J. Nonlinear Var. Anal. 8 (2024), No. 6, pp. 909-934
Available online at http://jnva.biemdas.com
https://doi.org/10.23952/jnva.8.2024.6.05

# FULLY POLYNOMIAL-TIME RANDOMIZED APPROXIMATION SCHEMES FOR GLOBAL OPTIMIZATION OF HIGH-DIMENSIONAL MINIMAX CONCAVE PENALIZED GENERALIZED LINEAR MODELS

CHARLES HERNANDEZ, HUNG-YI LEE, JINDONG TONG, HONGCHENG LIU*

*Department of Industrial and Systems Engineering, University of Florida, Gainesville 32611, USA*

**Abstract.** Global solutions to high-dimensional sparse estimation problems with a folded concave penalty (FCP) have been shown to be statistically desirable but are strongly NP-hard to compute which implies the non-existence of pseudo-polynomial time global optimization schemes in the worst case. This paper shows that, with high probability, a global solution to generalized linear models with minimax concave penalty (MCP), a specific type of FCP, coincides with a stationary point characterized by the significant subspace second order necessary conditions ($S^3$ONC). Given that the desired $S^3$ONC solution admits a fully polynomial-time randomized approximation scheme (FPRAS), we are able to demonstrate the existence of an FPRAS for this strongly NP-hard problem. We further demonstrate two versions of the FPRAS for generating the desired $S^3$ONC solutions. One follows the paradigm of an interior point trust region algorithm and the other is the well-studied local linear approximation (LLA). Our analysis thus provides new techniques for global optimization of certain NP-Hard problems and new insights on the effectiveness of LLA.

**Keywords.** Fully polynomial-time randomized approximation schemes; Generalized linear model; Minimax concave penalty; Significant subspace second order necessary conditions.

## 1. Introduction

This paper concerns global optimization of a folded concave penalized formulation of high-dimensional learning generalized linear models, which belongs to statistical/machine learning problems such that the number of dimensions (or number of fitting parameters) $p$ is (much) larger than the number of samples $n$. This type of problem has recently become very common in a variety of engineering and scientific applications [10, 8] including computational biology, speech recognition and image processing [15, 1, 31, 26, 2, 27]. Globally minimal solutions to such a nonconvex learning formulation have been shown effective to guarantee desirable statistical performance in order to address high dimensionality [33]. Nonetheless, generating a global solution admits no pseudo polynomial-time algorithm, unless "P = NP"; Indeed, global optimality is shown strongly NP-hard to achieve by [6, 16] while [5] shows similar results for several related problems in regularized minimization. In contrast to the existing pessimistic

*Corresponding author.
E-mail address: liu.h@ufl.edu (H. Liu).
Received 15 June 2024; Accepted 12 July 2024; Published online 25 October 2024.

result, we derive herein a fully polynomial-time randomized approximation scheme (FPRAS, as defined in 3.5) that theoretically ensures global minimality to 1.1 with high probability.

Specifically, consider a high-dimensional generalized linear model (GLM) as follows. Let $X = (x_1, \ldots, x_n)^\intercal$ be the $n \times p$ design matrix with $x_i = (x_{i1}, \ldots, x_{ip})^\intercal$, $i = 1, \ldots, n$, and $Y = (y_1, \ldots, y_n)^\intercal$ be the $n$-dimensional response vector. We will assume the design matrix $X$ is fixed, while the mean of the response is given by $\mathbb{E}[y_i] = \psi'(x_i^\top \beta^{true})$ for some known link function $\psi : \Theta \to \mathbb{R}$, where $\Theta \subseteq \mathbb{R}$ and $\beta^{true} = (\beta_1^{true}, \ldots, \beta_p^{true})$ is the unknown vector of true parameters of the model. Such a setup can be seen as a generalization of linear regression models with the link function allowing for nonlinear transformations that enable a more flexible approach to model estimation. The high-dimensional regression problem is to estimate $\beta^{true}$ given knowledge of $X$, $Y$, and $\psi$ in the undesirable scenario where $p \gg n > 0$. To that end, traditional statistical learning schemes often resort to the following formulation:

$$\mathscr{L}(\beta) = \sum_{i=1}^{n} \ell(y_i, x_i, \beta) = \frac{1}{n} \sum_{i=1}^{n} [\psi(x_i^\intercal \beta) - y_i x_i^\intercal \beta],$$

which, according to traditional statistical theories, would result in over fitting in general under the high-dimensional setting.

To resolve over fitting, modern statistical theories favor a modified formulation as below:

$$\min_{\beta} \left[ \mathscr{Q}(\beta) := \mathscr{L}(\beta) + \sum_{j=1}^{p} P_\lambda(|\beta_j|) \right], \tag{1.1}$$

where $P_\lambda(|\cdot|)$ is sparsity-inducing regularization term that penalizes any nonzero dimensions in the minimizer, and $\lambda > 0$ is a tuning parameter. Under the assumption that the true regression parameter $\beta^{true}$ is sparse, a global optimizer to (1.1) has been shown effective to address over fitting for many choices of specific regularization functions $P_\lambda$. Indeed, one of the most successful choices of $P_\lambda$ is the much studied Lasso-based regularized [28], aka, the $\ell_1$(-norm) penalty, which was demonstrated to entail desirable statistical properties [4, 25]. Another admirable property of the Lasso is that, especially when applied to least squares linear regression, it yields an extremely tractable problem via a variety of algorithms [12, 13]. However, per [35, 7], Lasso is not selection consistent without a strong irrepresentable condition and may sometimes introduce non-trivial estimation bias.

As a popular alternative to Lasso, the folded concave penalty (FCP) was first introduced by [7]. There are mainstream examples of FCP functions, including the SCAD by [7] and MCP by [32]. This paper focuses on the MCP, defined as $P_\lambda(|t|) = \int_0^{|t|} \frac{(a\lambda - s)_+}{a} ds$ for some fixed parameter $a > 0$. In contrast to the Lasso, the FCP regulaizations achieve variable selection consistency non-contingent on the irrepresentable condition and is demonstrated to be unbiased [7]. Furthermore, Zhang and Zhang [33] demonstrated that the global solution to the FCP-regularized formulation leads to desirable recovery of the oracle solution.

Nonetheless, FCP problems are significantly harder to solve than Lasso, the new penalty term moves the problem outside the realm of convex optimization, Chen et al. [6] even showed that any estimation problem with convex loss and folded concave regularization to be strongly NP-hard, ruling out the possibility of a pseudo-polynomial-time global optimization algorithm. Liu, Yao and Li [19] maybe the first to propose a global approach to the problem called MIPGO which reformulates the problem into a mixed integer program. Yet, the worst-case complexity of MIPGO is in exponential time.

Perhaps for this reason, current literature tends to focus on local algorithms for the FCP-regularized learning problems. The local quadratic approximation algorithm by [7] is an example of a majorization minimization algorithm, an approach which is also related to the local linear approximation (LLA) algorithm proposed by [37]. LLA was further explored by [11] showing the oracle property can be obtained with high probability despite the local approach. In [24, 9], it was demonstrated that coordinate optimization approaches for FCP while [30] used an approximate regularization path-following algorithm to obtain the optimal convergence rate to statistically desirable local solution. Wang, Kim and Li [29] analyzed the CCCP algorithm and prove, under certain conditions, that it asymptotically finds the oracle estimator. Liu et al. [20] took an algorithm agnostic approach by analyzing local solutions satisfying second order KKT conditions and showed desirable statistical properties like recovering the oracle solution and sparisty. These results discussed above primarily relate to FCP-regularized linear regression, a special case of GLM where $\psi$ is specifically the identity function. For analyses which encompass GLMs with FCP regularizers, Fan and Lv [9] demonstrated that GLMs, even in ultra high dimensional variable selection problems, have oracle properties when using FCP regularization and demonstrated a coordinate optimization algorithm for finding local solutions. In the area of M-estimators, which is a further generalization of our estimation method beyond even GLMs, In [21, 22], it was proved that under certain conditions all local solutions must be within statistical precision of the true parameter and its support while a two-step algorithm involving composite gradient descent to find a local solution was investigated in [23]. Bian and Chen [3] demonstrated a optimality conditions for a class of nonconvex optimization problems using nonlipscitz regularization.

From the numerous results pertaining to local solution schemes above, our research question is why local solutions are repetitively successful. In other words, are there certain geometric properties of the learning formulation (1.1) with FCP that allow all local schemes to perform well independent of the specific designs of the algorithmic procedures? Our answer to this question is affirmative; we show herein that all local solutions within an efficiently achievable sub-level set are actually globally optimal. Those local solutions are characterized by the significant subspace second-order necessary conditions (S$^3$ONC) provable admit FPRAS's. The S$^3$ONC are weaker conditions than the standard second-order KKT conditions. As per this result, all S$^3$ONC-guaranteeing algorithms (which include a large spectrum of local algorithms) belong to the class of FPRAS's for global optimization of the strongly NP-hard FCP-based learning problem. We subsequently develop theories for two specific algorithms of this type: one gradient-based method and the other is the same as the LLA.

It is worth noting that [32] provided conditions to establish the uniqueness of local solutions to FCP-based learning. When local solutions are unique, then any local optimization algorithms would ensure global optimality. However, a few critical assumptions are necessary to achieve the uniqueness result and, furthermore, many report numerical experiments, e.g., those in [11, 20, 19, 7] indicate the non-uniqueness of local solutions, instead. In contrast, our results in this paper imposes only standard assumptions commonly shared by a flexible set of high-dimensional GLMs and are applicable even if the local solutions are non-unique.

To our knowledge, this is the first geometric proof that global solutions coincide with pseudo-polynomial-time computable local solutions in an FCP-based regression formulation with high probability. The resulting algorithms are the first few FPRAS's to this problem.

Two works with notable relations to our own are [20] and [18]. The first applied a similar analytical framework to linear regression problems, however, our generalization to GLMs adds significant flexibility and it was unknown for their result that the oracle solution implies global optimality since it was only as of [33] that global optimality was known to potentially imply the oracle solution. Further, it is nontrivial to extend their existing result to global optimal results. On the other hand, [18] is a more general setup than our own though the tradeoff is that our rate is better and $S^3$ONC solutions do not ensure global optimality to the in-sample training error for their setup.

The rest of this paper is organized as follows. Section 2 goes through specific problem assumptions and explains the $S^3$ONC. Section 3 contains our main result for global optimality and uses it to make additional claims for LLA. Section 4 contains numerical results to verify our theoretical findings. Section 5 provides concluding remarks.

In this paper, we use $\|\cdot\|_0$ to denote the number of nonzero entries, $|\cdot|$ to denote the $\ell_1$-norm if the argument is a vector, or cardinality if the argument is a set, $\|\cdot\|$ to denote the $\ell_2$-norm, $\|\cdot\|_{max}$ to denote the maximum norm and $\|\cdot\|_{min}$ to denote the absolute value of the entry with the smallest magnitude. $(\cdot)_+$ is used equivalently to $\max(0,\cdot)$. For any vector $v$, $v_{\mathscr{Q}}$ is intended as $(v_j : j \in \mathscr{Q})$. For any set $\mathscr{Q}$, we denote the complement as $\mathscr{Q}^c$. In particular, let $S$ be the true support set, that is, $\mathscr{S} := \{j : \beta_j^{true} \neq 0\}$ and its complement is $\mathscr{S}^c$. We occasionally use the term the "oracle solution" to refer to the solution $\beta^{oracle}$ defined as

$$\beta^{oracle} \in \underset{\beta:\beta_j=0,\forall j\notin\mathscr{S}}{\arg\min} \mathscr{L}(\beta).$$

The oracle solution is a hypothetical solution which assumes prior knowledge of the true support set $S$ and thus can be considered a theoretical benchmark.

## 2. Setups, Preliminaries, and Assumptions

2.1. **Setups and assumptions.** Our analysis focuses on sparse GLMs that have a fixed design matrix and satisfy the following assumptions:

(A1) Assume that
   (i) $b_u \geq \psi''(x_i^\intercal \beta) \geq b_l > 0$ for all $x_i^\intercal \beta \in \Theta$;
   (ii) There exists $K > 0$ such that the design matrix satisfies $\frac{1}{n}\left\|X_j\right\|^2 < K$ for all $j \in [p]$. Let the tuning parameter $a$ in $P_\lambda$ satisfy $K < (b_u a)^{-1}$.

(A2) The vector of residuals $W \in \mathbb{R}^n$ such that $W := y - \mathbb{E}[y|X]$ is subgaussian($\sigma$) which means it satisfies that
   $\mathbb{P}[|\langle W, v\rangle| \geq t] \leq 2\exp(-t^2/2\sigma^2)$, for all $v : \|v\| = 1$ and any $t > 0$;

(A3) There exists a sequence $\{r_d \geq 0 : d = 1,2,...,p\}$ 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 $d : 1 \leq d \leq p$ and $\beta \in \mathbb{R}^p, \|\beta\|_0 \leq d$, it holds that $n^{-1}\|X\beta\|^2 \geq r_d \|\beta\|^2$.

**Remark 2.1.** Part (i) of (A1) states that our link function is both strongly convex and continuously differentiable; that is, the gradient being Lipschitz continuous. Many types of traditional

GLM problems satisfy this constraint including those for normal (linear regression), categorical (logistic regression), binomial, gamma and Poisson distributions, although in some cases the mild assumption on the boundedness of $\Theta$ has to be made. Even though the original domain of the link function can be unbounded, one may still consider its bounded subset given that it contains the vector of true parameters. Part (ii) of (A1) can be assumed without loss of generality by normalizing the design matrix columns.

**Remark 2.2.** (A2) is a common assumption in the literature, such as by [25] and [29].

**Remark 2.3.** Both (i) and (A2) are satisfied by a number of GLM setups, one example is linear regression. In such a setup, the response Y takes a gaussian distribution, while the gradient of the link function (encoding $\mathbb{E}[Y|X]$) is the identity. Note, it is difficult to have one without the other, since the loss formulation 1 is simply a log-likelihood maximization applied to a distribution within the class of exponential dispersion models [17]. Another classic example is logistic regression which is used for a Bernoulli or binomial distributed $Y$ along with a logit link function. It should be noted that we treat the matrix $X$ as fixed, so its generative distribution is not important to the analysis outside of whether it satisfies the assumptions and constraints mentioned. In the numerical experiments in Section 4, we use i.i.d. gaussian generation method since, as discussed for Definition 2.1, it means our design matrices will satisfy (A3) with high probability.

**Remark 2.4.** Assumption (A3) can be understood to be a lower bound on the eigenvalues for principal sub-matrices of $X^{\intercal}X$ of dimension $d \times d$ for all $d \in [p]$. For every $d : d \leq \tilde{p}^*$, the lower bounds are positive, meaning that the smallest eigenvalues of the $d \times d$ principal sub-matrices are assumed positive.

According to [20], Assumption (A3), for certain parameters, is provably a weaker condition than the restricted eigenvalue (RE) condition, as defined in Definition 2.1 below and first introduced by [4] as a plausible assumption to allow for the desired recovery quality of Lasso. The RE condition is a common assumption in the high-dimensional learning literature, such as [34] and [11].

**Definition 2.1.** (RE condition [34]) The matrix $X \in \mathbb{R}^{n\times p}$ is said to satisfy the RE condition if, for some $r_e > 0$, it holds that $\frac{1}{n}\|X\delta\|^2 \geq r_e\|\delta\|^2$ for all $\delta \in \bigcup_{|\hat{S}|=s}\mathfrak{C}(\hat{S})$ where $\mathfrak{C}(\hat{S}) := \{\delta := (\delta_i) \in \mathbb{R}^p : |\delta_{\hat{S}^c}| \leq 3|\delta_{\hat{S}}|\}, \delta_{\hat{S}^c} := (\delta_j : j \in \hat{S}^c)$, and $\delta_{\hat{S}} := (\delta_j : j \in \hat{S})$. Furthermore, the largest possible $r_e$ is said to be the restricted eigenvalue constant of $X$.

Random design matrices with with i.i.d. rows generated following subgaussian distributions as in (A2) have been shown to satisfy the RE condition with high probability [36] while proposition 1 in [21] includes a proof that with high probability, restricted strong convexity (RSC) is satisfied for a setup equivalent to our own. Note that satisfaction RSC implies the RE condition above. Thus, within our setup, (A3) is also satisfied with high probability for our setting.

2.2. **Preliminaries on S$^3$ONC.** Our results focus on the S$^3$ONC solutions, which has been formerly introduced by [20] in the special case of high-dimensional linear regression as a relaxation of the standard second-order KKT conditions. The definition of S$^3$ONC depends on the notion of first order necessary conditions (FONC) as below.

**Definition 2.2** (FONC)**.** A solution $\beta^*$ satisfies the first order necessary conditions (FONC) if

$$0 \in 1/n \sum_{i=1}^{n} \left[\psi'(x_i^\intercal \beta^*) - y_i\right] x_i + P'_\lambda(|\beta_j^*|)\partial(|\beta_j^*|), 1 \le j \le p,$$

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

**Definition 2.3** (S$^3$ONC)**.** A solution $\beta^*$ satisfies the significant subspace second-order necessary condition (S$^3$ONC) if it satisfies FONC and for all $j \in \{j : \beta_j^* \neq 0\}$,

$$\left.\frac{\partial^2 \mathscr{Q}(\beta)}{(\partial \beta_j)^2}\right|_{\beta=\beta^*} \ge 0$$

if the second derivative exists.

**Remark 2.5.** The S$^3$ONC can be intuited as the second order necessary condition applied only to the dimensions where $\beta_j \neq 0$, i.e., the significant dimensions. Since the S$^3$ONC is weaker than the standard second-order KKT conditions, any algorithm that guarantees the second-order KKT conditions can be used to obtain an S$^3$ONC solution, by requiring a more stringent optimality condition, may be slower than necessary. One specifically S$^3$ONC guaranteeing approach, presented in [18], utilizes an interior point trust region algorithm in order to guarantee an S$^3$ONC solution in polynomial time. This is the scheme which will be used later in Section 4.

## 3. Main Results

We now present our theoretical results for global optimization of FCP penalized GLMs. All proofs can be found in the appendix. We will make use of a short-hand notation:

$$\beta^{Lasso} \in \arg\min \mathscr{L}(\beta) + \lambda|\beta|. \tag{3.1}$$

**Theorem 3.1.** *Suppose assumptions (A1), (A2), and (A3) with any $\tilde{p}^* : \tilde{p}^* \ge 2|\mathscr{S}|$. Let $\beta^*$ be an arbitrary S$^3$ONC solution to (1.1) with $P_\lambda$ specified as the MCP. Assume that $\mathscr{Q}(\beta^*) \le \mathscr{Q}(\beta^{true}) + \Gamma$ for an arbitrary $\Gamma \ge 0$. (i) Let the sub-optimality gap satisfy $\Gamma < P_\lambda(a\lambda) - \frac{\sigma^2}{b_l n}(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t)$; (ii) Choose $P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1 + 2\sqrt{t'} + 2t') + \frac{\frac{\sigma^2}{n}|\mathscr{S}|(1+2\sqrt{t'}+2t')+\Gamma b_l}{b_l(\tilde{p}^*-2|\mathscr{S}|+1)}$, and (iii) Assume that the minimal signal strength satisfy*

$$\|\beta_{\mathscr{S}}^{true}\|_{\min} > \sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) + \frac{8}{r_{\tilde{p}} b_l}\min\left\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathscr{S}|, P_\lambda(a\lambda)|\mathscr{S}| + \Gamma\right\}}.$$

*Then the following two statements hold:*

(a) *$\beta^*$ is an oracle solution with probability at least* $1 - \exp\left(-t + \tilde{p}^* \ln\left(\frac{pe}{\tilde{p}^*}\right)\right)$ $- \exp\left(-\left(\tilde{p}^* + 1\right)\left(t' - \ln p\right)\right) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}$.

(b) *$\beta^*$ is both an oracle solution and an globally optimal solution to (1.1) with probability* $1 - 2\exp\left(-t + \tilde{p}^* \ln\left(\frac{pe}{\tilde{p}^*}\right)\right) - 2\exp\left(-\left(\tilde{p}^* + 1\right)\left(t' - \ln p\right)\right) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}$.

**Remark 3.1.** Theorem 3.1 (especially in the second statement) is perhaps the first result that establishes a set of conditions for any S$^3$ONC solution to be globally optimal with high probability. Further, this result is algorithm independent which allows for greater flexibility compared to most existing results as in [23] and [9] which rely on a specific algorithm choice.

**Remark 3.2.** The second part follows quite easily from the first due to the uniqueness of $\beta^{oracle}$ as well as the fact that $\beta^{opt}$ must also be an S$^3$ONC solution. Thus by applying the first part of the Theorem to $\beta^{opt}$ we are able to show that both our arbitrary $\beta^*$ and $\beta^{opt}$ coincide with the unique $\beta^{oracle}$.

**Remark 3.3.** The above constraints on $\Gamma$, $P_\lambda(a\lambda)$ and $\left\|\beta^{true}_{\mathscr{S}}\right\|_{min}$ may initially seem disparate but can all be converted to constraints on the sample size $n$ as is shown in Corollary 3.1 below. This is possible because $\Gamma$ can be bounded by some function of $n^{-\gamma}$ for some $\gamma > 0$. Given that, it can be seen that the lesser side of inequalities (i),(ii) and (iii) go to 0 as $n$ grows. Further discussion of how this is achieved for Corollary 3.1 can be found in Remark 3.7.

**Corollary 3.1.** *Assume* $\ln p \geq 1$, $b_l \leq 1$, *and* $s \geq 1$. *Let* $\beta^*$ *be an S$^3$ONC solution to* (1.1). *Let assumptions (A1), (A2), and the RE condition as defined in Definition 2.1 hold. Assume that* $\mathscr{Q}(\beta^*) \leq \mathscr{Q}(\beta^{Lasso})$ *almost surely, where* $\beta^{Lasso}$ *is the optimal solution to the Lasso problem with penalty coefficient* $\lambda^{Lasso} = \sigma\sqrt{\frac{lnp}{n^{1-\gamma/2}}}$, *where* $\gamma \in [0,1]$ *is an arbitrary scalar. Let* $\lambda = \frac{\sigma}{r_e}\sqrt{\frac{\ln p}{n^{\gamma/2}}}$ *and* $a \in [0.8, 1)$. *There exist problem-independent constants* $C_1 > 0$, $C_2 > 0$ *and* $C_3 > 0$ *such that if*

$$n > \max\left\{\frac{C_1}{b_l}, \left[C_2\frac{s}{b_l}\right]^{\frac{2}{1-\gamma}}, \left[C_3\frac{s\sigma^2\ln p}{\|\beta^{true}_{\mathscr{S}}\|^2_{min}b_l^2 r_e^4}\right]^{2/\gamma}\right\},$$

*then* $\beta^*$ *is the global solution to 1.1 with probability at least* $1 - C_4\exp\left(-C_5 s n^{\gamma/2}\ln p\right) - C_6\exp\left(-C_7 b_u n^{\gamma/2}\ln(p)\right)$ *for problem independent constants* $C_4, C_5, C_6$, *and* $C_7$.

**Remark 3.4.** Corollary 3.1 indicates that for $\gamma > 0$, the global optimal solution coincides with computable S$^3$ONC solution with overwhelming probability given that the sample size meets certain requirements. It should specifically be noted that the relationship between $n$ and $p$ require only $\frac{\ln p}{n^{\gamma/2}} = O(1)$, which ensures the applicability to the high-dimensional setting even if $n \ll p$.

**Remark 3.5.** Liu and Ye [18] derived a gradient-based algorithm that provably ensures an S$^3$ONC solution at pseudo-polynomial-time complexity. When $n$ is properly large, this pseudo-polynomial-time algorithm enables a straightforward design of an FPRAS for generating the global optimal solution as follows.

---

**FPRAS:** A pseudo-polynomial-time algorithm that generates global optimum at high probability

---

**Step 1:** Initialize the parameters $\delta, \lambda, a, \hat{a}, k = 0$ and $\beta^{Lasso}$ by solving (3.1).
**Step 2:** If Case 1: $|\beta_j^k| \in (0, a\lambda)$ for some $j = 1, ..., p$, then choose an arbitrary $\iota \in \{j : |\beta^k| \in (0, a\lambda)\}$ and solve

$$\beta_\iota^{k+1} \in \arg\min_\beta [\nabla\mathscr{L}(\beta^k)]_\iota \cdot \beta + P_\lambda(|\beta|)$$

$$s.t. \quad (\beta - \beta_\iota^k)^2 \leq \delta^2$$

and let $\beta_j^{k+1} = \beta_j^k$, for all $j \neq \iota$. Go to Step 3.

Else Case 2: If $|\beta_j^k| \notin (0, a\lambda)$ for all $j = 1, ..., p$ then for all $j = 1, ..., p$:

— If $\beta_j^k = 0$ then $\beta_j^{k+1} = \hat{a} \cdot \left[|[\nabla\mathscr{L}(\beta^k)]_j| - \lambda\right]_+ \cdot sign(-[\nabla\mathscr{L}(\beta^k)]_j)$.

— If $|\beta_j^k| \geq a\lambda$, then $\beta_j^{k+1} = \beta_j^k - \hat{a} \cdot [\nabla\mathscr{L}(\beta^k)]_j$. Go to Step 3.

**Step 3:** Algorithm stops if $|\beta_j^k| \notin (0, a\lambda)$ and $\left\|\beta_j^k - \beta_j^{k+1}\right\| < \delta$. Otherwise, let $k := k+1$ and go to Step 2.

---

**Remark 3.6.** Here, the above algorithm has iteration complexity of

$$\mathscr{O}\Big(\left(\mathscr{Q}\left(\beta^{Lasso}\right) - \mathscr{Q}\left(\beta^{opt}\right)\right) \cdot \max\left\{(1/(2a) - b_u/2)^{-1}, 2b_u^{-1}, (1/a - b_u/2)^{-1}\right\} \cdot 1/\delta^2\Big)$$

for any $\gamma$-accuracy S$^3$ONC solution. In this iteration complexity, all the quantities are verifiably upper bounded by a polynomial function of dimensionality $p$ and the desired accuracy $1/\gamma$. Furthermore, $\beta^{Lasso}$ is a solution to a convex problem, which be generated within polynomial time and the per–iteration problem admits a closed form, whose complexity is strongly in polynomial-time. Therefore, this algorithm is an FPRAS in generating an S$^3$ONC (global) solution.

**Remark 3.7.** We are able to remove $\Gamma$ from the result by bounding the performance difference between $\beta^{true}$ and $\beta^{lasso}$ using similar techniques as in [4]. In order to use this bound for our S$^3$ONC solution, we require that $\mathscr{Q}(\beta^*) \leq \mathscr{Q}(\beta^{Lasso})$. However, this can generally be obtained by initializing any S$^3$ONC guaranteeing algorithm with $\beta^{Lasso}$ in a similar fashion to [11] for LLA. The FPRAS above follows the same initialization scheme.

**Remark 3.8.** The above specification of values for $a, \lambda$ and $\lambda^{Lasso}$ can be thought of as examples rather than strict requirements. A closer examination of the proof for Corollary 3.1 will reveal that the values for $\lambda$ and $\lambda^{Lasso}$ can be chosen in a much more flexible fashion, though the corresponding values of $C_1$ through $C_7$ may be different for different combinations of $\lambda$ and $\lambda^{Lasso}$.

The techniques used in the proof of Theorem 3.1 can be used to provide insights into other optimization schemes. As an example, we can apply the same analysis to the state-of-the-art FCP-based algorithm, LLA, using the framework in [11] as a starting point.

---

**LLA:** local linear approximation.

---

**Step 1.:** Set $k = 0$. Initialize the algorithm with $\beta^0 = \beta^{Lasso}$, where $\beta^{Lasso}$ is generated by solving (3.1). Let $N$ be the maximal iteration number.

**Step 2.:** For all $k = 1, ..., N$, solve the following convex program to generate $\beta^{k+1}$:

$$\beta^{k+1} \in \arg\min_{\beta} \mathscr{L}(\beta) + \sum_{j\in[p]} P'_\lambda(|\beta_j^k|) \cdot |\beta_j|,$$

where $P'_\lambda$ is the first derivative of $P_\lambda$. Let $k := k+1$.

---

We can show that in fact the LLA is another FPRAS that achieves the global optimal solution. The proof of this can be found in the appendix.

**Corollary 3.2.** *For problem* (1.1). *If the RE condition in Definition 2.1 holds,* $\left\|\beta_{\mathscr{S}}^{true}\right\|_{min} >$ $(a+1)\lambda > (a+1)\max\left\{\frac{3\lambda^{Lasso}s^{1/2}}{b_l r_e}, \frac{4\sigma\sqrt{s+2\sqrt{st_1}+2t_1}}{b_l(an/b_u)^{1/2}}, \frac{2\sigma\sqrt{s+2\sqrt{st_2}+2t_2}}{b_l\sqrt{nr_e}}\right\}$, *then*

(a) *The LLA algorithm initialized with $\beta^{Lasso}$ converges to the oracle solution in two iterations with probability $1-\phi_0-\phi_1-\phi_2$, where*

$$\phi_0 := \mathbb{P}\left(\left\|\beta^{Lasso}-\beta^{true}\right\|_{max} > \lambda\right) \leq 2p\exp\left(\frac{-\left(\lambda^{Lasso}\right)^2 nb_u a}{8\sigma^2}\right),$$

$$\phi_1 := \mathbb{P}\left(\left\|\nabla_{S^c_{\tilde{p}}}\ell_n(\beta^{oracle})\right\|_{max} \geq \lambda\right) \leq \left(\tfrac{pe}{s}\right)^s \exp(-t_1)+2\exp\left(\tfrac{-\lambda^2 ab_u n}{8\sigma^2}\right),$$

$$\phi_2 := \mathbb{P}\left(\left\|\beta^{oracle}_{\mathscr{S}}\right\|_{\min} \leq a\lambda\right) \leq \left(\tfrac{pe}{s}\right)^s \exp(-t_2).$$

(b) *If in addition (A1) and (A2) holds, while the parameters of $(a,\lambda)$ satisfy that $P_\lambda(a\lambda) > \max\left\{\frac{\sigma^2}{2nb_l}(1+2\sqrt{t_4}+2t_4)+\frac{\frac{\sigma^2}{n}|\mathscr{S}|(1+2\sqrt{t_4}+2t_4)b_l}{b_l(\tilde{p}^*-2|\mathscr{S}|+1)}, \frac{\sigma^2}{b_l n}\left(\tilde{p}^*+2\sqrt{\tilde{p}^* t_3}+2t_3\right)\right\}$ and $\left\|\beta^{true}_{\mathscr{S}}\right\|_{\min} > \sqrt{\frac{8\sigma^2}{r_{\tilde{p}}b_l^2 n}\left(\tilde{p}^*+2\sqrt{\tilde{p}^*t_3}+2t_3\right)+\frac{8}{r_{\tilde{p}}b_l}\min\left\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathscr{S}|, P_\lambda(a\lambda)|\mathscr{S}|\right\}}$ then the LLA algorithm initialized by $\beta^{Lasso}$ converges to the global solution in two iterations with probability at least $1-\phi_0-\phi_1-\phi_2-\phi_3$, where*

$$\phi_3 := \mathbb{P}(\beta^{oracle} \neq \beta^{opt})$$
$$\leq \exp\left(-t_3+\tilde{p}^*\ln\left(\frac{pe}{\tilde{p}^*}\right)\right)+\exp\left(-\left(\tilde{p}^*+1\right)\left(t_4-\ln p\right)\right)\cdot\frac{1-\exp\left(-\left(p-\tilde{p}^*\right)\left(t_4-\ln p\right)\right)}{1-\exp\left(-t_4+\ln p\right)},$$

*and $t_1,t_2,t_3,t_4>0$ are arbitrary constants.*

**Remark 3.9.** Since each iteration of the LLA solves a convex program, which can be done within polynomial-time. When $n$ is properly large, the above theorem then indicates that the LLA is another FPRAS in globally optimizing the FCP-based nonconvex formulation.

## 4. Numerical Experiments

4.1. **Experimental setup.** We focus our tests on sparse logistic regression. Our problem and data are implemented in a similar way as [11]. We construct $\beta^{true}$ as below: Firstly, $\beta^{true}_{\mathscr{S}}$ is constructed randomly by choosing 10 elements of $\beta$ and choosing the magnitude of each to be a uniform value within $[1,2]$. Each value is chosen to be negative with probability 0.5. Then, the remaining entries $\beta^{true}_{\mathscr{S}^c}$ are set to be 0. The design matrix $X\in\mathbb{R}^{n\times p}$ is constructed by generating $n$ iterations of $x_i\sim N_p(0,\Sigma)$ where $\Sigma=(0.5^{|j-j'|})_{p\times p}$. We then generate Y using a Bernoulli distribution where $\mathbb{P}(y_i=1)=(1+e^{-x_i^\intercal\beta^{true}})^{-1}$. With this data, we train a logistic regression model by invoking Algorithm 1 in solving (1.1) with MCP for S$^3$ONC solutions initialized with Lasso implemented in Python 3. The tuning parameters $\lambda$ and $a$ are obtained by cross validation following [11].

We would like to ascertain whether our FCP classifier, obtained using $S^3ONC$ methods, is actually the global optimal solution. We do this by taking each element of the FCP classifier and perturbing it to find a new potential solution. Each element's perturbation is independent and generated by a $N(0,1/p^{1/2})$-random variable. We then check if this perturbed classifier has better FCP regularized performance on the training data than the FCP classifier. If not, we repeat until either a better solution is found, or until 2000 perturbations have been tried.

Additionally, we compare our solution's statistical performance to those of other popular regularization methods. Using the data generation method above, we obtain two sets of data, both with 100 samples. One set is for training the model, and the other is the test set for out-of-sample tests. We repeat the above process for 100 times to generate 100 training-and-test instances, each with 100 samples. We compare those trained using the method described above with Lasso solutions generated by the global minimizer to (3.1) and an estimator generated by solving (1.1) when $P_\lambda$ is substantiated by an $\ell_2$ penalty. The Lasso and $\ell_2$ classifiers are solved using the *scikit learn* python library.

We compare the above estimators in terms of statistical performance for both $\ell_1$ loss: $|\beta^* - \beta^{true}|$ and $\ell_2$ loss: $\|\beta^* - \beta^{true}\|$.

TABLE 1. Percent of time FCP beat all perturbations

| | $n = 100$ $p = 500$ | $n = 100$ $p = 1000$ | $n = 100$ $p = 1500$ | $n = 100$ $p = 2000$ |
|---|---|---|---|---|
| **% Best FCP** | 100% | 100% | 100% | 100% |

TABLE 2. Statistical performance of the four classifiers.

| | | $n = 100,\ p = 1000$ | | $n = 100,\ p = 1500$ | | $n = 100,\ p = 2000$ | |
|---|---|---|---|---|---|---|---|
| **Classifier** | **Measure** | **Mean** | **Std. dev** | **Mean** | **Std. dev** | **Mean** | **Std. dev** |
| MCP | $\ell_1$ loss | 13.909907 | 1.471911 | 14.818059 | 1.698191 | 14.506226 | 1.480686 |
| | $\ell_2$ loss | 4.108019 | 0.320061 | 4.304993 | 0.374453 | 4.489184 | 0.399441 |
| Lasso | $\ell_1$ loss | 15.015975 | 1.039529 | 15.882654 | 1.29422 | 17.079414 | 1.545309 |
| | $\ell_2$ loss | 4.3255 | 0.25996 | 4.397969 | 0.326336 | 4.433467 | 0.362707 |
| $\ell_2$ penalty | $\ell_1$ loss | 22.211963 | 0.791955 | 26.026067 | 0.966091 | 28.485075 | 0.993699 |
| | $\ell_2$ loss | 4.734209 | 0.241683 | 4.738025 | 0.296726 | 4.755959 | 0.296746 |

4.2. **Numerical results.** Table 1 contains the numbers from optimality analysis. This technique did not yield a single perturbed solution that could beat the FCP classifier obtained from the FPRAS in any of our thousands of iterations.

Table 2 shows the numerical results for the statistical performance measurements. We show the two performance measures for each of the three classifiers for tphree different problem types.

As expected, the FCP classifier generally outperformed the lasso and $\ell_2$ classifiers. The margins are fairly thin between FCP and lasso, especially compared to the standard deviation. Other values of $n$ and $p$ were tried but the results generally followed the same pattern.

As a result we tentatively conclude that our numerical results align with our theoretical results though further testing of the global optimality probability would be valuable.

## 5. CONCLUSIONS

This paper investigated both the theoretical and empirical performance of FPRAS's on MCP regularized GLMs. Despite such a problem being strongly NP-Hard, we demonstrated two FPRAS schemes that achieve global optimality. To our knowledge this is the first probability bound for global optimization of FCP regularized GLMs using an FPRAS. Further, the same technique can be used to extend other results in order to obtain global optimization bounds for a wide variety of problems. While this paper focuses on GLMs, further exploration will focus on the question whether similar results can be found for more general problem classes under weaker assumptions. High-dimensional M-estimation problems could potentially be a future avenue of investigation.

**Acknowledgments**
On this very special occasion of Professor Panos Pardalos's 70th birthday, we wish to extend our heartfelt best wishes to him. His immense contributions to the field of optimization have been continuously inspiring all of us. Happy birthday, Professor Pardalos! This work was partially supported by NSF CMMI 2016571 and NSF CMMI 2213459.

## REFERENCES

[1] Z. Y. Algamal, M.H. Lee, Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification, Expert Systems with Applications 42 (2015), 9326–9332.
[2] A. Bagaria, V. Jaravine, Y. J. Huang, G. T. Montelione, P. Güntert, Protein structure validation by generalized linear model root-mean-square deviation prediction, Protein Science 21 (2012), 229–238.
[3] W. Bian, X. Chen, Optimality conditions and complexity for non-Lipschitz constrained optimization problems, Preprint, Corpus ID: 18376023, 2014.
[4] P. J. Bickel, Y. Ritov, A.B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist. 37 (2009), 1705–1732.
[5] X. Chen, D. Ge, Z. Wang, Y. Ye, Complexity of unconstrained L2-Lp minimization, Mathematical Programming 143 (2014), 371–383.
[6] Y. Chen, D. Ge, M. Wang, Z. Wang, Y. Ye, H. Yin, Strong NP-hardness for sparse optimization with concave penalty functions, In: Proceedings of the 34th International Conference on Machine Learning 70 (2017), 740–747.
[7] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001), 1348–1360.
[8] J. Fan, R. Li, Statistical challenges with high dimensionality: Feature selection in knowledge discovery, arXiv preprint math/0602133 (2006).
[9] J. Fan, J. Lv, Nonconcave penalized likelihood with NP-dimensionality, IEEE Trans. Info. Theory 57 (2011), 5467–5484.
[10] J. Fan, F. Han, H. Liu, Challenges of big data analysis, Natl. Sci. Rev. 1 (2014), 293–314.
[11] J. Fan, L. Xue, H. Zou, Strong oracle optimality of folded concave penalized estimation, Ann. Statist. 42 (2014), 819.
[12] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (2008), 432–441.
[13] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010), 1.
[14] D. Hsu, S. Kakade, T. Zhang, A tail inequality for quadratic forms of subgaussian random vectors, Electronic Communications in Probability 17 (2012).
[15] H.-H. Huang, X.-Y. Liu, Y. Liang, Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2+ 2 regularization, PloS One 11 (2016), e0149675.
[16] X. Huo, J. Chen, Complexity of penalized likelihood estimation, J. Stat. Comput. Simul. 80 (2010), 747–759.
[17] B. Jørgensen, Exponential dispersion models, J. R. Stat. Soc. B 49 (1987), 127–145.
[18] H. Liu, Y. Ye, High-Dimensional Learning under Approximate Sparsity: A Unifying Framework for Non-smooth Learning and Regularized Neural Networks, arXiv preprint arXiv:1903.00616 (2019).
[19] H. Liu, T. Yao, R. Li, Global solutions to folded concave penalized nonconvex learning, Ann. Statist. 44 (2016), 629.

[20] H. Liu, T. Yao, R. Li, Y. Ye, Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory for local solutions, Math. Program. 166 (2017), 207–240.
[21] P.-L. Loh, M.J. Wainwright, Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima, Adv. Neural Inf. Process. Syst. 2013 (2013), 476–484.
[22] P.-L. Loh, M.J. Wainwright, et al., Support recovery without incoherence: A case for nonconvex regularization, Ann. Statist. 45 (2017), 2455–2482.
[23] P.-L. Loh, et al., Statistical consistency and asymptotic normality for high-dimensional robust *m*-estimators, Ann. Statist. 45 (2017), 866–896.
[24] R. Mazumder, J.H. Friedman, T. Hastie, Sparsenet: Coordinate descent with nonconvex penalties, J. Amer. Statist. Assoc. 106 (2011), 1125–1138.
[25] S. N. Negahban, P. Ravikumar, M.J. Wainwright, B. Yu, et al., A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizers, Stat. Sci. 27 (2012), 538–557.
[26] D. D. Palmer, M. Ostendorf, Improving information extraction by modeling errors in speech recognizer output, Proc. 1st Int. Conf. Hum. Lang. Tech. Res. (2001), 1–5. https://doi.org/10.3115/1072133.1072186
[27] G. Paul, J. Cardinale, I.F. Sbalzarini, Coupling image restoration and segmentation: a generalized linear model/Bregman perspective, Int. J. Comput. Vision 104 (2013), 69–93.
[28] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B 58 (1996), 267–288.
[29] L. Wang, Y. Kim, R. Li, Calibrating non-convex penalized regression in ultra-high dimension, Ann. Statist. 41 (2013), 2505.
[30] Z. Wang, H. Liu, T. Zhang, Optimal computational and statistical rates of convergence for sparse nonconvex learning problems, Ann. Statist. 42 (2014), 2164.
[31] E. P. Xing, M. I. Jordan, R. M. Karp, Feature selection for high-dimensional genomic microarray data, In: Proceedings of the Eighteenth International Conference on Machine Learning 1 (2001), pp. 601–608.
[32] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2010), 894–942.
[33] C.-H. Zhang, T. Zhang, A general theory of concave regularization for high-dimensional sparse estimation problems, Stat. Sci. 27 (2012), 576–593.
[34] Y. Zhang, M. J. Wainwright, M. I. Jordan, Lower bounds on the performance of polynomial-time algorithms for sparse linear regression, In: JMLR: Workshop and Conference Proceedings 35 (2014), 921–948.
[35] P. Zhao, B. Yu, On model selection consistency of Lasso, J. Mach. Learn. Res. 7 (2006), 2541–2563.
[36] S. Zhou, Restricted eigenvalue conditions on subgaussian random matrices, arXiv preprint arXiv:0912.4045 (2009).
[37] H. Zou, R. Li, One-step sparse estimates in nonconcave penalized likelihood models, Ann. Statist. 36 (2008), 1509.

## Appendix A. Appendix

The Appendix is organized as below: Section A.1 presents the proofs for the main results, Sections A.2 and A.3 present central lemmata to be useful in Section A.1.

### A.1. Proof of main results.

A useful relationship in our proofs is that, for an S$^3$ONC solution $\beta^*$ within $\{\beta^* : \mathscr{Q}(\beta^*) \leq \mathscr{Q}(\beta^{true}) + \Gamma\}$ for any $\Gamma \geq 0$, we have the following useful inequality under Assumption (A1):

$$\frac{b_l}{2n}\|X\delta^*\|^2 - \frac{1}{n}W^\intercal X\delta^* + \sum_{j\in\mathscr{S}} P_\lambda(|\beta_j^*|) \leq \sum_{j\in\mathscr{S}} P_\lambda(|\beta_j^{true}|) + \Gamma,$$

where $\delta^* = \beta^* - \beta^{true}$. This is obtained by invoking the strong convexity of $\psi$, which leads to $\psi(x_i^\intercal\beta^*) \geq \psi(x_i^\intercal\beta^{true}) + \psi'(x_i^\intercal\beta^{true})(x_i^\intercal\beta^* - x_i^\intercal\beta^{true}) + 0.5\cdot b_l(x_i^\intercal\beta^* - x_i^\intercal\beta^{true})^2$.

*Proof of Theorem 3.1.* First, given our assumption that (A1) holds, that is, (i) $\tilde{p}^* \geq 2|\mathscr{S}|$, (ii) $\beta^*$ is $S^3ONC$ satisfying $\mathscr{Q}(\beta^*) \leq \mathscr{Q}(\beta^{true}) + \Gamma$ for some $\Gamma \geq 0$, and (iii)

$$P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1 + 2\sqrt{t'} + 2t') + \frac{\frac{\sigma^2}{n}|\mathscr{S}|(1 + 2\sqrt{t'} + 2t') + \Gamma b_l}{b_l(\tilde{p}^* + 1 - 2|\mathscr{S}|)},$$

we can apply Lemma A.5 with $\tilde{p} = \tilde{p}^*$. This means that $\|\beta^* - \beta^{true}\|_0 \leq \tilde{p}^*$ with probability at least

$$1 - \exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \frac{1 - \exp(-(p - \tilde{p}^*)(t' - \ln p))}{1 - \exp(-t' + \ln p)}.$$

In view of the additional assumption that (A3) holds, we can apply the second part of Lemma A.4 with $\tilde{p} = \tilde{p}^*$ to obtain that, for any $t > 0$,

$$\begin{aligned}&\frac{1}{n}\left\|X(\beta^* - \beta^{true})\right\|^2 \\ &\leq \frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) + \frac{8}{b_l}\min\{\lambda^2(|\mathscr{S}| - \|\beta^*\|_0)r_{\tilde{p}^*}^{-1}, P_\lambda(a\lambda)\cdot(|\mathscr{S}| - \|\beta^*\|_0) + \Gamma\}\end{aligned}$$

holds with probability at least $1 - \exp(-t + \tilde{p}^*\ln(\frac{pe}{\tilde{p}^*}))$. Given that for 2 arbitrary sets $A$ and $B$,

$$\begin{aligned}\mathbb{P}(A\cap B) = \mathbb{P}(B)\mathbb{P}(A|B) &= (1 - \mathbb{P}(B^c))(1 - \mathbb{P}(A^c|B)) \\ &= 1 - \mathbb{P}(A^c|B) - \mathbb{P}(B^c) + \mathbb{P}(B^c)\mathbb{P}(A^c|B) \\ &= 1 - \mathbb{P}(A^c|B) - \mathbb{P}(B^c)(1 - \mathbb{P}(A^c|B)) \geq 1 - \mathbb{P}(A^c|B) - \mathbb{P}(B^c).\end{aligned}$$

Thus they hold simultaneous with probability at least

$$1 - \exp(-t + \tilde{p}\ln(\frac{pe}{\tilde{p}^*})) - \exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \frac{1 - \exp(-(p - \tilde{p}^*)(t' - \ln p))}{1 - \exp(-t' + \ln p)}.$$

The same sequence of arguments can be used to show that $\beta^{opt}$ also satisfies $\|\beta^{opt} - \beta^{true}\| \leq \tilde{p}^*$ and

$$\begin{aligned}&\frac{1}{n}\left\|X(\beta^{opt} - \beta^{true})\right\|^2 \\ &\leq \frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) + \frac{8}{b_l}\min\left\{\lambda^2(|\mathscr{S}| - \|\beta^*\|_0)r_{\tilde{p}^*}^{-1}, P_\lambda(a\lambda)\cdot(|\mathscr{S}| - \|\beta^*\|_0) + \Gamma\right\}\end{aligned}$$

with the same probability. Using again the union bound and De Morgan's law, we say that $\beta^*$ and $\beta^{opt}$ satisfy the above conditions simultaneously with probability

$$1-2\exp(-t+\tilde{p}^*\ln(\frac{pe}{\tilde{p}^*}))-2\exp(-(\tilde{p}^*+1)(t'-\ln p))\cdot\frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}.$$

With this, our $\Gamma$ assumption, and our minimal signal strength assumption, we can apply Lemma A.6 to show that $\beta^*=\beta^{opt}$ with probability at least

$$1-2\exp(-t+\tilde{p}^*\ln(\frac{pe}{\tilde{p}^*}))-2\exp(-(\tilde{p}^*+1)(t'-\ln p))\cdot\frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}.$$

$\square$

*Proof of Corollary 3.1.* First we need to bound $\Gamma$. In order to do this, we use the lasso problem

$$\mathscr{Q}^{lasso}(\beta)=\sum_{i\in\mathscr{N}}\ell(\beta,x_i,y_i)+\sum_{j\in\mathscr{P}}\lambda^{lasso}|\beta_j|$$

as well as the concavity of MCP over positive values to obtain the following 2 inequalities

$$\begin{aligned}&\mathscr{Q}^{lasso}(\beta^{lasso})\le\mathscr{Q}^{lasso}(\beta^{true})\\&\sum_{i\in\mathscr{N}}\ell(\beta_j^{lasso},x_i,y_i)-\ell(\beta_j^{true},x_i,y_i)\le\sum_{j\in\mathscr{P}}\lambda^{lasso}(|\beta_j^{true}|-|\beta_j^{lasso}|)\\&\le\sum_{j\in\mathscr{P}}\lambda^{lasso}|\beta_j^{lasso}-\beta_j^{true}|\end{aligned}$$

and

$$\sum_{j\in\mathscr{P}}P_\lambda(\beta_j^{true})-\sum_{j\in\mathscr{P}}P_\lambda(\beta_j^{lasso})\le\sum_{j\in\mathscr{P}}P'_\lambda(\beta_j^{lasso})(|\beta_j^{true}|-|\beta_j^{lasso}|)\le\sum_{j\in\mathscr{P}}\lambda|\beta_j^{lasso}-\beta_j^{true}|.$$

We also need 2 results from the proof for $\phi_0$ in Corollary 3.2, which shows that both $|\delta^\ell_{\mathscr{S}_c}|\le 3|\delta^\ell_{\mathscr{S}}|$ and $\frac{b_l}{n}\left\|X\delta^\ell\right\|^2\le 3\lambda^{lasso}|\delta^\ell_{\mathscr{S}}|$ are conditional on $\mathscr{A}$, where $\delta^\ell=\beta^{lasso}-\beta^{true}$. Given our restricted eigenvalue assumption $\frac{\|X\delta^\ell\|^2}{n\|\delta^\ell\|^2}\ge r_e$, this can be used to show

$$\begin{aligned}|\delta^\ell|\le 4|\delta^\ell_{\mathscr{S}}|\le 4\sqrt{s}\frac{\|\delta^\ell_{\mathscr{S}}\|^2}{\|\delta^\ell_{\mathscr{S}}\|}\le 4\sqrt{s}\frac{\|\delta^\ell\|^2}{\|\delta^\ell_{\mathscr{S}}\|}\\ \le\frac{4\sqrt{s}}{r_en}\frac{\|X\delta^\ell\|^2}{\|\delta^\ell_{\mathscr{S}}\|}\le\frac{4\sqrt{s}}{r_e}\frac{3\lambda^{lasso}|\delta^\ell_{\mathscr{S}}|}{b_l\|\delta^\ell_{\mathscr{S}}\|}\le\frac{4\sqrt{s}}{r_e}\frac{3\lambda^{lasso}\sqrt{s}\|\delta^\ell_{\mathscr{S}}\|}{b_l\|\delta^\ell_{\mathscr{S}}\|},\end{aligned}$$

which means $|\delta^\ell|\le\frac{12\lambda^{lasso}s}{b_lr_e}$ with conditional on $\mathscr{A}$ which occurs with probability at least $1-2p\exp(\frac{-(\lambda^{lasso})^2nb_ua}{8\sigma^2})$.

Finally, we are able to bound gamma by combining the above

$$\begin{aligned}\Gamma &\leq \mathscr{Q}(\beta^*) - \mathscr{Q}(\beta^{true}) \leq \mathscr{Q}(\beta^{lasso}) - \mathscr{Q}(\beta^{true}) \\ &\leq \sum_{i\in\mathscr{N}} \ell(\beta_j^{lasso}, x_i, y_i) - \sum_{j\in\mathscr{P}} P_\lambda(\beta_j^{lasso}) - [\sum_{i\in\mathscr{N}} \ell(\beta_j^{true}, x_i, y_i) - \sum_{j\in\mathscr{P}} P_\lambda(\beta_j^{true})] \\ &\leq \sum_{j\in\mathscr{P}} (\lambda^{lasso}|\beta_j^{lasso} - \beta_j^{true}| + \lambda|\beta_j^{lasso} - \beta_j^{true}|) \\ &\leq (\lambda^{lasso} + \lambda)|\delta^\ell| \leq (\lambda^{lasso} + \lambda)\frac{12\lambda^{lasso}s}{b_l r_e}. \end{aligned} \tag{A.1}$$

Next, we consider the conditions necessary to apply Theorem 3.1. We have assumptions (A1), (A2), and (A3) per our assumption that the RE condition holds combined with A.7. That leaves the 3 requirements on $\Gamma$, $P_\lambda(a\lambda)$, and $\|\beta_{\mathscr{S}}^{true}\|_{min}$. We will convert each of these to inequalities on $n$. Utilizing (A.1) and substituting $\lambda = \frac{Q\sigma}{r_e}\sqrt{\frac{\ln p}{n^{\gamma/2}}}$ and $\lambda^{lasso} = \varepsilon\sigma\sqrt{\frac{\ln p}{n^{1-\gamma/2}}}$, where $Q, \varepsilon > 0$ are arbitrary constants, and setting $\tilde{p}^* = 4s$, $t = \tilde{p}^* n^{\gamma/2}\ln p$, $t' = n^{\gamma/2}\ln p$, we obtain

$$P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1 + 2\sqrt{t'} + 2t') + \frac{\frac{\sigma^2}{2}s(1 + 2\sqrt{t'} + 2t') + \Gamma b_l}{b_l(\tilde{p}^* - 2s + 1)} n > \frac{8 + 12\varepsilon^2 + 12\varepsilon Q}{b_l a Q^2} = C_1/b_l,$$

$$\Gamma < P_\lambda(a\lambda) - \frac{\sigma^2}{b_l n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) n > \left[\left(\frac{12\varepsilon}{aQ} + \sqrt{\frac{20 + 12\varepsilon^2}{aQ^2}}\right)\frac{s}{b_l}\right]^{\frac{2}{1-\gamma}} = \left[C_2\frac{s}{b_l}\right]^{\frac{2}{1-\gamma}},$$

$$\|\beta_{\mathscr{S}}^{true}\|_{\min} > \sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) + \frac{8}{r_{\tilde{p}} b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathscr{S}|, P_\lambda(a\lambda)|\mathscr{S}| + \Gamma\}},$$

and

$$n > \left[(160 + 8Q^2)\frac{s\sigma^2\ln p}{(\|\beta_{\mathscr{S}}^{true}\|_{min} r_{4s} b_l r_e)^2}\right]^{2/\gamma} = \left[C_3\frac{s\sigma^2\ln p}{(\|\beta_{\mathscr{S}}^{true}\|_{min} r_{4s} b_l r_e)^2}\right]^{2/\gamma}$$

for some constants $C_1, C_2$ and $C_3 > 0$. We can then apply Theorem 3.1 (conditional on $\mathscr{A}$). We substitute our values and simplify to obtain that $\beta^*$ is the global solution with probability at least

$$\begin{aligned} &1 - 2\exp(-t + \tilde{p}^*\ln(\frac{pe}{\tilde{p}^*})) - 2\exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \left[\frac{1 - \exp(-(p - \tilde{p}^*)(t' - \ln p))}{1 - \exp(-t' + \ln p)}\right] \\ \geq &1 - 2\exp(-(n^{\gamma/2} - 1)4s\ln p) - 2\left[\sum_{k=1}^{p-\tilde{p}^*}\exp(-(\tilde{p}^* + k)(n^{\gamma/2} - 1)\ln p)\right] \\ \geq &1 - 2\exp(-(n^{\gamma/2} - 1)4s\ln p) - 2\exp(-[(4s + 1)(n^{\gamma/2} - 1) - 1]\ln p) \\ \geq &1 - C_4\exp(-C_5 s n^{\gamma/2}\ln p). \end{aligned}$$

We then use the same technique as in Theorem 3.1 to combine this number with the probability of $\mathscr{A}$ to obtain the final non-conditional probability that $\beta^*$ is the global solution with probability at least

$$\geq 1 - C_4 \exp(-C_5 s n^{\gamma/2} \ln p) - 2p \exp(\frac{-(\lambda^{lasso})^2 n b_u a}{8\sigma^2})$$
$$\geq 1 - C_4 \exp(-C_5 s n^{\gamma/2} \ln p) - 2 \exp(\frac{-(\varepsilon^2 b_u a n^{\gamma/2} - 8) \ln p}{8})$$
$$\geq 1 - C_4 \exp(-C_5 s n^{\gamma/2} \ln p) - C_6 \exp(-C_7 b_u n^{\gamma/2} \ln(p)),$$

for some constants $C_4$, $C_5$ $C_6$ and $C_7 > 0$. Note that these constants, as well as $C_1$, $C_2$, and $C_3$, are dependent only on the value of $a$ $Q$ and $\varepsilon$, as far as problem dependencies are concerned. Thus given that $a$ $Q$ and $\varepsilon$ are chosen to be any positive constant value, as in the statement of Corollary 3.1, $C_1$ through $C_7$ are problem independent, which is the desired result. □

*Proof of Corollary 3.2*. The first result is simply Corollary 2 in [11]. If we initialize the LLA algorithm with $\beta^{lasso}$, the solution to LASSO using $\lambda^{lasso}$ as the LASSO constant, then the LLA algorithm converges to the oracle solution in 2 iterations with probability $1 - \phi_0 - \phi_1 - \phi_2$. However, we still need to solve for the actual values of $\phi_0, \phi_1, \phi_2$ for GLM.

First, we consider $\phi_0 = \mathbb{P}(\left\|\beta^{lasso} - \beta^{true}\right\|_{max} > a_0\lambda)$. Similar to Lemma B.1. in [4], to bound this, we start by noticing that, for the lasso penalized loss function

$$\mathscr{Q}^{lasso}(\beta) = \sum_{i\in\mathscr{N}} l(\beta, x_i, y_i) + \lambda^{lasso} \sum_{j\in\mathscr{P}} |\beta_j|,$$

we have $\mathscr{Q}^{lasso}(\beta^{lasso}) \leq \mathscr{Q}^{lasso}(\beta^{true})$. If we then let $\delta^\ell = \beta^{lasso} - \beta^{true}$, we can use the same tactic as in the derivation of A.1 to obtain $\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 - \frac{1}{n}W^\intercal X\delta^\ell \leq \lambda^{lasso}\sum_{j\in\mathscr{P}}|\beta_j^{true}| - |\beta_j^{lasso}|$, which can then be rearranged to obtain

$$\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 - \frac{1}{n}\sum_{j\in\mathscr{P}}|W^\intercal X_j||\delta_j'| \leq \lambda^{lasso}\sum_{j\in\mathscr{P}}|\beta_j^{true}| - |\beta_j^{lasso}|.$$

Next, let $\mathscr{A} = \bigcap_{j\in\mathscr{P}}\{|\frac{1}{n}W^\intercal X_j| \leq \lambda^{lasso}/2\}$. We can combine this with A.1 to see that $\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 + \lambda^{lasso}/2\sum_{j\in\mathscr{P}}|\beta_j^{lasso} - \beta_j^{true}| \leq \lambda^{lasso}\sum_{j\in\mathscr{P}}|\beta_j^{lasso} - \beta_j^{true}| + \lambda^{lasso}\sum_{j\in\mathscr{P}}|\beta_j^{true}| - |\beta_j^{lasso}|$ conditional on $\mathscr{A}$. From this, we notice that the right term goes to zero when $\beta_j^{true} = 0$ so we then have that $\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 + \lambda^{lasso}/2\sum_{j\in\mathscr{P}}|\beta_j^{lasso} - \beta_j^{true}| \leq \lambda^{lasso}\sum_{j\in\mathscr{S}}|\beta_j^{lasso} - \beta_j^{true}| + |\beta_j^{true}| - |\beta_j^{lasso}|$. Using the triangle inequality and the definition of $\delta^\ell$, we can simplify this to

$$\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 + \frac{\lambda^{lasso}}{2}|\delta^\ell| \leq 2\lambda^{lasso}|\delta^\ell_{\mathscr{S}}|$$

conditional on $\mathscr{A}$. By relaxing different parts of the equation, this can be further simplified to both $\frac{b_l}{n}\left\|X\delta^\ell\right\|^2 \leq 3\lambda^{lasso}|\delta^\ell_{\mathscr{S}}| \leq 3\lambda^{lasso}s^{1/2}||\delta^\ell_{\mathscr{S}}||_2$ and $|\delta^\ell_{\mathscr{S}_c}| \leq 3|\delta^\ell_{\mathscr{S}}|$. Note that the second of these shows that $\delta^\ell$ satisfies the constraint for the RE condition 2.1. Therefor $\frac{\left\|X\delta^\ell\right\|^2}{n\left\|\delta^\ell\right\|^2} \geq r_e$. If this is combined with the first of the two equations, we can obtain that $\frac{1}{n^{1/2}}\left\|X\delta^\ell\right\| \leq \frac{3\lambda^{lasso}s^{1/2}}{b_l(r_e)^{1/2}}$ conditional on $\mathscr{A}$. Observe $\left\|\delta^\ell\right\|_{max} \leq \left\|\delta^\ell\right\| \leq \left\|X\delta^\ell\right\|_2^2/(\left\|\delta^\ell\right\| n r_e) \leq \frac{3\lambda^{lasso}s^{1/2}}{b_l r_e} < a_0\lambda$ if $\lambda > \frac{3\lambda^{lasso}s^{1/2}}{b_l a_0 r_e}$. This is the inverse of the condition that defines $\phi_0$. Thus, we can bound $\phi_0$ with $\phi_0 \leq \mathbb{P}(\mathscr{A}^c) = \mathbb{P}(\bigcup_{j\in\mathscr{P}}|\frac{1}{n}W^\intercal X_j| > \lambda^{lasso}/2) = \mathbb{P}(\bigcup_{j\in\mathscr{P}}|W^\intercal X_j|/\left\|X_j\right\| > n\lambda^{lasso}/(2\left\|X_j\right\|)) \leq$

$p\mathbb{P}(|\langle W,v\rangle| > \frac{\lambda^{lasso}n}{2\|X_j\|}) \leq p\mathbb{P}(|\langle W,v\rangle| > \frac{\lambda^{lasso}(nb_u a)^{1/2}}{2}) \leq 2p\exp\frac{-(\lambda^{lasso})^2 nb_u a}{8\sigma^2}$, which uses both (A1)(ii) and (A2) as long as $\lambda > \frac{3\lambda^{lasso}s^{1/2}}{b_l a_0(r_e)}$ per (A2).

Next, we consider $\phi_1 = \mathbb{P}(\left\|\nabla_{S^c_{\tilde{p}}}\ell_n(\beta^{oracle}\right\|_{max} \geq a_1\lambda)$,

$$\begin{aligned}
\phi_1 =& \mathbb{P}(\left\|\nabla_{S^c_{\tilde{p}}}\ell_n(\beta^{oracle}\right\|_{max} \geq a_1\lambda) \\
=& \mathbb{P}(\exists j \in \mathscr{P} : |\nabla_j \ell_n(\beta^{oracle})| \geq a_1\lambda) \\
=& \mathbb{P}(\exists j \in \mathscr{P} : |\frac{1}{n}\sum_{i\in\mathscr{N}}[\psi'(x_i^\intercal\beta^{oracle})x_{i,j} - y_i x_{i,j}]| \geq a_1\lambda) \\
=& \mathbb{P}(\exists j \in \mathscr{P} : |\frac{1}{n}\sum_{i\in\mathscr{N}}[\psi'(x_i^\intercal\beta^{oracle})x_{i,j} - \psi'(x_i^\intercal\beta^{true})x_{i,j} + W_i x_{i,j}]| \geq a_1\lambda) \\
\leq& \mathbb{P}(\frac{1}{n}|X_j^\intercal(\psi'(X\beta^{oracle}) - \psi'(X\beta^{true}) + W)| \geq a_1\lambda) \\
\leq& \mathbb{P}(\frac{1}{n}|X_j^\intercal(\psi'(X\beta^{oracle}) - \psi'(X\beta^{true}))| + |W^\intercal X_j| \geq a_1\lambda) \\
\leq& \mathbb{P}(\frac{1}{n}\|X_j\|\left\|\psi'(X\beta^{oracle}) - \psi'(X\beta^{true})\right\| + |W^\intercal X_j| \geq a_1\lambda) \\
\leq& \mathbb{P}(\frac{1}{n}\left\|\psi'(X\beta^{oracle}) - \psi'(X\beta^{true})\right\| + |W^\intercal X_j|/\|X_j\| \geq a_1\lambda\|X_j\|^{-1}) \\
\leq& \mathbb{P}(\left\|\psi'(X\beta^{oracle}) - \psi'(X\beta^{true})\right\| + |W^\intercal X_j|/\|X_j\| \geq (ab_u n)^{1/2}a_1\lambda) \\
\leq& \mathbb{P}(b_u\left\|X\beta^{oracle} - X\beta^{true}\right\| + |W^\intercal v| \geq (ab_u n)^{1/2}a_1\lambda) \\
\leq& \mathbb{P}(b_u\|X\delta^o\| + |W^\intercal v| \geq (ab_u n)^{1/2}a_1\lambda),
\end{aligned}$$

where $v \in \mathbb{R}^n$ is some vector with $\|v\| = 1$ as indicated in (A2) and $\delta^o = \beta^{oracle} - \beta^{true}$. From this, using De Morgan's law and the union bound, we notice that $\mathbb{P}(A+B \geq C) \leq \mathbb{P}(A \geq C/2) + \mathbb{P}(B \geq C/2)$ which can be used to further simplify

$$\begin{aligned}
\phi_1 &\leq \mathbb{P}(b_u\|X\delta^o\| + |W^\intercal v| \geq a_1\lambda(ab_u n)^{1/2}) \\
&\leq \mathbb{P}(\|X\delta^o\| \geq (1/2)a_1\lambda(an/b_u)^{1/2}) + \mathbb{P}(|W^\intercal v| \geq (1/2)a_1\lambda(ab_u n)^{1/2}).
\end{aligned}$$

We can simplify both terms individually. For the first term, $\mathbb{P}(b_u\|X\delta^o\| \geq (1/2)a_1\lambda(ab_u n)^{1/2})$, given the fact that the oracle solution and true solution have the same support, the oracle solution must be in the $\Gamma = 0$ level set of the true solution. Using similar arguments to Lemma A.5, we have that $\frac{b_l}{2n}\|X\delta^o\|^2 \leq \frac{1}{n}W^\intercal X\delta^o$. Thus Lemma A.2 can be applied since we know $\left\|\beta^{oracle} - \beta^{true}\right\|_0 \leq s$. With some simplification, one has $\|X\delta^o\| \leq \frac{2}{b_l}\left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=s}\left\|\tilde{U}^\intercal_{S_{\tilde{p}}}W\right\|\right)$. Utilizing Lemma A.3 with $s$ in place of $\tilde{p}$ shows that

$$\mathbb{P}\left[\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=s}\frac{2}{b_l}\left\|\tilde{U}^\intercal_{S_{\tilde{p}}}W\right\| \geq \frac{2}{b_l}\sigma\sqrt{s+2\sqrt{st_1}+2t_1}\right] \leq \left(\frac{pe}{s}\right)^s\exp(-t_1).$$

This is the first half of $\phi_1$ as long as $(1/2)a_1\lambda(an/b_u)^{1/2} \geq \frac{2}{b_l}\sigma\sqrt{s+2\sqrt{st}+2t}$, which is equivalent to the assumed condition $\lambda \geq \frac{4\sigma\sqrt{s+2\sqrt{st_1}+2t_1}}{b_l a_1 (an/b_u)^{1/2}}$. Next, the second term can be easily bounded using (A2): $\mathbb{P}(|W^\intercal v| \geq (1/2)a_1\lambda(ab_u n)^{1/2}) \leq 2\exp(-\frac{a_1^2\lambda^2 ab_u n}{8\sigma^2})$, therefor $\phi_1 \leq (\frac{pe}{s})^s\exp(-t_1)+2\exp(\frac{-a_1^2\lambda^2 ab_u n}{8\sigma^2})$.

Next, we consider $\phi_2 = \mathbb{P}(\|\beta_{\mathscr{S}}^{oracle}\|_{\min} \leq a\lambda)$. First, given the assumption $\left\|\beta^{true]}\right\|_{min} > (a+1)\lambda$, we can see that

$$\phi_2 = \mathbb{P}(\left\|\beta_{\mathscr{S}}^{oracle}\right\|_{\min} \leq a\lambda) \leq \mathbb{P}(\left\|\beta_{\mathscr{S}}^{oracle} - \beta_{\mathscr{S}}^{true}\right\|_{\max} > \lambda) \leq \mathbb{P}(\left\|\beta^{oracle} - \beta^{true}\right\|_2 > \lambda)$$
$$= \mathbb{P}(\|\delta^o\|_2 > \lambda).$$

Since the support of $\beta^{oracle}$ and $\beta^{true}$ is $\mathscr{S}$, we know that $|\delta^o_{\mathscr{S}_c}| = 0 \leq 3|\delta^o_{\mathscr{S}}|$, which is the constraint for the RE condition. Thus $\frac{\|X\delta^o\|^2}{n\|\delta^o\|^2} \geq r_e$. With this and a similar line of argument as in $\phi_1$, we see that $\phi_2 \leq \mathbb{P}(\|\delta^o\| > \lambda) \leq \mathbb{P}(\|X\delta^o\| > \lambda\sqrt{nr_e}) \leq \mathbb{P}(\frac{2}{b_l}\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=s}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| > \lambda\sqrt{nr_e}) = \mathbb{P}(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=s}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| > \lambda\frac{b_l\sqrt{nr_e}}{2} \geq \sigma\sqrt{s+2\sqrt{st_2}+2t_2}) \leq (\frac{pe}{s})^s\exp(-t_2)$ assuming that $\lambda\frac{b_l\sqrt{nr_e}}{2} \geq \sigma\sqrt{s+2\sqrt{st}+2t}$, which is equivalent to the condition $\lambda \geq \frac{2\sigma\sqrt{s+2\sqrt{st_2}+2t_2}}{b_l\sqrt{nr_e}}$. This, combined with the fact that for MCP, $a_0 = a_1 = a_2 = 1$ shows the first result.

The second result can be seen by first noting all the assumptions of Theorem 3.1 part 2 are satisfied, where (A3) with $r_{4s}$ is implied by A.7.

Thus using the same arguments as in Theorem 3.1 part 2 shows that the oracle solution is unique and that the global solution is the oracle solution with some probability, since the global solution is almost surely $S^3ONC$ with $\Gamma = 0$. If $t = t_3$ and $t' = t_4$, we obtain the probability that the global solution is not the oracle solution as $\phi_3 \leq \exp(-t_3 + \tilde{p}\ln(\frac{pe}{\tilde{p}})) + \exp(-(\tilde{p}^*+1)(t_4 - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t_4-\ln p))}{1-\exp(-t_4+\ln p)}$. This combined with the first result shows that the LLA algorithm converges to the global solution in 2 iterations with probability $1-\phi_0-\phi_1-\phi_2-\phi_3$, which is the second result. □

## A.2. Central lemmas and their proofs.

**Lemma A.1.** *Let $\beta^*$ be a $S^3ONC$ solution to 1.1. If assumption (A1) holds, then* $\mathbb{P}[|\beta_j^*| \notin (0,a\lambda), \forall j \in \{1,2,...,p\}] = 1$.

*Proof of Lemma A.1.* First, define events $\gamma_j$ and $\delta_j$ as

$$\gamma_j := \left\{ \frac{\partial^2 \mathscr{Q}(\beta)}{(\partial\beta_j)^2}\bigg|_{\beta=\beta^*} \geq 0 \right\}$$
$$\delta_j := \left\{|\beta_j^*| \in (0,a\lambda)\right\}.$$

For any given $j \in \mathscr{P}$, we solve for $\mathbb{P}[\gamma_j \cap \delta_j]$ given our assumptions. We can start with $\frac{\partial^2\mathscr{Q}(\beta)}{(\partial\beta_j)^2}\Big|_{\beta=\beta^*} \geq 0$ which gives us $1/n\sum_{i=1}^{n}\psi''\left(x_i^\intercal\beta^*\right)x_{i,j}^2 + P_\lambda''(|\beta_j^*|) \geq 0$. We can rearrange this to obtain $b_u \sum_{i=1}^{n} x_{i,j}^2 \geq \sum_{i=1}^{n} \psi''(x_i^\intercal \beta^*) x_{i,j}^2 \geq -nP''_\lambda(|\beta^*|) = n/a$ where we get the leftmost inequality from assumption (A1) part (i) and the rightmost equality from the definition of MCP. More concisely, $b_u \left\|X_j\right\|^2 \geq n/a$ which contradicts (A1) part (ii). Thus $\mathbb{P}[\gamma_j \cap \delta_j] = 0$. It should also be noted that $\mathbb{P}[\gamma_j^c] = 0$ since $\beta^*$ satisfies $S^3ONC$ conditions. By applying De Morgan's law and the union bound, it can be obtained that

$$0 = \mathbb{P}[\gamma_j \cap \delta_j] = 1 - \mathbb{P}[\gamma_j^c \cup \delta_j^c] \geq 1 - \mathbb{P}[\gamma_j^c] - \mathbb{P}[\delta_j^c] = 1 - \mathbb{P}[\delta_j^c] = \mathbb{P}[\delta_j].$$

We can then apply this result to all indices to obtain that $\mathbb{P}[\delta_j] = 0$ for all $j \in \{1, 2, ..., p\}$, which is the desired result. □

**Lemma A.2.** *Consider an arbitrary $S^3ONC$ solution $\beta^*$ to (1.1) with MCP. Given the event that for some integer $\tilde{p}$ : $\|\beta^* - \beta^{true}\|_0 \leq \tilde{p}$, $|W^\intercal X\delta^*| \leq \left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right) \|X\delta^*\|$, a.s. where*

$$(\tilde{U}_{S_{\tilde{p}}})_{i,j} := \begin{cases} U_{S_{\tilde{p}}}, & \textit{if } j \in S_{\tilde{p}} \\ 0, & \textit{else} \end{cases}$$

*and $U_{S_{\tilde{p}}} \in \mathbb{R}^{n\times\tilde{p}}$ is defined as in the following Thin SVD: $X_{S_{\tilde{p}}} = U_{S_{\tilde{p}}} D_{S_{\tilde{p}}} V_{S_{\tilde{p}}}$.*

*Proof.* Denote $\delta^* := (\delta_j^*) = \beta^* - \beta^{true}$, $S_{\tilde{p}} := (j : \delta_j^* \neq 0) \subseteq \mathscr{P}$, $\delta_{S_{\tilde{p}}}^* := (\delta_j^* : j \in S_{\tilde{p}})$, and $X_{S_{\tilde{p}}} := (x_{ij} : i \in \mathscr{N}, j \in S_{\tilde{p}})$. By assumption, we know that $\|\delta^*\|_0 \leq |S_{\tilde{p}}| = \tilde{p}$. First decompose $X_{S_{\tilde{p}}}$ using Thin SVD to obtain $X_{S_{\tilde{p}}} = U_{S_{\tilde{p}}} D_{S_{\tilde{p}}} V_{S_{\tilde{p}}}$, where $U_{S_{\tilde{p}}} \in \mathbb{R}^{n\times\tilde{p}}$. Since $U_{S_{\tilde{p}}}^\intercal U_{S_{\tilde{p}}} = I$, we have that, for any $\upsilon \in \mathbb{R}^{\tilde{p}}$, $\left\|D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \upsilon\right\|^2 = (D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \upsilon)^\intercal I (D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \upsilon) = \upsilon^\intercal V_{S_{\tilde{p}}}^\intercal D_{S_{\tilde{p}}}^\intercal U_{S_{\tilde{p}}}^\intercal U_{S_{\tilde{p}}} D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \upsilon = \upsilon^\intercal X_{S_{\tilde{p}}}^\intercal X_{S_{\tilde{p}}} \upsilon = \left\|X_{S_{\tilde{p}}} \upsilon\right\|^2$. It follows that

$$\begin{aligned} |W^\intercal X\delta^*| &= |W^\intercal X_{S_{\tilde{p}}} \delta_{S_{\tilde{p}}}^*| \leq \left\|W^\intercal U_{S_{\tilde{p}}}\right\| \left\|D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \delta_{S_{\tilde{p}}}^*\right\| \\ &= \left\|U_{S_{\tilde{p}}}^\intercal W\right\| \left\|X_{S_{\tilde{p}}} \delta_{S_{\tilde{p}}}^*\right\| \leq \left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right) \|X\delta^*\|, \quad a.s. \end{aligned}$$

where

$$(\tilde{U}_{S_{\tilde{p}}})_{i,j} := \begin{cases} U_{S_{\tilde{p}}}, & \text{if } j \in S_{\tilde{p}}, \\ 0, & \text{else.} \end{cases}$$

□

**Lemma A.3.** *Consider an arbitrary $S^3ONC$ solution $\beta^*$ to 1.1 with MCP. If (A2) holds, then for some integer $\tilde{p} \leq p$, $\mathbb{P}\left[\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| \leq \sigma\sqrt{\tilde{p} + 2\sqrt{\tilde{p}t} + 2t}\right] \geq 1 - (\frac{pe}{\tilde{p}})^{\tilde{p}} \exp(-t)$, where*

$$(\tilde{U}_{S_{\tilde{p}}})_{i,j} := \begin{cases} U_{S_{\tilde{p}}}, & \textit{if } j \in S_{\tilde{p}}, \\ 0, & \textit{else,} \end{cases}$$

*and $U_{S_{\tilde{p}}} \in \mathbb{R}^{n\times\tilde{p}}$ is defined as in the following Thin SVD: $X_{S_{\tilde{p}}} = U_{S_{\tilde{p}}} D_{S_{\tilde{p}}} V_{S_{\tilde{p}}}$.*

*Proof.* We attempt to bound $\left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right)$. Given that we now have $W$ multiplied by a square matrix, we can apply Lemma A.9. In the Lemma, let $\Sigma_u = \tilde{U}_{S_{\tilde{p}}} \tilde{U}_{S_{\tilde{p}}}^\intercal$. The fact that $\Sigma_u \Sigma_u = \Sigma_u$ means that $\Sigma_u$ is an idempotent matrix with $\|\Sigma_u\| \leq 1$ and $Tr(\Sigma_u) = rank(\Sigma_u) \leq$

$rank(\tilde{U}_{S_{\tilde{p}}}) \leq rank(U_{S_{\tilde{p}}}) \leq \tilde{p}$. Lemma A.9 then states that $\mathbb{P}\left[\left\|\tilde{U}_{S_{\tilde{p}}}^{\intercal} W\right\| \leq \sigma\sqrt{\tilde{p}+2\sqrt{\tilde{p}t}+2t}\right] \geq 1-\exp(-t)$. From this, we can show that

$$\mathbb{P}\left[\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}_{S_{\tilde{p}}}^{\intercal} W\right\| \leq \sigma\sqrt{\tilde{p}+2\sqrt{\tilde{p}t}+2t}\right] \geq 1-\binom{p}{\tilde{p}}\exp(-t) \geq 1-(\frac{pe}{\tilde{p}})^{\tilde{p}}\exp(-t).$$

Here the first inequality holds by noting the following fact. If $\eta_k \in \mathbb{R}^k$ is a sequence of i.i.d random variables and $\theta \in \mathbb{R}$ is a scalar, by applying De Morgan's Law and then using the union bound, it can be obtained that $\mathbb{P}[\max_{k\in K}\eta_k \leq \theta] = \mathbb{P}[\bigcap_{k\in K}\eta_k \leq \theta] = 1-\mathbb{P}[\bigcup_{k\in K}\eta_s \geq \theta] \geq 1-\sum_{k\in K}\mathbb{P}[\eta_k \geq \theta] = 1-|K|(1-\mathbb{P}[\eta_k \leq \theta])$, which yields the same inequality as in A.2. This accomplishes the desired result. □

**Lemma A.4.** *Consider an arbitrary $S^3ONC$ solution $\beta^*$ to 1.1 with MCP. Let Assumptions (A1) and (A2) hold. Given the simultaneous occurrence of (i), the event that $\mathcal{Q}(\beta^*) \leq \mathcal{Q}(\beta^{true})+\Gamma$ holds for some $\Gamma \geq 0$; (ii) the event that for some integer $\tilde{p}: \|\beta^*-\beta^{true}\|_0 \leq \tilde{p}$. Then, for any $t>0$,*

$$\begin{aligned}&\frac{1}{n}\left\|X(\beta^*-\beta^{true})\right\|^2\\ &\leq \frac{4\sigma^2}{b_l^2 n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t)+\frac{8}{b_l}\min\left\{\sum_{j\in S}P_\lambda'(|\beta_j^*|)|\beta_j^*|, P_\lambda(a\lambda)(|\mathscr{S}|-\|\beta^*\|_0)+\Gamma\right\}\end{aligned}$$

*holds with probability at least $1-\exp(-t+\tilde{p}\ln(\frac{pe}{\tilde{p}}))$. If, in addition, (A3) holds with $\tilde{p}^* \geq \tilde{p}$, then*

$$\begin{aligned}&\frac{1}{n}\left\|X(\beta^*-\beta^{true})\right\|^2\\ &\leq \frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{b_l}\min\left\{\lambda^2(|\mathscr{S}|-\|\beta^*\|_0)r_{\tilde{p}}^{-1}, P_\lambda(a\lambda)\dot{(}|\mathscr{S}|-\|\beta^*\|_0)+\Gamma\right\}\end{aligned}$$

*holds where $r_{\tilde{p}}>0$ for any $t>0$ with probability at least $1-\exp\left(-t+\tilde{p}\ln\left(\frac{pe}{\tilde{p}}\right)\right)$.*

*Proof.* First, we denote $\delta^* := (\delta_j^*) = \beta^*-\beta^{true}$, $S_{\tilde{p}} := (j : \delta_j^* \neq 0) \subseteq \mathscr{P}$, $\delta_{S_{\tilde{p}}}^* := (\delta_j^* : j \in S_{\tilde{p}})$, and $X_{S_{\tilde{p}}} := (x_{ij} : i \in \mathscr{N}, j \in S_{\tilde{p}})$. Observe that $\|\delta^*\|_0 \leq |S_{\tilde{p}}| = \tilde{p}$. Further, let us denote

$$\mathscr{T}_1 := \min\left\{\sum_{j\in S}P_\lambda'(|\beta_j^*|)|\beta_j^{true}|,\ \sum_{j\in S}P_\lambda'(|\beta_j^*|)|\beta_j^*-\beta_j^{true}|,\ P_\lambda(a\lambda)(|\mathscr{S}|-\|\beta^*\|_0)+\Gamma\right\}.$$

We now start to define the desired bound by applying the second part of Lemma A.8. The result simplified using the above definitions becomes $\frac{b_l}{2n}\|X\delta^*\|^2 \leq \frac{1}{n}W^\intercal X\delta^*+\mathscr{T}_1$ a.s. It follows that

$$\frac{b_l}{2n}\|X\delta^*\|^2 \leq \frac{1}{n}\left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}_{S_{\tilde{p}}}^{\intercal} W\right\|\right)\|X\delta^*\|+\mathscr{T}_1.$$

We can then complete the square

$$\frac{1}{\sqrt{n}}\|X\delta^*\| \leq \frac{1}{b_l\sqrt{n}} \max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| + \sqrt{\left(\frac{1}{b_l\sqrt{n}} \max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right)^2 + \frac{2}{b_l}\mathscr{T}_1}$$

$$\leq 2\sqrt{\left(\frac{1}{b_l\sqrt{n}} \max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right)^2 + \frac{2}{b_l}\mathscr{T}_1},$$

where the last inequality holds due to the value inside the square root being larger than the term outside. Squaring both sides gives us

$$\frac{1}{n}\|X\delta^*\|^2 \leq \frac{4}{b_l^2 n}\left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right)^2 + \frac{8}{b_l}\mathscr{T}_1.$$

Finally, we have

$$\frac{1}{n}\|X\delta^*\|^2 \leq \frac{4\sigma^2}{b_l^2 n}\left(\tilde{p} + 2\sqrt{\tilde{p}t} + 2t\right) + \frac{8}{b_l}\mathscr{T}_1,$$

with probability at least $1-(\frac{pe}{\tilde{p}})^{\tilde{p}}\exp(-t)$. Thus by the definition of $\mathscr{T}_1$, the first result of the lemma has been shown.

For the second part, we look to bound the central term of $\mathscr{T}_1$. We first notice (a) that, since assumption (A1) holds, Lemma A.1 indicates that if $\beta_j^* \neq 0 \Rightarrow |\beta_j^*| \geq a\lambda$ for all $j \in \mathscr{P}$ with probability one; (b) that for this range of $\beta_j^*$, $P'_\lambda(|\beta_j^*|) = 0$; (c) that per the definition of MCP $0 \leq P'_\lambda(|\beta_j^*|) \leq \lambda$ for any $\beta_j^* \in \Re$. If we combine these observations with A.2 and the definition of $\delta^*$, we see that

$$\mathscr{T}_1 \leq \sum_{j\in\mathscr{S}} P'_\lambda(|\beta_j^*|)|\delta^*| \leq \lambda\sqrt{|\mathscr{S}| - \left\|\beta_{\mathscr{S}}^*\right\|_0} \cdot \|\delta^*\|.$$

Since (A3) holds with $\tilde{p}^* \geq \tilde{p}$, and $r_{\tilde{p}} \geq r_{\tilde{p}^*} \geq 0$, we can use (A3) part (iii) to show that $\mathscr{T}_1 \leq \lambda\sqrt{|\mathscr{S}| - \left\|\beta_{\mathscr{S}}^*\right\|_0} \cdot \frac{\|X\delta^*\|}{\sqrt{nr_{\tilde{p}}}}$. Since this holds almost surely, it can then be combined with A.2 to obtain

$$\frac{b_l}{2n}\|X\delta^*\|^2 \leq \frac{1}{n}\left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right)\|X\delta^*\| + \lambda\sqrt{|\mathscr{S}| - \left\|x_{\mathscr{S}}^*\right\|_0} \cdot \frac{\|X\delta^*\|}{\sqrt{nr_{\tilde{p}}}}.$$

We can then multiply by $2\sqrt{n}/b_l\|X\delta^*\|$ to get

$$\frac{1}{\sqrt{n}}\|X\delta^*\| \leq \frac{2}{b_l\sqrt{n}} \max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| + \frac{2\lambda}{b_l\sqrt{r_{\tilde{p}}}}\sqrt{|\mathscr{S}| - \left\|x_{\mathscr{S}}^*\right\|_0}.$$

We then square both sides and use the rule that $(A+B)^2 \leq 2A^2 + 2B^2$ to get

$$\frac{1}{n}\|X\delta^*\|^2 \leq \left[\frac{2}{b_l\sqrt{n}} \max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| + \frac{2\lambda}{b_l\sqrt{r_{\tilde{p}}}}\sqrt{|\mathscr{S}| - \left\|x_{\mathscr{S}}^*\right\|_0}\right]^2$$

$$\leq \frac{8}{b_l^2 n} \max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|^2 + \frac{8\lambda^2}{b_l^2 r_{\tilde{p}}}\left(|\mathscr{S}| - \|x_{\mathscr{S}}^*\|_0\right).$$

Combining this with A.2 yields that

$$\frac{1}{n}\|X\delta^*\|^2 \leq \frac{8}{b_l^2 n} \max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|^2 + \frac{8}{b_l}\min\left\{\frac{\lambda^2}{r_{\tilde{p}}}\left(|\mathscr{S}| - \|x_{\mathscr{S}}^*\|_0\right), \mathscr{T}_1\right\}.$$

Note from (A.2) that $\mathscr{T}_1 \leq P_\lambda(a\lambda)(|\mathscr{S}| - \|\beta^*\|_0) + \Gamma$. It follows that

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

with probability at least $1 - (\frac{pe}{\tilde{p}})^{\tilde{p}}\exp(-t)$, which is the desired result. □

**Lemma A.5.** *Let Assumptions (A1) and (A2) hold. Consider a solution $\beta^*$ satisfying $S^3ONC$ of* (1.1)*. Assume that $\mathscr{Q}(\beta^*) \leq \mathscr{Q}(\beta^{true}) + \Gamma$ holds for an arbitrary $\Gamma > 0$. For any integer $\tilde{p} : 2|\mathscr{S}| \leq \tilde{p} \leq p$ if the penalty parameters $(a, \lambda)$ satisfy $P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1 + 2\sqrt{t} + 2t) + \frac{\frac{\sigma^2}{n}|\mathscr{S}|(1+2\sqrt{t}+2t)+\Gamma b_l}{b_l(\tilde{p}+1-2|\mathscr{S}|)}$, for an arbitrary $t > 0$, then $\|\beta^* - \beta^{true}\|_0 \leq \tilde{p}$ 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.* We start from the useful inequality defined in A.1

$$\frac{b_l}{2n}\|X\delta^*\|^2 - \frac{1}{n}W^\intercal X\delta^* + \sum_{j\in\mathscr{S}} P_\lambda(|\beta_j^*|) \leq \sum_{j\in\mathscr{S}} P_\lambda(|\beta_j^{true}|) + \Gamma,$$

where $\delta^* = \beta^* - \beta^{true}$. Next, conditioning on the fact (i) that $\beta^*$ is $S^3ONC$, (ii) that all the assumptions for Lemma A.1 are satisfied (which implies that $P_\lambda(|\beta_j^*|) \in \{0, P_\lambda(a\lambda)\}$ with probability one) and (iii) that $P_\lambda(|\beta_j^{true}|) \leq P_\lambda(a\lambda)$, we have that

$$\frac{b_l}{2n}\|X\delta^*\|^2 - \frac{1}{n}W^\intercal X\delta^* + \|\beta^*\|_0 \cdot P_\lambda(a\lambda) \leq |\mathscr{S}| \cdot P_\lambda(a\lambda) + \Gamma.$$

Now, we consider an event $\mathscr{E}_1 := \{\|\beta^* - \beta^{true}\|_0 = \tilde{p} + k\}$ for an arbitrary integer $k : 1 \leq k \leq p - \tilde{p}$. Conditioning on this event, we may denote and $S_{\tilde{p}+k} \subseteq \mathscr{P}$ such that $\delta_j^* \neq 0$ for all $j \in S_{\tilde{p}+k}$. By assumption we can ensure that $|S_{\tilde{p}+k}| = \tilde{p}+k$. Also denote by $X_{S_{\tilde{p}+k}} = (x_{ij} : i \in \mathscr{N}, j \in S_{\tilde{p}+k})$ and let $\delta^*_{S_{\tilde{p}+k}} := (\delta_j^* : j \in S_{\tilde{p}+k})$. Note that conditional on $\mathscr{E}_1$, the first part of the lemma (using $\tilde{p}+k$ in place of $\tilde{p}$) can be used to bound $W^\intercal X\delta^*$ in A.2. Additionally, by definition $\|\beta^{true}\|_0 = |\mathscr{S}|$ and conditional on $\mathscr{E}_1$, we can apply the substitution $\|\beta^*\|_0 \geq \tilde{p}+k-|\mathscr{S}|$. This gives us

$$\frac{b_l}{2}\left\|\frac{X\delta^*}{\sqrt{n}}\right\|^2 - \frac{1}{\sqrt{n}}\left(\max_{S_{\tilde{p}+k}:|S_{\tilde{p}+k}|=\tilde{p}+k}\left\|\tilde{U}^\intercal_{S_{\tilde{p}+k}}W\right\|\right)\left\|\frac{X\delta^*}{\sqrt{n}}\right\| \leq -(\tilde{p}+k-2|\mathscr{S}|)\cdot P_\lambda(a\lambda) + \Gamma.$$

In order to make this equation to be feasible, we know that the quadratic formula must have real roots. Therefore

$$\left(\max_{S_{\tilde{p}+k}:|S_{\tilde{p}+k}|=\tilde{p}+k}\left\|\frac{\tilde{U}^\intercal_{S_{\tilde{p}+k}}W}{\sqrt{n}}\right\|\right)^2 - 4[b_l/2][(\tilde{p}+k-2|\mathscr{S}|)\cdot P_\lambda(a\lambda) - \Gamma] \geq 0.$$

Now, we consider another event $\mathscr{E}_2(t) := \left\{\max_{|S_{\tilde{p}+k}|=\tilde{p}+k}\|U^\intercal_{S_{\tilde{p}+k}}W\| \leq \sigma\sqrt{\tilde{p}+k}\cdot\sqrt{1+2\sqrt{t}+2t}\right\}$ for an arbitrary $t > 0$. Conditioning on $\mathscr{E}_1 \cap \mathscr{E}_2(t)$, we can show, using first $\mathscr{E}_2(t)$ and then A.2, that $\frac{\sigma^2(\tilde{p}+k)}{n}\cdot(1+2\sqrt{t}+2t) \geq \left(\max_{|S_{\tilde{p}+k}|=\tilde{p}+k}\frac{\left\|\tilde{U}^\intercal_{S_{\tilde{p}+k}}W\right\|}{\sqrt{n}}\right)^2 \geq 2b_l[(\tilde{p}+k-2|\mathscr{S}|)\cdot P_\lambda(a\lambda) - \Gamma]$

almost surely, which contradicts the assumption on the parameters $(a,\lambda)$. This can be seen by starting from our original assumption that

$$P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1+2\sqrt{t}+2t) + \frac{\frac{\sigma^2}{n}|\mathscr{S}|(1+2\sqrt{t}+2t)+\Gamma b_l}{b_l(\tilde{p}+1-2|\mathscr{S}|)}$$
$$\geq \frac{\sigma^2}{2nb_l}(1+2\sqrt{t}+2t) + \frac{\frac{\sigma^2}{n}|\mathscr{S}|(1+2\sqrt{t}+2t)+\Gamma b_l}{b_l(\tilde{p}+k-2|\mathscr{S}|)}.$$

We can then multiply both (outer) sides by $2b_l(\tilde{p}+k-2|\mathscr{S}|)$ and rearrange to get $\frac{\sigma^2}{n}(\tilde{p}+k)\cdot(1+2\sqrt{t}+2t) < 2b_l[(\tilde{p}-2|\mathscr{S}|+k)\cdot P_\lambda(a\lambda)-\Gamma]$. Given this contradiction, we know that $\mathbb{P}[\mathscr{E}_1\cap\mathscr{E}_2(t)] = 0$. Therefore, using the union bound combined with DeMorgan's law again, we get that $\mathbb{P}[\mathscr{E}_1\cap\mathscr{E}_2(t)] \geq 1-\mathbb{P}[\mathscr{E}_1^c]-\mathbb{P}[\mathscr{E}_2(t)^c]$, which can be simplified to $\mathbb{P}[\mathscr{E}_2(t)^c] \geq \mathbb{P}[\mathscr{E}_1]$. Since all the assumptions of the Lemma A.3 are satisfied, we can next use it to bound $\mathbb{P}[\mathscr{E}_2(t)^c]$. For some $t'$, we see that

$$\mathbb{P}\left[\max_{S_{\tilde{p}+k}:|S_{\tilde{p}+k}|=(\tilde{p}+k)}\left\|\tilde{U}^\intercal_{S_{\tilde{p}+k}}W\right\| \geq \sigma\sqrt{(\tilde{p}+k)+2\sqrt{(\tilde{p}+k)t'}+2t'}\right] \leq (\frac{pe}{\tilde{p}+k})^{\tilde{p}+k}\exp(-t')$$
$$\leq p^{\tilde{p}+k}\exp(-t').$$

Letting $t' = (\tilde{p}+k)t$, we obtain

$$\mathbb{P}\left[\max_{S_{\tilde{p}+k}:|S_{\tilde{p}+k}|=(\tilde{p}+k)}\left\|\tilde{U}^\intercal_{S_{\tilde{p}+k}}W\right\| \geq \sigma\sqrt{\tilde{p}+k}\cdot\sqrt{1+2\sqrt{t}+2t}\right]$$
$$\leq p^{\tilde{p}+k}\exp\left(-(\tilde{p}+k)t\right).$$

Thus $p^{\tilde{p}+k}\exp(-(\tilde{p}+k)t) \geq \mathbb{P}[\mathscr{E}_2(t)^c]$, which implies that

$$p^{\tilde{p}+k}\exp(-(\tilde{p}+k)t \geq \mathbb{P}[\|\beta^*-\beta^{true}\|_0 = \tilde{p}+k] \quad \forall k\in\mathbb{Z}: 1\leq k\leq p-\tilde{p}.$$

With this, we can solve for our desired value

$$\mathbb{P}\left[\|\beta^*-\beta^{true}\|_0 \leq \tilde{p}\right] = 1-\mathbb{P}\left[\|\beta^*-\beta^{true}\|_0 \geq \tilde{p}+1\right] = 1-\sum_{k=1}^{p-\tilde{p}}\mathbb{P}\left[\|\beta^*-\beta^{true}\|_0 = \tilde{p}+k\right]$$
$$\geq 1-\sum_{k=1}^{p-\tilde{p}}\exp\left((\tilde{p}+k)(\ln p - t)\right)$$
$$= 1-\exp(-(\tilde{p}+1)(t-\ln p))\cdot\frac{1-\exp(-(p-\tilde{p})(t-\ln p))}{1-\exp(-t+\ln p)},$$

which is the desired result. □

**Lemma A.6.** *Consider an arbitrary $S^3$ONC solution $\beta^*$ to (1.1) with MCP. Let Assumptions (A1) and (A3) with $\tilde{p}^* \geq \tilde{p}$ hold. Assume the satisfaction of $\|\beta^*-\beta^{true}\|_0 \leq \tilde{p}$ and Event $\mathscr{E}_a(\tilde{p}) := \{\frac{1}{n}\|X(\beta^*-\beta^{true})\|^2 \leq \frac{8\sigma^2}{b_l^2 n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t)+\frac{8}{b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}(|\mathscr{S}|-\|\beta^*\|_0), P_\lambda(a\lambda)\cdot(|\mathscr{S}|-\|\beta^*\|_0)+\Gamma\}$. Assume that the sub-optimality gap satisfies $\Gamma < P_\lambda(a\lambda)-\frac{\sigma^2}{b_l n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t)$.*

*If the minimum signal strength satisfies*

$$\|\beta_{\mathscr{S}}^{true}\|_{\min} > \sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{r_{\tilde{p}} b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathscr{S}|, P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\}},$$

*then* $\beta^*$ *is the oracle solution to* 1.1*. If, in addition,* $\|\beta^{opt}-\beta^{true}\| \le \tilde{p}$ *and*

$$\mathscr{E}_b(\tilde{p}) := \{\frac{1}{n}\left\|X(\beta^{opt}-\beta^{true})\right\|^2 \le \frac{8\sigma^2}{b_l^2 n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t)$$
$$+\frac{8}{b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}(|\mathscr{S}|-\|\beta^*\|_0), P_\lambda(a\lambda)\cdot(|\mathscr{S}|-\|\beta^*\|_0)+\Gamma\}\},$$

*then* $\beta^*$ *is both the oracle solution and the global solution to* (1.1)*.*

*Proof.* First, let us denote $\beta^*-\beta^{true}=\delta^*$. We start by combining $\mathscr{E}_\alpha(\tilde{p})$ and (A3) iii, which is possible due to our assumption $\|\beta^*-\beta^{true}\|_0 \le \tilde{p}$. This gives us

$$\frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{b_l}\min\left\{\frac{\lambda^2}{r_{\tilde{p}}}(|\mathscr{S}|-\|\beta^*\|_0), P_\lambda(a\lambda)\cdot(|\mathscr{S}|-\|\beta^*\|_0)+\Gamma\right\}$$
$$\ge \frac{1}{n}\|X\delta^*\|^2 \ge r_{\tilde{p}}\|\delta^*\|^2 \ a.s. \tag{A.2}$$

If we relax $|\mathscr{S}|-\left\|\beta_{\mathscr{S}}^*\right\|_0$ to just $|\mathscr{S}|$, the definition of $\delta^*$ and note that $\|\delta^*\| \ge \left\|\delta_j^*\right\|$, we can obtain

$$\sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{r_{\tilde{p}} b_l}\min\left\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathscr{S}|, P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\right\}}$$
$$\ge \left\|\beta_j^*-\beta_j^{true}\right\| \ge |\beta_j^{true}|-|\beta_j^*|,$$

almost surely. From this, we can bound $|\beta_j^*|$ by using the square root term and $|\beta_j^{true}|$, so we know that if $|\beta_j^{true}|-\sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t)+\frac{8}{r_{\tilde{p}} b_l}\min\left\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathscr{S}|, P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\right\}} > 0$, then $|\beta_j^*|>0$. From this, we can obtain

$$\|\beta_{\mathscr{S}}^*\|_0 \ge \sum_{j\in\mathscr{S}} \mathbb{I}\left(|\beta_j^{true}|-\sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{r_{\tilde{p}} b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathscr{S}|, P_\lambda(a\lambda)|\mathscr{S}|+\Gamma\}} > 0\right)$$

almost surely. We can then combine this with our minimum signal strength assumption to get $\left\|\beta_{\mathscr{S}}^*\right\|_0 = |\mathscr{S}|$ a.s. We can combine this with (A.2), by focusing on the second part of the minimum term and noting the right side is always positive, to get

$$\frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{b_l}(-P_\lambda(a\lambda)\|\beta_{\mathscr{S}^c}^*\|_0+\Gamma) \ge 0 \ a.s.$$

which can be simplified into $\frac{\sigma^2}{b_l n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t)+\Gamma \ge P_\lambda(a\lambda)\left\|\beta_{\mathscr{S}^c}^*\right\|_0$ a.s. Thus it can be seen that if $P_\lambda(a\lambda) > \frac{\sigma^2}{b_l n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t)+\Gamma$, then $1 > \left\|\beta_{\mathscr{S}^c}^*\right\|_0 = 0$. This is satisfied by the assumption that $P_\lambda(a\lambda)-\frac{\sigma^2}{b_l n}(\tilde{p}+2\sqrt{\tilde{p}t}+2t) > \Gamma$.

Finally, because $\beta^*$ is an $S^3ONC$ solution, it has to satisfy FONC. Per 2.2, this means that $\beta^* \in \arg\inf\left\{\frac{1}{n}\sum_{i\in\mathscr{N}}\ell(\beta,x_i,y_i)+\sum_{j\in\mathscr{P}}P_\lambda'(|\beta_j^*|)|\beta_j| : \beta\in\mathbb{R}^p\right\}$. Due to Lemma A.1 we know that the penalty term goes to 0 almost surely since either $\beta_j^* = 0$ or $P'(|\beta_j^*|) = P'(|a\lambda|) = 0$ with probability one. Further we know that $\beta_j^* = 0$ for all $j \in \mathscr{S}^c$. Thus

$$\beta^* \in \arg\inf \left\{ \frac{1}{n} \sum_{i \in \mathscr{N}} \ell(\beta, x_i, y_i) : \beta \in \mathbb{R}^p, \beta_j = 0, \forall j \in \mathscr{S}^c \right\} \ a.s.$$

Given that the expression on the right is the definition of the oracle solution, we have shown the first result.

Next, we consider $\beta^{opt}$, which is the global optimal solution to (1.1). Given that the $S^3ONC$ conditions are necessary, $\beta^{opt}$ must be an $S^3ONC$ solution. With this fact and the assumption of $\mathscr{E}_b(\tilde{p})$, we have the same set of assumptions for $\beta^{opt}$ as we had for $\beta^*$. Thus the same sequence of arguments can be used to show that

$$\beta^{opt} \in \arg\inf \left\{ \frac{1}{n} \sum_{i \in \mathscr{N}} \ell(\beta, x_i, y_i) : \beta \in \mathbb{R}^p, \beta_j = 0, \forall j \in \mathscr{S}^c \right\} \ a.s.$$

Finally, per the strict convexity of our loss function as implied by (A1), we can see that the infimum of the above problem is unique. Therefore

$$\beta^* = \arg\inf \left\{ \frac{1}{n} \sum_{i \in \mathscr{N}} \ell(\beta, x_i, y_i) : \beta \in \mathbb{R}^p, \beta_j = 0, \forall j \in \mathscr{S}^c \right\} = \beta^{opt} \ a.s.$$

which is the second result. □

A.3. **Additional lemmas.**

**Lemma A.7.** *The RE condition in 2.1 implies (A3) with $r_{4s} \geq r_e > 0$ and $\tilde{p}^* \geq 4s$.*

*Proof.* As the Lemma 1 in [20]. □

**Lemma A.8.** *Let $\beta^*$ be a $S^3ONC$ solution to 1.1 If (A1) and $\mathscr{Q}(\beta^*) \leq \mathscr{Q}(\beta^{true}) + \Gamma$ hold for some $\Gamma \geq 0$, then*

$$\frac{b_l}{2n}\|X\delta^*\|^2 - \frac{1}{n}W^\intercal X\delta^*$$
$$\leq \min\left\{ \sum_{j \in S} P'_\lambda(|\beta_j^*|)|\beta_j^{true}|, \ \sum_{j \in S} P'_\lambda(|\beta_j^*|)|\beta_j^* - \beta_j^{true}|, \ P_\lambda(a\lambda)(|\mathscr{S}| - \|\beta^*\|_0) + \Gamma \right\}, \quad a.s.$$

*Proof.* First, we know that $\beta^* \in \arg\min_\beta \left\{ \sum_{i=1}^n \ell(\beta, x_i, y_i) + \sum_{j=1}^p P'_\lambda(|\beta^*|)|\beta_j| \right\}$ because the KKT conditions are the same as FONC which $\beta^*$ satisfies. This gives us

$$\sum_{i=1}^n \ell(\beta^*, x_i, y_i) + \sum_{j=1}^p P'_\lambda(|\beta^*|)|\beta_j^*| \leq \sum_{i=1}^n \ell(\beta^{true}, x_i, y_i) + \sum_{j=1}^p P'_\lambda(|\beta^*|)|\beta_j^{true}|,$$

which can be used along the same lines as the level set inequality in the derivation for A.1 to get

$$\frac{b_l}{2n}\|X\delta^*\|^2 - \frac{1}{n}W^\intercal X\delta^* \leq \sum_{j=1}^p P'_\lambda(\beta_j^*)(|\beta_j^{true}| - |\beta_j^*|).$$

The first two terms of the min function are easily obtained from this. The last term can be obtained by noting that $\beta_j^* \notin (0, a\lambda)$ for all $j$ with probability one and that $P_\lambda(a\lambda) = P_\lambda(\beta) \quad \forall \beta \geq a\lambda$. This gives us that $\frac{b_l}{2n}\|X\delta^*\|^2 - \frac{1}{n}W^\intercal X\delta^* \leq P_\lambda(a\lambda)(\mathscr{S} - \|\beta^*\|_0) + \Gamma$, which is the final term to complete the desired result.

□

**Lemma A.9.** *Consider a subgaussian ñ-dimensional random vector $\tilde{W} \in \mathbb{R}^{\tilde{n}}$ as defined in (A2). Then, for any $V \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$ and $\Sigma_v = V^\intercal V$, $\mathbb{P}[\left\|V\tilde{W}\right\|^2 \leq \sigma^2 \cdot (Tr(\Sigma_v) + 2\sqrt{Tr(\Sigma_v^2)t} + 2\|\Sigma_v\|t)] \geq 1 - \exp(-t)$ for any $t > 0$, where $Tr(\cdot)$ denotes the trace of a matrix.*

*Proof.* We apply [14, Theorem 2.1] where our $\tilde{W}, V$ and $\Sigma_v$ are equivalent to their $x$, $A$, and $\Sigma$. Note their expectation condition is equivalent to our (A2) with $\mu = \mathbb{E}[W] = 0$. This gives us that, for all $t > 0$,

$$\mathbb{P}\Big[\|VW\|^2 > \sigma^2 \cdot (Tr(\Sigma_v) + 2\sqrt{Tr(\Sigma_v)t} + 2\|\Sigma_v\|t) + Tr(\Sigma_v\mu\mu^\intercal)\cdot(1 + 2(\frac{\|\Sigma_v\|^2}{Tr(\Sigma_v^2)}t)^{1/2})\Big] \leq \exp(-t).$$

Given that $\mu = 0$, the term involving $Tr(\Sigma_v\mu\mu^\intercal)$ goes to zero. Thus the statement in A.9 can be obtained by taking the complement of the probability bound. □