Sparse learning with concave regularization: relaxation of the irrepresentable condition

Sparse learning with concave regularization: relaxation of the irrepresentable condition

V. Cerone, S. M. Fosson*, D. Regruto, A. Salam

***Abstract*— Learning sparse models from data is an important task in all those frameworks where relevant information should be identified within a large dataset. This can be achieved by formulating and solving suitable sparsity promoting optimization problems. As to linear regression models, Lasso is the most popular convex approach, based on an ℓ_1-norm regularization. In contrast, in this paper, we analyse a concave regularized approach, and we prove that it relaxes the irrepresentable condition, which is sufficient and essentially necessary for Lasso to select the right significant parameters. In practice, this has the benefit of reducing the number of necessary measurements with respect to Lasso. Since the proposed problem is non-convex, we also discuss different algorithms to solve it, and we illustrate the obtained enhancement via numerical experiments.**

I. Introduction

Sparse learning is the science of building parsimonious models from data. The main motivation for sparse learning is the concrete need of extracting relevant information from large collections of data, which nowadays are commonly available in many scientific fields. This task prevents drawbacks such as overfitting, redundancies, numerical complexity, and scarce understanding of the physical behavior of systems; we refer the reader to [1], [2], [3] for a comprehensive illustration of these issues. Recent applications of sparse learning can be found in identification of linear and non-linear systems, see [4], [5] and [6], [3], respectively; in model predictive control, see [7]; in neural networks and deep learning, see [8], [9]. In signal processing, the exploitation of sparsity has a long history, provided that many signals (e.g., images) admit sparse representations in opportune bases. In this context, the compressed sensing (CS) theory has been developed in the last fifteen years, which states that a sparse vector $x \in \mathbb{R}^n$ can be recovered from compressed, possibly noisy, linear measurements, see [10], [11]. Beyond signal processing, the CS paradigm has been applied, e.g., to linear/non-linear system identification, see [12], [13], [4], [3].

Often, sparse learning is formulated as an optimization problem where sparsity is promoted by a suitable regularization term. In principle, the ℓ_0-norm, i.e., the number of non-zero components of a vector, is the correct pseudonorm to represent the sparsity level; nevertheless, the ℓ_0-norm is non-convex and leads to NP-hard optimization. As an alternative, the ℓ_1-norm has been studied and proven to be the best convex approximation of the ℓ_0-norm. In the linear regression setting, the Lasso problem, introduced in [14], is very popular, and consists in the minimization of an ℓ_1 regularized, least squares cost functional.

The reliability of the Lasso estimator has been largely studied in the literature, in particular for what concerns its variable-selection consistency (VSC), i.e., the ability of identifying the support, which is the set of non-zero/significant components of the unknown vector. Finding the support is the most important task in sparse learning and CS, since its knowledge is sufficient to recover the complete vector. As illustrated in [15], [16], [17], Lasso enjoys the VSC if the so-called irrepresentable condition holds. More precisely, as we illustrate in the next section, this condition is sufficient and essentially necessary, see [18].

In this paper, we analyse an alternative approach to Lasso for sparse learning, based on a concave, semi-algebraic regularization. Recently, the use of non-convex regularizers has been gaining attention in the literature, see, e.g., [19], [20], [21], [22]. The rationale is that the shape of a non-convex regularizer is closer to the ℓ_0-norm than the ℓ_1-norm. Numerical experiments show that the non-convex approach is usually more effective than Lasso. Nevertheless, theoretical results are still missing, in particular in the compressed framework.

The goal of this paper is to prove that the considered non-convex estimator is more effective than Lasso, in the sense that it enjoys the VSC under a relaxed irrepresentable condition, even in a CS setting. This result is obtained by exploiting a restricted eigenvalue property and a boundedness assumption. Under these hypotheses, a minimum of the proposed functional has the correct support; sufficient conditions to evaluate whether this minimum is local or global are provided. In practice, the relaxation of the irrepresentable condition implies a reduction of the number of necessary measurements to select the right parameters. Since the problem is semi-algebraic, the Lasserre's approach can be used to compute the minimum [23]. On the other hand, iterative algorithms can be used for local minimization and are effective in many cases. This is analysed through numerical simulations.

The paper is organized as follows. In Section II, we present the problem and we illustrate the considered cost functional. Section III is devoted to the theoretical analysis of the VSC of the considered approach, which is the core of the contribution. In Section IV, we discuss the minimization algorithms and show numerical results. Finally, we draw some conclusions in Section V.

* Corresponding author. The authors are with the Dipartimento di Automatica e Informatica, Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino, Italy; e-mail: sophie.fosson@polito.it.

II. Problem statement and background

In this paper, we consider the following sparse optimization problem: we aim to estimate a k-sparse vector $\widetilde{x} \in \mathbb{R}^n$ from noisy linear measurements:

$$y = A\widetilde{x} + \eta \tag{1}$$

where $A \in \mathbb{R}^{m,n}$, and $\eta \in \mathbb{R}^m$ is a measurements noise. In particular, we focus on the CS case $m < n$, which is more challenging. We call $S \subset \{1, \dots, n\}$ the support of $\widetilde{x}$, and $\bar{S}$ its complementary. For any $v \in \mathbb{R}^n$, we denote by $v_S \in \mathbb{R}^k$ the vector that contains the components of v restricted to S. Similarly, A_S contains the columns of A indexed in S.

In the following, we consider this assumption.

Assumption 1: The non-zero parameters are bounded, i.e., $\widetilde{x}_i \in [-d, d]$ for each $i \in S$, where $d > 0$ is known.

This assumption is realistic, since in real-world applications a prior estimate on the maximum magnitude is usually available. Due to noise and compression, the general approach to this problem is a regularized least squares. In particular, Lasso [14] reads as follows:

$$\textbf{Lasso}: \quad \min_{x \in [-d,d]^n} \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1 \tag{2}$$

where $\lambda > 0$ is a design parameter which can be assessed based on prior information on k, see [11]. In the literature, the VSC of Lasso is analysed by using the irrepresentable condition.

Definition 1: **Irrepresentable condition (IRR)**, see [2, Section 11.4.1]. We say that A satisfies (ω, S)-IRR if there exists $\omega > 0$ such that

$$\max_{j \in \bar{S}} \left\| A_j^T A_S (A_S^T A_S)^{-1} \right\|_1 \leq \frac{1}{\omega}. \tag{3}$$

For Lasso, strong (ω, S)-IRR (i.e., $\omega > 1$) and weak (ω, S)-IRR (i.e., $\omega = 1$) respectively are sufficient and necessary for the VSC in the asymptotic case $m \to \infty$, see [24], [16]. In the non-asymptotic case, the VSC of Lasso is proven with large probability, under IRR and sufficiently large λ, see, e.g., [2, Theorem 11.3].

In this paper, we analyse a concave alternative to ℓ_1 regularization. In the last years, different concave penalties have been proposed in sparse optimization, in particular, $\log(x)$, see [19]; ℓ_p, with $p \in (0, 1)$, see [25], and minimax concave penalties (MCP, [26]). In some cases, the use of such regularizers is associated with the exploitation of ℓ_1-reweighting techniques for minimization, see [19]. The key idea behind the use of concave penalties is that they are closer to the ℓ_0-norm with respect to the ℓ_1-norm, therefore they are expected to promote sparsity more accurately. We refer the reader to [20], [27] for their use in CS.

In this paper, we propose an MCP-based shrinkage and selection method, denoted as MCPS^2, which reads as follows.

$$\begin{aligned} \textbf{MCPS}^2: \quad & \min_{x \in [-d,d]^n} \mathcal{F}(x) := \frac{1}{2} \|y - Ax\|_2^2 + \lambda r(x) \\ & \text{where } r(x) := d \|x\|_1 - \frac{1}{2} \|x\|_2^2, \quad \lambda > 0. \end{aligned} \tag{4}$$

MCPS^2 is partially studied in [28], limited to CS of finite-valued signals, while in [22] an MCP cost functional is tested for the recovery of non-negative signals from low-precision data. Differently from those works, in this paper we extend the analysis to any real signal and we prove novel theoretical results.

For our analysis, we exploit the restricted eigenvalue condition. Let us define the cone

$$\mathcal{C}(\alpha, S) := \{x \in \mathbb{R}^n : \|x_{\bar{S}}\|_1 \leq \alpha \|x_S\|_1\}. \tag{5}$$

Definition 2: **Restricted eigenvalue condition (RE)**, see [2, Section 11.2.2]). We say that A satisfies (α, ϕ, S)-RE if there exist $\alpha > 0$ and $\phi > 0$ such that

$$\frac{\|Av\|_2^2}{\|v\|_2^2} \geq \phi \text{ for any } v \in \mathcal{C}(\alpha, S) \setminus \{0\}. \tag{6}$$

We remark that RE is generally weaker than IRR and than the restricted isometry property (RIP, see, e.g., [2, Definition 10.2]), the latter being largely used in CS; we refer the reader to [29] for a thorough study of the relationships between the different conditions considered in CS and sparse optimization. Moreover, while RIP is proven to hold for random, independent matrices, RE is proven to hold for a wider class of random, correlated matrices, see [30]. Thus, RE matches with a larger number of applications, such as autoregressive models.

We also remark that RE is used to evaluate the ℓ_2 error of Lasso (but not its VSC): Theorem 11.1 in [2] states that if $(3, \phi, S)$-RE holds and $\|\eta\|_\infty \leq m\lambda$, then $\|z - \widetilde{x}\|_2 \leq \frac{3\omega}{\omega - 1} \sqrt{k} \lambda$, where z is the Lasso solution. As explained in [2, Section 11.2.2], RE originates from the observation that $\|Ax - y\|_2^2$ is not strongly convex in CS, since $A^T A$ has not full rank, and the corresponding quadratic form is positive semi-definite. For this motivation, one looks for a strong convexity restricted to a significant subset, which for Lasso is the cone $\mathcal{C}(3, S)$.

III. Theoretical analysis

In this section, we analyse the VSC of MCPS^2 defined in (4) in a CS setting ($m < n$). We prove that a vector $x^\star$ with the same support of $\widetilde{x}$ is a local minimum of $\mathcal{F}$ in (4), under a relaxed (ω, S)-IRR (i.e., ω may be smaller than 1) and if (α, ϕ, S)-RE holds, provided that noise is sufficiently small. Therefore, a descent algorithm that starts sufficiently close to $x^\star$ can recover the right support. Furthermore, under additive conditions, $x^\star$ is proven to be the global minimum.

A. Local minimum

In the following results, we assume that $A_S^T A_S$ is positive definite ($A_S^T A_S \succ 0$); this is realistic in CS, since usually $k \leq m$ and the columns of A_S are linearly independent, see, e.g., [11], [2] for details.

Theorem 1: Let $y = A\widetilde{x} + \eta$, where $\widetilde{x} \in [-d, d]^n$ has support S, with $|S| = k$. We assume $A_S^T A_S \succ 0$, with minimum eigenvalue $\mathrm{a} > \lambda$. Let $\Omega := A_S^T A_S - \lambda I \succ 0$, $\mu := \frac{1}{d} \min_{i \in S} |\widetilde{x}_i| \in (0, 1]$, and $\gamma := \lambda d(1 - \mu) + \|A_S^T \eta\|_\infty$.

We define $x^\star \in [-d,d]^n$ as follows:

$$\begin{aligned} x^\star_S &:= \widetilde{x}_S + \Omega^{-1}\left[\lambda\widetilde{x}_S - \lambda d\text{sign}(\widetilde{x}_S) + A_S^T\eta\right] \\ x^\star_{\bar{S}} &:= 0. \end{aligned} \tag{7}$$

If the following condition holds:

C0. $\gamma\sqrt{k} < \mu d(\mathrm{a}-\lambda)$
then $\text{sign}(\widetilde{x}_S) = \text{sign}(x^\star_S)$.

Furthemore, given an arbitrarily small $\varepsilon > 0$, if the following additive conditions hold:

C1. A satisfies (ω, S)-IRR with $\omega > 1-\mu$,

C2. A satisfies (α, ϕ, S)-RE with $\phi > \lambda$,

C3. $\lambda d - \omega^{-1}\left(1+\frac{\lambda\sqrt{k}}{a-\lambda}\right)\gamma - \|A_{\bar{S}}^T\eta\|_\infty - \lambda\varepsilon\frac{1+\alpha}{\alpha} > 0$,
then, $x^\star$ is a local minimum of $\mathcal{F}(x)$, $x \in [-d,d]^n$.

Proof: A sufficient condition to obtain $\text{sign}(\widetilde{x}_S) = \text{sign}(x^\star_S)$ is $\|\widetilde{x}_S - x^\star_S\|_\infty < \mu d = \min_{i\in S}|\widetilde{x}_i|$. Moreover,

$$\begin{aligned} \|\widetilde{x}_S - x^\star_S\|_\infty &\le \|\Omega^{-1}\|_\infty\|\lambda\widetilde{x}_S - \lambda d\text{sign}(\widetilde{x}_S) + A_S^T\eta\|_\infty \\ &\le \sqrt{k}\|\Omega^{-1}\|_2\gamma \le \frac{\sqrt{k}}{a-\lambda}\gamma. \end{aligned}$$

Then, condition C0 can be obtained by upper bounding the last expression by μd.

To prove that $x^\star$ is a local minimum, we show that $\mathcal{F}(x^\star + h) > \mathcal{F}(x^\star)$ for a small $h \in \mathbb{R}^n \setminus \{0\}$, $x^\star + h \in [-d,d]^n$. In particular, we assume that $\|h\|_\infty \le \varepsilon$ for an arbitrarily small $\varepsilon > 0$. We have:

$$\begin{aligned} \mathcal{F}(x^\star + h) - \mathcal{F}(x^\star) =& \frac{1}{2}\|Ah\|_2^2 + h^TA^T(Ax^\star - y) + \\ &+ \lambda r(x^\star + h) - \lambda r(x^\star). \end{aligned} \tag{8}$$

Now, we assess the different terms in (8). First of all, we have $r(x^\star+h) - r(x^\star) = d\|x^\star + h\|_1 - d\|x^\star\|_1 - \frac{\|h\|_2^2}{2} - h^Tx^\star = d\|x^\star_S + h_S\|_1 + d\|h_{\bar{S}}\|_1 - d\|x^\star_S\|_1 - \frac{\|h\|_2^2}{2} - h_S^Tx^\star_S$. For each $i \in S$, we can assume that $|h_i| < |x^\star_i|$, so that $\text{sign}(x^\star_i + h_i) = \text{sign}(x^\star_i)$. Then, $\|x^\star_S + h_S\|_1 - \|x^\star_S\|_1 = h_S^T\text{sign}(x^\star_S)$ and $r(x^\star + h) - r(x^\star) = d\|h_{\bar{S}}\|_1 + h_S^T[d\text{sign}(x^\star_S) - x^\star_S] - \frac{\|h\|_2^2}{2}$. Furthermore, since $Ah = A_Sh_S + A_{\bar{S}}h_{\bar{S}}$,

$$\begin{aligned} &\mathcal{F}(x^\star + h) - \mathcal{F}(x^\star) = \\ &= \frac{1}{2}\|Ah\|_2^2 + h_{\bar{S}}^TA_{\bar{S}}^T(Ax^\star - y) + \lambda d\|h_{\bar{S}}\|_1 + \\ &+ h_S^T\left[\lambda d\text{sign}(x^\star_S) - \lambda x^\star_S + A_S^T(Ax^\star - y)\right] - \frac{\lambda}{2}\|h\|_2^2. \end{aligned}$$

From (7), the quantity within the square parenthesis is null. In fact, since $\text{sign}(x^\star_S) = \text{sign}(\widetilde{x}_S)$ and $Ax^\star - y = A_Sx^\star_S - A_S\widetilde{x}_S - \eta$, we obtain $\lambda d\text{sign}(x^\star_S) - \lambda x^\star_S + A_S^T(Ax^\star - y) = \lambda d\text{sign}(\widetilde{x}_S) + \Omega x^\star_S - A_S^TA_S\widetilde{x}_S - A_S^T\eta$. Then, by summing and subtracting $\lambda\widetilde{x}_S$, we obtain (7). Thus, we conclude that

$$\begin{aligned} &\mathcal{F}(x^\star + h) - \mathcal{F}(x^\star) = \\ &= \frac{1}{2}\|Ah\|_2^2 - \frac{\lambda}{2}\|h\|_2^2 + h_{\bar{S}}^TA_{\bar{S}}^T(Ax^\star - y) + \lambda d\|h_{\bar{S}}\|_1. \end{aligned} \tag{9}$$

Now, we compute a lower bound for $h_{\bar{S}}^TA_{\bar{S}}^T(Ax^\star - y)$, by exploiting the Hölder inequality.

$$\begin{aligned} |h_{\bar{S}}^TA_{\bar{S}}^T(Ax^\star - y)| &\le \|h_{\bar{S}}\|_1\left\|A_{\bar{S}}^T(Ax^\star - y)\right\|_\infty \\ &\le \|h_{\bar{S}}\|_1\left\|A_{\bar{S}}^TA_S(x^\star_S - \widetilde{x}_S)\right\|_\infty + \|h_{\bar{S}}\|_1\left\|A_{\bar{S}}^T\eta\right\|_\infty. \end{aligned} \tag{10}$$

Since $A_{\bar{S}}^TA_S(x^\star_S - \widetilde{x}_S) = A_{\bar{S}}^TA_S(A_S^TA_S)^{-1}(A_S^TA_S)(x^\star_S - \widetilde{x}_S)$, we elaborate on $\tau := A_S^TA_S(x^\star_S - \widetilde{x}_S)$ to exploit (ω, S)-IRR. In particular, we have $\|\tau\|_\infty = \|A_S^TA_S(x^\star_S - \widetilde{x}_S)\|_\infty = \left\|A_S^TA_S\Omega^{-1}[\lambda\widetilde{x}_S - \lambda d\text{sign}(\widetilde{x}_S) + A_S^T\eta]\right\|_\infty \le \left\|A_S^TA_S\Omega^{-1}\right\|_\infty\gamma \le \sqrt{k}\left(1+\frac{1}{a-\lambda}\right)\gamma$. Hence, by (ω, S)-IRR, we get

$$\begin{aligned} &\left\|A_{\bar{S}}^TA_S(A_S^TA_S)^{-1}\tau\right\|_\infty = \max_{j\in\bar{S}}\left\|A_j^TA_S(A_S^TA_S)^{-1}\tau\right\|_1 \\ &\le \max_{j\in\bar{S}}\left\|A_j^TA_S(A_S^TA_S)^{-1}\right\|_1\|\tau\|_\infty \\ &\le \frac{1}{\omega}\|\tau\|_\infty \le \frac{1}{\omega}\sqrt{k}\left(1+\frac{1}{a-\lambda}\right)\gamma. \end{aligned}$$

In conclusion, the following lower bound holds for (9):

$$\mathcal{F}(x^\star + h) - \mathcal{F}(x^\star) \ge \frac{1}{2}\|Ah\|_2^2 - \frac{\lambda}{2}\|h\|_2^2 + q_1\|h_{\bar{S}}\|_1$$

where $q_1 = \lambda d - \frac{1}{\omega}\sqrt{k}\left(1+\frac{1}{a-\lambda}\right)\gamma - \left\|A_{\bar{S}}^T\eta\right\|_\infty$.

Given that (α, ϕ, S)-RE holds, with $\phi > \lambda$, we distinguish two cases. If $h \in \mathcal{C}(\alpha, S) \setminus \{0\}$, then $\frac{1}{2}\|Ah\|_2^2 - \frac{\lambda}{2}\|h\|_2^2 \ge 0$. Thus, $\mathcal{F}(x^\star + h) > \mathcal{F}(x^\star)$ for any $h \in \mathcal{C}(\alpha, S) \setminus \{0\}$ whenever $q_1 > 0$.

Otherwise, if $h \notin \mathcal{C}(\alpha, S) \setminus \{0\}$, then $\|h_S\|_1 \le \frac{\|h_{\bar{S}}\|_1}{\alpha}$, which implies

$$\|h\|_2^2 \le \varepsilon\|h\|_1 = \varepsilon\|h_S\|_1 + \varepsilon\|h_{\bar{S}}\|_1 \le \varepsilon\|h_{\bar{S}}\|_1\frac{\alpha+1}{\alpha}.$$

Therefore, $\mathcal{F}(x^\star + h) - \mathcal{F}(x^\star) \ge \|h_{\bar{S}}\|_1 \ge q_1 - \lambda\varepsilon\frac{1+\alpha}{\alpha}$.

As $\|h_S\|_1 \le \frac{\|h_{\bar{S}}\|_1}{\alpha}$, if $h \neq 0$, then $h_{\bar{S}} \neq 0$. Thus, if $h \neq 0$, then $\mathcal{F}(x^\star + h) > \mathcal{F}(x^\star)$ whenever C3 holds. ∎

This result yields some considerations.

Remark 1: The condition $\omega - 1 + \mu > 0$ does not require $(1, S)$-IRR, which instead is necessary for the VSC of the Lasso. In fact, we just require (ω, S)-IRR with $\omega > 1 - \mu$, where $1 - \mu < 1$. In other terms, MCPS^2 requires a relaxed IRR which is tuned based on the minimum non-zero magnitude in $\widetilde{x}$. Interestingly, in the limit case where $\mu = 1$, no IRR is required, see Section III-C for details.

Remark 2: Theorem 1 states the possibility of achieving the desired $x^\star$ defined in (7) in the sense that $x^\star$ is a local minimum of the proposed functional, therefore we can reach it by a descent algorithm, given a suitable starting point.

Remark 3: The value of α is not assessed. Actually, since ε is arbitrarily small, in C3 we could neglect the term $\varepsilon(\alpha+1)$, thus remove α. For Lasso, (α, ϕ, S)-RE is useful when $\alpha = 3$, see [2, Theorem 11.1]. Moreover, the bound on the ℓ_2 error in [2, Theorem 11.1] is controlled by $\frac{3\lambda\sqrt{k}}{\phi}$, which suggests that $\phi \gg \lambda$ is necessary to have a small error. Instead, to prove Theorem 1, we only require $\phi \ge \lambda$.

Remark 4: Condition C3 and C0 determine the interplay between η and λ. Specifically, C0 requires for sufficiently small η and λ, while C3 states that λ must belong to an interval depending on η. The key idea is that if some

information about the maximum noise is given, along with the values of the system's parameters k, μ, a, ω, one can use them to assess the design paramater λ; the details, that can obtained by solving C3, are omitted for brevity. In particular, if $\eta = 0$ and $\mu = 1$, C3 reduces to $\lambda \in (0, \mathrm{a})$, which is the initial hypothesis of Theorem 1.

B. Global minimum

In this section, we provide sufficient conditions such that $x^\star$ defined in Theorem 1 is not only a local minimum, but also the global minimum of $\mathcal{F}$. From Theorem 1, let us assume that $x^\star$ is the unique minimum over $\mathcal{B}_\varepsilon(x^\star) := \{x^\star + h, \text{ for all } h \in \mathbb{R}^n \text{ with } \|h\|_\infty \leq \varepsilon\}$ for some $\varepsilon > 0$. Then, if the global minimum lies in $\mathcal{B}_\varepsilon(x^\star)$, it necessarily corresponds to $x^\star$. In the following proposition, we provide sufficient conditions for this occurrence.

Proposition 1: Let $\theta = \|(y - Ax^\star)^T A\|_\infty$. Let us assume $\theta < \lambda d$. If A satisfies (α, ϕ, S)-RE with $\alpha \geq \frac{2\lambda d + \theta}{\lambda d - \theta}$, $\phi > \lambda$, and $\frac{2(\theta + 2\lambda d)\sqrt{k}}{\phi - \lambda} \leq \varepsilon$, then, $x^\star$ defined in (7) is the global minimum of $\mathcal{F}$.

Proof: Let $x^{\star\star}$ be the global minimum of $\mathcal{F}$, and $\nu := x^{\star\star} - x^\star$. Then, $\mathcal{F}(x^\star + \nu) = \mathcal{F}(x^{\star\star}) \leq \mathcal{F}(x^\star)$. This is equivalent to $\frac{1}{2}\|A\nu + Ax^\star - y\|_2^2 + \lambda d\|x^\star + \nu\|_1 - \frac{\lambda}{2}\|x^\star + \nu\|_2^2 \leq \frac{1}{2}\|Ax^\star - y\|_2^2 + \lambda d\|x^\star\|_1 - \frac{\lambda}{2}\|x^\star\|_2^2$. Further simplifications lead to $\frac{1}{2}\|A\nu\|_2^2 + (Ax^\star - y)^T A\nu \leq \lambda d\|x^\star\|_1 - \lambda d\|x^\star + \nu\|_1 + \frac{\lambda}{2}\|\nu\|_2^2 + \lambda \nu_S^T x_S^\star$. Now, we remark that

$$\begin{aligned}\|x^\star\|_1 - \|x^\star + \nu\|_1 &= \|x_S^\star\|_1 - \|x_S^\star + \nu_S\|_1 - \|\nu_{\bar{S}}\|_1 \\ &\leq \|\nu_S\|_1 - \|\nu_{\bar{S}}\|_1.\end{aligned}$$

Moreover, since $x_S^\star = x_S^{\star\star} - \nu_S$, and $\|x^{\star\star}\|_\infty \leq d$,

$$\begin{aligned}\|\nu\|_2^2 + 2\nu_S^T x_S^\star &= \|\nu_S\|_2^2 + \|\nu_{\bar{S}}\|_2^2 + 2\nu_S^T x_S^{\star\star} - 2\|\nu_S\|_2^2 \\ &\leq -\|\nu_S\|_2^2 + \|\nu_{\bar{S}}\|_2^2 + 2d\|\nu_S\|_1.\end{aligned}$$

Then,

$$\frac{1}{2}\|A\nu\|_2^2 \leq (\theta + \lambda d)\|\nu_S\|_1 + \theta\|\nu_{\bar{S}}\|_1 + \lambda r(v_S) - \lambda r(v_{\bar{S}}).$$

If A satisfies (α, ϕ, S)-RE, then $\frac{\phi+\lambda}{2}\|\nu_S\|_2^2 + \frac{\phi-\lambda}{2}\|\nu_{\bar{S}}\|_2^2 \leq (\theta + 2\lambda d)\|\nu_S\|_1 + (\theta - \lambda d)\|\nu_{\bar{S}}\|_1$. If $\phi > \lambda$, then $0 \leq \frac{\phi+\lambda}{2}\|\nu_S\|_2^2 + \frac{\phi-\lambda}{2}\|\nu_{\bar{S}}\|_2^2$, and $(\lambda d - \theta)\|\nu_{\bar{S}}\|_1 \leq (\theta + 2\lambda d)\|\nu_S\|_1$. Therefore, if $\theta < \lambda d$, then $\nu \in \mathcal{C}\left(\frac{2\lambda d + \theta}{\lambda d - \theta}, S\right)$. Moreover,

$$\begin{aligned}&\frac{\phi - \lambda}{2}\|\nu\|_2^2 \leq (\theta + 2\lambda d)\|\nu_S\|_1 \leq (\theta + 2\lambda d)\sqrt{k}\|\nu\|_2 \\ &\quad \Rightarrow \|\nu\|_2 \leq \frac{2(\theta + 2\lambda d)\sqrt{k}}{\phi - \lambda}.\end{aligned} \tag{11}$$

If $\|\nu\|_2 \leq \varepsilon$, then $x^{\star\star} \in \mathcal{B}_\varepsilon(x^\star)$. Since $x^{\star\star}$ is the global minimum, while $x^\star$ is the global minimum limited to $\mathcal{B}_\varepsilon(x^\star)$, then $\|\nu\|_2 \leq \varepsilon$ implies $x^{\star\star} = x^\star$. According to (11), a sufficient condition to have $\|\nu\|_2 \leq \varepsilon$ is $\frac{2(\theta+2\lambda d)\sqrt{k}}{\phi-\lambda} \leq \varepsilon$. ■

The bound in (11) might be refined with further computations, which will be proposed in a future extended work. However, even though not perfectly tight, it well illustrates the robustness to noise: if η decreases, and as a consequence λ and θ can be proportionally decreased, then also $\|\nu\|_2$ decreases, which increases the probability that $x^\star$ is the global minimum.

C. Noise-free, limit case

To conclude, we illustrate the particular case with no noise, i.e., $\eta = 0$, and non-zeros concentrated on the boundaries of $[-d, d]^n$, i.e., $\mu = 1$. From Theorem 1 and Proposition 1, we derive the following result.

Corollary 1: Let us assume $\eta = 0$ and $\mu = 1$. If A satisfies (ω, S)-IRR, and A satisfies (α, ϕ, S)-RE with $\phi > \lambda$, and $\alpha = 2$, then $x^\star = \widetilde{x}$ is a local minimum of $\mathcal{F}$. More precisely, $x^\star = \widetilde{x}$ is the unique minimum in $\mathcal{B}_\varepsilon(x^\star)$, with $\varepsilon = \frac{2}{3}d\omega$. Moreover, if $\lambda < \frac{\omega\phi}{6\sqrt{k}+\omega}$, then $\widetilde{x}$ is the global minimum of $\mathcal{F}$.

The proof, omitted for brevity, can be straightforwardly obtained by replacing $\eta = 0$ and $\mu = 1$ in Theorem 1 and Proposition 1. Corollary 1 shows that in the favorable case without noise and with extreme non-zero values, the proposed approach is always effective, since it is sufficient to set a sufficiently small λ to obtain that the true $\widetilde{x}$ is the global minimum of $\mathcal{F}$, without bias. For more details on CS with discrete-valued signals, see [28].

D. Discussion

The proposed analysis is theoretical, as some parameters, such as ω and ϕ, are not a priori known, and therefore the choice of the design parameter λ is not precisely determined. The main contribution of this analysis is to state that MCPS^2, differently from Lasso, does not require the classical IRR. Similarly to classical CS, in future work, conditions on A will be studied that are provable in practice, at least for matrices' ensembles, such as the restricted isometry property. From a practical viewpoint, we expect that the relaxation of IRR leads to a reduction of the number m of measurements needed for a perfect recovery of the support; this is illustrated in numerical simulations in the next section.

Finally, we remark that finding the global minimum of $\mathcal{F}$ may be not straightforward, since $\mathcal{F}$ is non-convex. In the next section, we test different algorithms to achieve the global minimum, and show their effectiveness through numerical simulations.

IV. Algorithms and numerical simulations

In this section, we test different algorithms to achieve the minimum of $\mathcal{F}$ as defined in (4); this task is challenging due to non-convexity. We propose two approaches: the semidefinite programming relaxation (SDR), supported by recent results on polynomial optimization [23], and the alternating direction method of multipliers (ADMM, [31]).

A. Semi-algebraic optimization

Since $\mathcal{F}$ in (4) is semi-algebraic, the theory developed in [32] can be applied to compute the global minimum. In a nutshell, given a polynomial or semi-algebraic optimization problem, a hierarchy of SDR's can be constructed, whose solutions converge to the global optimal solution. The hierarchy

generically has finite convergence, see, e.g., [33]. Therefore, the global minimum can be achieved by solving an SDR of sufficiently large order. A shortcoming of the SDR approach is the numerical complexity, which is of order $\mathcal{O}(n^\zeta)$, n being the number of variables and ζ the relaxation order. For this motivation, in this paper, we consider only the SDR of order $\zeta = 1$, which corresponds to the Shor's relaxation.

B. ADMM

ADMM is an iterative algorithm, widely used in convex optimization for its fast convergence and simplicity of implementation, see [31]. In the non-convex setting, the convergence of ADMM to a local minimum has been proven only for some classes of functionals. In particular, in [34], convergence is proven for non-convex functionals that can be split into the sum of a non-convex, smooth term and of a convex, not necessarily smooth, term. $\mathcal{F}$ in (4) has this property, i.e., we can write it as

$$\min_{x,z\in[-d,d]^n} \frac{1}{2}\|y - Ax\|_2^2 - \frac{\lambda}{2}\|x\|_2^2 + \lambda d\|z\|_1 \quad \text{s. t.} \quad z = x$$

with The associated augmented Lagrangian is: $\mathcal{L}(x,z) = \frac{1}{2}\|y - Ax\|_2^2 - \frac{\lambda}{2}\|x\|_2^2 + \lambda d\|z\|_1 + u^T(x-z) + \frac{\rho}{2}\|x-z\|_2^2$ where $u \in \mathbb{R}^n$ is the dual variable, and $\rho > 0$. Then, we apply ADMM as explained in [34, Section 2], which consists in iteratively minimizing $\mathcal{L}$ with respect to x and to z, and updating u. This procedure is summarized in Algorithm 1, where P denotes the operator that projects onto $[-d,d]^n$, and $\mathbb{S}_a$ is the soft thresholding operator.

Algorithm 1 ADMM for MCPS^2

Input: $A, y = A\widetilde{x} + \eta$, $\lambda > 0$, $\rho > 0$

Output: $x_{T_{stop}}$ = estimate of $\widetilde{x}$

1: Initialize $z_0, w_0 \in \mathbb{R}^n$

2: **for all** $t = 1, \dots, T_{stop}$ **do**

3: $x_t = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \mathcal{L}(x, z_{t-1})$
$= \left[A^TA + (\rho - \lambda)I\right]^{-1}\left(A^Ty + \rho z_{t-1} - u_{t-1}\right)$

4: $z_t = \underset{z\in[-d,d]^n}{\operatorname{argmin}} \mathcal{L}(x_t, z) = P\left(\mathbb{S}_{\frac{\lambda d}{\rho}}\left(x_t + \frac{u_{t-1}}{\rho}\right)\right)$

5: $u_t = u_{t-1} + \rho(x_t - z_t)$

6: **end for**

C. Numerical results

We illustrate some numerical simulations with SDR and ADMM approaches to solve MCPS^2. In particular, we compare them to Lasso, solved with ADMM. The considered setting is as follows. The vector $\widetilde{x}$ that we aim to recover has length $n = 100$ and sparsity $k = 5$: its support is generated uniformly at random, and its non-zero entries have random magnitudes in $[\frac{1}{2}, 1]$. The available measurements are $y = A\widetilde{x} + \eta$, where $A \in \mathbb{R}^{m,n}$ has Gaussian independent entries $\mathcal{N}(0, \frac{1}{m})$, $m \in [10, 60]$. The measurement noise $\eta \in \mathbb{R}^m$ is Gaussian as well, and we consider a signal-to-noise-ratio SNR $= 25$ dB. The parameter λ is set to 10^{-2} via cross-validation. The considered approaches are compared in terms of VSC, i.e., the number of experiments where the support is exactly recovered, and false positive/negative rate, that is the rate of zeros estimated as non-zeros, and vice-versa. Finally, we compare the runtimes. All the algorithms are implemented in C++; for SDR, we use the Mosek C++ Fusion API, see [35].

The results, averaged over 200 runs, are shown in Figure 1. We see that MCPS^2 performs better than Lasso in terms of VSC, as expected from the proposed theoretical results in Section III. The enhancement is particularly evident for $m \in [20, 35]$. For example, for $m = 30$, Lasso is not reliable, with the 60% of successful support recovery, while MCPS^2 attains the 90%. Furthermore, we can see that MCPS^2 outperforms Lasso in terms of false positive rate, while the false negative rate of SDR is better than that of ADMM. Finally, The runtime is sufficiently fast for all the considered methods, SDR being a bit slower than ADMM.

As to MCPS^2, SDR and ADMM algorithms both suboptimal: on the one hand, SDR can be improved by increasing the relaxation order, now set to 1 to minimize the runtime; on the other hand, for ADMM, suitable initial conditions could be investigated to achieve the desired minimum. Moreover,in the proposed experiments, we observe that when k is correctly estimated, then the achieved point corresponds to $x^\star$. This suggests that $x^\star$ might be the unique minimum with sparsity k; in case, given the knowledge of k, this could be used to verify whether the achieved point is $x^\star$, and, if not, ADMM should be run again with different initial conditions to search $x^\star$. These points will be studied in future extended work.

V. Conclusions

In this work, we analyse MCPS^2, a non-convex, semi-algebraic optimization problem for sparse learning. Specifically, we study its variable-selection consistency in a compressed sensing framework, and we prove that, differently from Lasso, MCPS^2 relaxes the irrepresentable condition. In practice, this implies that MCPS^2 requires less measurements than Lasso. Future work will be oriented to provide conditions guaranteeing the variable-selection consistency of MCPS^2 that can be a priori verified. and to develop optimal strategies to achieve the global minimum.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer New York Inc., 2009.

[2] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, 2nd ed. CRC press, 2015.

[3] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.

[4] A. Y. Carmi, "Compressive system identification: Sequential methods and entropy bounds," *Dig. Signal Process.*, vol. 23, no. 3, pp. 751–770, 2013.

[5] S. Fattahi and S. Sojoudi, "Data-driven sparse system identification," in *Proc. Allerton Conf. Commun. Control Comput.*, 2018, pp. 462–469.

[6] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *PNAS*, vol. 113, no. 15, pp. 3932–3937, 2016.

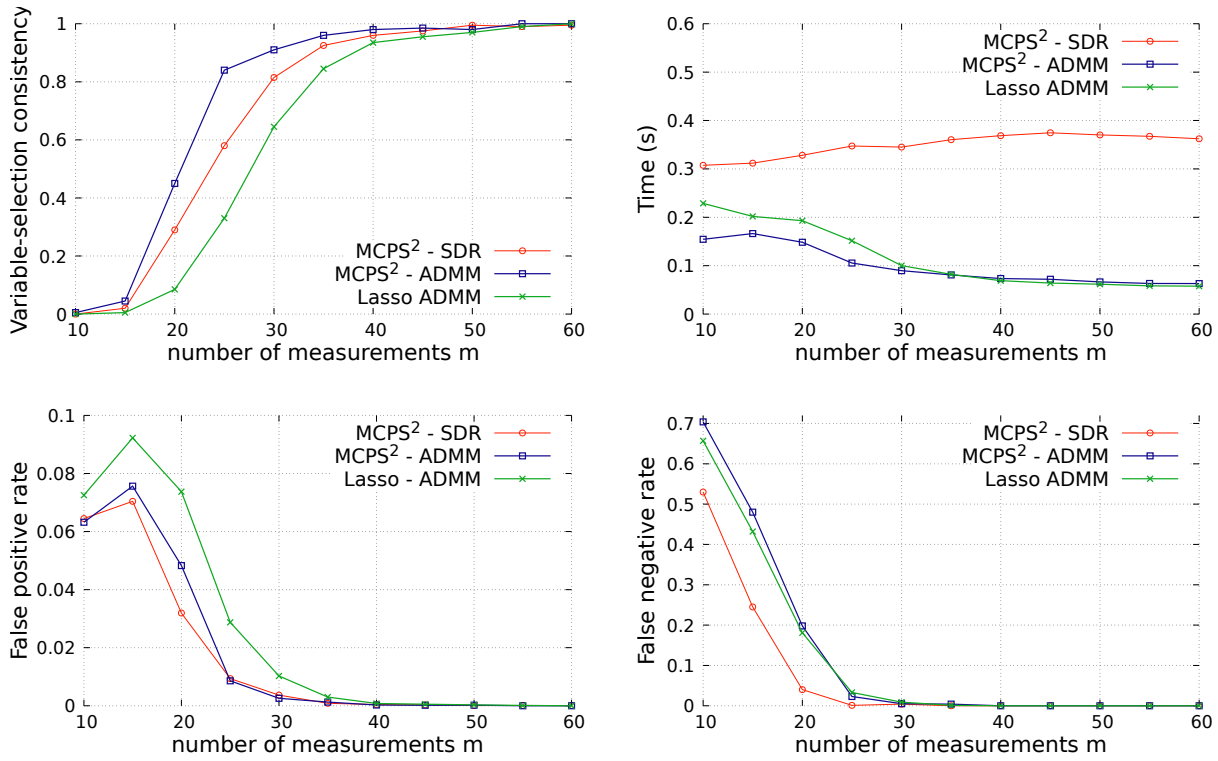

Fig. 1: Numerical results: comparison between the proposed MCPS^2 optimization problem, solved via SDR and ADMM, and the Lasso, solved via ADMM.

[7] M. Gallieri, *Lasso-MPC - Predictive Control with ℓ_1-Regularised Least Squares*. Springer Int., 2016.

[8] E. Tartaglione, S. Lepsøy, A. Fiandrotti, and G. Francini, "Learning sparse neural networks via sensitivity-driven regularization," in *Proc. Int. Conf. Neural Inf. Process. Sys. (NIPS)*, 2018, pp. 3882–3892.

[9] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, "Optimal approximation with sparsely connected deep neural networks," *SIAM J. Math. Data Sci.*, vol. 1, no. 1, pp. 8–45, 2019.

[10] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[11] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. New York: Springer, 2013.

[12] R. Tóth, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Compressive system identification in the linear time-invariant framework," in *Proc. IEEE Conf. Decis. Control (CDC)*, 2011, pp. 783–790.

[13] B. M. Sanandaji, T. L. Vincent, M. B. Wakin, and R. Tóth, "Compressive system identification of lti and ltv arx models," in *Proc. IEEE Conf. Decis. Control (CDC)*, 2011, pp. 783–790.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Series B*, vol. 58, pp. 267–288, 1996.

[15] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

[16] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541 – 2563, 2006.

[17] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[18] P. Bühlmann and S. van de Geer, *Variable selection with the Lasso*. Springer Berlin Heidelberg, 2011, pp. 183–247.

[19] E. J. Candès, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5-6, pp. 877–905, 2008.

[20] J. Woodworth and R. Chartrand, "Compressed sensing recovery via nonconvex shrinkage penalties," *Inverse Problems*, vol. 32, no. 7, pp. 75 004–75 028, 2016.

[21] I. Selesnick, "Sparse regularization via convex analysis," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4481–4494, 2017.

[22] V. Cerone, S. M. Fosson, and D. Regruto, "Sparse linear regression with compressed and low-precision data via concave quadratic programming," in *Proc. Conf. Decis. Control (CDC)*, 2019, pp. 6971–6976.

[23] J.-B. Lasserre, *An introduction to polynomial and semi-algebraic optimization*. Cambridge University Press, 2015.

[24] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Stat.*, vol. 34, pp. 1436–1462, 2006.

[25] S. Foucart and M.-J. Laui, "Sparsest solutions of underdetermined linear systems via ℓ_q minimization for $0 < q \leq 1$," *Appl. Comput. Harmon. Anal.*, vol. 26, pp. 395–407, 2009.

[26] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.

[27] S. M. Fosson, "A biconvex analysis for lasso ℓ_1 reweighting," *IEEE Signal Process. Lett.*, vol. early access, no. nn, pp. 1 – 1, 2018.

[28] ——, "Non-convex Lasso-kind approach to compressed sensing for finite-valued signals," *arxiv.org/abs/1811.03864v2*, 2018.

[29] S. A. van de Geer and P. Bühlmann, "On the conditions used to prove oracle results for the Lasso," *Electron. J. Stat.*, vol. 3, pp. 1360–1392, 2009.

[30] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *J. Machine Learn. Res.*, vol. 11, pp. 22 641–2259, 2010.

[31] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1 – 122, 2010.

[32] J. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM J. Optim.*, vol. 11, no. 3, pp. 796–817, 2001.

[33] J. Nie, "Optimality conditions and finite convergence of lasserre's hierarchy," *Math. Program.*, vol. 146, no. 1, pp. 97–121, 2014.

[34] M. Hong, Z. Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, 2016.

[35] MOSEK, *Fusion API for C++, version 9.1.5*, 2019. [Online]. Available: www.mosek.com/documentation/