Published in final edited form as:
Math Program. 2017 November ; 166(1-2): 207–240. doi:10.1007/s10107-017-1114-y.

Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory for local solutions

Hongcheng Liu,
Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park PA 16802, USA

Tao Yao,
Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park PA 16802, USA

Runze Li, and
Department of Statistics and the Methodology Center, The Pennsylvania State University, University Park PA 16802, USA

Yinyu Ye
Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA

Abstract

This paper concerns the *folded concave penalized sparse linear regression* (FCPSLR), a class of popular sparse recovery methods. Although FCPSLR yields desirable recovery performance when solved globally, computing a global solution is NP-complete. Despite some existing statistical performance analyses on local minimizers or on specific FCPSLR-based learning algorithms, it still remains open questions whether local solutions that are known to admit fully polynomial-time approximation schemes (FPTAS) may already be sufficient to ensure the statistical performance, and whether that statistical performance can be non-contingent on the specific designs of computing procedures. To address the questions, this paper presents the following threefold results: (i) Any local solution (stationary point) is a sparse estimator, under some conditions on the parameters of the folded concave penalties. (ii) Perhaps more importantly, any local solution satisfying a *significant subspace second-order necessary condition* (S^{3}ONC), which is weaker than the second-order KKT condition, yields a bounded error in approximating the true parameter with high probability. In addition, if the minimal signal strength is sufficient, the S^{3}ONC solution likely recovers the oracle solution. This result also explicates that the goal of improving the statistical performance is consistent with the optimization criteria of minimizing the suboptimality gap in solving the non-convex programming formulation of FCPSLR. (iii) We apply (ii) to the special case of FCPSLR with *minimax concave penalty* (MCP) and show that under the *restricted eigenvalue* condition, any S^{3}ONC solution with a better objective value than the Lasso solution entails the strong oracle property. In addition, such a solution generates a *model error* (ME)

Mathematics Subject Classification (2010) 90C26 · 90C90 · 62J05 · 62J07 · 68Q25

comparable to the optimal but exponential-time sparse estimator given a sufficient sample size, while the worst-case ME is comparable to the Lasso in general. Furthermore, to guarantee the S^{3}ONC admits FPTAS.

Keywords

Sparse recovery; Non-convex programming; NP-completeness; Folded concave penalty; Lasso

1 Introduction

Consider a linear regression model $b_j = \mathbf{a}_{j\cdot}^\top \mathbf{x}^{true} + \varepsilon_j$, $j = 1, \cdots, n$. Denote $\mathbf{A} := (\mathbf{a}_{1\cdot}, \ldots, \mathbf{a}_{n\cdot})^\top \in \Re^{n\times p}$, $\mathbf{b} := (b_1, \ldots, b_n)^\top$. and $W := (\varepsilon_1, \ldots, \varepsilon_n)^\top$ be the design matrix, response vector and error vector, respectively. Our target is to reconstruct the *true parameter* $\mathbf{x}^{true}$ given only finitely many observations of data ($\mathbf{A}$, $\mathbf{b}$), when the problem dimension p is allowed to be (much) larger than the sample size n but $\mathbf{x}^{true}$ is assumed to be sparse.

Following the literature (e.g., [6,43,27,31]), we quantify recovery quality by using *model error* (ME), *absolute deviation* (AD, i.e., ℓ_1 loss), and ℓ_2 *loss*: ME:

$$\text{ME:}\ \ \frac{1}{n}\|\mathbf{A}(\mathbf{x}-\mathbf{x}^{true})\|^2; \quad \text{AD:}\ |\mathbf{x}-\mathbf{x}^{true}|; \quad \ell_2\ \text{loss:}\ \|\mathbf{x}-\mathbf{x}^{true}\|. \tag{1}$$

Here |·| and ||·|| denote the ℓ_1-norm and ℓ_2-norm, respectively. People also considers the presence of the (strong) oracle property an important performance index ([14,15,17]). In [14], an estimator is said to have the *oracle property* if its asymptotic distribution is the same as the *oracle estimator* (*oracle solution*). In [15], an estimator is said to have the *strong oracle property* if with overwhelming probability, the estimator equals the following oracle solution

$$\mathbf{x}^{oracle} \in \underset{\mathbf{x}\in\Re^p:\ x_i=0,\ \forall i\in\mathscr{S}^c}{\arg\inf} \frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2, \tag{2}$$

where $\mathscr{S} := \{i : x_i^{true} \neq 0\}$ is called the true support set and its complement $\mathscr{S}^c := \{i : x_i^{true} = 0\}$. Let $|\mathscr{S}|$ be the cardinality of $\mathscr{S}$. Throughout the paper it is assumed that $|\mathscr{S}| \ll n \ll p$ and $n >$ 1. As in [14], the oracle solution is a statistically desirable solution that assumes *a priori* knowledge on $\mathscr{S}$. Explicit bounds on the recovery quality of the oracle solution in terms of ME, AD, and ℓ_2 loss can be obtained from theory of least squares estimator ([21]). Since $\mathscr{S}$ is unknown in practice, the use of (2) is merely for theoretical purposes.

Statisticians ([32,14]) use penalized least squares method to recover $\mathbf{x}^{true}$:

$$\inf_{\mathbf{x}:=(x_i:1\le i\le p)\in\Re^p}\left[f(\mathbf{x}):=\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\sum_{i=1}^{p}P_\lambda(|x_i|)\right]. \quad (3)$$

The objective function in (3) is the sum of the least squares function and a nonnegative penalty function P_λ that encourages sparsity. The choices of the penalty functions have been studied in literature ([16]). The penalized least squares with the ℓ_1 penalty, one of the most popular penalties, yield the Lasso [32]:

$$\mathbf{x}^{lasso}\in\arg\inf_{\mathbf{x}\in\Re^p}\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\lambda_{lasso}|\mathbf{x}|. \quad (4)$$

For FCPSLR problem, P_λ is set to be a *folded concave penalty* (FCP) satisfying the following properties for given $a, \lambda \in \Re_{++}$: (i) $P_\lambda(t)$ is non-decreasing and concave in $t \in \Re_+$ with $P_\lambda(0) = 0$ and $P_\lambda(t) > 0$ if $t > 0$; (ii) $P_\lambda(t)$ is differentiable at any $t \in \Re_+$; (iii) the first derivative $P_\lambda'(t)=0$ for any $t \ge a\lambda$; (iv) $0\le P_\lambda'(t)\le\lambda$ for any $t \ge 0$.

This paper will focus on two commonly used FCPs: the *smoothly clipped absolute deviation* (SCAD, [14]), given as

$P_{\lambda,SCAD}(t):=\lambda t\mathbb{I}(0\le t\le\lambda)+\frac{1}{a-1}\left(-\frac{\lambda^2}{2}+a\lambda t-\frac{1}{2}t^2\right)\mathbb{I}(\lambda<t\le a\lambda)+\frac{1}{2}(a+1)\lambda^2\mathbb{I}(t>a\lambda)$; and the *minimax concave penalty* (MCP, [41]), given as

$P_{\lambda,MCP}(t):=(\lambda t-\frac{t^2}{2a})\mathbb{I}(0\le t\le a\lambda)+\frac{1}{2}a\lambda^2\mathbb{I}(t>a\lambda)$, where $a > 1$ for SCAD and $a > 0$ for MCP. Here $\mathbb{I}(\cdot)$ is an indicator function, and $(\cdot)_+$: = max{0, ·}.

As shown in [17], the FCP entails desirable properties, including "unbiasedness", "sparsity", and "continuity". Thus, FCP may be intuitively more preferable than the Lasso and $\ell_{\mathbf{p}}$-penalties in general ($0 \le \mathbf{p} \le \infty$). Furthermore, under some conditions, global solutions to FCPSLR have the oracle property ([44]), while the Lasso does not have the oracle property.

Nonetheless, the FCP renders (3) non-convex, and thus there are limited optimization theories to analyze this problem. Existing solution techniques are also scarce to solve this problem globally. [24] proposes perhaps the first global scheme called MIPGO, which reformulates (3) into a mixed integer linear program (MIP), allowing FCPSLR to be solved with theoretically ascertained global optimality. Still, the theoretical worst-case complexity of MIP grows exponentially in the problem scale in general, although admittedly many MIPs can be solved with reasonable overhead in practice and there has been successful applications of MIP to least quantile regression problems by [3].

Such computational complexity is not surprising in theory, as FCPSLR is claimed to be NP-hard by [44] and [38]. [23] provides a formal proof for the NP-hardness of sparse linear regression with SCAD and some other penalty functions. In a more general case, [4] and [18] show that the minimization of a sparse regression problem with a "concave" and "monotone" penalty function is strongly NP-hard. [24] reformulates FCPSLR into an indefinite quadratic program. Since indefinite quadratic programs are in NP according to [34], we know that FCPSLR is in fact NP-complete.

In view of the NP-hardness in global minimization, several studies seek to solve the FCPSLR locally (see, e.g., [14,17,37,38]). Some existing studies, such as [14] and [15], show the existence of *local minimizers* that have the oracle property. Other reported theoretical findings, such as those by [17], [38], and [37], study specific FCPSLR-based learning algorithms in the form of local optimization procedures. Simulation studies in [37,17] imply local solutions[1] of FCPSLR outperform the Lasso.

In this paper, we consider local solutions to FCPSLR that satisfy a second-order necessary condition, called *significant subspace second-order necessary condition* (S^{3}ONC), which is weaker than the second-order KKT condition. We show that, at those S^{3}ONC solutions, the sound recovery quality is an intrinsic property, regardless of the choice of solution procedures. The S^{3}ONC relaxes the condition of *local minimality* in [14] and [15], and admits a fully polynomial-time approximation scheme (FPTAS, whose complexity is polynomial in both dimension and solution accuracy, but not necessarily polynomial in the bit length of accuracy). In contrast to [17,38,37], our analysis is algorithm-independent.

Specifically, inspired by [12,13], we present conditions on the choice of parameters a and λ to ensure the desired sparsity of local solutions based on a first-order necessary condition (FONC) and the S^{3}ONC. With either of these two conditions, we show that any dimension of a local solution is necessarily zero once its magnitude is smaller than an explicit threshold and that the total number of non-zero variables at a local solution is bounded from above. Our results imply that, under our conditions, any local solution is sparse.

Perhaps more interestingly, if the random error vector W is sub-Gaussian and $\mathbf{A}$ satisfies the *restricted eigenvalue (RE)* condition ([6]), we show that any solution satisfying the S^{3}ONC for FCPSLR may yield a bounded error in approximating the true parameter with high probability and even exactly recover the oracle solution. Furthermore, the statistical performance of FCPSLR is related with the optimization quality in minimizing the FCPSLR formulation. More precisely, the aforementioned error of an S^{3}ONC solution improves polynomially when the suboptimality gap decreases.

We apply the above findings to the S^{3}ONC solutions that have smaller objective values than a Lasso solution, namely, the S^{3}ONC solutions in the sub-level set $\{\mathbf{x}: f(\mathbf{x}) \le f(\mathbf{x}^{lasso})\}$, in the case of FCPSLR with MCP. Under the RE condition, we show that those local solutions have the strong oracle property, while, in contrast, the Lasso does not have the oracle property. Furthermore, when the sample size is above a certain threshold polynomial in ln p, those S^{3}ONC solutions can achieve an ME comparable to the optimal but exponential-time estimator of the form

[1] Throughout this paper, a “local solution” refers to a solution that at least satisfies the first-order KKT condition, and may or may not satisfy a second-order necessary condition.

$$\mathbf{x}^{exp} \in \underset{\mathbf{x}\in\Re^p:\|\mathbf{x}\|_0\leq|\mathscr{S}|}{\arg\inf} \frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|, \qquad (5)$$

which is shown by [43,31] to outperform the Lasso and possibly all the polynomial-time estimators in terms of ME. In the meantime, the worst-case ME of those S^{3}ONC solutions is comparable to the Lasso.

Our results based on the S^{3}ONC have some important differences from the analyses by [26], which shows that all solutions satisfying the FONC share the same upper bound on their distances from the true parameter under a *restricted strong convexity* (RSC) condition. Since our finding differentiates local solutions by their sub-level sets, our results may better explain the variation in performance among local solutions achieved by the same solution technique with different initial points, as observed from simulations by [17] and [37].

To ensure the statistical properties, we impose the RE condition on **A**. It is shown by [33] that RE is a relaxation of the restricted isotropy property (RIP, introduced by [8]) for some choices of parameters and is considered as one of the weakest conditions on the design matrices to ensure statistical performance for the Lasso as per [30]. Also under RE condition, [17] and [37] show respectively that FCPSLR solved specifically by local linear approximation (LLA) approach and by ConCave Convex procedure (CCCP), if initialized with the Lasso solution, results in the (strong) oracle property. We should note that, although both the RIP and the RE are established on fixed design matrices, some literature focuses on the probability for a random design matrix to satisfy the RIP or the RE condition. When design matrices are random, an isotropicity condition is often necessary for the RIP. To ensure the RE condition in random design matrices, the isotropicity condition can be dropped for Gaussian or subgaussian random designs [46] and [29]. Some special cases without subgaussian assumptions are presented by [30].

One remaining question is how to compute a solution satisfying S^{3}ONC at a reasonable overhead. Hereafter, an algorithmis said to be *S^{3}ONC-guaranteeing*, if it generates a solution that satisfies S^{3}ONC at convergence or at termination. Any algorithm that ensures the second-order KKT condition is S^{3}ONC-guaranteeing, since S^{3}ONC is weaker than the second-order KKT condition. To our knowledge, scarcely is there any discussion on S^{3}ONC-guaranteeing computing procedures in the statistics literature in solving FCPSLR, despite the several studies on solving the problem locally [14,26,17,38]. Instead, existing techniques yield local solutions that are only known to satisfy the first-order KKT condition. Nonetheless, in the optimization literature, algorithms ensuring a second-order KKT condition in non-convex optimization have been much studied by, for instance, [39,40,5,28,11]. In particular, the interior point algorithm by [5] is an FPTAS in achieving a second-order KKT solution with ε inaccuracy. In this paper, we elect to employ a S^{3}ONC-guaranteeing potential reduction (PR) method as an adaptation of [39] and [40].

The rest of the paper is organized as follows. Section 2 formally states the S^{3}ONC and presents the conditions for any local solution to be a sparse estimator. Section 3 presents our theoretical results on statistical properties at any solution satisfying the S^{3}ONC. Section 4 presents proofs and auxiliary lemmas for the results in Section 3. Section 5 briefly discusses the PR algorithm and summarizes some preliminary numerical results. Details of both the algorithm and the test results are presented in the online supplement [25]. Finally, Section 6 concludes this paper with final remarks.

Throughout this paper, we will denote by $\|\cdot\|_{\mathbf{p}}$ the $\ell_{\mathbf{p}}$-norm, except that $\|\cdot\|_0$ denotes the number of non-zero entries. For a finite set, $|\cdot|$ denotes the cardinality. Considering an arbitrary vector $\mathbf{x}$, we denote that $\mathbf{x}_{\mathscr{S}}: = (x_i: i \in \mathscr{S})$ and $\mathbf{x}_{\mathscr{S}^c}: = (x_i: i \in \mathscr{S}^c)$, which are subvectors of $\mathbf{x}$. For any index set $\hat{S}$, we denote by $\hat{S}^c$ the complement of $\hat{S}$ with respective to $\{1,\ldots, p\}$. We will also use the abbreviation “a.s.” for “almost surely”. When we present results indifferent between the SCAD and the MCP cases, we will refer to both FCPSLR with SCAD and FCPSLR with MCP as FCPSLR for convenience. Accordingly, we will use P_λ to denote both $P_{\lambda,SCAD}$ and $P_{\lambda,MCP}$. Otherwise, we may use FCPSLR-SCAD and FCPSLR-MCP to differentiate the two.

2 Necessary Optimality Conditions and Their Implications to Sparsity

This section first presents in Section 2.1 the necessary optimality conditions, including the FONC and the S^{3}ONC. Then, as implications of those necessary conditions, in Sections 2.2.1 and 2.2.2, we provide some sparsity properties of both FCPSLR-SCAD and -MCP, inspired by [12,13]: we show that each dimension of a local solution is necessarily zero once its magnitude is smaller than an explicit threshold. Such a threshold differentiates between solutions satisfying the FONC and those satisfying the S^{3}ONC. Utilizing that threshold, we derive the upper bounds on $\|\mathbf{x}^*\|_0$ of a local solution $\mathbf{x}^*$. These bounds are useful to estimate the magnitude and the number of the non-zero dimensions of a local solution using information that is computationally cheap to acquire. Denote by $\mathbf{a}_{\cdot i}$ the i-th column of $\mathbf{A}$ for $i = 1, \ldots, p$ throughout this paper.

2.1 Necessary conditions

The results in this section rely heavily on the following necessary conditions for a local minimal solution to (3).

First-order necessary condition (FONC): Solution $\mathbf{x}^*$ satisfies:

$$\exists \mathscr{D}(\mathbf{x}^*) \in \frac{1}{n}\mathbf{A}^\top(\mathbf{A}\mathbf{x}^* - \mathbf{b}) + (P_\lambda'(x_i^*)\partial|x_i^*| : 1 \le i \le p) \quad \mathrm{s.\ t.} \qquad \mathscr{D}(\mathbf{x}^*) = 0, \tag{6}$$

where $\partial|\cdot|$ denotes the subdifferential of $|\cdot|$.

Significant subspace second-order necessary condition (S^{3}ONC): Solution $\mathbf{x}^*$ satisfies FONC. Furthermore, for all $i \in \{i : x_i^* \neq 0\}$,

$$\frac{\partial^2 f(\mathbf{x})}{(\partial x_i)^2}\Big|_{\mathbf{x}=\mathbf{x}^*} \geq 0 \qquad (7)$$

if the second-order derivative exists.

The S^{3}ONC is based on the fact that a local minimal solution in the entire space must be a local minimizer in the subspace that considers only a single non-zero variable (See also [12]). Apply this observation to each of the significant (i.e., non-zero) dimensions, we obtain the second-order necessary condition in (7).

2.2 Sparsity at local solutions

In the subsequent, we present a set of bounds on the magnitude and the number of the non-zero dimensions at any local solution satisfying either the FONC or the S^{3}ONC. Specifically, Theorem 1 presents some general sparsity results for solutions satisfying the FONC. Corollaries 1 and 2 then apply Theorem 1 to the special cases of SCAD and MCP, respectively. In Subsection 2.2.2, Theorem 2 is another general result on the sparsity of solutions satisfying the S^{3}ONC. Following that are three Corollaries 3, 4, and 5 providing more details than Theorem 2 in the special cases of SCAD and MCP.

2.2.1 First-order bounds for non-zero entries—This subsection studies the first set of the promised thresholds and bounds based on the FONC. We start with a relatively general theorem that applies to both SCAD and MCP.

Theorem 1: *Let* $\mathbf{x}^*:=(x_i^*:1\leq i\leq p)\in\Re^p$ *be a solution satisfying FONC to* (3) *and let* $\mathbf{x}^0 \in \Re^p$ *be an arbitrary feasible solution. Assume* $f(\mathbf{x}^*) \leq f(\mathbf{x}^0)$. *If* $x_i^* \neq 0$, *then* $\|\mathbf{a}_{\cdot i}\|\sqrt{2f(\mathbf{x}^0)} \geq \sqrt{n}P'_\lambda(|x_i^*|)$.

Proof: We first notice that

$$\begin{aligned}&\|\mathbf{a}_{\cdot i}^\top(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|^2 \leq \|\mathbf{a}_{\cdot i}^\top\|^2\|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|^2 \leq \|\mathbf{a}_{\cdot i}^\top\|^2\left(\|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|^2+2n\sum_{i=1}^{p}P_\lambda(|x_i^*|)\right)\\&=2n\|\mathbf{a}_{\cdot i}^\top\|^2 f(\mathbf{x}^*) \leq 2n\|\mathbf{a}_{\cdot i}\|^2 f(\mathbf{x}^0)\end{aligned} \qquad (8)$$

Suppose that $|x_i^*|>0$. The FONC at $\mathbf{x}^*$ for the *i*-th dimension yields $\mathbf{a}_{\cdot i}^\top(\mathbf{A}\mathbf{x}^*-\mathbf{b})+nP'_\lambda(|x_i^*|)\cdot\mathrm{sign}(x_i^*)=0$, which, combining with (8), gives us $nP'_\lambda(|x_i^*|) \leq n|P'_\lambda(|x_i^*|)| \leq \|\mathbf{a}_{\cdot i}\|\sqrt{2nf(\mathbf{x}^0)}$. This completes the proof.

The above theorem has direct implications to the special cases of SCAD and MCP, as detailed in the following corollaries.

Corollary 1: *Consider the case of SCAD. Let* $\mathbf{x}^*$ *be a solution satisfying FONC to* (3) *and* $\mathbf{x}^0 \in \Re^p$ *a feasible solution. Assume* $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$.

a. *For any* i: $1 \le i \le p$, *if* $x_i^* \neq 0$ *and if*

$$\lambda > \|\mathbf{a}_{\cdot i}\| \sqrt{2f(\mathbf{x}^0)} / \sqrt{n} \qquad (9)$$

then $|x_i^*| > \lambda$ *and* $|x_i^*| \ge a\lambda - \frac{a-1}{\sqrt{n}} \|\mathbf{a}_{\cdot i}\| \sqrt{2f(\mathbf{x}^0)}$.

b. *Assume that* (9) *is satisfied for all* i: $1 \le i \le p$, *then we have*

$P_\lambda(a\lambda - \max_{i:1 \le i \le p} n^{-\frac{1}{2}}(a-1)\|\mathbf{a}_{\cdot i}\| \sqrt{2f(\mathbf{x}^0)}) > 0$ *and*

$$\|\mathbf{x}^*\|_0 \le \frac{f(\mathbf{x}^0)}{P_\lambda\left(a\lambda - \max_{i:1 \le i \le p} \frac{(a-1)\|\mathbf{a}_{\cdot i}\|}{\sqrt{n}} \sqrt{2f(\mathbf{x}^0)}\right)}.$$

Proof: By Theorem 1, we have that, if $x_i^* \neq 0$, then

$$\|\mathbf{a}_{\cdot i}\| \sqrt{2f(\mathbf{x}^0)} \ge \sqrt{n}\lambda[\mathbb{I}(|x_i^*| \le \lambda) + (a\lambda - |x_i^*|)_+ (a-1)^{-1} \lambda^{-1} \mathbb{I}(|x_i^*| > \lambda)] \ge \sqrt{n}\lambda[\mathbb{I}(|x_i^*| \le \lambda) + (a\lambda - |x_i^*|)(a - 1)^{-1} \lambda^{-1} \mathbb{I}(|x_i^*| > \lambda)]$$

. Combining with (9), the above inequality is satisfied if and only $|x_i^*| > \lambda$ and $|x_i^*| \ge a\lambda - \frac{a-1}{\sqrt{n}} \|\mathbf{a}_{\cdot i}\| \sqrt{2f(\mathbf{x}^0)}$ both hold, This completes the proofs of Part (a).

As to Part (b), we notice that P_λ vanishes at zero and is positive and non-decreasing on $\Re_{++}$.

Combining with Part (a), we have $|x_i^*| \ge a\lambda - \frac{a-1}{\sqrt{n}} \|\mathbf{a}_{\cdot i}\| \sqrt{2f(\mathbf{x}^0)}$ if $x_i^* \neq 0$ and

$$f(\mathbf{x}^0) \ge f(\mathbf{x}^*) \ge \sum_{i=1}^p P_\lambda(|x_i^*|) = \sum_{i:x_i^* \neq 0} P_\lambda(|x_i^*|) \ge \sum_{i:x_i^* \neq 0} P_\lambda \left(a\lambda - \max_{\hat{i}:1 \le \hat{i} \le p} \frac{(a-1)\|\mathbf{a}_{\cdot \hat{i}}\|}{\sqrt{n}} \sqrt{2f(\mathbf{x}^0)}\right) = \|x^*\|_0 P$$

$$-n^{-1/2}(a - 1)\max_{\hat{i}:1 \le \hat{i} \le p} \|\mathbf{a}_{\cdot \hat{i}}\| \sqrt{2f(\mathbf{x}^0)})$$

. Multiplying both sides of (9) by (a − 1), we have $a\lambda - \lambda > \frac{a-1}{\sqrt{n}} \|\mathbf{a}_{\cdot i}\| \sqrt{2f(\mathbf{x}^0)}$ for all i: $1 \le i \le p$. Therefore, $a\lambda - \frac{a-1}{\sqrt{n}} \|\mathbf{a}_{\cdot i}\| \sqrt{2f(\mathbf{x}^0)} > \lambda > 0$ for all i: $1 \le i \le p$. We obtain Part (b).

Corollary 2: *Consider the case of MCP. Let* $\mathbf{x}^*$ *be a solution satisfying FONC to* (3) *and by* $\mathbf{x}^0 \in \Re^p$ *a feasible solution. Assume* $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$.

a. *For any* i: $1 \le i \le p$, *if* $x_i^* \neq 0$, *then* $|x_i^*| \ge a\lambda - \frac{a\|\mathbf{a}_{\cdot i}\|}{\sqrt{n}} \sqrt{2f(\mathbf{x}^0)}$.

b. *Assume that* $\lambda > \frac{\|\mathbf{a}_{\cdot i}\|}{\sqrt{n}}\sqrt{2f(\mathbf{x}^0)}$ *for all i*: $1 \leq i \leq p$*, then* $P_\lambda(a\lambda - n^{-1/2}\max_{i:1\leq i\leq p} a\|\mathbf{a}_{\cdot i}\|\sqrt{2f(\mathbf{x}^0)}) > 0$ *and*

$$\|\mathbf{x}^*\|_0 \leq \frac{f(\mathbf{x}^0)}{P_\lambda\left(a\lambda - \max_{i:1\leq i\leq p}\frac{a\|\mathbf{a}_{\cdot i}\|}{\sqrt{n}}\sqrt{2f(\mathbf{x}^0)}\right)}.$$

Proof: Using Theorem 1 and definition of the MCP, this corollary can be shown by similar arguments to those in the proof of Corollary 1.

2.2.2 Second-order bounds for non-zero entries—

This subsection studies a different set of thresholds and bounds for the non-zero entries of S^3ONC solutions. These bounds are in general sharper than the results from the FONC. We will, again, start with a general theorem.

Theorem 2: *Let* $\mathbf{x}^*$ *be a solution satisfying* S^3*ONC to* (3) *and* $h(\cdot)$ *be the second-order derivative of* $P_\lambda(|\cdot|)$ *when it is twice differentiable. For any i*: $1 \leq i \leq p$*, if* $P_\lambda(|\cdot|)$ *is twice differentiable and concave at* x_i^**, then* $\|\mathbf{a}_{\cdot i}\|^2 \geq n|h(x_i^*)|$.

Proof: Per S^3ONC, if $P_\lambda(|\cdot|)$ is twice differentiable at the i-th dimension of $\mathbf{x}^*$ denoted x_i^*, then $\mathbf{a}_{\cdot i}^\top \mathbf{a}_{\cdot i} + nh(x_i^*) \geq 0$. Notice that $h(x_i^*) \leq 0$ per concavity of $P_\lambda(|\cdot|)$ at x_i^*. Therefore, $\|\mathbf{a}_{\cdot i}\| = \mathbf{a}_{\cdot i}^\top \mathbf{a}_{\cdot i} \geq -nh(x_i^*) = n|h(x_i^*)|$

Corollaries 3 and 4 below are direct applications of Theorem 2 for SCAD and MCP

Corollary 3: *Consider the case of SCAD. Let* $\mathbf{x}^*$ *be a* S^3*ONC solution to* (3).

a. *If* $\|\mathbf{a}_{\cdot i}\|^2 < \frac{n}{a-1}$*, then either* $|x_i^*| \geq a\lambda$ *or* $|x_i^*| \leq \lambda$ *is satisfied.*

b. *For feasible solution any* $\mathbf{x}^0$ *satisfying* $f(\mathbf{x}^*) \leq f(\mathbf{x}^0)$*, if* $\|\mathbf{a}_{\cdot i}\|^2 < \frac{n}{a-1}$*, and* $\lambda > \frac{\|\mathbf{a}_{\cdot i}\|\sqrt{2f(\mathbf{x}^0)}}{\sqrt{n}}$*, then either* $x_i^* = 0$ *or* $|x_i^*| \geq a\lambda$ *is satisfied.*

Proof: Invoking Theorem 2, if $\lambda < |x_i^*| < a\lambda$, then $h|h(x_i^*)| = n\frac{1}{a-1} \leq \|\mathbf{a}_{\cdot i}\|^2$. Therefore, if $n|h(x_i^*)| = n\frac{1}{a-1} > \|\mathbf{a}_{\cdot i}\|^2$, then $|x_i^*| \leq \lambda$, or $|x_i^*| \geq a\lambda$, which is immediately Part (a) of this corollary.

Further invoking Corollary 1, and noticing that a solution satisfying S^3ONC is also a solution satisfying FONC, we have the desired results in Part (b).

Corollary 4: *Consider the case of MCP. Let* $\mathbf{x}^*$ *be a* S^3*ONC solution to* (3). *If* $\|\mathbf{a}_{\cdot i}\|^2 < \frac{n}{a}$*, then either* $|x_i^*| > a\lambda$ *or* $x_i^* = 0$ *is satisfied.*

Proof: This corollary follows by using Theorem 2, definition of the MCP and techniques used in the proof of the last corollary.

Corollary 5: *Let $\mathbf{x}^*$ be a solution satisfying S^3ONC to* (3) *and* $\mathbf{x}^0 \in \Re^p$ *an arbitrary feasible solution. Assume* $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$.

a. *Consider the case of SCAD. If* $\lambda > \frac{\|\mathbf{a}_{\cdot i}\| \sqrt{2f(\mathbf{x}^0)}}{\sqrt{n}}$ *for all* $i : 1 \le i \le p$ *and* $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$, *then* $\|\mathbf{x}^*\|_0 \le f(\mathbf{x}^0)/P_\lambda(a\lambda)$.

b. *Consider the case of MCP. Then* $\|\mathbf{x}^*\|_0 \le f(\mathbf{x}^0)/P_\lambda(a\lambda)$.

Proof: To show (a): Per Part (b) of Corollary 3, if $x_i^* \neq 0$, then $|x_i^*| \ge a\lambda$ for all $i : 1 \le i \le p$. Combining with the fact that $P_\lambda(\cdot)$ is non-decreasing, $P_\lambda(0) = 0$, we have

$$f(\mathbf{x}^0) \ge f(\mathbf{x}^*) \ge \sum_{i=1}^{p} P_\lambda(|x_i^*|) = \sum_{i:x_i^* \neq 0} P_\lambda(|x_i^*|) \ge \sum_{i:x_i^* \neq 0} P_\lambda(a\lambda) = \|x^*\|_0 P_\lambda(a\lambda),$$

which, combining with the fact that $P_\lambda(a\lambda) > 0$, immediately implies the desired result in (a).

Part (b) is evident by following the same argument as in (a). Yet we will invoke Corollary 4 instead of Corollary 3.

Remark 1: Both the first-order and the second-order bounds are dependent on an arbitrary feasible solution $\mathbf{x}^0$. We may let $\mathbf{x}^0 = \mathbf{x}^*$ in all the results above to obtain the bounds for a local solution. Perhaps more interestingly, $\mathbf{x}^0$ can also be some solutions that are more easily available, such as the all-zero vector or a solution generated by any warm-starting procedure. If $\mathbf{x}^*$ is computed with a descent algorithm that starts at $\mathbf{x}^0$, we can ensure the satisfaction of the stipulated inequality $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$. Then, one can use the aforementioned bounds to estimate the sparsity of $\mathbf{x}^*$ using information only on $\mathbf{x}^0$. Such information is often computationally cheap.

Remark 2: The sparsity results in this section only serve as an estimate on the magnitude and the number of (non-)zero variables of a local solution. They do not provide any guarantee in approximating $|\mathscr{S}|$, the cardinality of the true support set. Nonetheless, in the following section, with some additional assumptions, our discussion covers the correctness in screening for the non-zero dimensions with the aid of the above analysis.

3 Statistical Accuracy of Local Solutions to FCPSLR

This section studies the statistical accuracy at a local solution satisfying the S^3ONC. Detailed settings and assumptions are discussed in Section 3.1. Then Section 3.2 presents the promised results.

We will denote by a_{ji} the entry at the i-th column and j-th row of $\mathbf{A}$. For a scalar $x \ge 0$, denote by $\lceil x \rceil$ (and $\lfloor x \rfloor$) the smallest (largest, rep.) integer greater (smaller) or equal to x. By definition of SCAD and MCP, we have the following fact will be used in our analysis: for all $x \in \Re$,

$$P_{\lambda,SCAD}(|x|) \le (a+1)\lambda^2/2; \quad P_{\lambda,MCP}(|x|) \le a\lambda^2/2. \qquad (10)$$

3.1 Setting and assumption

We will restrict our discussions to linear regression with fixed design matrices and random error terms. Our results rely on the following assumptions:

Assumption A

A.1 The vector of errors $W = (\varepsilon_j) \in \Re^n$ satisfies that $Prob[|\langle W, \upsilon\rangle| \geq t] \leq 2\exp(-t^2/2\sigma^2)$ for any $\upsilon \in \Re^n : ||\upsilon|| = 1$ and any $t > 0$.

A.2 The design matrix $\mathbf{A}$ satisfies a column normalization condition, i.e., $n^{-1}||\mathbf{a}_{\cdot i}||^2 \leq K$ for some $K > 0$ for all $i = 1, \ldots, p$.

A.3 There exists a sequence $\{r_d \geq 0 : d = 1, \ldots, p\}$ such that the following are satisfied: (i) For any $d_1, d_2 : 1 \leq d_1 \leq d_2 \leq p$, we have $r_{d_1} \geq r_{d_2}$; (ii) There exists some $\tilde{p}^* : 2|\mathscr{S}| \leq \tilde{p}^* \leq p$ such that $r_{\tilde{p}^*} > 0$; (iii) For all $d : 1 \leq d \leq p$, it holds that $n^{-1}||\mathbf{A}\delta\mathbf{x}||^2 \geq r_d||\delta\mathbf{x}||^2$ for any $\delta\mathbf{x} \in \Re^p : ||\delta\mathbf{x}||_0 \leq d$.

Assumptions A.1 and A.2 are commonly used conditions in the literature (see, e.g., [27,37]). Assumption A.1 holds if W follows an isotropic Gaussian distribution as in [6,43]. Assumption A.2 can be ensured via normalization. It is satisfied with high probability by random design matrices under sub-exponential or even weaker assumptions according to [35].

Assumption A.3 is the most critical one. Intuitively, for any $d : 1 \leq d \leq p$, the scalar r_d is the lower bound on the smallest eigenvalue of all the principle sub-matrices of $\mathbf{A}^\top\mathbf{A}$ with a size $d \times d$. Thus, by $r_{\tilde{p}^*} > 0$, it essentially means that any principal sub-matrix of $\mathbf{A}^\top\mathbf{A}$ with a size smaller or equal to $\tilde{p}^* \times \tilde{p}^*$ is positive definite. Regarding this assumption, we think it worthwhile to mention the following observation: When $\tilde{p}^* = 4|\mathscr{S}|$, Assumption A.3 is a critical condition for the Lasso to ensure recovery quality – the *restricted eigenvalue (RE)* condition, which is first introduced by [6] (see its definition taken from [43] in Definition 1 below). We illustrate this relationship between the RE condition and Assumption A.3 in Lemma 1. Under the RE condition, [6] shows the recovery quality of the Lasso. [46] provides conditions and probability lower bounds for RE condition to hold. Although the RE condition with a more general setting of parameters is discussed by [6] and [33], the performance of the Lasso is unknown under the more general setting.

Additionally, since the RE condition is also equivalent to the *restricted strong convexity* (RSC) condition in a linear regression model with some choices of parameters according to [27], therefore Assumption A.3 is also potentially weaker than the RSC condition discussed in [27].

Definition 1 (RE condition [43]): The matrix $\mathbf{A} \in \Re^{n \times p}$ is said to satisfy the RE condition if, for some $\varkappa(\mathbf{A}) > 0$, it holds that $\frac{1}{n}||\mathbf{A}\delta\mathbf{x}||^2 \geq \varkappa(\mathbf{A})||\delta\mathbf{x}||^2$ for all $\delta\mathbf{x} \in \cup_{|\hat{S}|=|\mathscr{S}|}\mathbb{C}(\hat{S})$ where $\mathbb{C}(\hat{S}) := \{\delta\mathbf{x} := (\delta x_i) \in \Re^p : |\delta\mathbf{x}_{\hat{S}^c}| \leq 3|\delta\mathbf{x}_{\hat{S}}|\}$, $\delta\mathbf{x}_{\hat{S}^c} := (\delta x_i : i \in \hat{S}^c)$, and $\delta\mathbf{x}_{\hat{S}} := (\delta x_i : i \in \hat{S})$. Furthermore, the largest possible $\varkappa(\mathbf{A})$ is said to be the restricted eigenvalue constant of $\mathbf{A}$.

Lemma 1: *(a) The RE condition in Definition 1 implies Assumption A.3 with* $r_{4|\mathscr{S}|} \geq \boldsymbol{\varkappa}(\mathbf{A}) > 0$ *and* $\tilde{p}^* \geq 4|\mathscr{S}|$. *(b) The reverse is not true.*

Proof: For Part (a), it suffices to show that for any $\delta\mathbf{x} = (\delta x_i) \in \Re^p : \|\delta\mathbf{x}\|_0 \leq 4|\mathscr{S}|$, there always exists an index set $\hat{S}'$: $|\hat{S}'| = |\mathscr{S}|$, such that $|\delta\mathbf{x}_{\hat{S}'^c}| \leq 3|\delta\mathbf{x}_{\hat{S}'}|$. Here $\delta\mathbf{x}_{\hat{S}'^c} := (\delta x_i : i \in \hat{S}'^c)$, and $\delta\mathbf{x}_{\hat{S}'} := (\delta x_i : i \in \hat{S}')$.

If $\|\delta\mathbf{x}\|_0 \leq |\mathscr{S}|$, the above is trivially true. Otherwise, if $\|\delta\mathbf{x}\|_0 > |\mathscr{S}|$, one can always pick $\hat{S}'$ to be the set of indices of the first $|\mathscr{S}|$ number of coordinates with the largest absolute value. As a result, $\min_{i\in\hat{S}'} |\delta x_i| \geq \max_{i\in\hat{S}'^c} |\delta x_i|$ and $|\hat{S}'| = |\mathscr{S}|$. We then know that

$$|\delta\mathbf{x}_{\hat{S}'}| \geq |\mathscr{S}| \cdot \min_{i\in\hat{S}'} |\delta x_i| \geq |\mathscr{S}| \cdot \max_{i\in\hat{S}'^c} |\delta x_i| = \frac{3|\mathscr{S}|}{3} \max_{i\in\hat{S}'^c} |\delta x_i| \geq \frac{\|\delta\mathbf{x}\|_0 - |\hat{S}'|}{3} \max_{i\in\hat{S}'^c} |\delta x_i| \geq \frac{1}{3} \sum_{i\in\hat{S}'^c \cap \{i:|\delta x_i| \neq 0\}} |\delta x_i| = \frac{1}{3} |\delta\mathbf{x}_{\hat{S}'^c}|$$

, which leads to the desired result in Part (a).

For Part (b), it suffices to show that for some $\delta\mathbf{x} \in \Re^p$, there exists an index set $\hat{S}'$: $|\hat{S}'| = |\mathscr{S}|$ such that $|\delta\mathbf{x}_{\hat{S}'^c}| \leq 3|\delta\mathbf{x}_{\hat{S}'}|$, but $\delta\mathbf{x}$ does not satisfy $\|\delta\mathbf{x}\|_0 \leq 4|\mathscr{S}|$. An example can be $\delta x_i = 1/|\mathscr{S}|$ for all $i \in \mathscr{S}$ and $\delta x_i = 1/(p-|\mathscr{S}|)$ for all $i \notin \mathscr{S}$. If we pick $\hat{S}' = \mathscr{S}$, the above is evident.

[46] and [29] show that the RE condition can be satisfied with high probability when the design matrix is generated following Gaussian and/or subgaussian distributions. Potentially more general settings for the RE condition can be obtained from the discussions by [30], [1] and [35]. Since Assumption A.3 can be more general than the RE condition, the former may be easier (in the sense of occurring with a better probability or of requiring weaker assumptions on the underlying distribution) to hold when the design matrix is random.

We will impose some conditions on the parameters of the FCP:

Condition B

(B1) For SCAD, $\|\mathbf{a}_{\cdot i}\|^2 < \frac{n}{a-1}$ for all $1 \leq i \leq p$ and $f(\mathbf{x}^0) < \min_{i:1\leq i\leq p} \frac{\lambda^2 n}{2\|\mathbf{a}_{\cdot i}\|^2}$ for a given initial solution $\mathbf{x}^0$.

(B2) For MCP, $\|\mathbf{a}_{\cdot i}\|^2 < \frac{n}{a}$ for all $1 \leq i \leq p$ and $\lambda > 0$.

The above condition ensures that the assumptions of Corollaries 3 and 4 hold for both SCAD and MCP. The stipulation on λ for the SCAD case is conceivably stronger than for the MCP case. For the former, a wise initial solution $\mathbf{x}^0$ that has a good solution quality may allow for more flexible choices of λ, while, for the latter, Condition B is non-restrictive on λ. Under Assumption A.2, the requirements of Condition B on parameter a is satisfied for any $a : a < 1 + K^{-1}$ in the SCAD case and for any $a : a < K^{-1}$ in the MCP case.

3.2 Major results

We now present our theoretical findings on the statistical performance of S^{3}ONC solutions. All proofs are postponed in Section 4, our main results can be summarized as following:

> Section 3.2.1 presents two "general" theorems. Theorem 3 establishes statistical performance bounds in terms of ME, AD, and ℓ_2 loss, for all S^{3}ONC solutions. These bounds imply the dependence of statistical performance on the optimization quality. Theorem 4 shows that the oracle solution (2) may be recovered by any S^{3}ONC solution under proper choices of parameters (a, λ), when the minimal signal strength $\min_{i\in\mathscr{S}}|x_i^{true}|$ is properly large.
>
> Sections 3.2.2 and 3.2.3 apply Theorems 3 and 4 to the case of FCPSLR-MCP and show that any S^{3}ONC solution which has a better objective value than the Lasso solution entails the strong oracle property (Corollary 6). Furthermore, Corollary 7 in Section 3.2.2 shows that those local solutions may incur a substantially better ME than the Lasso (4), if the sample size is above a certain threshold polynomial in ln p. Otherwise, the worst-case ME of FCPSLR-MCP is comparable to the Lasso.

We remark that, since we only wish to provide theoretical insights here, we anticipate that the constants used in our results may not be optimal.

3.2.1 Statistical accuracy of an arbitrary S^{3}ONC solution—

This subsection seeks to present a "general" result on the statistical performance at an arbitrary S^{3}ONC solution $\mathbf{x}^*$ within the sub-level set $\{\mathbf{x} : f(\mathbf{x}) \le \inf_{\mathbf{x}} f(\mathbf{x}) + \Gamma\}$ for an arbitrary $\Gamma \ge 0$. To this end, we consider a slightly larger sub-level set, $\{\mathbf{x} : f(\mathbf{x}) \le f(\mathbf{x}^{true}) + \Gamma\}$. Because $f(\mathbf{x}^{true}) \ge \inf_{\mathbf{x}} f(\mathbf{x})$, it holds that $\{\mathbf{x} : f(\mathbf{x}) \le \inf_{\mathbf{x}} f(\mathbf{x}) + \Gamma\} \subseteq \{\mathbf{x} : f(\mathbf{x}) \le f(\mathbf{x}^{true}) + \Gamma\}$.

To aid our presentation of the results, we recall the closed-form of $P_\lambda(a\lambda)$ given as in (10). We will also make use of some short-hand notations:

$$\mathbf{P}^*(t,\tilde{p}){:=}1{-}\exp\{-\tilde{p}(t{-}\ln p)\}{-}\exp\{-(\tilde{p}{+}1)(t{-}\ln p)\}\times\{1{-}\exp\left(-(p{-}\tilde{p})\,(t{-}\ln p)\right)\}/\{1{-}\exp\left(-t{+}\ln p\right)\}. \tag{11}$$

When $r_{\tilde{p}^*} > 0$, where $\tilde{p}^*$ is defined in Assumption A.3,

$$T_{a,\lambda,n,\mathbf{x}^{ture},\mathbf{A}}(t){:=}\{8\sigma^2\tilde{p}^*/n\}\left(1{+}2\sqrt{t}{+}2t\right){+}8\min\{\{\lambda^2(|\mathscr{S}|{-}\|\mathbf{x}^*_{\mathscr{S}}\|_0)\}/r_{\tilde{p}^*},\ P_\lambda(a\lambda)|\mathscr{S}|{+}\Gamma\}. \tag{12}$$

Theorem 3: *Denote* $\tilde{p}_{\Gamma,a,\lambda}{:=}\left\lfloor\{2|\mathscr{S}|\cdot P_\lambda(a\lambda){+}\Gamma\}/\{P_\lambda(a\lambda){-}\frac{\sigma^2}{2n}(1{+}2\sqrt{t}{+}2t)\}\right\rfloor$*. Consider an arbitrary S*^{3}*ONC solution* $\mathbf{x}^*$ *to FCPSLR* (3) *with either SCAD or MCP. Assume the* *simultaneous occurrence of (i) the event that Condition B is satisfied with any initial solution* $\mathbf{x}^0$*; and (ii) the event that* $f(\mathbf{x}^*) \le \min\{f(\mathbf{x}^0), f(\mathbf{x}^{true}) + \Gamma\}$ *holds for any* $\Gamma \ge 0$*. For any* $t > 0$*, assume that parameters* (a, λ) *of penalty* P_λ *satisfy* $P_\lambda(a\lambda) > \sigma^2(1+2\sqrt{t}+2t)/(2n)$*. Then under Assumption A.1, the following holds:*

1. *For any integer* $\tilde{p}^*_{\Gamma,a,\lambda}$: $\min\{\tilde{p}_{\Gamma,a,\lambda}, p\} \le \tilde{p}^*_{\Gamma,a,\lambda} \le p$*, the ME is bounded by*

$$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^{true})\|^2 \le \frac{4\sigma^2}{n} \cdot \tilde{p}^*_{\Gamma,a,\lambda} \cdot \left(1+2\sqrt{t}+2t\right) + 8\min\{\lambda|\mathscr{S}| \cdot \|\mathbf{x}^{true}\|_\infty, P_\lambda(a\lambda)|\mathscr{S}| + \Gamma\}. \tag{13}$$

with probability at least $\mathbf{P}^*(t, \tilde{p}^*_{\Gamma,a,\lambda})$ *(as in* (11)*).*

2. *If, in addition, Assumption A.3 holds and* (a, λ) *satisfy that*

$$\tilde{p}_{\Gamma,a,\lambda} \le \tilde{p}^*, \tag{14}$$

where $\tilde{p}^*$ *is defined in Assumption A.3. Then with probability greater or equal to* $\mathbf{P}^*(t, \tilde{p}^*)$ *(as in* (11)*), the following holds simultaneously:*

$$\frac{\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^{true})\|^2}{n} \le T_{a,\lambda,n,\mathbf{x}^{true},\mathbf{A}}(t); \quad \|\mathbf{x}^* - \mathbf{x}^{true}\|^2 \le \frac{T_{a,\lambda,n,\mathbf{x}^{true},\mathbf{A}}(t)}{r_{\tilde{p}^*}},$$
$$|\mathbf{x}^* - \mathbf{x}^{true}| \le \sqrt{\frac{\tilde{p}^*}{r_{\tilde{p}^*}} \cdot T_{a,\lambda,n,\mathbf{x}^{true},\mathbf{A}}(t)}; \quad \|\mathbf{x}^* - \mathbf{x}^{true}\|_0 \le \tilde{p}^* \tag{15}$$

and

$$|\mathscr{S}| - \|\mathbf{x}^*_{\mathscr{S}}\|_0 \le \sum_{i \in \mathscr{S}} \mathbb{I}$$
$$\left(|x_i^{true}| - \left[\frac{8\sigma^2\tilde{p}^*}{nr_{\tilde{p}^*}}\left(1+2\sqrt{t}+2t\right) + \frac{8}{r_{\tilde{p}^*}} \cdot \min\left\{\frac{\lambda^2|\mathscr{S}|}{r_{\tilde{p}^*}}, P_\lambda(a\lambda)|\mathscr{S}| + \Gamma\right\}\right]^{1/2} \le 0\right). \tag{16}$$

where $T_{a,\lambda,n,\mathbf{x}^{true},\mathbf{A}}(t)$ *is defined as in* (12).

Inequalities in (15) provide upper bounds to the statistical errors under different measures. In addition to the error bounds above, we show in the following that the S^{3}ONC solutions can exactly recover the oracle solution under some additional assumptions on the minimal signal strength, $\min_{i \in \mathscr{S}} |x_i^{true}|$.

Theorem 4: *Suppose Assumptions A.1 and A.3 with* $\tilde{p}^* \geq 2|\mathscr{S}|$ *hold. Consider an arbitrary* S^3*ONC solution* $\mathbf{x}^* \in \Re^p$ *to FCPSLR* (3) *with arbitrarily either SCAD or MCP. Assume the simultaneous occurrence of (i) the event that Condition B is satisfied with an arbitrary initial solution* $\mathbf{x}^0$*; and (ii) the event that* $f(\mathbf{x}^*) \leq \min\{f(\mathbf{x}^0), f(\mathbf{x}^{true}) + \Gamma\}$ *holds for an arbitrary* $\Gamma \geq 0$. *Let the parameters* (a, λ) *satisfy that* $\lambda > \sigma a^{-1}\sqrt{\frac{8\tilde{p}^*}{nr_{\tilde{p}^*}}(1+2\sqrt{t}+2t)}$ *and*

$$P_\lambda(a\lambda) > \frac{\sigma^2}{2n}(1+2\sqrt{t}+2t) + \frac{\frac{\sigma^2}{n}|\mathscr{S}|\cdot(1+2\sqrt{t}+2t)+\Gamma}{\tilde{p}^*-2|\mathscr{S}|+1}; \qquad (17)$$

and let minimal signal strength satisfy that

$$\min_{i\in\mathscr{S}}|x_i^{true}| > \sqrt{\frac{8\sigma^2\tilde{p}^*}{nr_{\tilde{p}^*}}\left(1+2\sqrt{t}+2t\right)+\frac{8}{r_{\tilde{p}^*}}\min\left\{\frac{\lambda^2|\mathscr{S}|}{r_{\tilde{p}^*}}, P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\right\}}, \qquad (18)$$

then $\mathbf{x}^*$ *equals the oracle solution, i.e.,* $\mathbf{x}^* = \mathbf{x}^{oracle} \in \underset{\mathbf{x}\in\Re^p:\ x_i=0,\,\forall i\in\mathscr{S}^c}{\arg\inf}\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2$*, with probability at least* $\mathbf{P}^*(t, \tilde{p}^*)$ *(as in* (11)*).*

Remark 3: Theorem 3 provides a set of upper bounds on the performance measures for any S^{3}ONC solution. Theorem 4 presents a set of conditions for any S^{3}ONC solution to recover the oracle solution. These results are algorithm independent. In contrast, the existing results in [17], [38], and [37] relay on specific choices of computing procedures.

Remark 4: If $\ln p \geq 1$ and $t = 2\ln p$, one may quickly verify that $\mathbf{P}^*(2\ln p, \tilde{p}^*) \geq 1 - O(1)\cdot\exp(-O(1)\cdot\tilde{p}^*\ln p)$, where we denote by $O(1)$ problem-independent constants. Recall that $n \ll p$. Theorem 3 implies that the bounds on statistical errors in (13) and (15) hold with high probability. Similarly, the recovery of oracle solution as in Theorem 4 holds with high probability.

Remark 5: For fixed (a, λ), Theorems 3 and 4 explicate the relationship between optimization quality in minimizing the non-convex formulation of FCPSLR and the statistical quality in approximating the true parameter. Specifically, the former theorem shows that the statistical performance of the S^{3}ONC solutions in terms of ME, ℓ_2 loss, and AD can all be written in parameterization of Γ, which is always an underestimate of the suboptimality gap. For a simple example, consider FCPSLR-MCP, if we (i) choose the parameters (a, λ) to be $\lambda = O(\sigma\sqrt{\frac{\ln p}{n}})$ and $a = O(1)$, (ii) let $t = 2\ln p$ (and assume $\ln p \geq 1$), and (iii) let $\mathbf{x}^0 = \mathbf{x}^*$, we obtain an upper bound on ME from (13) given as

$$O\left(\frac{\sigma^2|\mathscr{S}|\ln p}{n}+\Gamma\right) \quad (19)$$

with probability lower bounded by $1 - O(1) \cdot \exp(-O(1)|\mathscr{S}| \ln p)$. We think it interesting to compare the ME in (19) with that of an optimal, but exponential-time estimator. [31] shows that, under a comparatively more critical assumption that W is isotropic Gaussian, the exponential-time sparse estimator (5), which is claimed to be the *optimal* estimator by [43], yields an ME of

$$\begin{aligned}&Prob\left[\frac{1}{n}\|\mathbf{A}(\mathbf{x}^{exp}-\mathbf{x}^{true})\|^2 \le O\left(\frac{\ln(p/|\mathscr{S}|)}{n}\sigma^2|\mathscr{S}|\right)\right]\\&\ge 1-O(1)\cdot\exp\left(-O(1)\cdot|\mathscr{S}|\ln(p/|\mathscr{S}|)\right).\end{aligned} \quad (20)$$

We see a comparable performance between (19) with (20) when $\Gamma = 0$, that is, when the FCPSLR-MCP is minimized globally. Meanwhile, the ME of FCPSLR-MCP deteriorates linearly with an increased Γ. Similarly, for the recovery of the oracle solution, Theorem 4 indicates that the requirement (18) on the minimal signal strength is increasingly demanding if Γ becomes larger. To our knowledge, this is the first explication on the relationship between statistical performance and optimization quality in a non-convex learning problem. Furthermore, in spite of the tendency that the statistical performance degrades with the increase of suboptimality gap, S^{3}ONC solutions may still recover the oracle solution when the minimal signal strength, namely, $\min_{i\in\mathscr{S}}|x_i^{true}|$, is large enough to satisfy (18).

3.2.2 Strong oracle property—

This subsection focuses on FCPSLR-MCP and show that any of its S^{3}ONC solutions within the sub-level set $\{\mathbf{x}: f(\mathbf{x}) \le f(\mathbf{x}^{lasso})\}$ entails the strong oracle property. This means that any descent, S^{3}ONC guaranteeing algorithm that starts from the solution to the convex formulation of the Lasso in (4) can output the oracle solution with overwhelming probability. Initializing computing schemes for FCPSLR with the Lasso has been discussed by [17] and [37]. Nonetheless, these analyses are all algorithm-specific. We present in the following an algorithm-independent analysis.

Corollary 6: *Assume* $\ln p \ge 1$*. Denote by* $\mathbf{x}^*$ *in* $\Re^p$ *an* S^3*ONC solution to FCPSLR-MCP. Let Assumptions A.1, A.2 with* $K = 1$*, and the RE condition (as defined in Definition 1) with* $\varkappa(\mathbf{A}) < 1$ *be satisfied. Then* $r_{4|\mathscr{S}|} \ge \varkappa(\mathbf{A}) > 0$*. Assume that* $f(\mathbf{x}^*) \le f(\mathbf{x}^{lasso})$ *almost surely, where* $\mathbf{x}^{lasso}$ *is defined in* (4) *with problem data* $(\mathbf{x}^{true}, \mathbf{A}, \mathbf{b})$ *and parameter* $\lambda_{lasso}=4\sigma\sqrt{\frac{\ln p}{n^{1-\gamma}}}$*, where* $\gamma \in [0, 1]$ *is an arbitrary scalar. Let* $\lambda=\frac{35}{\varkappa(\mathbf{A})}\sigma\sqrt{\frac{\ln p}{n^{1-\gamma}}}$*, and* $a \in [0.8, 1)$*. There exists a problem-independent constant* c_3 *such that if*

$$\min_{i\in\mathscr{S}} |x^{true}|^2 \geq c_3 \cdot \frac{\sigma^2 |\mathscr{S}| \ln p}{n^{1-\gamma} \cdot \min\{r_{4|\mathscr{S}|}, r^2_{4|\mathscr{S}|}[\varkappa(\mathbf{A})]^2\}} \quad (21)$$

then the S^3*ONC solution* $\mathbf{x}^*$ *equals the oracle solution, i.e.,*

$$\mathbf{x}^* = \mathbf{x}^{oracle} \in \underset{\mathbf{x}\in\Re^p:\ x_i=0,\,\forall i\in\mathscr{S}^c}{\arg\inf} \frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2, \quad (22)$$

with probability at least $1-c_1\exp(-c_2 n^{\gamma}\ln p) - c_4\exp(-c_5|\mathscr{S}|n^{\gamma}\ln p)$ *for some problem independent constants* $c_1, c_2, c_4, c_5 > 0$.

Remark 6: Even though we only consider the undesirable case where the principle sub-matrices of $\mathbf{A}^\top\mathbf{A}$ is ill-conditioned in the sense that $\varkappa(\mathbf{A}) < 1$, which is the same setting as in [43], our results can be easily extended to the case of $\varkappa(\mathbf{A}) \geq 1$ by choosing a different value for λ.

Remark 7: It is worth mentioning that the choice of (λ, a) can be much more flexible than the above corollary presents. In fact, one may follow the same proof for this Corollary to verify that for a reasonably wide range of $\lambda = O(\frac{1}{\varkappa(\mathbf{A})}\sigma\sqrt{\frac{\ln p}{n^{1-\gamma}}})$ and $a = O(1)$, the above result remains to be true. In practice, one may choose λ via data-driven procedures such as the cross validation.

Remark 8: When $\gamma \in (0, 1]$, Corollary 6 indicates the overwhelming probability of recovering the oracle solution. Thus Corollary 6 implies that any S^{3}ONC solution that has a better objective value than the Lasso (with proper choice of parameter λ_{lasso}) entails the strong oracle property. In contrast, the Lasso in (4) does not have the oracle property according to [14] regardless of the choice of $\lambda_{lasso} > 0$. Therefore, the strong oracle property of an S^{3}ONC solution already indicates a possible outperformance of FCPSLR over the Lasso.

Remark 9: Corollary 6 is actually a direct implication of Theorem 4 to the case of FCPSLR-MCP. It is possible to also obtain some oracle property results for FCPSLR-SCAD with the Lasso initialization by applying Theorem 4. Nonetheless, due to the additional stipulations for λ to satisfy Condition B in the SCAD case, the determination of the penalty parameter λ is more involving. We will leave the simplification of Condition B for SCAD to future research. Nonetheless, in practice, a proper choice of λ for SCAD can also be determined via cross validation.

3.2.3 Comparison with Lasso in terms of ME—Apart from the comparison in terms of the oracle property between FCPSLR and the Lasso in Section 3.2.2, the result in this subsection may provide a second reason why local solutions to FCPSLR can potentially outperform the Lasso. [43] provides a set of intriguing comparisons between an optimal but exponential-time estimator and all the polynomial-time computable estimators, including the Lasso. Those comparisons indicate a non-trivial gap in ME between these two types of estimators. Motivated by that result, we are particularly interested in how FCPSLR-MCP compares with both the optimal estimator and the Lasso under the same criterion of performance, namely, ME. Again, we focus on the undesirable case where $\varkappa(\mathbf{A}) < 1$, as in [43].

Corollary 7: *Let* $\ln p \geq 1$ *and* $\varkappa(\mathbf{A}) < 1$. *Denote by* $\mathbf{x}^*$ *an* S^3*ONC solution to FCPSLR-MCP. Suppose that (i) Assumptions A.1, A.2 with* $K = 1$ *hold; and (ii) the RE condition (in Definition 1) is satisfied, then* $r_{4|\mathscr{S}|} \geq \varkappa(\mathbf{A}) > 0$. *Assume in addition* $f(\mathbf{x}^*) \leq f(\mathbf{x}^{lasso})$ *a.s., where* $\mathbf{x}^{lasso}$ *is defined in* (4) *with the problem data* $(\mathbf{x}^{true}, \mathbf{A}, \mathbf{b})$ *and parameter* $\lambda_{lasso}=4\sigma\sqrt{\frac{\ln p}{n}}$.

1. *If we let* $\lambda=4\sigma\sqrt{\frac{\ln p}{n}}$ *and* $a \in [0.8, 1)$, *then there exists a problem-independent constant* $c_6 > 0$ *such that:*

   $$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \leq \frac{c_6\sigma^2|\mathscr{S}|}{\varkappa(\mathbf{A})}\frac{\ln p}{n} \qquad (23)$$

   with probability at least $1 - c_1 \exp(-c_2 \ln p) - c_4 \exp(-c_5|\mathscr{S}| \ln p)$ *for some problem independent constants* $c_1, c_2, c_4, c_5 \in \Re_{++}$;

2. *If we let* $\lambda=\frac{35}{\varkappa(\mathbf{A})}\sigma\sqrt{\frac{\ln p}{n}}$ *and* $a \in [0.8, 1)$, *then there exists a problem-independent constant* c_7: $0 < c_7 < \infty$ *such that for any* $n > 1$ *that satisfies*

   $$n \geq c_7 \cdot \frac{\sigma^2|\mathscr{S}|\ln p}{\psi(\mathbf{x}^{true}) \cdot \min\left\{r_{4|\mathscr{S}|}, r_{4|\mathscr{S}|}^2 \cdot [\varkappa(\mathbf{A})]^2\right\}}, \qquad (24)$$

   where $\psi: \Re^p \to \Re_+$ *is defined as*

   $$\psi(\mathbf{x}^{true}):=\max_{\mathscr{S}_{sub}\subseteq\mathscr{S}:|\mathscr{S}_{sub}|\leq\max\left\{1,\lceil|\mathscr{S}|-r_{4|\mathscr{S}|}\cdot(\varkappa(\mathbf{A}))^2|\mathscr{S}|\rceil\right\}}\min_{i\in\mathscr{S}_{sub}}|x_i^{true}|^2,$$

   the ME is bounded by

   $$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \leq \frac{c_6\sigma^2}{n}|\mathscr{S}|\ln p \qquad (25)$$

   with probability at least $1 - c_1 \exp(-c_2 \ln p) - c_4 \exp(-c_5|\mathscr{S}| \ln p)$.

Remark 10: We would like to compare the S^{3}ONC solutions to FCPSLR-MCP with both the optimal but exponential-time estimator in (5) and the Lasso, a polynomial-time estimator as in (4). From [6] the Lasso achieves an ME of

$$Prob\left[\frac{1}{n}\|\mathbf{A}(\mathbf{x}^{lasso}-\mathbf{x}^{true})\|^2\leq O\left(\frac{\ln p}{n}\frac{\sigma^2|\mathscr{S}|}{\varkappa(\mathbf{A})}\right)\right]\geq 1-O(1)\cdot\exp\left(-O(1)\cdot\ln(p)\right). \tag{26}$$

Comparing the ME of $\mathbf{x}^{exp}$ in (20) and that of $\mathbf{x}^{lasso}$ in (26), one may anticipate a significant gap in the performance when $\varkappa(\mathbf{A})$ ∈(0, 1) is small. It is also shown by [43] that, for any polynomial-time sparse estimator, the gap incurred by small $\varkappa(\mathbf{A})$ cannot be reduced without compromising the rate in n. In contrast, the results presented in Corollary 7 indicates that the performance of FCPSLR-MCP may resemble either $\mathbf{x}^{exp}$ or $\mathbf{x}^{lasso}$ in two different modes:

> **Mode 1.** *The Lasso-comparable mode:* From the first part of Corollary 7, we see that the ME of FCPSLR-MCP as presented in (23) is comparable to the Lasso as in (26) in the worst-case, when the sample size is small.
>
> **Mode 2.** *The optimal estimator-comparable mode:* When the sample size is greater than a threshold (24), which is linear in ln p and polynomial in some other problem dependent numbers, properly tuning parameter λ allows FCPSLR-MCP to enter a substantially enhanced mode that incurs an ME presented as in (25). The resulting ME is comparable to the optimal estimator, as in (20), in terms of dependency. Upon entering this mode, the ME of the S^{3}ONC solution is no longer dependent on $\varkappa(\mathbf{A})$ and may dominate the Lasso especially when $\varkappa(\mathbf{A})$ is small.

In fact, one may choose a fairly flexible range of $\lambda=O(\sigma\sqrt{\frac{\ln p}{n}})$ and $a = O(1)$ to achieve the same rate as in (23), and of $\lambda=O(\frac{\sigma}{\varkappa(\mathbf{A})}\sqrt{\frac{\ln p}{n}})$ and $a = O(1)$ to achieve the rate in (25). Practitioners can tune λ through in-sample cross validation, which is indeed the commonly adopted practice. Therefore, it may be unnecessary to keep in mind the aforementioned rules in determining λ, but use whichever value that works best according to the in-sample trials.

Remark 11: The function $\psi(\mathbf{x}^{true})$ measures the signal strength of the majority of the nonzero signals, or more precisely, the largest "$\max\{1, \lceil|\mathscr{S}| - r_{4|\mathscr{S}|}\cdot(\varkappa(\mathbf{A}))^2|\mathscr{S}|\rceil\}$"-many nonzero dimensions (in terms of their absolute values), in the true parameter. One may observe that the inequality $\psi(\mathbf{x}^{true})\geq\min_{i\in\mathscr{S}}|x_i^{true}|$ always holds. This indicates that FCPSLR-MCP may enter the optimal estimator-comparable mode even if $\min_{i\in\mathscr{S}}|x_i^{true}|$ is very close to zero, given that the majority of the nonzero signals are strong enough.

Remark 12: Corollary 7 is in fact a direct implication of Theorem 3 to the case of FCPSLR-MCP. One may also apply Theorem 3 to obtain a bound for ME of FCPSLR-SCAD. However, admittedly, Condition B for FCPSLR-SCAD is more restrictive than that for the MCP case, resulting in a possibly less desirable theoretical performance estimate. Nonetheless, we later will show in Section 5 that the empirical performance of FCPSLR-SCAD and that of FCPSLR-MCP appear to be quite alike. We therefore think that our results may have underestimated the power of SCAD. Yet we will leave improving the analysis for FCPSLR-SCAD to future research.

4 Technical Proofs

We prove our major results in Subsection 4.1, while some auxiliary results are presented in Subsection 4.2.

4.1 Proof of major results

4.1.1 Proof of Theorem 3—Firstly, under the simultaneous occurrence of both (a) the event that Condition B holds with initial solution $\mathbf{x}^0$ and (b) the event that $f(\mathbf{x}^*) \leq \min\{f(\mathbf{x}^0), f(\mathbf{x}^{true})+\Gamma\}$ is satisfied, for any $t > 0$, invoke Lemma 5 with any (a, λ) and $\tilde{p}$: $p \geq \tilde{p} \geq 2|\mathscr{S}|$ such that $P_\lambda(a\lambda)>\frac{\sigma^2}{2n}(1+2\sqrt{t}+2t)+\frac{\frac{\sigma^2}{n}|\mathscr{S}|\cdot(1+2\sqrt{t}+2t)+\Gamma}{\tilde{p}-2|\mathscr{S}|+1}$, we have that $\|\mathbf{x}^*-\mathbf{x}^{true}\|_0 \leq \tilde{p}$ holds with probability at least $1-\exp(-(\tilde{p}+1)(t-\ln p))\cdot\frac{1-\exp(-(p-\tilde{p})(t-\ln p))}{1-\exp(-t+\ln p)}$. One obtains via some algebra that the above inequality is satisfied if $\tilde{p}$: $p \geq \tilde{p} \geq \min\{p, \tilde{p}_{\Gamma,a,\gamma}\} \geq 2|\mathscr{S}|$ for arbitrarily fixed $(a, \lambda){:}P_\lambda(a\lambda)>\frac{\sigma^2}{2n}(1+2\sqrt{t}+2t)$. This means that for any integer $\tilde{p}^*_{\Gamma,a,\lambda}{:}\min\{\tilde{p}_{\Gamma,a,\lambda},p\} \leq \tilde{p}^*_{\Gamma,a,\lambda} \leq p$. with probability lower bounded by $1-\exp\left(-(\tilde{p}^*_{\Gamma,a,\lambda}+1)\,(t-\ln p)\right)\cdot\frac{1-\exp\left(-(p-\tilde{p}^*_{\Gamma,a,\lambda})(t-\ln p)\right)}{1-\exp(-t+\ln p)}$.

We next provide a probabilistic bound on the ME for any S^{3}ONC solution $\mathbf{x}^*$ satisfying $\|\mathbf{x}^*-\mathbf{x}^{true}\|_0 \leq \tilde{p}^*_{\Gamma,a,\lambda}$. To do so, we invoke Lemma 2 for an arbitrary $t > 0$, where we let $\tilde{p}=\tilde{p}^*_{\Gamma,a,\lambda}$ within that lemma. Conditioning on the event that $\|\mathbf{x}^*-\mathbf{x}^{true}\|_0 \leq \tilde{p}^*_{\Gamma,a,\lambda}$, we have $\frac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \leq \frac{4\sigma^2}{n}\cdot\tilde{p}^*_{\Gamma,a,\lambda}\cdot(1+2\sqrt{t}+2t)+8\min\{\sum_{i\in\mathscr{S}}P'_\lambda(|x^*_i|)|x^{true}_i|, P_\lambda(a\lambda)\cdot(|\mathscr{S}|-\|\mathbf{x}^*\|_0)+\Gamma\}$ with probability at least $1-\exp\left(-\tilde{p}^*_{\Gamma,a,\lambda}(t-\ln p)\right)$. (Notice that here we also let t only within Lemma 2 to be rescaled into $\tilde{p}^*_{\Gamma,a,\lambda}t$.) By the union bound and by the facts that (a) $\|\mathbf{x}^*\|_0 \geq 0$ surely; and that (b) $\sum_{i\in\mathscr{S}}P'_\lambda(|x^*_i|)|x^{true}_i| \leq \lambda|\mathscr{S}|\|\mathbf{x}^{true}\|_\infty$ surely, This completes the proof of Theorem 3 Part 1.

To show part 2, under the additional assumption of (14), we may follow almost the same argument as in the first part and obtain

$$\|\mathbf{x}^*-\mathbf{x}^{true}\|_0 \leq \tilde{p}^* \qquad (27)$$

with probability lower bounded by $1-\exp\left(-(\tilde{p}^*+1)\,(t-\ln p)\right)\cdot\frac{1-\exp(-(p-\tilde{p}^*)(t-\ln p))}{1-\exp(-t+\ln p)}$, which is the claimed result in the fourth inequality of (15). We also recall the notation of $T_{a,\gamma,n,\mathbf{x}^{true},\mathbf{A}}(t)$ as in (12). If we invoke the second part of Lemma 2 by letting $\tilde{p}:=\tilde{p}^*$ (and rescale t only within Lemma 2 into $\tilde{p}^* t$), together with the fact that $\|\mathbf{x}^*\|_0 \geq 0$ surely, we obtain by the union bound that

$$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \leq \frac{8\sigma^2\tilde{p}^*}{n}\left(1+2\sqrt{t}+2t\right) \\ +8\min\left\{\frac{\lambda^2(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0)}{r_{\tilde{p}^*}}, P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\right\}=T_{a,\lambda,n,\mathbf{x}^{true},\mathbf{A}}(t) \tag{28}$$

holds with probability greater or equal to $\mathbf{P}^*(t, \tilde{p}^*)$ as defined in (11), which is immediately the first inequality in the claimed results (15). Then by Lemma 3 (where we let $\tilde{p} = \tilde{p}^*$), combined with (28) and (27), we have the rest of the inequalities in (15) and (16) hold as desired in Theorem 3.

4.1.2. Proof of Theorem 4—By assumption, (17) holds, which implies both

$P_\lambda(a\lambda)-\frac{\sigma^2}{2n}(1+2\sqrt{t}+2t)>0$ and $\tilde{p}_{\Gamma,a,\lambda}=\left|\frac{2|\mathscr{S}|\cdot P_\lambda(a\lambda)+\Gamma}{P_\lambda(a\lambda)-\frac{\sigma^2}{2n}(1+2\sqrt{t}+2t)}\right| \leq \tilde{p}^*$. Consider the following two events: $\{\|\mathbf{x}^*-\mathbf{x}^{true}\|_0 \leq \tilde{p}^*\}$ and $\mathfrak{E}_a(\tilde{p}^*):=\{n^{-1}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \leq 8\sigma^2\tilde{p}^*n^{-1}(1+2\sqrt{t}+2t)\}+8\min\{\lambda^2 r_{\tilde{p}^*}^{-1}\cdot(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0),\ P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\}\}$. By Lemma 4 with $\tilde{p} = \tilde{p}^*$, we have that under the assumptions of Theorem 4, the simultaneous occurrence of the above two events leads to the desired result. The probability for their simultaneous occurrence is lower bounded by $\mathbf{P}^*(t, \tilde{p}^*)$ due to the second part of Theorem 3, which completes the proof.

4.1.3 Proof of Corollary 6—By invoking Lemma 6 under the assumption that $f(\mathbf{x}^*) \leq f(\mathbf{x}^{lasso})$ almost surely, one has that

$$f(\mathbf{x}^*)-f(\mathbf{x}^{true}) \leq (\lambda_{lasso}+\lambda)\left|\mathbf{x}^{lasso}-\mathbf{x}^{true}\right|, \quad \text{a. s.} \tag{29}$$

Using Corollaries 1 and 2 of [27], under the RE condition and Assumptions A.1 and A.2 with $K = 1$, we have that for $\lambda_{lasso} > 0$, it holds that $\left|\mathbf{x}^{lasso}-\mathbf{x}^{true}\right| \leq \frac{6\lambda_{lassso}}{\varkappa(\mathbf{A})}|\mathscr{S}|$, with probability at least $1-c_1'\exp(-c_2'n\lambda_{lasso}^2/\sigma^2)$ for some $c_1', c_2'>0$. Since $\lambda_{lasso}=4\sigma\sqrt{\frac{\ln p}{n^{1-\gamma}}}$ for some arbitrarily chosen $\gamma \in [0, 1]$, combined with (29), we have

$$f(\mathbf{x}^*)-f(\mathbf{x}^{true}) \leq \frac{96\sigma^2}{\varkappa(\mathbf{A})}|\mathscr{S}|\frac{\ln p}{n^{1-\gamma}}+\frac{24\sigma\lambda}{\varkappa(\mathbf{A})}|\mathscr{S}|\sqrt{\frac{\ln p}{n^{1-\gamma}}}, \tag{30}$$

holds with probability $1-c_1\exp(-c_2n^\gamma\ln p)$. Invoking Lemma 1, we know that the RE condition implies Assumption A.3 with $r_{4|\mathscr{S}|} \geq \varkappa(\mathbf{A}) > 0$. Under the assumption that $\ln p \geq 1$, $n > 1$, and $\varkappa(\mathbf{A}) < 1$, one may easily verify that, with the given set of parameters, i.e., $\lambda=\frac{35}{\varkappa(\mathbf{A})}\sigma\sqrt{\frac{\ln p}{n^{1-\gamma}}}$, $a \in [0.80, 1)$, both $\lambda>\sigma a^{-1}\sqrt{\frac{32|\mathscr{S}|}{nr_{4|\mathscr{S}|}}(1+2\sqrt{2n^\gamma\ln p}+4n^\gamma\ln p)}$ and

$$P_\lambda(a\lambda)>\left[\frac{\sigma^2}{2n}(1+2\sqrt{t}+2t)+\frac{\frac{\sigma^2}{n}|\mathscr{S}|\cdot(1+2\sqrt{t}+2t)+\frac{96\sigma^2}{\varkappa(\mathbf{A})}|\mathscr{S}|\frac{\ln p}{n}+\frac{24\sigma\lambda}{\varkappa(\mathbf{A})}|\mathscr{S}|\sqrt{\frac{\ln p}{n}}}{\tilde{p}-2|\mathscr{S}|+1}\right]_{\substack{t=2n^\gamma\ln p\\ \tilde{p}=4|\mathscr{S}|}}$$

are satisfied. Invoking Theorem 4 with $\mathbf{x}^0 = \mathbf{x}^*$ (which implies $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$ surely), $t = 2n^\gamma \ln p$, $\tilde{p}^* = \tilde{p} = 4|\mathscr{S}|$, $r_{\tilde{p}^*} = r_{4|\mathscr{S}|}$, and $\Gamma=\frac{96\sigma^2}{\varkappa(\mathbf{A})}|\mathscr{S}|\frac{\ln p}{n^{1-\gamma}}+\frac{24\sigma\lambda}{\varkappa(\mathbf{A})}|\mathscr{S}|\sqrt{\frac{\ln p}{n^{1-\gamma}}}$, to show the strong oracle property of $\mathbf{x}^*$, it suffices to show that both Condition B and (18) holds.

According to Assumption A.2 with $K = 1$, Condition B holds if $a < 1$, as assumed. Then (18) holds because of the following: Given $n^{1-\gamma} \ge c_3\cdot\frac{\sigma^2|\mathscr{S}|\ln p}{\min_{i\in\mathscr{S}}|x^{true}|^2\cdot\min\{r_{4|\mathscr{S}|},r^2_{4|\mathscr{S}|}[\varkappa(\mathbf{A})]^2\}}$ by assumption, then $\left(32\sigma^2(nr_{4|\mathscr{S}|})^{-1}|\mathscr{S}|(1+2\sqrt{2n^\gamma\ln p}+4n^\gamma\ln p)+8r^{-2}_{4|\mathscr{S}|}\cdot\lambda^2|\mathscr{S}|\right)^{1/2}\le\sqrt{\frac{\tilde{c}_3}{c_3}}\cdot\min_{i\in\mathscr{S}}|x^{true}|$ for some problem-independent constant $\tilde{c}_3$. We may as well let $\tilde{c}_3/c_3 < 1$. As a result, condition (18) is satisfied. Combining Theorem 4 with (30) by the union bound, we have $\mathbf{x}^* = \mathbf{x}^{oracle}$, with probability at least $\mathbf{P}^*(2n^\gamma \ln p, 4|\mathscr{S}|) - c_1 \exp(-c_2 n^\gamma \ln p)$. Further invoking (11) and the assumption that $n > 1$ and $\ln p \ge 1$, we have $\mathbf{P}^*(2n^\gamma \ln p, 4|\mathscr{S}|)-c_1 \exp(-c_2 n^\gamma \ln p) \ge 1-c_4 \exp(-c_5|\mathscr{S}|n^\gamma \ln p)- c_1 \exp(-c_2 n^\gamma \ln p)$ for some problem independent constants c_1, c_2, c_4, $c_5 > 0$.

4.1.4 Proof of Corollary 7—

It follows by Lemma 1 that $r_{4|\mathscr{S}|} \ge \varkappa(\mathbf{A}) > 0$. Using Lemma 6 under the assumption that $f(\mathbf{x}^*) \le f(\mathbf{x}^{lasso})$ a.s., one has that

$$f(\mathbf{x}^*)-f(\mathbf{x}^{true})\le(\lambda_{lasso}+\lambda)\left|\mathbf{x}^{lasso}-\mathbf{x}^{true}\right|,\quad \text{a. s.} \tag{31}$$

Now we may invoke a well-known result on the recovery quality of the Lasso in the form of (4): Invoking Corollary 2 by [27], under the RE condition and Assumptions A.1 and A.2 with $K = 1$ we have that when $\lambda_{lasso}=4\sigma\sqrt{\frac{\ln p}{n}}$, it holds that $\left|\mathbf{x}^{lasso}-\mathbf{x}^{true}\right|\le\frac{24\sigma}{\varkappa(\mathbf{A})}|\mathscr{S}|\sqrt{\frac{\ln p}{n}}$ with probability at least $1 - c_1 \exp(-c_2 \ln p)$ for some $c_1, c_2 > 0$. Now combining this with (31) yields $f(\mathbf{x}^*)-f(\mathbf{x}^{true})\le(\lambda_{lasso}+\lambda)\left|\mathbf{x}^{lasso}-\mathbf{x}^{true}\right|\le\frac{96\sigma^2|\mathscr{S}|}{\varkappa(\mathbf{A})}\frac{\ln p}{n}+\frac{24\sigma\lambda|\mathscr{S}|}{\varkappa(\mathbf{A})}\sqrt{\frac{\ln p}{n}}$, with probability at least $1 - c_1 \exp(-c_2 \ln p)$.

For MCP, Condition B is satisfied by Assumptions A.2, $K = 1$, and $a < 1$.

To show Part 1 of the corollary: Since we have assumed that $\lambda=4\sigma\sqrt{\frac{\ln p}{n}}$, and $a \ge 0.8$, in the MCP case of consideration, $P_\lambda(a\lambda)=\frac{a\lambda^2}{2}=\frac{8a\sigma^2\ln p}{n}\ge 6.4\cdot\frac{\sigma^2\ln p}{n}$, which ensures that $2P_\lambda(a\lambda)>\frac{\sigma^2}{n}(1+2\sqrt{2\ln p}+4\ln p)$, given $\ln p \ge 1$ by assumption. Therefore, we may invoke (a) the first part of Theorem 3 with $\mathbf{x}^0 = \mathbf{x}^*$ (which implies $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$ almost surely),

$\varGamma = \frac{96\sigma^2}{\varkappa(\mathbf{A})}|\mathscr{S}|\frac{\ln p}{n} + \frac{24\sigma\lambda}{\varkappa(\mathbf{A})}|\mathscr{S}|\sqrt{\frac{\ln p}{n}}$, and $\tilde{p}_{\varGamma,a,\lambda} := \left\lfloor \frac{2|\mathscr{S}|\cdot P_\lambda(a\lambda) + \frac{96\sigma^2}{\varkappa(\mathbf{A})}|\mathscr{S}|\frac{\ln p}{n} + \frac{24\sigma\lambda}{\varkappa(\mathbf{A})}|\mathscr{S}|\sqrt{\frac{\ln p}{n}}}{P_\lambda(a\lambda) - \frac{\sigma^2}{2n}(1+2\sqrt{2\ln p}+4\ln p)} \right\rfloor \le c_5\frac{|\mathscr{S}|}{\varkappa(\mathscr{A})}$; for some problem-independent constant c_5, and (b) the union bound, to obtain that

$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^{true})\|^2 \le \frac{4\sigma^2}{n}\cdot c_5\frac{|\mathscr{S}|}{\varkappa(\mathscr{A})}\cdot\left(1+2\sqrt{2\ln p}+4\ln p\right) + 4a\lambda^2|\mathscr{S}| + \frac{768\sigma^2}{\varkappa(\mathbf{A})}|\mathscr{S}|\frac{\ln p}{n} + \frac{192\sigma\lambda}{\varkappa(\mathbf{A})}|\mathscr{S}|\sqrt{\frac{\ln p}{n}} \le \frac{c_6\sigma^2|\mathscr{S}|}{\varkappa(\mathbf{A})}\frac{\ln p}{n}$ for some problem-independent constant c_6, with probability at least $\mathbf{P}^*(2\ln p, \min\{p, c_5|\mathscr{S}|\cdot[\varkappa(\mathscr{A})]^{-1}\}) - c_1\exp(-c_2\ln p)$. Notice that $\mathbf{P}^*$ is defined as in (11). Observing that, since $\ln p \ge 1$, one has $1/(1-\exp(-\ln p)) < 2$ and that $\varkappa(\mathscr{A}) < 1$. Therefore,

$$\mathbf{P}^*(2\ln p, \min\{p, c_5|\mathscr{S}|[\varkappa(\mathscr{A})]^{-1}\}) \ge 1 - c_4\exp(-c_5|\mathscr{S}|\ln p),$$

where c_4 is some problem-independent constant.

Now, consider the second part of the Corollary, where $\lambda = \frac{35}{\varkappa(\mathbf{A})}\sigma\sqrt{\frac{\ln p}{n}}$, and $a \in [0.8, 1)$, under the assumption that $\ln p \ge 1$. Since $\varkappa(\mathbf{A}) < 1$, one may easily verify that, with the given set of parameters, we obtain $\tilde{p}_{\varGamma,a,\lambda} \le 4|\mathscr{S}|$. Recall that we have shown $r_{4|\mathscr{S}|} \ge \varkappa(\mathbf{A}) > 0$ at the beginning of this proof. Then we may invoke the second part of Theorem 3 with $\tilde{p}^* = 4|\mathscr{S}|$, $t = 2\ln p$, $\mathbf{x}^0 = \mathbf{x}^*$ (which implies $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$ almost surely), $\varGamma = \frac{96\sigma^2}{\varkappa(\mathbf{A})}|\mathscr{S}|\frac{\ln p}{n} + \frac{24\sigma\lambda}{\varkappa(\mathbf{A})}|\mathscr{S}|\sqrt{\frac{\ln p}{n}}$, and obtain that

$$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^{true})\|^2 \le \frac{32\sigma^2|\mathscr{S}|}{n}\left(1+2\sqrt{2\ln p}+4\ln p\right) + 8\frac{\lambda^2(|\mathscr{S}| - \|\mathbf{x}^*_{\mathscr{S}}\|_0)}{r_{4|\mathscr{S}|}} \quad (32)$$

holds with probability at least $\mathbf{P}^*(2\ln p, 4|\mathscr{S}|) - c_1\exp(-c_2\ln p) \ge 1 - c_4\exp(-c_5|\mathscr{S}|\ln p) - c_1\exp(-c_2\ln p)$, where

$$|\mathscr{S}| - \|\mathbf{x}^*_{\mathscr{S}}\|_0 \le \sum_{i\in\mathscr{S}}\mathbb{I}\left(|x_i^{true}| - \sqrt{\frac{32\sigma^2|\mathscr{S}|}{nr_{4|\mathscr{S}|}}\left(1+2\sqrt{2\ln p}+4\ln p\right) + \frac{8}{r_{4|\mathscr{S}|}}\cdot\frac{\lambda^2|\mathscr{S}|}{r_{4|\mathscr{S}|}}} \le 0\right).$$

(33)

Recall that

$$\lambda = \frac{35}{\varkappa(\mathbf{A})}\sigma\sqrt{\frac{\ln p}{n}} \Rightarrow \lambda^2 = \frac{1225}{[\varkappa(\mathbf{A})]^2}\sigma^2\frac{\ln p}{n}. \quad (34)$$

If it holds that $n \geq c_7 \cdot \frac{\sigma^2|\mathscr{S}|\ln p}{\psi(\mathbf{x}^{true})\cdot\min\{r_{4|\mathscr{S}|}, r^2_{4|\mathscr{S}|}\cdot[\varkappa(\mathbf{A})]^2\}}$, then $\left(32(nr_{4|\mathscr{S}|})^{-1}\sigma^2|\mathscr{S}|(1+2\sqrt{2\ln p}+4\ln p)+8r^{-2}_{4|\mathscr{S}|}\cdot\lambda^2|\mathscr{S}|\right)^{1/2} \leq \sqrt{\frac{\tilde{c}_7}{c_7}}\cdot\sqrt{\psi(\mathbf{x}^{true})}$ for some problem-independent constant $\tilde{c}_7$. We also recall that

$$\sqrt{\psi(\mathbf{x}^{true})} = \max_{\mathscr{S}_{sub}\subseteq\mathscr{S}:|\mathscr{S}_{sub}|\leq\max\left\{1,\lceil|\mathscr{S}|-r_{4|\mathscr{S}|}\cdot[\varkappa(\mathbf{A})]^2|\mathscr{S}|\rceil\right\}} \min_{i\in\mathscr{S}_{sub}} |x_i^{true}|^2.$$

which is the “max $\{1, \lceil|\mathscr{S}| - r_{4|\mathscr{S}|} \cdot [\boldsymbol{\varkappa}(\mathbf{A})]^2|\mathscr{S}|\rceil\}$”-th largest non-zero dimension of $\mathbf{x}^{true}$. Now, we let $\tilde{c}_7/c_7 < 1$. Combined with (33),

$$|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0 \leq \sum_{i\in\mathscr{S}} \mathbb{I}\left(|x_i^{true}|-\sqrt{\tilde{c}_7\psi(\mathbf{x}^{true})}\leq 0\right) \leq |\mathscr{S}|-\max\{1,\lceil|\mathscr{S}|-r_{4|\mathscr{S}|}[\varkappa(\mathbf{A})]^2|\mathscr{S}|\rceil\} \leq r_{4|\mathscr{S}|}[\varkappa(\mathbf{A})]^2|$$

. This with (32) and (34) completes the proof.

4.2 Auxiliary results

Lemma 2: *Consider an arbitrary* S^3 *ONC solution* $\mathbf{x}^*$ *to FCPSLR* (3) *with either SCAD or MCP. For any integer* $\tilde{p}: 0 \leq \tilde{p} \leq p$, *let Assumptions A.1 and A.3 with* $\tilde{p}^* \geq \tilde{p}$, *i.e.,* $r_{\tilde{p}^*} > 0$, *hold. Assume the simultaneous occurrence of (i) the event that Condition B is satisfied with an arbitrary initial solution* $\mathbf{x}^0$*; (ii) the event that* $f(\mathbf{x}^*) \leq \min\{f(\mathbf{x}^0), f(\mathbf{x}^{true}) + \Gamma\}$ *holds for an arbitrary* $\Gamma \geq 0$*; (iii) the event that* $\tilde{p} \geq \|\mathbf{x}^* - \mathbf{x}^{true}\|_0$ *obtains for some integer* $\tilde{p}$*. Then for any* $t > 0$,

$$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \leq \frac{4\sigma^2}{n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t)+8\min\{\sum_{i\in\mathscr{S}} P_\lambda^{'}(|x_i^*|)|x_i^{true}|,\ P_\lambda(a\lambda)\cdot(|\mathscr{S}|-\|\mathbf{x}^*\|_0)+\Gamma\}$$

holds with probability at least $1 - \exp(-t + \tilde{p}\ln p)$.

If, in addition, Assumption A.3 holds with $\tilde{p}^* \geq \tilde{p}$, *then*

$$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \leq \frac{8\sigma^2}{n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t)+8\min\left\{\frac{\lambda^2(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0)}{r_{\tilde{p}}}, P_\lambda(a\lambda)\cdot(|\mathscr{S}|-\|\mathbf{x}^*\|_0)+\Gamma\right\}$$

holds with probability at least $1 - \exp(-t + \tilde{p}\ln p)$*, where* $r_{\tilde{p}} > 0$.

Proof: Denote $\delta\mathbf{x}^* := (\delta x_i^*) = \mathbf{x}^* - \mathbf{x}^{true}$ and $S_{\tilde{p}} \subseteq \{1, \ldots, p\}$ such that $\delta x_i^* = 0$ for all $i \notin S_{\tilde{p}}$. By assumption, we can ensure that $\|\delta\mathbf{x}^*\|_0 \leq |S_{\tilde{p}}| = \tilde{p}$.

Denote by $\mathbf{A}_{S_{\tilde{p}}}$ the sub-matrix of $\mathbf{A}$ with the largest size such that $\mathbf{A}_{S_{\tilde{p}}} = (\boldsymbol{a}_{ji} : j = 1, \ldots, n, i \in S_{\tilde{p}})$. Also denote $\delta\mathbf{x}^*_{S_{\tilde{p}}} := (\delta x_i^* : i \in S_{\tilde{p}})$. Following the argument in Lemma 8 of [31], $\mathbf{A}_{S_{\tilde{p}}}$ admits a singular value decomposition with $\mathbf{A}_{S_{\tilde{p}}} = V_{S_{\tilde{p}}} D_{S_{\tilde{p}}} U_{S_{\tilde{p}}}$, for some matrix $V_{S_{\tilde{p}}} \in \Re^{n\times\tilde{p}}$ with orthonormal columns, that is, $V_{S_{\tilde{p}}}^\top V_{S_{\tilde{p}}} = I$, where I is an identity matrix. By such a construction, we have, for any $\upsilon \in \Re^{\tilde{p}}$, $\|\mathbf{A}_{S_{\tilde{p}}}\upsilon\| = \|D_{S_{\tilde{p}}} U_{S_{\tilde{p}}}\upsilon\|$. Therefore, by the assumed event $\{\|\mathbf{x}^* - \mathbf{x}^{true}\|_0 \leq \tilde{p}\}$, one obtains that

$$\begin{aligned}|W^\top \mathbf{A}\delta\mathbf{x}^*|=&\left|W^\top \mathbf{A}_{S_{\tilde p}}\delta\mathbf{x}^*_{S_{\tilde p}}\right| \le \|W^\top V_{S_{\tilde p}}\|\|D_{S_{\tilde p}}U_{S_{\tilde p}}\delta\mathbf{x}^*_{S_{\tilde p}}\| \\ =&\|V_{S_{\tilde p}}^\top W\|\|\mathbf{A}_{S_{\tilde p}}\delta\mathbf{x}^*_{S_{\tilde p}}\| \le \inf_{\substack{S_{\tilde p}:|S_{\tilde p}|=\tilde p,\\ S_{\tilde p}\subseteq\{1,\ldots,p\}}} \|V_{S_{\tilde p}}^\top W\|\|\mathbf{A}\delta\mathbf{x}^*\|, \quad a.s.\end{aligned} \tag{35}$$

We want to make use of Lemma 8, which follows Theorem 2.1 in [22], to find an upper bound to (35). In that lemma, considering a fixed $S_{\tilde p}$, we let $\sum_v = V_{S_{\tilde p}}V_{S_{\tilde p}}^\top$. Noticing that $(\Sigma_v)^2 = \Sigma_v\Sigma_v = \Sigma_v$. This means that Σ_v is an idempotent matrix. Then $\|\Sigma_v\| \le 1$ and $\mathbf{Tr}(\Sigma_v) = rank(\Sigma_v)$. Here $\mathbf{Tr}(\cdot)$ and $rank(\cdot)$ are the trace and the rank of a matrix. Notice that $V_{S_{\tilde p}}$ has orthonormal columns. Therefore, $V_{S_{\tilde p}}V_{S_{\tilde p}}^\top$ is a projection matrix onto the span of the column vectors of $V_{S_{\tilde p}}$, which is at most $\tilde p$ dimensional. Then $\mathbf{Tr}(\sum_v^2)=\mathbf{Tr}(\sum_v)=rank(\sum_v) \le \tilde p$. Now, we may invoke Lemma 8 with the settings discussed above and obtain $Prob\left[\|V_{S_{\tilde p}}^\top W\| \le \sigma\cdot\sqrt{\tilde p+2\sqrt{\tilde p t}+2t}\right] \ge 1-\exp(-t)$. This inequality can be easily extended to yield

$$Prob\left[\sup_{S_{\tilde p}:|S_{\tilde p}|=\tilde p,\ S_{\tilde p}\subseteq\{1,\ldots,p\}}\|V_{S_{\tilde p}}^\top W\| > \sigma\cdot\sqrt{\tilde p+2\sqrt{\tilde p t}+2t}\right] \le \begin{pmatrix} p \\ \tilde p\end{pmatrix}\cdot\exp(-t) \le p^{\tilde p}\cdot\exp(-t) \tag{36}$$

In this last inequality, we have not used a potentially tighter bound $\begin{pmatrix} p \\ \tilde p\end{pmatrix}\cdot\exp(-t) \le \left(\frac{pe}{2\tilde p}\right)^{\tilde p}\cdot\exp(-t)$ for the sake of notational simplicity. We shall find (36) useful soon in the subsequent.

Now we start to derive the claimed bound in the first part of the lemma. For convenience, let us denote that

$$\mathscr{T}_1 := \min\left\{\sum_{i\in\mathscr{S}} P_\lambda^{'}(|x_i^*|)|x_i^{true}|, \sum_{i\in\mathscr{S}} P_\lambda^{'}(|x_i^*|)|x_i^*-x_i^{true}|,\ P_\lambda(a\lambda)\cdot(|\mathscr{S}|-\|\mathbf{x}^*\|_0)+\Gamma\right\} \tag{37}$$

Invoking the second part of Lemma 7, we obtain

$$\frac{1}{2n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le \frac{1}{n}W^\top\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})+\mathscr{T}_1, \quad a.s. \qquad (38)$$

If one combines (38) with (35), conditioning on the assumed event that $\|\mathbf{x}^* - \mathbf{x}^{true}\|_0 \le \tilde{p}$ holds,

$$\frac{1}{2n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le \sup_{\substack{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p},\\ S_{\tilde{p}}\subseteq\{1,\dots,p\}}} \frac{1}{n}\|V_{S_{\tilde{p}}}^\top W\|\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|+\mathscr{T}_1, \quad a.s. \qquad (39)$$

By solving the inequality for $\frac{1}{\sqrt{n}}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|$ we obtain

$$\frac{1}{\sqrt{n}}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\| \le \frac{1}{\sqrt{n}}\sup_{\substack{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p},\\ S_{\tilde{p}}\subseteq\{1,\dots,p\}}} \|V_{S_{\tilde{p}}}^\top W\|+\sqrt{\left(\frac{1}{\sqrt{n}}\sup_{\substack{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p},\\ S_{\tilde{p}}\subseteq\{1,\dots,p\}}} \|V_{S_{\tilde{p}}}^\top W\|\right)^2 +2\mathscr{T}_1$$

, almost surely, which implies

$$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le \frac{4}{n}\sup_{\substack{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p},\\ S_{\tilde{p}}\subseteq\{1,\dots,p\}}} \|V_{S_{\tilde{p}}}^\top W\|^2+8\mathscr{T}_1 \quad ,$$

almost surely. Invoking (36), we know that

$$\frac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le \frac{4\sigma^2}{n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+8\mathscr{T}_1. \qquad (40)$$

with probability at least $1-p^{\tilde{p}}\cdot\exp(-t) = 1-\exp(-t+\tilde{p}\ln p)$. This completes the proof for the first part of the lemma by the definition of $\mathscr{T}_1$ in (37).

To show the second part, we notice that (a) by Corollaries 3 and 4 under Condition B and the assumed event that $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$, we have $x_i^* \ne 0 \Rightarrow |x_i^*| \ge a\lambda$ for all $i = 1, \dots, p$; (b) per properties of SCAD and MCP, for any $x \in \Re$ such that $|x| \ge a\lambda$, one has $P_\lambda'(|x|)=0$; (c) per properties of SCAD and MCP again, $0 \le P_\lambda'(|x|) \le \lambda$ for any $x \in \Re$. Combining these observations yields $\sum_{i\in\mathscr{S}} P_\lambda'(|x_i^*|)|x_i^*-x_i^{true}| \le \lambda\sqrt{|\mathscr{S}|-\|x_{\mathscr{S}}^*\|_0}\cdot\|\mathbf{x}^*-\mathbf{x}^{true}\|$ almost surely. This combined with (39) and (37) implies

$$\frac{1}{2n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le \sup_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p},S_{\tilde{p}}\subseteq\{1,\dots,p\}} \frac{1}{n}\|V_{S_{\tilde{p}}}^\top W\|\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|+\lambda\sqrt{|\mathscr{S}|-\|x_{\mathscr{S}}^*\|_0}\cdot\|\mathbf{x}^*-\mathbf{x}^{true}\|$$

almost surely. Now by Assumption A.3 with $\tilde{p}^* \ge \tilde{p}$, we know that $r_{\tilde{p}} \ge r_{\tilde{p}^*} > 0$. Thus, we may continue from the above inequality to obtain

$$\frac{1}{2n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le \sup_{\substack{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p},\\ S_{\tilde{p}}\subseteq\{1,\dots,p\}}} \frac{1}{n}\|V_{S_{\tilde{p}}}^\top W\|\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|+\sqrt{|\mathscr{S}|-\|\mathbf{x}_{\mathscr{S}}^*\|_0}\cdot\frac{\lambda\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|}{\sqrt{nr_{\tilde{p}}}}$$

almost surely, which immediately leads to

$$\tfrac{1}{\sqrt{n}}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\| \le \sup_{\substack{S_{\tilde p}:|S_{\tilde p}|=\tilde p,\\ S_{\tilde p}\subseteq\{1,\dots,p\}}} \tfrac{2}{\sqrt{n}}\|V_{S_{\tilde p}}^{\top}W\|+\tfrac{2\lambda\sqrt{|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0}}{\sqrt{r_{\tilde p}}}$$ almost surely.

Further invoking (36), one has

$$\tfrac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le \left[2n^{-\frac{1}{2}}(\sigma\sqrt{\tilde p+2\sqrt{\tilde p t}+2t})+2\lambda\sqrt{|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0}\cdot r_{\tilde p}^{-\frac{1}{2}}\right]^2 \le \tfrac{8}{n}\sigma^2(\tilde p+2\sqrt{\tilde p t}+2t)+\tfrac{8\lambda^2(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0)}{r_{\tilde p}}$$

holds with probability at least $1-\exp(-t+\tilde p\ln p)$. Now, we recall that (40) holds almost surely conditioning on the same event. Therefore, with the same probability, one has

$$\tfrac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le \tfrac{8}{n}\sigma^2(\tilde p+2\sqrt{\tilde p t}+2t)+\min\left\{\tfrac{8\lambda^2(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0)}{r_{\tilde p}},8\mathscr{T}_1\right\},$$ which is immediately the desired result if we recall again the definition of $\mathscr{T}_1$ in (37).

<u>Lemma 3:</u> *Consider an arbitrary S^3ONC solution $\mathbf{x}^*\in\Re^p$ to FCPSLR* (3) *with arbitrarily either SCAD or MCP. Assume the simultaneous occurrence of (i) the event that $\tilde p\ge\|\mathbf{x}^*-\mathbf{x}^{true}\|_0$ obtains for some integer $\tilde p$; and (ii) Event $\mathscr{E}_a(\tilde p)$ defined as*

$$\mathscr{E}_a(\tilde p):=\{n^{-1}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le 8n^{-1}\sigma^2\tilde p(1+2\sqrt{t}+2t)+8\min\{r_{\tilde p}^{-1}\lambda^2(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0),P_\lambda(a\lambda)\cdot|\mathscr{S}|+\Gamma\}\}$$

. If Assumption A.3 holds for $\tilde p^*\ge\tilde p$, then $r_{\tilde p}>0$ and the following inequalities simultaneously hold a.s.:

$$\|\mathbf{x}^*-\mathbf{x}^{true}\|^2 \le \frac{8\sigma^2\tilde p}{nr_{\tilde p}}\left(1+2\sqrt{t}+2t\right)+\frac{8}{r_{\tilde p}}\cdot\min\left\{\frac{\lambda^2(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0)}{r_{\tilde p}},P_\lambda(a\lambda)\cdot|\mathscr{S}|+\Gamma\right\}; \tag{41}$$

$$|\mathbf{x}^*-\mathbf{x}^{true}|^2 \le \frac{8\tilde p\sigma^2}{nr_{\tilde p}}\left(1+2\sqrt{t}+2t\right)+\frac{8\tilde p}{r_{\tilde p}}\cdot\min\left\{\frac{\lambda^2(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0)}{r_{\tilde p}},P_\lambda(a\lambda)\cdot|\mathscr{S}|+\Gamma\right\}; \tag{42}$$

and

$$|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0 \le \sum_{i\in\mathscr{S}} \mathbb{I}\left(|x_i^{true}|-\left[\frac{8\sigma^2\tilde p}{nr_{\tilde p}}\left(1+2\sqrt{t}+2t\right)+\frac{8}{r_{\tilde p}}\cdot\min\left\{\frac{\lambda^2|\mathscr{S}|}{r_{\tilde p}},P_\lambda(a\lambda)\cdot|\mathscr{S}|+\Gamma\right\}\right]^{1/2}\le 0\right). \tag{43}$$

Proof: Per Assumption A.3, $r_{\tilde p}\ge r_{\tilde p^*}>0$. Combined with the assumed event, $\mathscr{E}_a(\tilde p)$, we immediately have

$$r_{\tilde{p}}\|\mathbf{x}^*-\mathbf{x}^{true}\|^2\le\frac{1}{n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2\le\frac{8\sigma^2\tilde{p}}{n}\left(1+2\sqrt{t}+2t\right)+8\cdot\min\left\{\frac{\lambda^2(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0)}{r_{\tilde{p}}},P_\lambda(a\lambda)\cdot|\mathscr{S}|+\Gamma\right\},\quad a.s.$$

(44)

Furthermore, by the assumed event that $\tilde{p}\ge\|\mathbf{x}^*-\mathbf{x}^{true}\|_0$ holds, we obtain from the above inequality that

$\frac{r_{\tilde{p}}}{\tilde{p}}|\mathbf{x}^*-\mathbf{x}^{true}|^2\le r_{\tilde{p}}\|\mathbf{x}^*-\mathbf{x}^{true}\|^2\le\frac{8\sigma^2\tilde{p}}{n}(1+2\sqrt{t}+2t)++8\cdot\min\left\{\frac{\lambda^2(|\mathscr{S}|-\|\mathbf{x}^*_{\mathscr{S}}\|_0)}{r_{\tilde{p}}},P_\lambda(a\lambda)\cdot|\mathscr{S}|+\Gamma\right\}$ a. s. . This along with (44) yields the results in (41) and (42), respectively.

To show (43), combining (44) with the fact that $\|\mathbf{x}^*_{\mathscr{S}}\|_0\ge0$, we know that, for all $i\in\mathscr{S}$, almost surely, $|x^*_i-x^{true}_i|\le(\frac{8\sigma^2\tilde{p}}{nr_{\tilde{p}}}(1+2\sqrt{t}+2t)+8r_p^{-1}\cdot\min\{\lambda^2r_{\tilde{p}}^{-1}|\mathscr{S}|,P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\})^{1/2}$. As an immediate result, almost surely

$$|x^*_i|\ge|x^{true}_i|-\sqrt{\frac{8\sigma^2\tilde{p}}{nr_{\tilde{p}}}\left(1+2\sqrt{t}+2t\right)+\frac{8}{r_{\tilde{p}}}\cdot\min\left\{\frac{\lambda^2|\mathscr{S}|}{r_{\tilde{p}}},P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\right\}}.$$

Therefore, $|x^*_i|>0$ if the right hand side of the above is strictly positive. As an immediate result, $\|\mathbf{x}^*_{\mathscr{S}}\|_0\ge\sum_{i\in\mathscr{S}}\mathbb{I}(|x^{true}_i|-[8\sigma^2\tilde{p}n^{-1}r_{\tilde{p}}^{-1}(1+2\sqrt{t}+2t)+8r_{\tilde{p}}^{-1}\cdot\min\left\{\lambda^2|\mathscr{S}|r_{\tilde{p}}^{-1},P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\}\right]^{1/2}>0)$ a. s. , which leads to (43).

Lemma 4: *Consider an arbitrary S^3ONC solution* $\mathbf{x}^*$ *to FCPSLR* (3) *with either SCAD or MCP. Let Assumption A.3 holds with $\tilde{p}^*\ge\tilde{p}$. Assume the simultaneous occurrence of (i) the event that Condition B is satisfied with any initial solution* $\mathbf{x}^0$*; (ii) the event that* $f(\mathbf{x}^*)\le f(\mathbf{x}^0)$ *holds; (iii) the event that* $\|\mathbf{x}^*-\mathbf{x}^{true}\|_0\le\tilde{p}$ *obtains for some integer $\tilde{p}$; (iv) Event $\mathscr{E}_a(\tilde{p})$. If for all $i\in\mathscr{S}$:*

$$|x^{true}_i|>\sqrt{\frac{8\sigma^2\tilde{p}}{nr_{\tilde{p}}}\left(1+2\sqrt{t}+2t\right)+\frac{8}{r_{\tilde{p}}}\cdot\min\left\{\frac{\lambda^2|\mathscr{S}|}{r_{\tilde{p}}},P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\right\}},\qquad(45)$$

then $r_{\tilde{p}}>0$ and it holds a.s. that $\|\mathbf{x}^*_{\mathscr{S}^c}\|_0\le\frac{1}{a^2\lambda^2}\cdot\frac{8\sigma^2\tilde{p}}{nr_{\tilde{p}}}(1+2\sqrt{t}+2t)$. *Furthermore, if* $\sigma a^{-1}\sqrt{\frac{8\tilde{p}}{nr_{\tilde{p}}}(1+2\sqrt{t}+2t)}<\lambda$*, then* $\mathbf{x}^*$ *equals the oracle solution, i.e.,* $\mathbf{x}^*=\mathbf{x}^{oracle}\in\underset{\mathbf{x}\in\Re^p:\ x_i=0,\ \forall i\in\mathscr{S}^c}{\arg\inf}\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2,\ a.s.$.

Proof: Per Assumption A.3, $r_{\tilde{p}} \geq r_{\tilde{p}}^{*} > 0$. Inequality (43) holds by invoking in Lemma 3 under the assumed events (iii) and (iv) and Assumption A.3. Combining (43) with the assumption of (45), we have that

$$\|\mathbf{x}_{\mathscr{S}}^{*}\|_{0}=|\mathscr{S}|\,\text{and}\,|x_{i}^{*}|>0\,\text{for all}\,i\in\mathscr{S},\ \text{a. s.} \qquad (46)$$

Therefore, invoking the assumed Event $\mathscr{E}_{a}(\tilde{p})$ again, one obtains that

$\|\mathbf{x}^{*}-\mathbf{x}^{true}\|^{2} \leq \frac{8\sigma^{2}\tilde{p}}{nr_{\tilde{p}}}(1+2\sqrt{t}+2t)+\frac{8}{r_{\tilde{p}}}\cdot\min\left\{\frac{\lambda^{2}(|\mathscr{S}|-\|\mathbf{x}_{\mathscr{S}}^{*}\|_{0})}{r_{\tilde{p}}},P_{\lambda}(a\lambda)|\mathscr{S}|+\Gamma\right\}=\frac{8\sigma^{2}\tilde{p}}{nr_{\tilde{p}}}(1+2\sqrt{t}+2t)\ \text{a. s.}$. Due to Corollaries 3 and 4 under S^{3}ONC and the assumed events (i) and (ii), we know that $x_{i}^{*}\neq 0 \Rightarrow |x_{i}^{*}|\geq a\lambda$ for all $i = 1, \cdots, p$ a.s.. Combining this with the above inequality, as well as the fact that $x_{i}^{true}=0$ for all $i \in \mathscr{S}^{c}$ by definition, results in $a^{2}\lambda^{2}\|\mathbf{x}_{\mathscr{S}^{c}}\|_{0} \leq \frac{8\sigma^{2}\tilde{p}}{nr_{\tilde{p}}}(1+2\sqrt{t}+2t)\ \text{a. s.}$, which is the first inequality in the claimed result.

Now if we let $\sigma a^{-1}\sqrt{\frac{8\tilde{p}}{nr_{\tilde{p}}}(1+2\sqrt{t}+2t)}<\lambda$, we know from the above inequality that $1 > \|\mathbf{x}_{\mathscr{S}^{c}}\|_{0} = 0$. Consider the satisfaction of S^{3}ONC by $\mathbf{x}^{*}$, which implies that $\mathbf{x}^{*}$ also satisfies FONC. Therefore,

$$x^{*}\in\arg\inf\left\{\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^{2}+\sum_{i=1}^{p}P_{\lambda}^{'}(|x_{i}^{*}|)|x_{i}|:\mathbf{x}\in\Re^{p}\right\},\qquad a.s. \qquad (47)$$

Further observe that

$\inf\{\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^{2}+\sum_{i=1}^{p}P_{\lambda}^{'}(|x_{i}^{*}|)|x_{i}|:\mathbf{x}\in\Re^{p}\}\leq\inf\{\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^{2}+\sum_{i=1}^{p}P_{\lambda}^{'}(|x_{i}^{*}|)|x_{i}|:\mathbf{x}\in\Re^{p},\ x_{i}=0,\ \forall i\in\mathscr{S}^{c}\}\ a$. Recall that $x_{i}^{*}\neq 0 \Rightarrow |x_{i}^{*}|\geq a\lambda$ for all $i = 1,\dots, p$ due to Corollaries 3 and 4 under S^{3}ONC, the assumed events (i) and (ii). Also notice that $|x_{i}^{*}|>0$ for all $i \in \mathscr{S}$ as shown in (46), we then may continue the above inequality as

$\inf\{\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^{2}+\sum_{i\in\mathscr{S}^{c}}P_{\lambda}^{'}(|x_{i}^{*}|)|x_{i}|:\mathbf{x}\in\Re^{p}\}\leq\inf\{\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^{2}:\mathbf{x}\in\Re^{p},\ x_{i}=0,\ \forall i\in\mathscr{S}^{c}\}$. Since we have shown that $\mathbf{x}^{*}$ satisfies (47) and $x_{i}^{*}=0$ for all $i \in \mathscr{S}^{c}$ almost surely, that is, $\mathbf{x}^{*}$ is a feasible solution to $\{\mathbf{x} \in \Re^{p} : x_{i} = 0, \forall i \in \mathscr{S}^{c}\}$ a.s. We then know $\mathbf{x}^{*}\in\arg\inf\{\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^{2}:\mathbf{x}\in\Re^{p},\ x_{i}=0,\ \forall i\in\mathscr{S}^{c}\}\ \text{a. s.}$, which is immediately the desired result.

Lemma 5: *Let Assumptions A.1 hold. Consider a solution* $\mathbf{x}^{*}$ *satisfying S*^{3}*ONC of FCPSLR* (3) *with either SCAD or MCP. Assume the simultaneous occurrence of (i) the event that Condition B with any initial solution* $\mathbf{x}^{0}$ *is satisfied; and (ii) the event that* $f(\mathbf{x}^{*}) \leq \min\{f(\mathbf{x}^{0}), f(\mathbf{x}^{true}) + \Gamma\}$ *holds for an arbitrary* $\Gamma \geq 0$. *For any integer* $\tilde{p} : 2|\mathscr{S}| \leq \tilde{p} \leq p$ *if the penalty parameters* (a, λ) *satisfy that* $P_{\lambda}(a\lambda)>\frac{\sigma^{2}}{2n}(1+2\sqrt{t}+2t)+\frac{\frac{\sigma^{2}}{n}|\mathscr{S}|\cdot(1+2\sqrt{t}+2t)+\Gamma}{\tilde{p}-2|\mathscr{S}|+1}$, *for an arbitrary t*

> 0, then $\|\mathbf{x}^* - \mathbf{x}^{true}\|_0 \leq \tilde{p}$ with probability at least $1-\exp(-(\tilde{p}+1)(t-\ln p)) \cdot \frac{1-\exp(-(p-\tilde{p})(t-\ln p))}{1-\exp(-t+\ln p)}$.

Proof: Conditioning on the event that $f(\mathbf{x}^*) \leq f(\mathbf{x}^{true}) + \Gamma$, we know that $\frac{1}{2n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})-W\|^2+\sum_{i=1}^{p} P_\lambda(|x_i^*|) \leq \frac{1}{2n}\|\mathbf{A}(\mathbf{x}^{true}-\mathbf{x}^{true})-W\|^2+\sum_{i=1}^{p} P_\lambda(|x_i^{true}|)+\Gamma$ almost surely. Therefore, combined with the fact that $P_\lambda(|x|) \leq P_\lambda(a\lambda)$ for all $x \in \Re$, it holds that

$\frac{1}{2n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2-\frac{W^\top \mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})}{n}+\frac{1}{2n}\|W\|^2+\sum_{i=1}^{p} P_\lambda(|x_i^*|) \leq \frac{1}{2n}\|W\|^2+|\mathscr{S}|\cdot P_\lambda(a\lambda)+\Gamma \text{ a. s.}$. Combining the satisfaction of S^{3}ONC by $\mathbf{x}^*$ with (a) Corollaries 3 and 4, conditioning on both the event that Condition B holds and the event that $f(\mathbf{x}^*) \leq f(\mathbf{x}^0)$ is satisfied, which imply that $|x^*| \geq a\lambda$ unless $|x^*| = 0$; and with (b) the property of P_λ that $P_\lambda(|x|) = P_\lambda(a\lambda)$ for all $x \in \Re : |x| \geq a\lambda$, one knows that $\sum_{i=1}^{p} P_\lambda(|x_i^*|)=\|\mathbf{x}^*\|_0 \cdot P_\lambda(a\lambda) \text{ a. s.}$ Therefore,

$$|\mathscr{S}|\cdot P_\lambda(a\lambda)+\Gamma \geq \frac{1}{2n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2-\frac{W^\top \mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})}{n}+\|\mathbf{x}^*\|_0 \cdot P_\lambda(a\lambda), \quad a.s. \tag{48}$$

Now consider an event $\mathscr{E}_1 := \{\|\mathbf{x}^* - \mathbf{x}^{true}\|_0 = \tilde{p} + k\}$ for an arbitrary integer $k : 1 \leq k \leq p-\tilde{p}$. Conditioning on this event, following the same argument as in Lemma 3 (inspired by Lemma 8 of [31]), we may denote $\delta\mathbf{x}^*:=(\delta x_i^*)=\mathbf{x}^*-\mathbf{x}^{true}$ and $S_{\tilde{p}+k} \subseteq \{1,..., p\}$ such that $\delta x_i^*=0$ for all $i \notin S_{\tilde{p}+k}$. By assumption, we can ensure that $|S_{\tilde{p}+k}| = \tilde{p} + k$. Also denote by $\mathbf{A}_{S_{\tilde{p}+k}}$ the sub-matrix of $\mathbf{A}$ of the largest size such that $\mathbf{A}_{S_{\tilde{p}+k}} = (a_{ji} : j = 1,..., n, i \in S_{\tilde{p}+k})$ and let $\delta\mathbf{x}^*_{S_{\tilde{p}+k}}:=(\delta x_i^*:i \in S_{\tilde{p}+k})$. Then, $\mathbf{A}_{S_{\tilde{p}+k}}$ admits a singular value decomposition with $\mathbf{A}_{S_{\tilde{p}+k}} = V_{S_{\tilde{p}+k}} D_{S_{\tilde{p}+k}} U_{S_{\tilde{p}+k}}$, for some matrix $V_{S_{\tilde{p}+k}} \in \Re^{n\times(\tilde{p}+k)}$ with orthonormal columns, i.e., $V_{S_{\tilde{p}+k}}^\top V_{S_{\tilde{p}+k}}=I$, where I is an identity matrix and for any $v \in \Re^{\tilde{p}+k}$, $\|\mathbf{A}_{S_{\tilde{p}+k}} v\| = \|D_{S_{\tilde{p}+k}} U_{S_{\tilde{p}+k}} v\|$. Therefore, one obtains that

$$\begin{aligned}\left|W^\top \mathbf{A}\delta\mathbf{x}^*\right| &= \left|W^\top \mathbf{A}_{S_{\tilde{p}+k}} \delta\mathbf{x}^*_{S_{\tilde{p}+k}}\right| \leq \|W^\top V_{S_{\tilde{p}+k}}\| \|D_{S_{\tilde{p}+k}} U_{S_{\tilde{p}+k}} \delta\mathbf{x}^*_{S_{\tilde{p}+k}}\| \\ &= \|V_{S_{\tilde{p}+k}}^\top W\| \|\mathbf{A}_{S_{\tilde{p}+k}} \delta\mathbf{x}^*_{S_{\tilde{p}+k}}\| \leq \inf_{|S_{\tilde{p}+k}|=\tilde{p}+k} \|V_{S_{\tilde{p}+k}}^\top W\| \|\mathbf{A}\delta\mathbf{x}^*\|, \quad a.s.\end{aligned} \tag{49}$$

Therefore, conditioning on Event $\mathscr{E}_1$ and observing that $\|\mathbf{x}^*\|_0 \geq \tilde{p}+k-|\mathscr{S}|$ almost surely (because of Event $\mathscr{E}_1$ and the fact that $\|\mathbf{x}^{true}\|_0 = |\mathscr{S}|$), one may continue from (48) to obtain $\frac{1}{2}\left(\frac{\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|}{\sqrt{n}}\right)^2-\sup_{|S_{\tilde{p}+k}|=\tilde{p}+k} \frac{\|V_{S_{\tilde{p}+k}}^\top W\|}{\sqrt{n}} \cdot \frac{\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|}{\sqrt{n}} \leq -(\tilde{p}+k-2|\mathscr{S}|)\cdot P_\lambda(a\lambda)+\Gamma$ almost surely. One may see the above as a quadratic inequality for $\frac{\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|}{\sqrt{n}}$. From the analysis above, such an inequality is feasible as long as $f(\mathbf{x}^*) \leq f(\mathbf{x}^{true}) + \Gamma$, which holds almost surely conditioning on the assumed events. This feasibility implies that

$$\left(\sup_{|S_{\tilde{p}+k}|=\tilde{p}+k}\frac{\|V_{S_{\tilde{p}+k}}^{\top}W\|}{\sqrt{n}}\right)^{2}-2[(\tilde{p}+k-2|\mathscr{S}|)\cdot P_{\lambda}(a\lambda)-\Gamma]\geq 0\quad a.s. \tag{50}$$

Now consider another event $\mathscr{E}_2(t){:=}\{\sup_{|S_{\tilde{p}+k}|=\tilde{p}+k}\|V_{S_{\tilde{p}+k}}^{\top}W\|\leq\sigma\sqrt{\tilde{p}+k}\cdot\sqrt{1+2\sqrt{t}+2t}\}$ for an arbitrary $t > 0$. Conditioning on the simultaneous occurrence of both $\mathscr{E}_1$ and $\mathscr{E}_2(t)$, we know from (50) that

$\frac{\sigma^2(\tilde{p}+k)}{n}\cdot(1+2\sqrt{t}+2t)\geq\left(\sup_{|S_{\tilde{p}+k}|=\tilde{p}+k}\frac{\|V_{S_{\tilde{p}+k}}^{\top}W\|}{\sqrt{n}}\right)^{2}\geq 2[(\tilde{p}+k-2|\mathscr{S}|)\cdot P_{\lambda}(a\lambda)-\Gamma]$ almost surely, which contradicts with the assumption on the parameters (a, λ) such that

$P_{\lambda}(a\lambda)>\frac{\sigma^2}{2n}(1+2\sqrt{t}+2t)+\frac{\frac{\sigma^2}{n}|\mathscr{S}|\cdot(1+2\sqrt{t}+2t)+\Gamma}{\tilde{p}-2|\mathscr{S}|+1}\geq\frac{\sigma^2}{2n}(1+2\sqrt{t}+2t)+\frac{\frac{\sigma^2}{n}|\mathscr{S}|\cdot(1+2\sqrt{t}+2t)+\Gamma}{\tilde{p}-2|\mathscr{S}|+k}$ which implies $\frac{\sigma^2}{n}(\tilde{p}+k)\cdot(1+2\sqrt{t}+2t)<2[(\tilde{p}-2|\mathscr{S}|+k)\cdot P_{\lambda}(a\lambda)-\Gamma]$. This means that $0 = Prob[\mathscr{E}_1 \cap \mathscr{E}_2(t)]$. Therefore, by the union bound,

$$0\geq 1-Prob[\overline{\mathscr{E}}_1]-Prob[\overline{\mathscr{E}}_2(t)] \tag{51}$$

where $\bar{\mathscr{E}}_1$ and $\bar{\mathscr{E}}_2(t)$ are the complements of $\mathscr{E}_1$ and $\mathscr{E}_2(t)$, respectively. Inequality (51) implies that $Prob[\bar{\mathscr{E}}_2(t)] \geq Prob[\mathscr{E}_1]$.

Recall the decomposition of $\mathbf{A}_{S_{\tilde{p}+k}}$ and the definition of $S_{\tilde{p}+k}$ as in (49). We want to use Lemma 8, which follows Theorem 2.1 in [22], to bound $Prob[\bar{\mathscr{E}}_2(t)]$. In that lemma, we let $\sum_v = V_{S_{\tilde{p}+k}}V_{S_{\tilde{p}+k}}^{\top}$. Noticing that Σ_V is an idempotent matrix, then $\|\Sigma_V\| \leq 1$ and $\mathbf{Tr}(\Sigma_V) = rank(\Sigma_V)$, where $\mathbf{Tr}(\cdot)$ and $rank(\cdot)$ are trace and rank of a matrix. Because $V_{S_{\tilde{p}+k}}V_{S_{\tilde{p}+k}}^{\top}$ is a projection matrix onto the span of the column vectors of $V_{S_{\tilde{p}+k}}$, which is at most $\tilde{p}+k$ dimensional. Then $\mathbf{Tr}(\sum_v^2)=\mathbf{Tr}(\sum_v)=rank(\sum_v)\leq\tilde{p}+k$. Now, we may invoke Lemma 8 with the settings discussed above and obtain, for a fixed $S_{\tilde{p}+k}$ and an arbitrary $t' > 0$,

$Prob\left[\|V_{S_{\tilde{p}+k}}^{\top}W\|\leq\sigma\cdot\sqrt{\tilde{p}+k+2\sqrt{(\tilde{p}+k)t'}+2t'}\right]\geq 1-\exp(-t')$. Further invoking the union bound, one obtains

$Prob[\sup_{|S_{\tilde{p}+k}|=\tilde{p}+k}\|V_{S_{\tilde{p}+k}}^{\top}W\|>\sigma\cdot\sqrt{\tilde{p}+k+2\sqrt{(\tilde{p}+k)t'}+2t'}\leq\begin{pmatrix}p\\ \tilde{p}+k\end{pmatrix}\cdot\exp(-t')\leq p^{(\tilde{p}+k)}\cdot\exp(-t')$. For notational simplicity we have not used a potentially tighter bound

$\begin{pmatrix}p\\ \tilde{p}+k\end{pmatrix}\cdot\exp(-t')\leq\left(\frac{pe}{2(\tilde{p}+k)}\right)^{(\tilde{p}+k)}\cdot\exp(-t')$ for the last inequality. By letting $t' = (\tilde{p} + k)t$, we immediately have $Prob[\bar{\mathscr{E}}_2(t)] \leq p^{(\tilde{p}+k)} \cdot \exp(-(\tilde{p} + k)t)$.

Notice that the above argument holds for any integer $k : 1 \leq k \leq p - \tilde{p}$. Combined with (51), one may obtain that $Prob\,[\|\mathbf{x}^* - \mathbf{x}^{true}\|_0 = \tilde{p} + k] \leq \exp\,(-(\tilde{p} + k)\,(t - \ln p))$ for all $k : 1 \leq k \leq$

$p - \tilde{p}$. Also notice that if $\|\mathbf{x}^* - \mathbf{x}^{true}\|_0 \geq \tilde{p} + 1$, then it must hold that $\|\mathbf{x}^* - \mathbf{x}^{true}\|_0 \in \{\tilde{p} + 1, \ldots, p\}$. Hence, invoking the union bound,

$$Prob[\|\mathbf{x}^* - \mathbf{x}^{true}\|_0 \geq \tilde{p}+1] \leq \sum_{k=1}^{p-\tilde{p}} Prob[\|\mathbf{x}^* - \mathbf{x}^{true}\|_0 = \tilde{p}+k] \leq \sum_{k=1}^{p-\tilde{p}} \exp(-(\tilde{p}+k)(t - \ln p)) = \exp(-(\tilde{p}+1)(t - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p})(t-\ln p))}{1-\exp(-t+\ln p)}$$

. The last equality is from the geometric sum. This immediately implies the desired result.

5 An S³ONC-Guaranteeing Algorithm and Numerical Results

To solve for a solution satisfying S^{3}ONC, we adopt the potential reduction (PR) algorithm based on [40]. PR converges to a second-order KKT solution, which implies the S^{3}ONC. See the online supplement [25] for more details. Alternative approaches such as interior point methods in [5] are theoretically guaranteed FPTAS to a second-order KKT solution. Since our analysis on the sparsity and statistical performance of FCPSLR are algorithm-independent, one may substitute PR by any S^{3}ONC-guaranteeing algorithms. We leave the comprehensive comparison among all alternative algorithms for future study.

We conduct two sets of numerical tests. We compare PR with several different initialization schemes in solving both FCPSLR-SCAD and FCPSLR-MCP. We also compare the first-order solutions and the S^{3}ONC solutions to FCPSLR-MCP. The detailed experiment setups are provided in [25].

In the first test set, we consider three different initialization schemes for both FCPSLR-SCAD and -MCP: (i) starting from the analytic center; (ii) starting from the all-zero solution; (iii) starting from the Lasso solution. The observations are summarized in the following: Firstly, starting from any of the three initial solutions, the PR can correctly recover the oracle solution in nearly all instances, while (iii) results in the best performance. Secondly, the recovery qualities are fairly insensitive to the choice between MCP and SCAD, while our theories presented formerly indicate that SCAD often may require more conditions to ensure statistical performance. This implies that our theoretical findings on SCAD can potentially be improved. Thirdly, when the minimal signal strength is stronger, recovering the oracle solutions becomes easier.

The second test focuses on FCPSLR-MCP and compares between an S^{3}ONC solution generated by PR and the FONC solution generated by local linear approximation (LLA). LLA is a learning algorithm proposed by [47] to solve sparse estimation problems including FCPSLR and is shown by [20] to converge to an FONC solution asymptotically. We compare PR with the state-of-the-art variant of LLA variant proposed by [17], namely, LLA initialized with the Lasso. The same paper shows the LLA variant entails the strong oracle property under the RE condition. With the same set of parameters and the same initialization scheme, we think that solutions generated by PR and by LLA are different at least in terms of whether the S^{3}ONC is ensured or only the FONC is guaranteed. Therefore, the comparison here may essentially represent the comparison between an S^{3}ONC solution and a (wisely determined) FONC solution. From the numerical results, we see that both LLA and PR yield good performance. Nonetheless, we observe noticeable outperformance of PR over LLA for some choices of parameters, thus showing that the S^{3}ONC solutions are more robust than the FONC solutions in statistical performance at least for some choices of parameters.

6 Conclusion

This paper studies the properties of an FCPSLR problem using SCAD or MCP for regularization. Despite that the global solution is shown to entail desirable recovery properties by [44], globally minimizing FCPSLR is NP-complete. This paper shows that the global optimality is in fact not necessarily stipulated to ensure the recovery quality. Specifically, we provide conditions for the parameters under which any local solution is a sparse estimator. More importantly, from an algorithm-independent point of view, we show the following results: (i) Any solution satisfying S^{3}ONC to FCPSLR may achieve bounded statistical errors. Furthermore, those local solutions may even exactly recover the oracle solution, given that the minimal signal strength is large enough. These results also reveal that the statistical performance improves polynomially with the reduction in suboptimality gap. (ii) In the MCP case, the S^{3}ONC solutions that have a lower objective value than the Lasso solution entail the strong oracle property. These local solutions may also dominate the Lasso in terms of ME when sample size is greater than a certain threshold, while the worst-case ME of those S^{3}ONC solutions is comparable to Lasso. To our knowledge, this is the first theoretical guarantee for the statistical performance at the S^{3}ONC solutions, disregarding the choice of computing procedures; it is also the first attempt to reveal the correlation between the optimization quality in solving the non-convex formulation of the learning problem and the statistical quality in sparse recovery. An S^{3}ONC solution admits FPTAS, such as the interior point methods proposed by [5].

We employ PR to generate an S^{3}ONC solution. Several predictions by our theory are verified by the numerical results. Meanwhile, they also indicate a potential gap between our theoretical results and the actual performance of FCPSLR-SCAD, which is an interesting question to pursue in future. Also of future interest is a comprehensive comparison among different S^{3}ONC-guaranteeing algorithms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the AE and referees for their valuable comments, which significantly improve the paper. This work was supported by Penn State Grace Woodward Collaborative Research Grant, NSF grants CMMI 1300638 and DMS 1512422, NIH grants P50 DA036107 and P50 DA039838, Marcus PSU-Technion Partnership grant, Air Force Office of Scientific Research grant FA9550-12-1-0396, and Mid-Atlantic University Transportation Centers grant. This work was also partially supported by NNSFC grants 11690014 and 11690015. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, the NIDA, the NIH, the AFOSR, the MAUTC or the NNSFC.

References

1. Adamczak R, Litvak A, Pajor A, Tomczak-Jaegermann N. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. J Amer Math Soc. 2010; 234:535–561.
2. Ahlswede R, Winter A. Strong converse for identification via quantum channels. IEEE Trans Inform Theory. 2002; 48(3):569–579.
3. Bertsimas D, Mazumder R. Least quantile regression via modern optimization. Ann Stat. 2014; 42:2494–2525.
4. Bian, W., Chen, X. Optimality conditions and complexity for non-Lipschitz constrained optimization problems. 2014. http://www.polyu.edu.hk/ama/staff/xjchen/OCT26
5. Bian W, Chen X, Ye Y. Complexity analysis of interior point algorithms for non-Lipschitz and non-convex minimization. Math Prog A. 149:301–327.
6. Bickel PJ, Ritov Y, Tsybakov AB. Simultaneous analysis of Lasso and Dantzig selector. Ann Stat. 2009; 37:1705–1732.
7. Bunea F, Tsybakov A, Wegkamp M. Aggregation for Gaussian regression. Ann Stat. 2007; 35(4): 1674–1697.
8. Candés E, Tao T. Decoding by linear programming. IEEE Trans Inf Theory. 2005; 51(12):4203–4215.
9. Candés E, Tao T. Near optimal signal recovery from random projections: Universal encoding strategies? IEEE Trans Inf Theory. 2006; 52(12):5406–5425.
10. Candés E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. Ann Stat. 2007; 35(6):2313–2351.
11. Cartis C, Gould NIM, Toint PI. Adaptive cubic regularization methods for unconstrained optimization. Part I: motivation, convergence and numerical results. Math Prog A. 2011; 127:245–295.
12. Chen X, Ge D, Wang Z, Ye Y. Complexity of unconstrained L_2-L_p minimization. Math Prog A. 2014; 143:371–383.
13. Chen X, Xu F, Ye Y. Lower bound theory of non-zero entries in solutions of L_2-Lp minimization. SIAM J Sci Comput. 2010; 32(5):2832–2852.
14. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc. 2001; 96:1348–1360.
15. Fan J, Lv J. Nonconcave penalized likelihood with NP-dimensionality. IEEE Trans Inform Theory. 2011; 57:5467–5484.
16. Fan J, Lv J, Qi L. Sparse high dimensional models in economics. Annu Rev Econom. 2011; 3:291–317. [PubMed: 22022635]
17. Fan J, Xue L, Zou H. Strong oracle optimality of folded concave penalized estimation. Ann Stat. 2014; 42(3):819–849. [PubMed: 25598560]
18. Ge, D., Wang, Z., Ye, Y., Yin, H. Strong NP-hardness result for regularized L_q-minimization problems with concave penalty functions. 2015. http://arxiv.org/pdf/1501.00622v1
19. Hillar, CJ., Lin, S., Wibisono, A. Tight bounds on the infinity norm of inverses of symmetric diagonally dominant positive matrices. 2014. http://www.msri.org/people/members/chillar/files/HLW_inv_positive_diag_dom
20. Hunter D, Li R. Variable selection using MM algorithms. Ann Stat. 2005; 33:1617–1642. [PubMed: 19458786]
21. Hsu, D., Kakade, SM., Zhang, T. Random design analysis of ridge regression. 2014. http://arxiv.org/pdf/1106.2363v2
22. Hsu D, Kakade SM, Zhang T. A tail inequality for quadratic forms of subgaussian random vectors. Electron Commun Probab. 2012; 17(52):1–6.
23. Huo X, Chen J. Complexity of penalized likelihood estimation. J Stat Comput Simul. 2010; 80(7): 747–759.
24. Liu H, Yao T, Li R. Global solutions for folded concave penalized nonconvex learning. Ann Stat. 2016; 44(2):629–659. [PubMed: 27141126]

25. Liu H, Yao T, Li R, Ye Y. Electronic Companion to: Folded Concave Penalized Sparse Linear Regression: Sparsity, Statistical Performance, and Algorithmic Theory for Local Solutions.
26. Loh P-L, Wainwright MJ. Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. J Mach Learn Res. 2015; 16:559–616.
27. Negahban SN, Ravikumar P, Wainwright MJ, Yu B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statist Sci. 2012; 27(4):538–557.
28. Nesterov, Yu, Polyak, BT. Cubic regularization of Newton's method and its global performance. Math Program. 2006; 108(1):177–205.
29. Raskutti G, Wainwright M, Yu B. Restricted nullspace and eigenvalue properties for correlated Gaussian designs. J Mach Learn Res. 2010; 11:2241–2259.
30. Rudelson M, Zhou S. Reconstruction from Anisotropic random measurements. IEEE Trans Inf Theory. 2013; 59(6):3434–3447.
31. Raskutti G, Wainwright MJ, Yu B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE Trans Inform Theory. 2011; 57(10):6976–6994.
32. Tibshirani R. Regression shrinkage and selection via the Lasso. J Royal Statist Soc B. 1996; 58(1): 267–288.
33. van de Geer SA, Bühlmann P. On the conditions used to prove oracle results for the Lasso. Electron J Statist. 2009; 3:1360–1392.
34. Vavasis SA. Quadratic programming is in NP. Inform Process Lett. 1990; 36:73–77.
35. Vershynin, R. How close is the sample covariance matrix to the actual covariance matrix. 2010. http://arxiv.org/pdf/1004.3484v2
36. Volkov YS, Miroshnichenko VL. Norm estimates for the inverses of matrices of monotone type and totally positive matrices. Sib Math J. 2009; 50(6):982–987.
37. Wang L, Kim Y, Li R. Calibrating non-convex penalized regressioni in ultra-high dimension. Ann Stat. 2013; 41(5):2505–2536. [PubMed: 24948843]
38. Wang Z, Liu H, Zhang T. Optimal computational and statistical rates of convergence for sparse non-convex learning problems. Ann Stat. 2014; 42(6):2164–2201. [PubMed: 25544785]
39. Ye Y. On affine scaling algorithms for non-convex quadratic programming. Math Prog. 1992; 56:285–300.
40. Ye Y. On the complexity of approximating a KKT point of quadratic programming. Math Prog. 1998; 80:195–211.
41. Zhang C. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010; 28:894–942.
42. Zhang CH, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. Ann Stat. 2008; 36:1567–1594.
43. Zhang Y, Wainwright MJ, Jordan MI. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. JMLR: Workshop and Conference Proceedings. 2014; 35:1–18.
44. Zhang C, Zhang T. A general theory of concave regularization for high dimensional sparse estimation problems. Statist Sci. 2012; 27(4):576–593.
45. Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc B. 2013; 76(1):217–242.
46. Zhou, S. Restricted Eigenvalue Conditions on Subgaussian Random Matrices. 2009. http://arxiv.org/pdf/0912.4045v2
47. Zou H, Li R. One-step sparse estimation in non-concave penalized likelihood models. Ann Stat. 2008; 36:1509–1533. [PubMed: 19823597]

A Some Useful Lemmas

Lemma 6

For any $\mathbf{x}^{true} \in \Re^p$, $\mathbf{A} \in \Re^{n\times p}$, $W \in \Re^n$, $\mathbf{b} = \mathbf{A}\mathbf{x}^{true} + W$, *consider f as defined in* (3) *with arbitrarily either* $P_\lambda = P_{\lambda,SCAD}$ *or* $P_\lambda = P_{\lambda,MCP}$. *Let* $\mathbf{x}^0 \in \Re^p$ *be a feasible solution to* (3). *If* $f(\mathbf{x}^0)$ *satisfies that* $f(\mathbf{x}^0) \le f(\mathbf{x}^{lasso})$, *where* $\mathbf{x}^{lasso}$ *is defined in* (4) *with the same problem data* $\mathbf{x}^{true}$, $\mathbf{A}$, *and* $\mathbf{b}$ *as* (3) *and with an arbitrary penalty parameter* $\lambda_{lasso} > 0$, *then* $f(\mathbf{x}^0) - f(\mathbf{x}^{true}) \le (\lambda_{lasso} + \lambda)\,|\mathbf{x}^{lasso} - \mathbf{x}^{true}|$.

Proof

Denote that $f_{lasso}(\mathbf{x}) = (2n)^{-1}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \sum_{i=1}^{p} \lambda_{lasso}|x_i|$ for any $\mathbf{x} = (x_i) \in \Re^p$.

Firstly, notice that by definition of $\mathbf{x}^{lasso}$ in (4), $f_{lasso}(\mathbf{x}^{lasso}) \le f_{lasso}(\mathbf{x}^{true})$. We then know that

$$(2n)^{-1}\|\mathbf{A}\mathbf{x}^{lasso} - \mathbf{b}\|^2 - (2n)^{-1}\|\mathbf{A}\mathbf{x}^{true} - \mathbf{b}\|^2 \le \sum_{i=1}^{p} \lambda_{lasso}|x_i^{true}| - \sum_{i=1}^{p} \lambda_{lasso}|x_i^{lasso}| \le \sum_{i=1}^{p} \lambda_{lasso}|x_i^{true} - x_i^{lasso}| = \lambda_{lasso}|\mathbf{x}^{true} - \mathbf{x}^{lasso}|$$

Secondly, due to the concavity and differentiability of $P_\lambda(\cdot)$ on $\Re_+$ and the fact that $0 \le P_\lambda'(|x|) \le \lambda$ for all $x \in \Re$,

$$\sum_{i=1}^{p} P_\lambda(|x_i^{lasso}|) - \sum_{i=1}^{p} P_\lambda(|x_i^{true}|) \le \sum_{i=1}^{p} P_\lambda'(|x_i^{true}|) \cdot (|x_i^{lasso}| - |x_i^{true}|) \le \sum_{i=1}^{p} P_\lambda'(|x_i^{true}|) \cdot |x_i^{lasso} - x_i^{true}| \le \lambda|\mathbf{x}^{lasso} - \mathbf{x}^{true}|.$$

Combining the above and the assumption that $f(\mathbf{x}^0) \le f(\mathbf{x}^{lasso})$, we know that

$$f(\mathbf{x}^0) - f(\mathbf{x}^{true}) \le f(\mathbf{x}^{lasso}) - f(\mathbf{x}^{true}) = (2n)^{-1}\|\mathbf{A}\mathbf{x}^{lasso} - \mathbf{b}\|^2 + \sum_{i=1}^{p} P_\lambda(|x_i^{lasso}|) - (2n)^{-1}\|\mathbf{A}\mathbf{x}^{true} - \mathbf{b}\|^2 - \sum_{i=1}^{p} P_\lambda(|x_i^{true}|) \le (\lambda_{lasso} + \lambda)|\mathbf{x}^{lasso} - \mathbf{x}^{true}|$$

, as claimed.

Lemma 7

Assume that Condition B holds with initial solution $\mathbf{x}^0 \in \Re^p$. For any $\mathbf{x}^{true} \in \Re^p$, $\mathbf{A} \in \Re^{n\times p}$, $W \in \Re^n$, $\mathbf{b} = \mathbf{A}\mathbf{x}^{true} + W$, *and for any* $\mathbf{x}^* = (x_i^*) \in \Re^p$ *that satisfies (i)* S^3 *ONC to* (3) *with arbitrarily either* $P_\lambda = P_{\lambda,SCAD}$ *or* $P_\lambda = P_{\lambda,MCP}$; *and (ii) the inequality that* $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$, *the following inequality holds:*

$$(2n)^{-1}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le n^{-1}W^\top\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})+\min\{\sum_{i\in\mathscr{S}}P_\lambda'(|x_i^*|)|x_i^{true}|, \sum_{i\in\mathscr{S}}P_\lambda'(|x_i^*|)|x_i^*-x_i^{true}|\}$$

. If, in addition, $f(\mathbf{x}^*) \le f(\mathbf{x}^{true})+\Gamma$ *for an arbitrary* $\Gamma \ge 0$, *then*

$$\frac{1}{2n}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le \frac{1}{n}W^\top\mathbf{A}(\mathbf{x}^* -\mathbf{x}^{true})+\min\{\sum_{i\in\mathscr{S}}P_\lambda'(|x_i^*|)|x_i^{true}|, \sum_{i\in\mathscr{S}}P_\lambda'(|x_i^*|)|x_i^*-x_i^{true}|, P_\lambda(a\lambda)\cdot(|\mathscr{S}|-\|\mathbf{x}^*\|_0)+\Gamma\}$$

.

Proof

Notice that $\mathbf{b} = \mathbf{A}\mathbf{x}^{true} + W$. Then for any

$$\mathbf{x}=(x_i)\in\Re^p: (2n)^{-1}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\sum_{i=1}^p P_\lambda'(|x_i^*|)|x_i| = (2n)^{-1}\|\mathbf{A}(\mathbf{x}-\mathbf{x}^{true})\|^2+(2n)^{-1}W^\top W-n^{-1}W^\top\mathbf{A}(\mathbf{x} -\mathbf{x}^{true})+\sum_{i=1}^p P_\lambda'(|x_i^*|)|x_i|$$

.

Since $\mathbf{x}^*$ satisfies S^{3}ONC, which implies FONC, we know that

$\mathbf{x}^* \in \arg\inf\{\frac{1}{2n}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\sum_{i=1}^p P_\lambda'(|x_i^*|)|x_i| : \mathbf{x}\in\Re^p\}$. Therefore,

$\frac{1}{2n}\|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|^2+\sum_{i=1}^p P_\lambda'(|x_i^*|)|x_i^*| \le \frac{1}{2n}\|\mathbf{A}\mathbf{x}^{true}-\mathbf{b}\|^2+\sum_{i=1}^p P_\lambda'(|x_i^*|)|x_i^{true}|$. Combining the above, we know that

$$(2n)^{-1}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2-n^{-1}W^\top\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})+\sum_{i=1}^p P_\lambda'(|x_i^*|)|x_i^*| \le \sum_{i=1}^p P_\lambda'(|x_i^*|)|x_i^{true}|$$

. Further invoking the definitions of $\mathbf{x}^{true}$ and $\mathscr{S}$ as well as triangular inequality and the fact that $P_\lambda'(|x|)\ge 0$ for any $x \in \Re$, we have

$$(2n)^{-1}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2 \le n^{-1}W^\top\mathbf{A}(\mathbf{x}^* -\mathbf{x}^{true})+\sum_{i\in\mathscr{S}}P_\lambda'(|x_i^*|)|x_i^{true}|-\sum_{i=1}^p P_\lambda'(|x_i^*|)|x_i^*| \le n^{-1}W^\top\mathbf{A}(\mathbf{x}^* -\mathbf{x}^{true})+\sum_{i\in\mathscr{S}}P_\lambda'(|x_i^*|)|x_i^{true} -x_i^*| -\sum_{i\in\mathscr{S}^c}P_\lambda'(|x_i^*|)|x_i^*|$$

. We then obtain the claimed result in the first part of the lemma.

To show the second part, by assumption, $f(\mathbf{x}^*) \le f(\mathbf{x}^{true})+\Gamma$, we know

$$(2n)^{-1}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2-n^{-1}W^\top\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})+(2n)^{-1}\|W\|^2+\sum_{i=1}^p P_\lambda(|x_i^*|) \le (2n)^{-1}\|W\|^2+\sum_{i=1}^p P_\lambda(|x_i^{true}|$$

. Noticing the fact that (i) $0 \le P_\lambda(|x|) \le P_\lambda(a\lambda)$ for any $x \in \Re$, (ii) $P_\lambda(|0|) = 0$, and (iii) by definition of $\mathscr{S}^c$, $x_i^{true}=0$ for all $i \in \mathscr{S}^c$, we hence know

$$(2n)^{-1}\|\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true})\|^2-n^{-1}W^\top\mathbf{A}(\mathbf{x}^*-\mathbf{x}^{true}) \le P_\lambda(a\lambda)\cdot|\mathscr{S}|-\sum_{i=1}^p P_\lambda(|x_i^*|)+\Gamma.$$

Invoking Corollaries 3 and 4 under Condition B and the assumption that $f(\mathbf{x}^*) \le f(\mathbf{x}^0)$, we know that $x_i^* \ne 0 \Rightarrow |x_i^*| \ge a\lambda$. Also notice that $P_\lambda(|x|) = P_\lambda(a\lambda)$ for all $x \in \Re$: $|x| \ge a\lambda$.

Therefore, the above implies $\sum_{i=1}^p P_\lambda(|x_i^*|)=P_\lambda(a\lambda)\cdot\|\mathbf{x}^*\|_0$ and $(2n)^{-1}\|\mathbf{A}(\mathbf{x}^* - \mathbf{x}^{true})\|^2 - n^{-1}W^\top \mathbf{A}(\mathbf{x}^* - \mathbf{x}^{true}) \le P_\lambda(a\lambda)\cdot(|\mathscr{S}| - \|\mathbf{x}^*\|_0) + \Gamma$. Combined with the results from the first part of this lemma, we have the claimed result in the second part.

Lemma 8

Consider a subgaussian ñ-dimensional random vector $\tilde{W}$ in $\Re^{\tilde{n}}$ that satisfies Prob$[|\langle \tilde{W}, \upsilon\rangle| \geq t] \leq 2\exp(-t^2(2\sigma^2)^{-1})$.*for any* $\upsilon \in \Re^{\tilde{n}}: \|\upsilon\| = 1$, *then for any* $V \in \Re^{\tilde{n}\times\tilde{n}}$ *and* $\Sigma_V = V^\top V$, $Prob[\|V\tilde{W}\|^2 \leq \sigma^2(\mathbf{Tr}(\sum_v)+2\sqrt{\mathbf{Tr}(\sum_v^2)t}+2\|\sum_v\|t)] \geq 1-\exp(t)$ *for any* $t > 0$, *where* $\mathbf{Tr}(\cdot)$ *denotes the trace of a matrix.*

Proof

Evident from Theorem 2.1 in [22].