > F_0(\mathbf{u}_{\mathbf{x}}^-)$, which prevents $\mathbf{x}$ from being a partial support CW point and we thus have $\mathcal{S}_{\mathrm{CW}} \subseteq \{\mathbf{0}_{\mathbb{R}^N}\}$.

Finally, let $\mathbf{x} = \mathbf{0}_{\mathbb{R}^N}$. To assert if it is a partial support CW point, we have to compare $F_0(\mathbf{0}_{\mathbb{R}^N})$ to $F_0(\mathbf{u}^+_{\mathbf{0}_{\mathbb{R}^N}})$,

$$F_0(\mathbf{0}_{\mathbb{R}^N}) - F_0(\mathbf{u}^+_{\mathbf{0}_{\mathbb{R}^N}}) = \beta(\mathbf{u}^+_{\mathbf{0}_{\mathbb{R}^N}}) - \lambda \quad (94)$$

$$< \beta(\mathbf{u}^+_{\mathbf{0}_{\mathbb{R}^N}}) - \lambda_\infty \underset{(90)}{\leq} 0 \quad (95)$$

This shows that $\mathbf{0}_{\mathbb{R}^N} \in \mathcal{S}_{\mathrm{CW}}$ and thus $\mathcal{S}_{\mathrm{CW}} = \{\mathbf{0}_{\mathbb{R}^N}\}$.

References

1. H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1):91–129, February 2013.
2. A. Beck and Y. Eldar. Sparsity Constrained Nonlinear Optimization: Optimality Conditions and Algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, January 2013.
3. A. Beck and N. Hallak. On the Minimization Over Sparse Symmetric Sets: Projections, Optimality Conditions, and Algorithms. *Mathematics of Operations Research*, 41(1):196–223, February 2016.
4. A. Beck and N. Hallak. Proximal Mapping for Symmetric Penalty and Sparsity. *SIAM Journal on Optimization*, 28(1):496–527, January 2018.
5. T Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, November 2009.
6. S. Bourguignon, J. Ninin, H. Carfantan, and M. Mongeau. Exact Sparse Approximation Problems via Mixed-Integer Programming: Formulations and Computational Performance. *IEEE Transactions on Signal Processing*, 64(6):1405–1419, March 2016.
7. L. Breiman. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4):373–384, 1995.
8. E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, February 2006.
9. E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing Sparsity by Reweighted ℓ_1 Minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, December 2008.
10. M. Carlsson, D. Gerosa, and C. Olsson. An unbiased approach to compressed sensing. *arXiv:1806.05283 [math]*, June 2018. arXiv: 1806.05283.
11. Marcus Carlsson. On Convex Envelopes and Regularization of Non-convex Functionals Without Moving Global Minima. *Journal of Optimization Theory and Applications*, 183(1):66–84, October 2019.
12. S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, March 1991.
13. E. Chouzenoux, A. Jezierska, J. Pesquet, and H. Talbot. A Majorize-Minimize Subspace Approach for $\ell_2 - \ell_0$ Image Regularization. *SIAM Journal on Imaging Sciences*, 6(1):563–591, January 2013.
14. W. Dai and O. Milenkovic. Subspace Pursuit for Compressive Sensing Signal Reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, May 2009.
15. D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, June 2006.
16. S. Durand and M. Nikolova. Stability of the Minimizers of Least Squares with a Non-Convex Regularization. Part I: Local Behavior. *Applied Mathematics and Optimization*, 53(2):185–208, March 2006.
17. S. Durand and M. Nikolova. Stability of the Minimizers of Least Squares with a Non-Convex Regularization. Part II: Global Behavior. *Applied Mathematics and Optimization*, 53(3):259–277, May 2006.
18. J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
19. S. Foucart. Hard Thresholding Pursuit: An Algorithm for Compressive Sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, January 2011.
20. S. Foucart and M-J. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0<q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, May 2009.
21. D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3):367–383, March 1992.
22. I. F. Gorodnitsky and B. D. Rao. Sparse signal reconstruction from limited data using FOCUSS: A reweighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, pages 600–616, 1997.
23. C. Herzet and A. Drémeau. Bayesian pursuit algorithms. In *2010 18th European Signal Processing Conference*, pages 1474–1478, August 2010.
24. P. Jain, A. Tewari, and I. S. Dhillon. Orthogonal Matching Pursuit with Replacement. In *Advances in Neural Information Processing Systems 24*, pages 1215–1223. Curran Associates, Inc., 2011.
25. S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, December 1993.
26. A. Marmin, M. Castella, and J. Pesquet. How to Globally Solve Non-convex Optimization Problems Involving an Approximate ℓ_0 Penalization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5601–5605, May 2019.
27. D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, May 2009.
28. T. T. Nguyen, C. Soussen, J. Idier, and E-H. Djermoune. NP-hardness of ℓ_0 minimization problems: revision and extension to the non-negative setting. In *International Conference on Sampling Theory and Applications (SampTa)*, Bordeaux, 2019.
29. M. Nikolova. Analysis of the Recovery of Edges in Images and Signals by Minimizing Nonconvex Regularized Least-Squares. *Multiscale Modeling & Simulation*, 4(3):960–991, January 2005.
30. M. Nikolova. Bounds on the Minimizers of (nonconvex) Regularized Least-Squares. In F. Sgallari, A. Murli, and N. Paragios, editors, *Scale Space and Variational Methods in Computer Vision*, Lecture Notes in Computer Science, pages 496–507. Springer Berlin Heidelberg, 2007.

31. M. Nikolova. Solve exactly an under determined linear system by minimizing least squares regularized with an ℓ_0 penalty. *Comptes Rendus Mathematique*, 349(21):1145–1150, November 2011.
32. M. Nikolova. Description of the Minimizers of Least Squares Regularized with ℓ_0-norm. Uniqueness of the Global Minimizer. *SIAM Journal on Imaging Sciences*, 6(2):904–937, January 2013.
33. M. Nikolova. Relationship between the optimal solutions of least squares regularized with ℓ_0-norm and constrained by k-sparsity. *Applied and Computational Harmonic Analysis*, 41(1):237–265, July 2016.
34. M. Nikolova and M. Ng. Analysis of Half-Quadratic Minimization Methods for Signal and Image Recovery. *SIAM Journal on Scientific Computing*, 27(3):937–966, January 2005.
35. P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock. On Iteratively Reweighted Algorithms for Nonsmooth Nonconvex Optimization in Computer Vision. *SIAM Journal on Imaging Sciences*, 8(1):331–372, January 2015.
36. Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44 vol.1, November 1993.
37. M. Pilanci, M. J. Wainwright, and L. El Ghaoui. Sparse learning via Boolean relaxations. *Mathematical Programming*, 151(1):63–87, June 2015.
38. A. Repetti, M. Q. Pham, L. Duval, É Chouzenoux, and J. Pesquet. Euclid in a Taxicab: Sparse Blind Deconvolution with Smoothed ℓ_1/ℓ_2 Regularization. *IEEE Signal Processing Letters*, 22(5):539–543, May 2015.
39. I. Selesnick. Sparse Regularization via Convex Analysis. *IEEE Transactions on Signal Processing*, 65(17):4481–4494, September 2017.
40. I. Selesnick and M. Farshchian. Sparse Signal Approximation via Nonseparable Regularization. *IEEE Transactions on Signal Processing*, 65(10):2561–2575, May 2017.
41. E. Soubies, L. Blanc-Féraud, and G. Aubert. A Continuous Exact ℓ_0 Penalty (CEL0) for Least Squares Regularized Problem. *SIAM Journal on Imaging Sciences*, 8(3):1607–1639, January 2015.
42. E. Soubies, L. Blanc-Féraud, and G. Aubert. A Unified View of Exact Continuous Penalties for ℓ_2-ℓ_0 Minimization. *SIAM Journal on Optimization*, 27(3):2034–2060, January 2017.
43. C. Soussen, J. Idier, D. Brie, and J. Duan. From Bernoulli–Gaussian Deconvolution to Sparse Signal Restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584, October 2011.
44. C. Soussen, J. Idier, J. Duan, and D. Brie. Homotopy Based Algorithms for ℓ_0-Regularized Least-Squares. *IEEE Transactions on Signal Processing*, 63(13):3301–3316, July 2015.
45. V. N. Temlyakov. Greedy approximation. *Acta Numerica*, 17:235–409, May 2008.
46. J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, October 2004.
47. J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, March 2006.
48. F. Wen, L. Chu, P. Liu, and R. C. Qiu. A Survey on Nonconvex Regularization-Based Sparse and Low-Rank Recovery in Signal Processing, Statistics, and Machine Learning. *IEEE Access*, 6:69883–69906, 2018.
49. C-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, April 2010.
50. N. Zhang and Q. Li. On optimal solutions of the constrained ℓ_0 regularization and its penalty problem. *Inverse Problems*, 33(2):025010, January 2017.
51. T. Zhang. Multi-stage Convex Relaxation for Learning with Sparse Regularization. In *Advances in Neural Information Processing Systems 21*, pages 1929–1936. Curran Associates, Inc., 2009.
52. H. Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, December 2006.