Submitted to Statistical Science

# A Unifying Framework of High-Dimensional Sparse Estimation with Difference-of-Convex (DC) Regularizations

**Shanshan Cao*, Xiaoming Huo* and Jong-Shi Pang†**

*Abstract.* Under the linear regression framework, we study the variable selection problem when the underlying model is assumed to have a small number of nonzero coefficients. Non-convex penalties in specific forms are well-studied in the literature for sparse estimation. A recent work [1] has pointed out that nearly all existing non-convex penalties can be represented as difference-of-convex (DC) functions, which are the difference of two convex functions, while itself may not be convex. There is a large existing literature on optimization problems when their objectives and/or constraints involve DC functions. Efficient numerical solutions have been proposed. Under the DC framework, directional-stationary (d-stationary) solutions are considered, and they are usually not unique. In this paper, we show that under some mild conditions, a certain subset of d-stationary solutions in an optimization problem (with a DC objective) has some ideal statistical properties: namely, asymptotic estimation consistency, asymptotic model selection consistency, asymptotic efficiency. Our assumptions are either weaker than or comparable with those conditions that have been adopted in other existing works. This work shows that DC is a nice framework to offer a unified approach to these existing works where non-convex penalties are involved. Our work bridges the communities of optimization and statistics.

*MSC 2010 subject classifications:* Primary 62J05, 62F12; secondary 62J12.
*Key words and phrases:* (Generalized) linear regression, high-dimensional sparse estimation, nonconvex regularization, difference of convex (DC) functions, DC algorithms, asymptotic optimality, model selection consistency.

## 1. INTRODUCTION

Sparse estimation under a linear regression model is a fundamental and classical problem in statistics. It continues to be highly active in the high-dimensional regime when the underlying parameter is believed to be sparse. Properties on the

*Georgia Institute of Technoloty
†University of South California

resulting estimators have been extensively studied with different penalties of the sparsity in [39, 30, 8, 9, 40, 14, 34, 37, 38], etc. However, most existing works focus on the properties of a specific solution to the possibly nonconvex objective function, which is used to derive a sparse estimation of the unknown parameter. The stationary solutions of other kinds might also be of interest and possess satisfying properties, such as the desired asymptotic estimation consistency, asymptotic model selection consistency, asymptotic efficiency. A unified framework for the penalized high-dimensional sparse estimation and the relation to a subfield of optimization problems, namely, the difference-of-convex (DC) programming are missing in the literature. We establish such a connection in this paper.

### 1.1 Sparsity induced penalties

We first present the formulation of high-dimensional sparse estimation in linear regression setting using sparsity-induced penalties. Consider observations $(y_1, x_1)$, $(y_2, x_2)$, ..., $(y_n, x_n)$, where we have the response $y_i \in \mathbb{R}$ and the predictor $x_i \in \mathbb{R}^p$ satisfy

$$y_i = \beta^{*T} x_i + \epsilon_i.$$

Here, $\beta^{*T}$ is the transpose of the vector $\beta^* \in \mathbb{R}^p$, which is the true however unknown underlying parameter to be estimated. We further assume that noises $\epsilon_i$'s are independently distributed, with 0 mean and equal variance $\sigma^2$ (which can be a sub-Gaussian distribution with variance parameter $\sigma^2$), and are independent of $x_i$'s. The above model is commonly written in the following matrix form:

$$y = X\beta^* + \epsilon, \tag{1.1}$$

where the vector $y = (y_1, \cdots, y_n)^T \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is the model matrix with rows being individual predictors, $x_1^T, \cdots, x_n^T$, and the random vector $\epsilon$ contains the noises.

In the high-dimensional regime where the number of the parameters, denoted by $p$, exceeds the sample size, denoted by $n$, one of the most important methods (according to many works such as [5, 3, 4]) is to estimate the parameter by using the LASSO [27] approach. It is interesting to note that a mathematically equivalent approach was proposed in [6] around the same time in the computational and applied mathematics literature. LASSO is defined through solving the following convex optimization problem:

$$\hat{\beta}^{lasso}(y, X; \lambda) = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{1.2}$$

The first term in the above objective is the goodness-of-fit measure (a.k.a., the residual sum of squares) in the linear regression model (1.1). The second term in the objective is a penalty function, which is the sum of absolute values: $\sum_{i=1}^k \lambda|\beta_i|$. We can further write the penalty in a more general form $\sum_{i=1}^k P_\lambda(\beta_i)$, where the univariate function $P_\lambda(x)$ takes the form $P_\lambda(x) = \lambda|x|$ in the LASSO approach. Many existing works, including [39] and [30] and others, have proved that with high probability (i.e., the probability goes to 1 as sample size goes to infinity), under some conditions on the design matrix and the choices for $\lambda$, the LASSO will be able to find the right signed support for the unknown parameter $\beta^*$. The cases that have been studied include (1) when the matrix can be fixed or random,

(2) the dimension of the unknown parameter is fixed or goes to infinity as the sample size increases, and (3) other interesting situations.

Despite the success of obtaining sparse estimators by using LASSO, it is also well known that the resulting estimator is biased. This can be readily seen by considering the special scenario, where the design matrix $X$ is orthonormal, consequently, the $L_1$ penalty leads to a soft-thresholding solution, which is biased from the true parameter $\beta^*$. The de-biasing procedure has been studied in [35, 15, 29, 16]. In the present paper, we decide to focus on the regularization (i.e., adding a penalty function) approach, partially because the de-biasing approach may require solving multiple optimization problems, therefore could be computationally disadvantageous. At the same time, we may explore other algorithmic-design approaches in the future.

An effective extension of the LASSO estimator is to replace the penalty function $P_\lambda(x)$ in (1.2) into some folded concave functions, which are non-convex. Some representitive works include SCAD [8, 9], MCP [34], adaptive LASSO [40], capped-$l_1$ [38], together with others. In general, this leads to an NP-hard problem; therefore no polynomial-time algorithm is known in finding the global optimal solution. Specifically, SCAD is proposed in [8, 9], in order to debias the estimation when the parameter is numerically relatively large, which gives a constant penalty as the parameter is large enough. Adaptive LASSO is studied in [40, 14], which is motivated by the fact that in the orthogonal design, the bias of the parameter estimation is approximately $\lambda$ in LASSO. The authors suggest giving different sizes of penalties to different parameters so that the variables with large coefficients have smaller weights in the $\ell_1$ penalty (depending on some consistent estimator $\hat{\beta}$ of $\beta^*$). Then they can reduce the estimation bias of the lasso while retaining its sparsity property. MCP is proposed in [34], which also gives a constant penalty as the parameter is large enough. Capped-$l_1$ in [37, 38] gives a penalty of truncated $l_1$ penalty to ensure a constant penalty when the estimation is large. Consistency results, including measuring the squared distance of the estimation, prediction, signed support recovery, for the previously mentioned formulations can be found in the original works.

### 1.2 Difference-of-convex (DC) unified penalties

Recently, it is pointed out by [1] that all the previously listed penalties can be written in a unified DC form. Especially, the first term is the $\ell_1$ penalty $\|\beta\|_1$. This leads to a DC formulation; i.e., solving the penalized least square problem with a generalized DC penalty function. In [1], they prove under some strong assumptions (strong convexity of the loss functions) that the d-stationary solutions found by a standard DC algorithm (i.e., DCA) are the global minimum. This result might not be surprising because, under their assumptions, the objective function (the DC-penalized loss functions) in fact can be strongly convex, which makes the solution unique. They also prove that the $\ell_0$ norm of the d-stationary solution has an upper bound, which doesn't shed light on the support recovery property. Our paper is inspired by [1]. Compared to [1], we relax the assumptions on strong convexity for the loss function and prove the existence of a class of d-stationary solutions, which have the oracle properties in the linear regression scenario. Our result indicates that the assumption on the strong convexity of the loss function within the domain is not necessary. In addition, the aforemen-

tioned work has an applied mathematics focus – their $\ell_0$ norm result does not imply statistical properties of the d-stationary solutions. In statistical literature, the distance between the d-stationary solution and the assumed ground truth is considered. Our result will be more formulated towards the statistical properties of the d-stationary solutions: namely model estimation consistency, asymptotic convergence rate in estimation, and model selection consistency. Despite the difference, it is interesting to point out that both works will require the restricted convexity assumption, which is assumed in nearly all related work. Besides, we also generalize the results to DC penalized general loss functions.

The use of DC functions offers a general framework for non-convex regularization. Some special cases are discussed in [18] and [32], although they don't explicitly mention the DC functions in their work. The first work [18] assumes that the penalty function $p_\lambda(\beta)$ is separable in parameter $\beta$ and each of the univariate penalty can be written as the difference of a convex function $p_{\lambda,\mu}(t)$ and a quadratic function $\frac{\mu}{2}t^2$, where $\mu$ is a known positive constant. Therefore one has $p_\lambda(t) = p_{\lambda,\mu}(t) - \frac{\mu}{2}t^2$. They restrict the feasible region to a bounded region containing the ground truth $\beta^*$. Under certain regularity conditions on the penalty, including differentiability, and restrictive strong convexity of the loss function, they give the optimal upper bound for the estimation error as well as for prediction error. Their assumption includes the popularly studied penalties like SCAD and MCP. On the other hand, They don't have results on the support recovery and they purposely eliminated possible stationary solutions outside the bounded feasible region they defined. The second work [32] mainly assumes the restricted strong convexity of the penalties and the loss functions. They mainly discuss the elliptical design regression, least-square loss, and logistic loss with SCAD, MCP penalties, which can be written as the summation of the $\ell_1$ penalty $\|\beta\|_1$ and a concave function $q_\lambda(\beta)$ with proper bound in the concavity. They argue that the local quadratic approximation algorithm they provided converges to a unique local minimum, which enjoys the oracle properties as if you've already known the support for the true parameter. They prove the estimation error upper bound. And they are able to prove the support recovery results for a linear regression model with the least square loss function. Both are under the assumption that the concavity of the function $q_\lambda(\beta)$ is bounded.

Both works [18, 32] assume the decreasing first-order derivative of the penalty function on the nonnegative real line, which is necessary for the unbiasedness for estimation of larger $\beta$. They both restrict the penalties such that the objective function is strongly convex within some regions where the local optimal solutions, as well as the true unknown parameters, are in the given convex set.

While in the current work, we solve the unconstrained problem and prove the asymptotic convergence results of the estimation for a class of local d-stationary solutions without using the assumption of the bounded concavity of the $q_\lambda(\cdot)$ function and constant penalty when the parameter is large enough, which allows us to include other penalties such as transformed $\ell_1$ [36, 19] and logarithmic [20], into our analysis. Equipped with the bounded convexity assumption, we further prove that the support recovery consistency for the class of d-stationary solutions we find near the ground truth.

From the computation perspective, there is rich literature on solving the penalized (also known as regularized) problem numerically. For example, Local Linear

Approximation (LLA) in [41] prove that a one-step estimator from LLA performs well in SCAD with penalized likelihood estimation. They also prove the asymptotic normality under some regularity conditions of the Fisher Information matrix. In this paper, we would like to explore the relationship between LLA and the popular DC Algorithm (DCA), which is often used in DC programming. It turns out that all the above-mentioned algorithms are special cases of DCA.

This paper builds a bridge between optimization, where people focused on solving the optimization problem efficiently, and statistics, where people mainly focused on inference (finding the estimation). The link here would be the DC programming and DCA. The DC programming enables us to generalize the classical penalized likelihood function to the DC penalized likelihood function, while DCA provides us efficient algorithms to solve the corresponding numerical problems. Borrowing strength from existing literature enables us to solve the optimization problem efficiently with convergence guarantees. We unify the existing algorithms in the literature for finding the local minima of non-convex optimization problems under the DCA framework.

### 1.3 Notations

For a real number $q \in [1, +\infty)$, the $\ell_q$ norm of a vector $\beta = (\beta_1, \beta_2, \cdots, \beta_p) \in \mathbb{R}^p$ is defined as $\|\beta\|_q = (\sum_{i=1}^p \beta_i^q)^{1/q}$. Specially, the $\ell_\infty$ norm is defined as $\|\beta\|_\infty = \max_{1\le i\le p}\{|\beta_i|\}$. The $\ell_0$ norm is defined as $\|\beta\|_0 = \text{card}\{\text{supp}(\beta)\}$, where we have $\text{supp}(\beta) = \{i : \beta_i \neq 0\}$ and $\text{card}\{\cdot\}$ is the cardinality of the set. We denote the cardinality of a set $S$ by $|S|$ and its complement by $S^c$. For $\beta = (\beta_1, \beta_2, \cdots, \beta_p) \in \mathbb{R}^p$, we let $\beta_S$ denote the sub-vector (of $\beta$) whose elements correspond to the set $S$; we let $X_S$ denote the sub-matrix (of $X$) whose columns indices correspond to the set $S$.

### 1.4 Organization

The rest of the paper is organized as follows. We review the basic properties of the DC programming and the DC functions in Section 2. We form a penalized least square problem with a generalized DC-penalty in Section 3. Under mild assumptions, we prove in Section 4 that a set of the d-stationary solutions are close to the ground truth. Furthermore, they are also support recovery consistent. We also extend our results to generalized loss functions, such as logistic loss, etc., in Section 5. We provide the connections among popular exiting algorithms in DC programming and statistics estimation with non-convex objective functions in Section 6. We finally conclude this work in Section 7. When possible, the technical proofs are relegated to the Appendix, which is included in an online supplementary file.

## 2. DC FUNCTIONS AND RELATED BASIC PROPERTIES

In this section, we first provide the necessary background as well as a definition of the Difference-of-Convex (DC) functions, before proceeding to our formulation (Section 3) and the main results (Section 4 and 5). We present the definition of DC functions and their known properties in Section 2.1. The directional derivatives are reviewed in Section 2.2. The class of DC functions that we are particularly interested in are reviewed in Section 2.3. We then define and study the directional stationarity (d-stationarity) that we focus on in this work in Section 2.4.

## 2.1 DC programming and DC functions

DC programming is pervasive nowadays in both optimization and statistics. The DC program has been introduced in the literature from the 1950's [2]. The paper [11] gives a wealth of basic properties of the DC functions, which are the functions that are used in the objectives and constraints in the DC programming. In particular, the DC programming has been intensively studied in the field of optimization in the early 1980's [13, 12, 25, 28]. The following gives a formal definition for a DC function.

DEFINITION 2.1. *A function, $p(x)$, is called a difference-of-convex (DC) function if we have*

$$p(x) = g(x) - h(x)$$

*where both $g(x)$ and $h(x)$ are convex functions.*

There are many known results regarding DC functions and DC programming. We summarize these properties of DC programming from the literature in Appendix A.

## 2.2 Directional derivative

To enable our description, we define the *directional derivative* in the following.

DEFINITION 2.2. *For a function $Q(\beta)$ that is defined on $\Omega$ where $\beta \in \Omega \subset \mathbb{R}$ or $\mathbb{R}^p$, for $\beta_0$, $\beta_1 \in \Omega$, the directional derivative at $\beta_0$ in the direction of $\beta_1 - \beta_0$ is defined as follows:*

$$Q'(\beta_0; \beta_1 - \beta_0) = \lim_{\tau \to 0+} \frac{Q(\beta_0 + \tau(\beta_1 - \beta_0)) - Q(\beta_0)}{\tau},$$

*where $\tau \in \mathbb{R}_+$.*

To compute the directional derivative, when a function $P(\beta)$ is differentiable in $\mathbb{R}^p$, the directional derivative with regard to $\beta$ at $\beta_0$ is given below:

$$P'(\beta_0; \beta - \beta_0) = \langle \nabla P(\beta_0), \beta - \beta_0 \rangle,$$

where $\nabla P(\beta_0)$ is the gradient of the function $P(\beta)$ at $\beta_0$, and $\langle \cdot, \cdot \rangle$ represents the inner product. When the function $P(\beta)$ is non-differentiable however convex in $\mathbb{R}^p$, given its sub-gradient set $\partial P(\beta_0)$ at $\beta_0$, the directional derivative with regard to $\beta$ at $\beta_0$ can be written as [23, Theorem 23.4]:

$$P'(\beta_0; \beta - \beta_0) = \max_{v \in \partial P(\beta_0)} \langle v, \beta - \beta_0 \rangle.$$

The recent papers [21] and [1] discuss the pervasiveness of the existence of the DC functions as well as its relation to statistics. Specifically, they have the following results considering the pervasiveness of DC functions.

LEMMA 2.1. *[21, Proposition 1] For any univariate continuous concave function $p$ that is defined on $\mathbb{R}_+$, the composite function $\theta(|t|) = p(|t|)$ is Difference-of-Convex (DC) on $\mathbb{R}$ if and only if $p'(0;+)$, the directional derivative of $p(t)$ at 0, which can be written as follows,*

$$p'(0;+) = \lim_{\tau \to 0+} \frac{p(\tau) - p(0)}{\tau},$$

*exists and is finite.*

The above lemma leads to the realization that nearly all the folded-concave penalties [8, 9, 34, 37, 38] in the sparsity study nowadays belong to the DC family. We articulate details in the subsequent subsection.

### 2.3 Relation to statistics

Based on the definition of the DC functions (Definition 2.1), it has been realized (e.g., [31],[32], and [1]) that many well-studied penalties, such as SCAD, MCP, capped $\ell_1$, transformed $\ell_1$, logarithmic, can be written as DC functions. We articulate details in the following. Throughout this paper, we consider penalties $P(\beta)$ that are separable in the (potentially multivariate) parameter $\beta$: $P(\beta) = \sum_{i=1}^{p} p(\beta_i)$, where $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$. We argue that function $p(x)$ is a DC function for the popular existing penalties in the literature; that is, we have $p(x) = g(x) - h(x)$, where functions $g$ and $h$ are convex. More specifically, for the penalties that are of interests to us and are widely used in statistical inference, we always have $g(x) = |x|$ (or $g(x) = \lambda|x|$, when an algorithmic parameter $\lambda$ is involved), however the function $h(x)$ varies for different penalties.

In the following, we describe how the popular penalty functions $p(x)$ that was mentioned previously can be decomposed as DC functions. For simplicity, without loss of generality, we set the tuning parameter as $\lambda = 1$.

1. In SCAD [8, 9], we have

$$p^{SCAD}(t) = |t| - h_\gamma^{SCAD}(t),$$

where

$$h_\gamma^{SCAD}(t) = \begin{cases} 0 & |t| \le 1 \\ \frac{(|t|-1)^2}{2(\gamma-1)} & 1 \le |t| \le \gamma \\ |t| - \frac{(\gamma+1)}{2} & |t| \ge \gamma \end{cases}$$

and the function $h_\gamma^{SCAD}(t)$ can be verified to be convex on the positive real line and have a continuous first order derivative.

2. In MCP [34], we have

$$p^{MCP}(t) = |t| - h_\gamma^{MCP}(t),$$

where

$$h_\gamma^{MCP}(t) = \begin{cases} \frac{|t|^2}{2\gamma} & |t| \le \gamma \\ |t| - \frac{\gamma}{2} & |t| \ge \gamma \end{cases}$$

and the function $h_\gamma^{MCP}(t)$ can be verified to be convex on the positive real line and have a continuous first order derivative.

3. In Capped $\ell_1$ [37, 38], we have

$$p^{\text{capped } \ell_1}(t) = |t| - \max\left\{0, \frac{2t}{\gamma} - 1, -\frac{2t}{\gamma} - 1\right\},$$

where one can verify that both $|t|$ and $\max\left\{0, \frac{2t}{\gamma} - 1, -\frac{2t}{\gamma} - 1\right\}$ are convex functions of $t$.

4. In Transformed $\ell_1$ [36, 19], we have

$$p^{\text{Transformed } \ell_1}(t) = \frac{a+1}{a}|t| - \left(\frac{a+1}{a}|t| - \frac{(a+1)|t|}{a+|t|}\right),$$

where one can verify that both $\frac{a+1}{a}|t|$ and $\left(\frac{a+1}{a}|t| - \frac{(a+1)|t|}{a+|t|}\right)$ are convex functions.

5. In Logarithmic [20], we have

$$p^{Log}(t) = \frac{1}{\epsilon}|t| - \left(\frac{|t|}{\epsilon} - \log(|t|+\epsilon) + \log\epsilon\right),$$

where $\epsilon$ is a given positive scalar; similarly, one can verify that both $\frac{\lambda}{\epsilon}|t|$ and $\left(\frac{|t|}{\epsilon} - \log(|t|+\epsilon) + \log\epsilon\right)$ are convex functions.

Table 1 summarizes the aforementioned DC decompositions. The penalty functions and their first order derivatives in terms of $|t|$ are plotted in Figure 1. One common point that most of the above penalties share is that their first order derivative goes to 0 as $|t| \to \infty$.

| Penalty | $p(t)$ | $g(t)$ | $h(t)$ |
|---|---|---|---|
| $\ell_1$ | $\lvert t\rvert$ | $\lvert t\rvert$ | 0 |
| SCAD | $\int_0^{\lvert t\rvert} 1 \wedge (1 - \frac{x-1}{\gamma-1})_+ dx$ | $\lvert t\rvert$ | $\frac{(\lvert t\rvert-1)^2}{2(\gamma-1)} I\{1 < \lvert t\rvert < \gamma\}$ $+(\lvert t\rvert - \frac{\gamma+1}{2}) I\{\lvert t\rvert \geq \gamma\}$ |
| MCP | $\int_0^{\lvert t\rvert} (1 - \frac{x}{\gamma})_+ dx$ | $\lvert t\rvert$ | $(\lvert t\rvert - \gamma/2) I\{\lvert t\rvert > \gamma\}$ $+\frac{t^2}{2\gamma} I\{\lvert t\rvert < \gamma\}$ |
| Capped-$\ell_1$ | $\min\{\gamma/2, \lvert t\rvert\}$ | $\lvert t\rvert$ | $\max\{0, \frac{2t}{\gamma} - 1\}$ |
| Transformed $\ell_1$ | $\frac{(a+1)\lvert t\rvert}{a+\lvert t\rvert}$ | $\frac{a+1}{a}\lvert t\rvert$ | $\frac{a+1}{a}\lvert t\rvert - \frac{(a+1)\lvert t\rvert}{a+\lvert t\rvert}$ |
| Logarithmic | $\log(\lvert t\rvert+\epsilon) - \log\epsilon$ | $\frac{\lvert t\rvert}{\epsilon}$ | $\frac{\lvert t\rvert}{\epsilon} - \log(\lvert t\rvert+\epsilon) + \log\epsilon$ |

TABLE 1

*The DC decompositions of some well-known penalty functions in statistical inference. The first column contains the name of the methodology. The second column describes the penalty function. The last two columns present the corresponding two convex functional components (i.e., $g$ and $h$) in the DC decomposition: $p(t) = g(t) - h(t)$.*

Although there are many other different DC decompositions, this one has the advantage of easy interpretation and correcting the penalty of LASSO, which in some sense, does a debiasing job for the LASSO estimator (by choosing $h_\lambda(t)$ to be linear with slope $\lambda$ when $|t|$ is large enough). It also shares common features with popular penalties in literature, like SCAD, MCP, capped $\ell_1$ penalties where the penalty is close to or equal to $\ell_1$ penalty when the solution is around the origin. Furthermore, the resulting penalty $p(x) = g(x) - h(x)$ is singular at 0, which makes it possible to achieve the condition of sparsity and continuity of the estimation [8]. The results in later sections can be applied to SCAD, MCP, capped-$\ell_1$, and many others.

## 2.4 Directional stationary points

Another important definition in this paper is the d(irectional)-stationary point, which is used to describe the set of stationary solutions we are interested in this paper. We give the definition of the d-stationary point in the following.

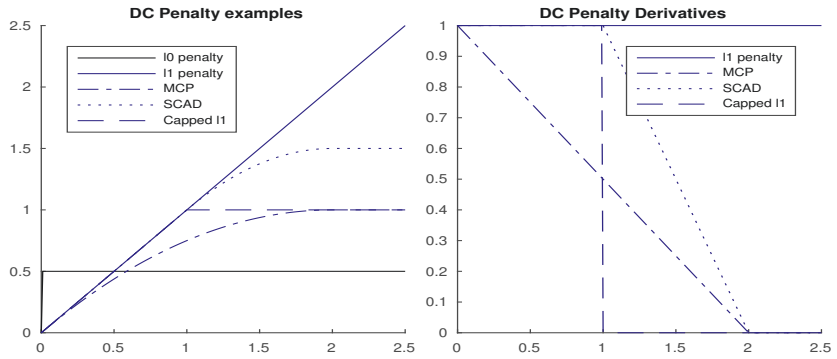


FIG 1. *Examples of famous DC penalties and their derivatives in the literature: the $\ell_0$ penalty is plotted in the solid black line; the $\ell_1$ penalty function is plotted in the solid blue line; the MCP penalty is plotted in the dotted blue line; the MCP penalty is plotted in the dash-dot line; the Capped-$\ell_1$ penalty is plotted in the dashed blue line.*

DEFINITION 2.3 (*d-stationary point*). *A vector $\hat{\beta} \in \Omega$ is a d-stationary point to a function $Q(\beta)$ if the directional derivative, which is defined as*

$$Q'(\hat{\beta}; \beta - \hat{\beta}) = \lim_{\tau \to 0+} \frac{Q(\hat{\beta} + \tau(\beta - \hat{\beta})) - Q(\hat{\beta})}{\tau},$$

*satisfies $Q'(\hat{\beta}; \beta - \hat{\beta}) \geq 0$ for all $\beta \in \Omega$.*

We prove later that under some proper conditions on the penalty function as well as on the design matrix (which in some general cases are about the loss functions), a set of d-stationary solutions to the DC programming problem are $\sqrt{n}$ consistent estimators with high probability (Theorem 4.1). Under further conditions, it also recovers the true support in the unknown parameter with high probability (Theorem 4.3).

A motivation of choosing the directional stationary solutions (which are the directional stationary points in the corresponding optimization problem) rather than stationary solutions of other kinds, such as that of a critical point for DC programs, is provided in [1]. The authors argue that the directional stationary solutions are the sharpest kind among stationary solutions of other kinds in the sense, a directional stationary solution must be stationary according to other definitions of stationarity. In the above sense, the d-stationary solutions possess minimizing properties that are not in general satisfied by stationary solutions of other kinds. We refer to the original paper for a more detailed discussion.

## 3. FORMULATION AND ASSUMPTIONS

In this section, we first give our detailed formulation in Section 3.1. We discuss the scale invariant properties of the formulation with some specific form of penalties in Section 3.2. We then list the assumptions on the penalty functions in Section 3.3, on the d-stationary solutions in Section 3.4, on the design matrix for our analytical study and corresponding justifications in Section 3.5.

### 3.1 Formulation

We present our problem formulation in the following. Recall that the widely-known SCAD [8] and MCP [34] choose their regularization term (i.e., the penalty

function) as a function in the form, $\lambda|t| - h_\lambda(t)$, where the second term $h_\lambda(t)$ has a continuous first order derivative. Motivated by the above, we propose to analyze the following parameter estimation approach:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 - h_\lambda(\beta), \tag{3.1}$$

where function $h_\lambda(\beta)$ is assumed to be convex and the model is specified in (1.1). As we have shown in Table 1, popular non-convex penalties in the literature can all be expressed in DC form. In our formulation, following the approaches in the main stream methodology, we focus on separable penalty, that is we have

$$P(\beta) = \lambda\|\beta\|_1 - h_\lambda(\beta) = \sum_{i=1}^{p} \lambda|\beta_i| - h_\lambda(\beta_i),$$

for $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$. Note that based on the context, function $h_\lambda(\cdot)$ can take both univariate and multivariate inputs. In our formulation, the univariate function $h_\lambda(\cdot)$ is assumed to be convex. Its properties are further specified later.

### 3.2 Scale-invariant property

In real world of processing data, programmers always perform rescaling on the raw data set. We can make our formulation scale-invariant by assuming the following format of the penalty function.

ASSUMPTION 3.1. $p_\lambda(t) = \lambda^2 p(\frac{t}{\lambda})$.

Suppose we scale the model in (1.1) by a scalar $c$, which leads to the following model:

$$cY = X(c\beta^*) + c\epsilon,$$

Let $F(\beta, \lambda) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 - h_\lambda(\beta)$ denote the objective function, corresponding to (1.1). Let $F(c\beta, c\lambda) = \frac{1}{2n}\|cY - X(c\beta)\|_2^2 + c\lambda\|c\beta\|_1 - h_{c\lambda}(c\beta)$ denote the objective function, corresponding to the scaled model. One can easily verify that for any given positive scalar $c$,

$$\min_{\beta\in\mathbb{R}^p} F(c\beta, c\lambda), \tag{3.2}$$

is equivalent to the original problem of $\min_{\beta\in\mathbb{R}^p} F(\beta, \lambda)$ in (3.1) with scale free penalties such as SCAD, $\ell_1$, MCP, capped-$\ell_1$, which have the common form stated in Assumption 3.1. One can verify that most of the functions in Table 1 satisfy this scale free condition except the logarithmic function and the transformed $\ell_1$. However, according to the unification DCA in Section 6, where in each iteration, by using only the linear approximation, we solve a re-weighted LASSO problem, which is scale invariant.

### 3.3 Assumptions on $h_\lambda(\cdot)$

We present the assumptions that we need in the analytical study in this section. Recall that our penalty function has the form $P(\beta) = \lambda\|\beta\|_1 - h_\lambda(\beta)$. Notice that the first term of the DC penalty is always the $\ell_1$ function. We specify our assumptions on the univariate function $h_\lambda(\beta)$, for $\beta \in \mathbb{R}$. We also require the regularity conditions on the design matrix $X$, which is articulated later.

The following assumptions are utilized in our analysis. We briefly discuss the assumptions and argue that our assumptions are equivalent or weaker to conditions in most existing work.

ASSUMPTION 3.2. *We have* $\sup_{t\in\mathbb{R}} |h'_\lambda(t)| \leq \lambda$.

ASSUMPTION 3.3. *We have* $h_\lambda(t)$ *is symmetric about 0.*

Both Assumption 3.2 on the non-negativity of the penalty and Assumption 3.3 on the symmetry of the penalty function are standard assumptions in the literature. Assumption 3.2 makes sure that the penalty function $P_\lambda(\beta_i)$ is nonnegative. In fact, we can even relax this condition to $\sup_{t\in\mathbb{R}} |h'_\lambda(t)| \leq \lambda$ as long as the first order derivative of the function $h_\lambda(t)$ is uniformly bounded in the real line.

ASSUMPTION 3.4. $h'_\lambda(t)$ *is monotonically increasing and there exist two nonnegative constants* $\eta^- \geq \eta^+ \geq 0$ *such that for any* $t_2 > t_1$*:*

$$0 \leq \eta^+ \leq \frac{h'_\lambda(t_2) - h'_\lambda(t_1)}{t_2 - t_1} \leq \eta^- \tag{3.3}$$

Regarding Assumption 3.4, the lower bound $\eta^+$ of the convexity of the function $h_\lambda(t)$ is usually assumed to be 0 in other works, such as in the SCAD and the MCP. The upper-bound of the convexity $\eta^-$ is used to control the convexity of the function $h_\lambda(t)$. If $h_\lambda(t)$ has a "lot" of convexity, we are not able to have the Restricted Strong Convexity of the objective function later. On the other hand, this can be regarded as requiring the first order derivative of $h_\lambda(t)$ to be continuous. The continuity assumption together with Assumption 3.2 and 3.6 ensure that Assumption 3.4 holds.

ASSUMPTION 3.5. *We have* $h_\lambda(0) = h'_\lambda(0) = 0$.

Assumption 3.5 is utilized to ensure the soft thresholding property of the penalty function [8], recalling that the singularity of the whole penalty function at 0.

ASSUMPTION 3.6. *For some positive* $\zeta$*, we have* $h'_\lambda(t) = \lambda$ *for all* $|t| \geq \zeta$.

Assumption 3.6 is based on the fact [8] that making sure $h'_\lambda(t) = \lambda$ for $t$ positive enough and $h'_\lambda(t) = -\lambda$ for $t$ negative enough help producing an unbiased estimator. Recall that one of the main reasons of considering a generalized version of the LASSO method is the bias of the estimation from LASSO.

Below, we make a table of the penalties discussed in Table 1 and presents the decomposition to $\lambda|t| - h_\lambda(t)$. We also listed the properties that each of the $h_\lambda(t)$ holds.

From Table 2, we can see that, SCAD and MCP penalty class satisfy all of the assumptions. While Capped-$\ell_1$ has discontinuous first order derivative, which violates the Assumption 3.4. In order to extend the theories in this work to Capped-$\ell_1$ penalty, performing smoothing around the non-differentiable point is enough. We re-scaled the linear term in Transformed $\ell_1$ to match the assumptions. $\epsilon$ in Logarithmic penalty is chosen to be 1 in the Table 2.

| Penalty | $h(t)$ | sgn(t)$h'(t)$ | Convexity measure | Assumptions |
|---|---|---|---|---|
| $\ell_1$ | 0 | 0 | 0 | All except 3.6 |
| Capped-$\ell_1$ | $\max\{0, \frac{2t}{\gamma} - 1\}$ | $\frac{2}{\gamma} I\{\lvert t\rvert > \gamma/2\}$ | $\infty$ | All except 3.4 |
| MCP | $(\lvert t\rvert - \gamma/2) I\{\lvert t\rvert > \gamma\}$ $+\frac{t^2}{2\gamma} I\{\lvert t\rvert < \gamma\}$ | $\min\{\frac{\lvert t\rvert}{\gamma}, 1\}$ | $\gamma^{-1}$ | All |
| SCAD | $\frac{(\lvert t\rvert-1)^2}{2(\gamma-1)} I\{1 < \lvert t\rvert < \gamma\}$ $+(\lvert t\rvert - \frac{\gamma+1}{2}) I\{\lvert t\rvert \geq \gamma\}$ | $\frac{\lvert t\rvert-1}{\gamma-1} I\{1 < \lvert t\rvert < \gamma\}$ $+I\{\lvert t\rvert \geq \gamma\}$ | $(\gamma-1)^{-1}$ | All |
| Transformed $\ell_1$ | $\lvert t\rvert - \frac{a\lvert t\rvert}{a+\lvert t\rvert}$ | $\frac{(a+\lvert t\rvert)^2 - a^2}{(a+\lvert t\rvert)^2}$ | $\frac{2}{a}$ | All except 3.6 |
| Logarithmic | $\lvert t\rvert - \log(\lvert t\rvert + 1)$ | $\frac{\lvert t\rvert}{\lvert t\rvert+1}$ | 1 | All except 3.6 |

TABLE 2

*The penalties in the sparse estimation literature and their properties with respect to our assumptions. The first column gives the name of the methods. The second column presents the h-function, which is the second component in the DC decomposition ($p = g - h$) of the corresponding penalty function. The third column contains their first derivatives on the positive axe. This is to verify Assumption 3.2. The fourth column computes for the quantities that are raised in Assumption 3.4. The last column summarizes the assumptions that the corresponding penalty satisfies.*

### 3.4 Assumptions on the d-stationary solution

We list some assumptions on the d-stationary solutions.

ASSUMPTION 3.7. *Let $\beta^*$ be the unknown true parameter, $\hat{\beta}$ be the d-stationary solution to problem (3.1), which satisfies the following condition:*

$$\frac{1}{n} X_j^T X(\beta - \hat{\beta}) sign(\hat{\beta}_j) \geq c\lambda, \text{ for all } j \in S^c, \ \beta = \beta^* \text{ with } c \in (0, 1).$$

REMARK 3.1. *The above Assumption 3.7 is no stronger than the assumptions used in LASSO estimator [30] to prove the desired statistical properties in later Sections. We show below that in the proof of LASSO estimator, it corresponds to when $c = \frac{1}{2}$. Let $\hat{\beta}^{lasso} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$. Recall that in [30], by the First Order Condition (FOC) at $\hat{\beta}^{lasso}$, we have*

$$-\frac{1}{n} X^T (Y - X\hat{\beta}^{lasso}) + \lambda \partial \|\hat{\beta}^{lasso}\|_1 = 0,$$

*where $\partial \|\hat{\beta}^{lasso}\|_1$ is a subgradient at $\hat{\beta}^{lasso}$ for $\|\beta\|_1$. Multiplying $(\beta^* - \hat{\beta}^{lasso})^T$ on both sides, we have*

$$-\frac{1}{n} (\beta^* - \hat{\beta}^{lasso})^T X^T (Y - X\hat{\beta}^{lasso}) + \lambda (\beta^* - \hat{\beta}^{lasso})^T \partial \|\hat{\beta}^{lasso}\|_1 = 0.$$

*Since*

$$\begin{aligned} & (\beta^* - \hat{\beta}^{lasso})^T \partial \|\hat{\beta}^{lasso}\|_1 \\ =& (\beta^* - \hat{\beta}^{lasso})_S^T \partial \|\hat{\beta}_S^{lasso}\|_1 + (\beta^* - \hat{\beta}^{lasso})_{S^c}^T \partial \|\hat{\beta}_{S^c}^{lasso}\|_1 \\ =& (\beta^* - \hat{\beta}^{lasso})_S^T \partial \|\hat{\beta}_S^{lasso}\|_1 - \|\hat{\beta}_{S^c}^{lasso}\|_1 \end{aligned} \tag{3.4}$$

*Plugging into the FOC, we have*

$$
\begin{aligned}
&\frac{1}{n}(\beta^* - \hat{\beta}^{lasso})^T X^T (Y - X\hat{\beta}^{lasso}) \\
=&\frac{1}{n}(\beta^* - \hat{\beta}^{lasso})^T X^T X(\beta^* - \hat{\beta}^{lasso}) + \frac{1}{n}(\beta^* - \hat{\beta}^{lasso})^T X^T \epsilon \\
=&\frac{1}{n}(\beta^* - \hat{\beta}^{lasso})^T X^T X(\beta^* - \hat{\beta}^{lasso}) + \frac{1}{n}(\beta^* - \hat{\beta}^{lasso})_S^T X_S^T \epsilon \\
&+ \frac{1}{n}(\beta^* - \hat{\beta}^{lasso})_{S^c}^T X_{S^c}^T \epsilon \\
=&\frac{1}{n}(\beta^* - \hat{\beta}^{lasso})^T X^T X(\beta^* - \hat{\beta}^{lasso}) + \frac{1}{n}(\beta^* - \hat{\beta}^{lasso})_S^T X_S^T \epsilon \\
&- \sum_{S^c} \frac{1}{n} |\hat{\beta}^{lasso}|_i X_i^T \epsilon sign(\hat{\beta}_i^{lasso}) \\
=&\lambda(\beta^* - \hat{\beta}^{lasso})_S^T \partial \|\hat{\beta}_S^{lasso}\|_1 - \lambda \|\hat{\beta}_{S^c}^{lasso}\|_1
\end{aligned} \tag{3.5}
$$

*where in the first equality, we plugged in $Y = X\beta^* + \epsilon$. If we have the condition that $\frac{1}{n} X_i^T X(\beta^* - \hat{\beta}^{lasso}) sign(\hat{\beta}_i^{lasso}) > c\lambda$ ($c = \frac{1}{2}$ in LASSO) for all $i \notin S$, we have*

$$
\begin{aligned}
&\frac{1}{n}(\beta^* - \hat{\beta}^{lasso})^T X^T X(\beta^* - \hat{\beta}^{lasso}) \\
&\leq \frac{3}{2}\lambda \|\hat{\beta}_S^{lasso}\|_1 - c\lambda \|\hat{\beta}_{S^c}^{lasso}\|_1
\end{aligned} \tag{3.6}
$$

*with high probability. Similarly, we made Assumption 3.7 in the generalized penalized regression to ensure good property of the solution.*

REMARK 3.2. *Since the condition in Assumption 3.7 cannot be verified directly, in real data analysis, we can use the following checkable conditions instead:*

$$\frac{1}{n} X_j^T (Y - X\hat{\beta}) sign(\hat{\beta}_j) \geq c\lambda, \text{ for all } j \in S^c \text{ such that } \hat{\beta}_j \neq 0, \ \beta = \beta^*,$$

*where $c$ is defined in Assumption 3.7. If the above holds, Assumption 3.7 holds with high probability using similar argument of sub-Gaussian random variables as in the proof of Theorem 4.2.*

### 3.5 Assumptions on the design matrix $\boldsymbol{X}$

DEFINITION 3.1. *The restricted strong convexity (RSC) condition on model matrix $X$ with respect to $\mathcal{C}$ is the following, there exists some constant $\gamma > 0$ such that:*

$$\frac{\frac{1}{N}\nu X^T X \nu}{\|\nu\|_2^2} \geq \gamma \text{ for all nonzero } \nu \in \mathcal{C}$$

here $\gamma$ is called the restricted eigenvalue bound with regard to $\mathcal{C}$.

ASSUMPTION 3.8. *Denote $C_{\mathcal{S}}$ by the diagonal matrix with $\{c_i, i \in \mathcal{S}\}$, $C_{\mathcal{S}^C}$ by the diagonal matrix with $\{c_i, i \in \mathcal{S}^C\}$, the restricted eigenvalues (RE) condition holds on the following set with some positive $c$ defined in Assumption 3.7:*

$$\mathcal{C} = \left\{ \nu \in \mathbb{R}^p \,\middle|\, \|C_{\mathcal{S}^C} \cdot \nu_{\mathcal{S}^C}\|_1 \leq \frac{5}{2c} \|C_{\mathcal{S}} \cdot \nu_{\mathcal{S}}\|_1 \right\},$$

*where $\cdot$ indicates the matrix-vector multiplication.*

We have $\mathcal{C} \subset \mathbb{R}^p$ strictly since it is of the form of cone.

The RSC (Assumption 3.8) is a standard assumption in the literature for proving the consistency results of regularized high-dimensional sparse estimation problems.

## 4. CONSISTENCY RESULTS FOR SOME D-STATIONARY SOLUTIONS

We prove our main results in this section. The non-asymptotic upper bound for estimation errors is derived in Section 4.1. As a corollary, we provide the upper bound for prediction errors as a byproduct in Section 4.2. We provide the results regarding the asymptotic consistency of the estimation in Section 4.3. The asymptotic consistency in support recovery is discussed in Section 4.4.

### 4.1 Non-asymptotic upper bound for estimation errors

In this section, we present our results on the non-asymptotic upper bound for the estimation error. We mainly use the assumptions on the model matrix to prove that the difference between the ground truth $\beta^*$ and the d-stationary solution $\hat{\beta}_n^\lambda$ will be in a cone-like set, where we have the restricted strong convexity (RSC) assumption hold (as defined in Assumption 3.8). Without Assumption 3.4 on the continuity of the first order derivative on the function $h_\lambda(t)$, we will be able to obtain the upper-bound of the $\ell_2$ distance between the ground truth and the d-stationary estimation.

THEOREM 4.1. *Suppose $h_\lambda(t)$ satisfies Assumptions 3.2, 3.3, 3.5, design matrix $X$ satisfies the restricted strong convexity with respect to $\mathcal{C}$, which is defined in Assumption 3.8 with $c_i = 1$ for $i = 1, \cdots, p$, with $\lambda \geq \frac{2\|X^T\epsilon\|_\infty}{n}$. If we further assume Assumption 3.7 holds at the d-stationary solution $\hat{\beta}_n^\lambda$, we will have the upper bound for estimation error on $\beta^*$ with the d-stationary estimation $\hat{\beta}_n^\lambda$:*

$$\|\hat{\beta}_n^\lambda - \beta^*\|_2 \leq \frac{5}{2\gamma}\lambda\sqrt{\|\beta^*\|_0}$$

The proof of the above Theorem 4.1 is in an online supplementary file. The results in Theorem 4.1 suggest that the d-stationary solution to Problem (3.1), under mild conditions, will be able to retrieve the information in the unknown parameter $\beta^*$ with error bounded within the order of $O(\frac{\lambda\sqrt{|S|}}{\gamma})$, which is optimal.

### 4.2 Non-asymptotic upper bound for prediction errors

From the proof of Theorem 4.1, we will be able to further give the upper bound for the prediction error below.

COROLLARY 4.1. *Under the assumptions of Theorem 4.1, we can further get the upper-bound for the prediction error as:*

$$\left\|\frac{1}{\sqrt{n}}X(\beta^* - \hat{\beta}_n^\lambda)\right\|_2^2 \leq (\frac{5}{2}\lambda)^2\frac{1}{\gamma}\sqrt{|S|}$$

The proof of Corollary 4.1 is straight forward according to the proof in Theorem 4.1 and is in an online supplementary file.

### 4.3 Asymptotic convergence rate

If we further assume that the errors are independent sub-Gaussian distributed, we will be able to bound the estimation error in the order of $\sqrt{\frac{|S|\log p}{n}}$ with high probability.

COROLLARY 4.2. *Under the assumptions of Theorem 4.1, if we further assume that the errors are from independent sub-Gaussian with variance parameter $\sigma^2$ and mean 0, we will have the following hold with probability at least $1-2\exp(-\frac{\tau-2}{2}\log p)$*

$$\|\hat{\beta}_n^{\lambda}-\beta^*\|_2 \lesssim \frac{5}{\gamma}\sigma\sqrt{\frac{\tau|S|\log p}{n}}.$$

We provide the proof of Corollary 4.2 in an online supplementary file.

REMARK 4.1. *All the results above are considering the problem in (3.1). However, the conclusions will still hold for the constrained version of problem (3.1) as long as the assumptions are satisfied. The results in this work assumes the d-stationary solution that satisfies our assumptions exists. We will justify the existence of the d-stationary solution satisfying our assumptions in Section 5.1.*

### 4.4 Support recovery

In this section, we will first provide the KKT conditions for d-stationary solutions, which says that the d-stationary condition in our work is equivalent to the first order condition in the case of no constraints. Then we prove the Restricted Strong Convexity for the objective function in Problem (3.1) under some regularity conditions. By usage of the oracle estimator defined later in Problem (4.3), we will be able to prove the support recovery consistency of some of the d-stationary solutions to Problem (3.1).

LEMMA 4.1. *Let $F(\beta)=L(\beta)+g(\beta)-h(\beta)$, where $L(\beta)$, $g(\beta)$, $h(\beta)$ are convex with $\beta\in\mathbb{R}^p$. Further assume that $L(\beta)$ and $h(\beta)$ have continuous first order derivative, $g(\beta)=\|\beta\|_1$. Let $\beta_0$ be a d(irectional)-stationary solution to $F(\beta)$, we have the following first order condition (FOC) hold at $\beta_0$. We will be able to get the following equivalent condition: $\beta_0$ is a d(irectional)-stationary solution to $F(\beta)$ if and only if there exists some $z\in\partial g(\beta_0)$, where $\partial g(\beta_0)$ is the set of subgradient of $g(\beta)$ at $\beta_0$, such that:*

$$\nabla L(\beta_0)+z-\nabla h(\beta_0)=0, \tag{4.1}$$

*where $\nabla L(\beta_0)$, $\nabla h(\beta_0)$ is the gradient of $L$, $h$ at $\beta_0$.*

The above Lemma 4.1 states the equivalence between d-stationary solution and first order condition (FOC) in the unconstrained case. While in constrained case, this does not necessarily hold. From the proof Lemma 4.1, we can derive similar conditions for "local maximals" for $\tilde{\beta}$. We obtain that as long as $\min_{i=1}^{p}\{\tilde{\beta}_i\}=0$, it will only satisfy the condition for "local" minimals and thus be a d-stationary solution. Thus, in order to find the d-stationary solution, we only need to find a $\beta_0$ such that, there exists a vector $z\in\partial\|\beta_0\|_1$, the subgradient of function $\|\beta\|_1$

at $\beta = \beta_0$, such that $\nabla L(\beta_0) - \nabla h(\beta_0) + z = 0$. Furthermore, if $\min_{i=1}^{p}\{z_i\} < 1$, which is known as the strict dual feasibility condition [30], it will be satisfying the condition for "local" minimals.

REMARK 4.2. *Generally, a d-stationary solution is not necessarily local minimal. For example, for a differentiable function $f(x, y) = x^2 - y^2$, where both the function $g(x, y)$ and $h(x, y)$ are differentiable (slightly different from the above situation in Lemma 4.1), at a saddle point $(0, 0)$, which is stationary with $\mathbf{0}$ gradient, the directional derivative at this point will all be 0, which makes it a d-stationary solution however not a local minimal. Another example would be $f(x, y) = |x| - y^2$ at the saddle point $(0, 0)$.*

REMARK 4.3. *The necessary condition to be a local minimal is being a d-stationary point in the feasible region.*

The following Lemma shows the RSC of the Problem (3.1).

LEMMA 4.2. *Under Assumption 3.8 with $h_\lambda$ satisfying Assumptions 3.2, 3.3, 3.4, 3.5, let $\beta_1$, $\beta_2 \in \mathbb{R}^p$ such that $\nu = \beta_1 - \beta_2 \in \mathcal{C}$, where $\mathcal{C}$ is defined in Assumption 3.8. Then $f_\lambda(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 - h_\lambda(\beta)$ will satisfy the restricted strong convexity given that $\gamma > \eta^-$:*

$$f_\lambda(\beta_2) \geq f_\lambda(\beta_1) + \nabla f_\lambda(\beta_1)^T(\beta_2 - \beta_1) + \frac{\gamma - \eta^-}{2}\|\beta_2 - \beta_1\|_2^2 \tag{4.2}$$

The proof can be found in an online supplementary file.

*4.4.1 Oracle estimator* The oracle estimator is defined as follows:

$$\beta^O = \arg\min_{\beta \in \mathbb{R}^p, \beta_{S^c}=0} \frac{1}{2n}\|Y - X\beta\|_2^2. \tag{4.3}$$

The oracle estimator is obtained as if there is an oracle telling the true support of the underlying unknown estimator. According to the definition of oracle estimator $\beta^O$, we are able to provide the following $\ell_\infty$ error bound between $\beta^O$ and $\beta^*$. We also demonstrate that $\beta^O$ is a d-stationary solution to the DC-penalized Problem (3.1), which we are interested in this paper. The following Theorem 4.2 and Lemma 4.3 also appeared in the work by Wang et al. [32]

THEOREM 4.2. *Under Assumption 3.8, the oracle estimator is the unique global minimizer of (4.3). If the noise is independent sub-Gaussian with variance parameter $\sigma^2$, the oracle estimator will satisfy the following $\ell_\infty$ error bound with high probability.*

$$\|\beta^O - \beta^*\|_\infty \leq C\sigma\sqrt{2/\gamma}\sqrt{\frac{\log s}{n}}.$$

The proof is in an online supplementary file.

LEMMA 4.3. *Under Assumption 3.8 with $h_\lambda$ satisfying Assumptions 3.2, 3.3, 3.4, 3.5, 3.6, let $\beta^O$ be the aforementioned oracle estimator. Assume further that for the ground truth $\beta^*$, we have $\min_{i=1}^{p}|\beta_i^*| > 2\zeta$, for $\zeta > 0$. There exists a*

*subgradient* $\xi^O \in \partial\|\beta^O\|_1$ *(where* $\partial\|\beta^O\|_1$ *stands for the subgradient of function* $\|\beta\|_1$ *at* $\beta = \beta^O$*) such that for any* $\beta \in \mathbb{R}^p$*:*

$$(\beta - \beta^O)^T \left(\nabla f_\lambda(\beta^O) + \lambda\xi^O\right) \geq 0 \tag{4.4}$$

The above Lemma 4.3 assumes that the penalty on the parameters will be a constant when the parameters are large. As it requires Assumption 3.6, the result is not applicable for transformed $\ell_1$ and logarithmic penalties. We provide the proof in an online supplementary file.

LEMMA 4.4. *Under the assumptions in Lemma 4.3, let* $\hat{\beta}$ *be a d-stationary solution to (3.1) satisfying Assumption 3.7, and* $\beta^O$ *be the oracle estimator. The following will hold with large probability:*

$$\nu = \hat{\beta} - \beta^O \in \mathcal{C}. \tag{4.5}$$

The proof is in an online supplementary file. Based on the previous results of the oracle estimator $\beta^O$ and properties of the d-stationary estimator $\hat{\beta}$, we will now be able to prove the support recovery consistency for our generalized DC-penalized model.

THEOREM 4.3. *Under the conditions of Lemma 4.4, we will have* $supp(\hat{\beta}) = supp(\beta^O) = supp(\beta^*)$ *with high probability.*

The proof is provided in an online supplementary file. We prove the support recovery consistency for a set of d-stationary solutions, it implies that a set of the convergence points (satisfying Assumption 3.7) from the DCA will converge to the oracle estimator, which is unique. In the work from Wang et al [32], they prove that the convergence point from each stage of the specific algorithm converges to the oracle estimator in the linear model setting. The above results also inform us how we should choose the penalty function such that the d-stationary solution will be support recovery consistent. The penalty needs to be a constant when the parameter gets larger (Assumption 3.6), so that the resulting oracle estimator will be a d-stationary solution to the original Problem (3.1). Assumption 3.4 is necessary for the restricted strong convexity in $\mathcal{C}$.

## 5. DC PENALTY WITH GENERALIZED LOSS FUNCTIONS

In the previous section, we mainly focus on the linear model scenario. Most of the analysis can be readily extended to generalized loss functions such as the logistic loss function, etc. Below, we will present the formulation of DC penalized likelihood and provide the statistical analysis regarding to the d-stationary solutions.

We begin with a brief review on the exponential family. Exponential family is a family of distributions with the probability density proportional to $P(Y|X, \beta^*) \propto \exp\{\frac{YX^T\beta^* - \psi(X^T\beta^*)}{c(\sigma)}\}$, where $c(\sigma)$ is a scaling parameter and $\psi(\cdot)$ is the cumulant function. According to [17], one standard property of exponential family is

$$\psi'(X^T\beta^*) = \mathbb{E}[Y|X, \beta^*, \sigma].$$

Given that $\psi(\cdot)$ is a univariate convex function, let $L(\beta) = \psi(X^T\beta) - YX^T\beta$ be the negative log likelihood function, $L_n(\beta) = \frac{1}{n}\sum_{i=1}^n(\psi(X_i^T\beta) - Y_iX_i^T\beta)$ be the sample average of the negative log likelihood function, one can easily check that $\mathbb{E}[\nabla L(\beta^*)] = 0$ and $\nabla^2 L_n(\beta) \geq 0$. This implies that $L_n(\beta)$ is convex. In the following, we might omit the subscript $n$ in the expression of $L_n(\beta)$ where no confusion will rise. In order to estimate the sparse ground truth $\beta^*$, we will solve the following DC penalized optimization problem:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n(\psi(X_i^T\beta) - Y_iX_i^T\beta) + \lambda\|\beta\|_1 - h_\lambda(\beta), \tag{5.1}$$

Below, we will state the assumptions on the generalized loss functions, which enable us to provide the analysis that the error between the d-stationary solution $\hat{\beta}$ and the ground truth $\beta^*$ is of the order $\mathbb{O}(\frac{17\lambda}{8(\gamma-\eta^-)}\sqrt{|S|})$.

ASSUMPTION 5.1. *Let $\beta^*$ represent the ground truth of the unknown parameter, $L(\beta)$ be the negative log likelihood function. Assume that the infinity norm of the gradient of the loss function at the ground truth $\|\nabla L(\beta^*)\|_\infty \leq \frac{\lambda}{8}$.*

ASSUMPTION 5.2. *Let $\beta^*$ be the unknown true parameter, $\hat{\beta}$ be the d-stationary solution to problem (3.1), which satisfies the following condition:*

$$\|\nabla h_\lambda(\hat{\beta}_{S^c})\|_\infty \leq (1-c)\lambda, \text{ with } c \in (0,1).$$

ASSUMPTION 5.3. *Let $\beta^*$ represent the ground truth of the unknown parameter, $L(\beta)$ be the negative log-likelihood function. Assume that the following restricted strong convexity holds on the set $\mathcal{C}$,*

$$\begin{aligned} \mathcal{C} &= \left\{\nu \in \mathbb{R}^p \,\middle|\, \|C_{\mathcal{S}^C}\nu_{\mathcal{S}^C}\|_1 \leq \frac{4+c}{c}\|C_{\mathcal{S}}\nu_{\mathcal{S}}\|_1\right\}, \\ L(\beta_1) &\geq L(\beta_2) + \nabla L(\beta_2)^T(\beta_1-\beta_2) + \frac{\gamma}{2}\|\beta_1-\beta_2\|_2^2, \end{aligned}$$

*for any $\beta_1$ and $\beta_2$ such that $\beta_1 - \beta_2 \in \mathcal{C}$.*

THEOREM 5.1. *Let $\hat{\beta}$ be the d-stationary solution to the penalized loss function in (5.1). Suppose $h_\lambda(t)$ satisfies Assumptions 3.2, 3.3, 3.5, assume further that Assumptions 5.1, 5.3 and 5.2 hold, if $\gamma > \eta^-$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n(\psi'(X_i^T\beta^*)X_i - Y_iX_i)\right\|_\infty \leq \frac{c}{2},$$

*we will have the following upper-bound for the estimation error of the d-stationary solution*

$$\|\beta^* - \hat{\beta}\|_2 \leq \frac{17\lambda}{8(\gamma-\eta^-)}\sqrt{|S|}$$

The proof is provided in an online supplementary file.

### 5.1 Existence of d-stationary solution

In this section, we will show the existence of the d-stationary solutions we studied above. It is easy to see that in the linear regression setting with square loss, the oracle estimator is a d-stationary solution under suitable conditions we stated in Lemma 4.3. For general settings with generalized loss functions, let $r_0 > 0$ be such that $h'_\lambda(r_0) = (1-c)\lambda$, consider the following constrained problem:

$$\min_{\|\beta-\beta^*\|_2 \le r} L_n(\beta) + \lambda\|\beta\|_1 - h_\lambda(\beta), \tag{5.2}$$

where $r = c\lambda\sqrt{|S|} \wedge r_0$. It is straightforward to check that the sationary solutions to Problem (5.2) satisfies all the assumptions of the d-stationary solution studied in Section 4 and Section 5, which verifies the existence of the wanted d-stationary solutions.

## 6. NUMERICAL APPROACH TO FIND THE D-STATIONARY POINTS

In this section, we will review the efficient algorithms in the DC literature, for finding the local optima in the statistics and optimization areas. This provides a comprehensive summary of solving DC programming. The most classic algorithm is the Difference-of-Convex Algorithm (DCA) that has been studied in [13, 25, 24], which iterates between the primal problem and the dual problem to find the local minima. Given the DC problem below:

$$\min_{x\in\mathbb{R}^n} f(x) = g(x) - h(x), \tag{6.1}$$

where $g(\cdot)$ and $h(\cdot)$ are convex functions. For a function $g : \mathbb{R}^n \to \mathbb{R}$, let $g^*(y)$ be its convex conjugate function, which is defined as $g^*(y) = \sup\{x^T y - g(x) : x \in \mathbb{R}^n\}$. We have

$$\begin{aligned}
&\inf_{x\in\mathbb{R}^n} f(x) = g(x) - h(x) \\
&= \inf_x\{g(x) - \sup_y\{x^T y - h^*(y)\}\} \\
&= \inf_x\{\inf_y\{g(x) + h^*(y) - x^T y\}\} \\
&= \inf_y\{-\sup_x\{x^T y - g(x)\} + h^*(y)\} \\
&= \inf_y\{h^*(y) - g^*(y)\}.
\end{aligned} \tag{6.2}$$

Thus, by iterating between the primal and the dual problems, the DCA will converge to a d-stationary solution. Below shows the DCA.

According to [26] in Section 2.5, DCA has linear convergence rate for general DC programmings. While in the statistics literature, Local Linear Approximation (LLA) in [41] is widely used for solving regularized estimation problems with non-convex penalties. The update at each iteration takes the LLA of the penalty function:

$$x^{k+1} = \arg\min\{g(x) - \partial h(x_k)^T x\},$$

which is exactly the same procedure as shown in the Algorithm 1.

In the setting of this paper, the objective is defined in (3.1), where we are minimizing the objective function over all $\beta \in \mathbb{R}^p$ with the first part of the DC

Choose the initial $x_0$
*loop*:
**for** $k \in \mathbb{N}$ **do**
  Choose $y_k \in \partial h(x_k)$.
  Choose $x_{k+1} \in \partial g^*(y_k)$.
  **if** $(\min\{|(x_{k+1} - x_k)_i|, |\frac{(x_{k+1}-x_k)_i}{(x_k)_i}|\} \leq \delta)$ **then**
    **return** $x_{k+1}$
  **end if**
**end for**

**Algorithm 1:** Difference-of-Convex Algorithm (DCA)

function as $g(\beta) = L_n(\beta) + \lambda\|\beta\|_1$, and the second part of DC function $h(\beta) = h_\lambda(\beta)$. The DCA can be simplified to Local Linear Approximation (LLA) in the general case as in [41], the detailed procedures can be found in [24]. Specifically, if $h(x)$ is differentiable, we will have the following equivalent algorithm as DCA:

Choose the initial $\beta_0$
*loop*:
**for** $k \in \mathbb{N}$ **do**
  Choose $z_k \in \nabla h(\beta_k)$.
  $\beta_{k+1} = \arg\min L_n(\beta) + \lambda\|\beta\|_1 - \langle \beta, \nabla h(\beta_k)\rangle$.
  **if** $(\min\{|(\beta_{k+1} - \beta_k)_i|, |\frac{(\beta_{k+1}-\beta_k)_i}{(\beta_k)_i}|\} \leq \delta)$ **then**
    **return** $\beta_{k+1}$
  **end if**
**end for**

**Algorithm 2:** DCA (LLA)

According to [33], DCA is exactly the formulation of Convex Concave Procedure (CCCP), which is also discussed in [22]. Thus, under proper conditions, all results in [33] can be applied to the problem studied here. Since our formulation (3.1) is a special form of the model considered in [22], which adopts the classical algorithm DCA (Difference-of-Convex Algorithm) in [13, 25, 24] and solves a strictly convex problem at each iteration, it is guaranteed to converge quickly to a d-stationary solution. Since the penalty is a function of the absolute value of the estimator, one minor change to the above algorithm would be solving the following transformed optimization problem within each iteration:

$$\beta_{k+1} = \arg\min L_n(\beta) + \sum_{i=1}^{p} (\lambda - h'(|\beta_{ki}|))|\beta_i|, \tag{6.3}$$

which is exactly the formulation of weighted LASSO estimator and can be solved efficient using the LARS algorithm in [7].

LEMMA 6.1. *By updating the parameter $\beta$ as in Procedure 6.3, the objective function $F(\beta)$ defined in 3.1 is monotonically decreasing.*

In the one-step LLA procedure [41], the authors prove that starting from the maximum likelihood estimator (MLE), after one step of the LLA update, the resulting estimator is consistent when SCAD penalty function is used. While in [10], they prove that from the LASSO initialization, with high probability that the LLA converges to the oracle estimator in 2 iterations. The above results can be similarly extended to our DC setting.

## 7. CONCLUSIONS

In this work, we close the gap between the statistics and optimization by finding a set of d-stationary solutions to the DC penalized loss functions. Specifically, we relax the assumptions used in [1] and provide stronger statistical results on the penalized estimation problem. We prove that a certain subset of d-stationary solutions in an optimization problem (with a DC objective) has the ideal statistical properties: asymptotic estimation consistency, asymptotic model selection consistency, asymptotic efficiency under the linear model and the GLM settings. We also provide the non-asymptotic upper bound for the estimation errors in both scenarios. We unify the framework of non-convex penalized high-dimensional sparse estimation problems and the existing popular algorithms to solve the problems in a DC framework.

Several open questions remain, which might be interesting directions for future research. Since in this work, we mainly consider the unconstrained DC programming, it is unclear whether a proper constraint, which might depend on specific problems, will ensure a better set of solutions or possibly a unique solution to the high-dimensional sparse estimation problem. Another direction would be more general loss functions. When the observations have outliers or missing values, it would be desirable to obtain theoretical guarantees on the sparse estimations with possibly non-convex loss functions, such as Huber loss, Cauchy loss, etc.

## APPENDIX A: PROPERTIES OF DC PROGRAMMING

The following are some known properties of the DC functions [25, 13].

1. Every DC function has a nonnegative DC decomposition; that is for a DC function $f$, there exists a decomposition, $f = g - h$, where both $g$ and $h$ are nonnegative and convex.
2. Every $C^1$ (i.e., functions with continuously first order derivatives) function with a Lipschitz gradient is a DC function.
3. Every twice continuously differentiable function is a DC function.
4. Every continuous function on a convex set is a limit of a sequence of uniformly converging DC functions.
5. Let $f_i$ be DC functions for $i = 1, \cdots, m$. The DC functions are *closed* under the following operations:
   - summation: $\sum_{i=1}^m \lambda_i f_i(x)$, for $\lambda_i \in \mathbb{R}$, $i = 1, \cdots, m$
   - maximization: $\max_{i=1,\cdots,m} f_i(x)$
   - minimization: $\min_{i=1,\cdots,m} f_i(x)$
   - product: $\prod_{i=1,\cdots,m} f_i(x)$
6. A locally DC function that is defined in $\mathbb{R}^n$ is a DC function.
7. The following statements about a DC program are equivalent:
   - $\sup\{f(x) : x \in C\}$, function $f$ and set $C$ are convex
   - $\inf\{g(x) - h(x) : x \in \mathbb{R}^n\}$, functions $g$ and $h$ are convex
   - $\inf\{g(x) - h(x) : x \in C, f_1(x) - f_1(x) \leq 0\}$, functions $g$, $h$, $f_1$, and $f_2$ and set $C$ are all convex

Regarding the optimal solutions in the DC programming, the following have been developed in the literature [25, 13].

DEFINITION A.1. **($\epsilon$-subdifferential)** *For a convex function $g(x)$, and $\epsilon > 0$, the $\epsilon$-subdifferential of function $g(x)$ at point $x_0$ is denoted by $\partial_\epsilon g(x_0)$ and is defined as follows:*

$$\partial_\epsilon g(x_0) = \{\nu \in \mathbb{R}^n | g(x) \geq g(x_0) + \langle x - x_0, \nu \rangle - \epsilon\}.$$

One can verify that the subgradient [23, Chapter 23] (which is denoted by $\partial g(x_0)$) of function $g(x)$ at $x_0$ is the 0-subdifferential (i.e., $\epsilon = 0$).

THEOREM A.1. **(Global optimality condition)** *A point $x^*$ is a global optimal if and only if (iff) $\partial_\epsilon h(x^*) \subset \partial_\epsilon g(x^*)$ for any $\epsilon > 0$.*

THEOREM A.2. **(Local optimality condition)** *A point $x^*$ is a local optimal if $\partial h(x^*) \subset int\ \partial g(x^*)$, where $int\ \partial g(x^*)$ represents the interior of the set $\partial g(x^*)$.*

## ACKNOWLEDGEMENTS

This project is partially supported by the Transdisciplinary Research Institute for Advancing Data Science (TRIAD), http://triad.gatech.edu, which is a part of the TRIPODS program at NSF and locates at Georgia Tech, enabled by the NSF grant CCF-1740776. The first two authors are also partially supported by the NSF grant DMS-1613152.

## REFERENCES

[1] AHN, M., PANG, J.-S. and XIN, J. (2017). Difference-of-convex learning: directional stationarity, optimality, and sparsity. *SIAM Journal on Optimization* **27** 1637–1665.
[2] ALEKSANDROV, A. (1950). Surfaces represented as a difference of two convex functions, Russian Acad. Sci. In *Dokl. Math* **1**.
[3] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* 1705–1732.
[4] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
[5] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when $p$ is much larger than $n$. *The Annals of Statistics* 2313–2351.
[6] CHEN, S. and DONOHO, D. L. (1995). Examples of basis pursuit. In *SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation* 564–574. International Society for Optics and Photonics.
[7] EFRON, B., HASTIE, T., JOHNSTONE, I., TIBSHIRANI, R. et al. (2004). Least angle regression. *The Annals of statistics* **32** 407–499.
[8] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360.
[9] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961.
[10] FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of statistics* **42** 819.
[11] HARTMAN, P. (1959). On functions representable as a difference of convex functions. *Pacific Journal of Mathematics* **9** 707–713.
[12] HIRIART-URRUTY, J.-B. (1985). Generalized Differentiability/Duality and Optimization for Problems Dealing with Differences of Convex Functions. In *Convexity and duality in optimization* 37–70. Springer.
[13] HORST, R. and THOAI, N. V. (1999). DC programming: overview. *Journal of Optimization Theory and Applications* **103** 1–43.
[14] HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* 1603–1618.

[15] Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory* **60** 6522–6554.
[16] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15** 2869–2909.
[17] Lehmann, E. and Casella, G. (1998). Theory of Point Estimation, Springer-Verlag. *New York*.
[18] Loh, P.-L. and Wainwright, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems* 476–484.
[19] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* 3498–3528.
[20] Mazumder, R., Friedman, J. H. and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* **106** 1125–1138.
[21] Nouiehed, M., Pang, J.-S. and Razaviyayn, M. (2017). On the Pervasiveness of Difference-Convexity in Optimization and Statistics. *arXiv preprint arXiv:1704.03535*.
[22] Pang, J.-S., Razaviyayn, M. and Alvarado, A. (2016). Computing B-stationary points of nonsmooth DC programs. *Mathematics of Operations Research* **42** 95–118.
[23] Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.
[24] Sriperumbudur, B. K. and Lanckriet, G. R. (2012). A proof of convergence of the concave-convex procedure using Zangwill's theory. *Neural computation* **24** 1391–1407.
[25] Tao, P. D. and An, L. T. H. (1997). Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Mathematica Vietnamica* **22** 289–355.
[26] Tao, P. D. et al. (2005). The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of operations research* **133** 23–46.
[27] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.
[28] Tuy, H. (1987). Global minimization of a difference of two convex functions. *Nonlinear Analysis and Optimization* 150–182.
[29] Van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42** 1166–1202.
[30] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using $l_1$-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory* **55** 2183–2202.
[31] Wang, L., Kim, Y. and Li, R. (2013). Calibrating non-convex penalized regression in ultra-high dimension. *Annals of statistics* **41** 2505.
[32] Wang, Z., Liu, H. and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics* **42** 2164.
[33] Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural computation* **15** 915–936.
[34] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics* **38** 894–942.
[35] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242.
[36] Zhang, S. and Xin, J. (2014). Minimization of Transformed L_1 Penalty: Theory, Difference of Convex Function Algorithm, and Robust Application in Compressed Sensing. *arXiv preprint arXiv:1411.5735*.
[37] Zhang, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research* **11** 1081–1107.
[38] Zhang, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli* **19** 2277–2293.
[39] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research* **7** 2541–2563.
[40] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American statistical association* **101** 1418–1429.
[41] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood

models. *Annals of statistics* **36** 1509.

## Online Supplementary File for the Paper

This online supplementary file contains all the proofs in our paper.

## APPENDIX B: PROOFS IN SECTION 4

### B.1 Proof of Theorem 4.1

PROOF. Since $\hat{\beta}_n^\lambda$ is d-stationary, the directional derivatives should be nonnegative in all directions, especially in the direction of $\beta^* - \hat{\beta}_n^\lambda$:

$$-\frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)^T X^T (Y - X\hat{\beta}_n^\lambda) + P'_\lambda(\hat{\beta}_n^\lambda; \beta^* - \hat{\beta}_n^\lambda) \geq 0,$$

where $P'_\lambda(\hat{\beta}_n^\lambda; \beta^* - \hat{\beta}_n^\lambda)$ is the directional derivative for the penalty function $P_\lambda = \lambda\|\beta\|_1 - h_\lambda(\beta)$ at $\hat{\beta}_n^\lambda$ in the direction of $\beta^* - \hat{\beta}_n^\lambda$.

According to Lemma 4.1, there exists a subgradient $z \in \partial g(\hat{\beta}_n^\lambda)$, where $\partial g(\hat{\beta}_n^\lambda)$ is the set of subgradient of $g(\beta) = \|\beta\|_1$ at $\hat{\beta}_n^\lambda$, such that:

$$\nabla L(\hat{\beta}_n^\lambda) + \lambda z - \nabla h_\lambda(\hat{\beta}_n^\lambda) = 0, \tag{B.1}$$

Multiplying by $(\beta^* - \hat{\beta}_n^\lambda)^T$ on both side and plugging in $Y = X\beta^* + \epsilon$, we have

$$-\frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)^T X^T X(\beta^* - \hat{\beta}_n^\lambda) - \frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)^T X^T \epsilon + P'_\lambda(\hat{\beta}_n^\lambda; \beta^* - \hat{\beta}_n^\lambda) = 0,$$

where without ambiguity, we let $P'_\lambda(\hat{\beta}_n^\lambda; \beta^* - \hat{\beta}_n^\lambda) = (\beta^* - \hat{\beta}_n^\lambda)^T(\lambda z - \nabla h_\lambda(\hat{\beta}_n^\lambda))$ since the true directional derivative for the penalty $P_\lambda = \lambda\|\beta\|_1 - h_\lambda(\beta)$ at $\hat{\beta}_n^\lambda$ in the direction of $\beta^* - \hat{\beta}_n^\lambda$ is greater than $(\beta^* - \hat{\beta}_n^\lambda)^T(\lambda z - \nabla h_\lambda(\hat{\beta}_n^\lambda))$. We thus will have

$$\frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)^T X^T X(\beta^* - \hat{\beta}_n^\lambda) = -\frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)^T X^T \epsilon + P'_\lambda(\hat{\beta}_n^\lambda; \beta^* - \hat{\beta}_n^\lambda), \tag{B.2}$$

which implies we have the following hold for $j \notin S$

$$\begin{aligned} &-\frac{1}{n}|(\hat{\beta}_n^\lambda)_j|\text{sign}((\hat{\beta}_n^\lambda)_j)X_j^T X(\beta^* - \hat{\beta}_n^\lambda) \\ =&\frac{1}{n}|(\hat{\beta}_n^\lambda)_j|\text{sign}((\hat{\beta}_n^\lambda)_j)X_j^T \epsilon - \lambda|(\hat{\beta}_n^\lambda)_j| + h'_\lambda((\hat{\beta}_n^\lambda)_j)|(\hat{\beta}_n^\lambda)_j|\text{sign}((\hat{\beta}_n^\lambda)_j) \\ \leq& - c\lambda|(\hat{\beta}_n^\lambda)_j|, \end{aligned} \tag{B.3}$$

where the "$\leq$" follows from Assumption 3.7.

For $j \in S$

$$\begin{aligned} &\frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)_j X_j^T X(\beta^* - \hat{\beta}_n^\lambda) \\ =& -\frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)_j X_j^T \epsilon + \lambda(\beta^* - \hat{\beta}_n^\lambda)_j z_j - h'_\lambda((\hat{\beta}_n^\lambda)_j)(\beta^* - \hat{\beta}_n^\lambda)_j \\ \leq& \frac{5}{2}\lambda|(\beta^* - \hat{\beta}_n^\lambda)_j|, \end{aligned} \tag{B.4}$$

where the "$\leq$" follows from $\lambda \geq \frac{2\|X^T\epsilon\|_\infty}{n}$ and Assumption 3.2.

Let $\nu = \beta^* - \hat{\beta}_n^\lambda$, we thus will have

$$\frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)^T X^T X(\beta^* - \hat{\beta}_n^\lambda) \leq -c\lambda\|(\nu)_{S^c}\|_1 + \frac{5}{2}\lambda\|(\nu)_S\|_1, \tag{B.5}$$

where the inequality follows from Assumption 3.2.

Since the left hand side of the above is nonnegative, we will have $\nu = \beta^* - \hat{\beta}_n^\lambda \in \mathcal{C}$. Under the restricted strong convexity condition, we will have

$$\gamma\|\nu\|_2^2 \leq \frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)^T X^T X(\beta^* - \hat{\beta}_n^\lambda) \leq \frac{5}{2}\lambda\|(\nu)_S\|_1 \tag{B.6}$$

Thus we will further have

$$\|\nu\|_2^2 \leq \frac{5}{2\gamma}\lambda\|(\nu)_S\|_1 \leq \frac{5}{2\gamma}\lambda\sqrt{|S|}\|(\nu)\|_2,$$

from which we will have the upper bound

$$\|\nu\|_2 \leq \frac{5}{2\gamma}\lambda\sqrt{|S|} \propto \frac{\lambda\sqrt{|S|}}{\gamma} \tag{B.7}$$

□

## B.2 Proof of Corollary 4.1

Proof. From the proof of Theorem 4.1, let $\nu = \beta^* - \hat{\beta}_n^\lambda$,we have

$$\begin{aligned}
&\|\frac{1}{\sqrt{n}}X(\beta^* - \hat{\beta}_n^\lambda)\|_2^2 \\
=&\frac{1}{n}(\beta^* - \hat{\beta}_n^\lambda)^T X^T X(\beta^* - \hat{\beta}_n^\lambda) \\
\leq& - c\lambda\|(\nu)_{S^c}\|_1 + \frac{5}{2}\lambda\|(\nu)_S\|_1 \\
\leq&(\frac{5}{2}\lambda)^2\frac{1}{\gamma}\sqrt{|S|}
\end{aligned} \tag{B.8}$$

□

## B.3 Proof of Corollary 4.2

Proof. Since $\epsilon_i$ for $i = 1, \cdots, n$ are from sub-Gaussian distribution with parameter $\sigma^2$,

$$\mathbb{P}\left(\frac{x_j^T\epsilon}{n} \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

By Bonferroni bound, we will have

$$\mathbb{P}(\frac{\|X^T\epsilon\|_\infty}{n} \geq t) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2} + \log p\right)$$

By setting $t = \sigma\sqrt{\frac{\tau\log p}{n}}$ for some $\tau \geq 2$, we will be able to have

$$\mathbb{P}(\frac{\|X^T\epsilon\|_\infty}{n} \geq t) \leq 2\exp\left(-\frac{\tau-2}{2}\log p\right).$$

Thus:

$$\|\hat{\beta}_n^\lambda - \beta^*\|_2 \lesssim \frac{5}{\gamma}\sigma\sqrt{\frac{\tau|S|\log p}{n}} \tag{B.9}$$

□

### B.4 Proof of Lemma 4.1

PROOF. We will first prove the necessity. Since $\beta_0$ is a d-stationary solution to the objective function $F(\beta)$, we will have

$$F'(\beta_0, \beta - \beta_0) \geq 0,$$

for any $\beta \in \mathbb{R}^p$, where $F'(\beta_0, \beta - \beta_0)$ denotes the directional derivative in the direction of $\beta - \beta_0$. For any $i = 1, \cdots, p$, let $\beta^{i+} = \beta_0 + e_i$, where $e_i \in \mathbb{R}^p$ denotes the unit vector with 1 in the ith position and 0 everywhere else. Let $\beta^{i-} = \beta_0 - e_i$. We will have:

$$F'(\beta_0, \beta^{i+} - \beta_0) \geq 0,$$
$$F'(\beta_0, \beta^{i-} - \beta_0) \geq 0,$$

which implies

- For $i$ such that $\beta_{0i} \neq 0$,

$$\nabla L(\beta_0)_i - \nabla h(\beta_0)_i + \text{sign}(\beta_{0i}) \geq 0,$$
$$-\nabla L(\beta_0)_i + \nabla h(\beta_0)_i - \text{sign}(\beta_{0i}) \geq 0,$$

  where $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = -1$ if $x < 0$.
- For $i$ such that $\beta_{0i} = 0$,

$$\nabla L(\beta_0)_i - \nabla h(\beta_0)_i + 1 \geq 0,$$
$$-\nabla L(\beta_0)_i + \nabla h(\beta_0)_i + 1 \geq 0.$$

We thus conclude that

- For $i$ such that $\beta_{0i} \neq 0$,

$$\nabla L(\beta_0)_i - \nabla h(\beta_0)_i + \text{sign}(\beta_{0i}) = 0,$$

- For $i$ such that $\beta_{0i} = 0$,

$$|\nabla L(\beta_0)_i - \nabla h(\beta_0)_i| \leq 1.$$

Thus for $z$ such that $z_i = \text{sign}(\beta_{0i})$ when $\beta_{0i} \neq 0$ and $z_i = -(\nabla L(\beta_0)_i - \nabla h(\beta_0)_i)$ when $\beta_{0i} = 0$ is what we need.

On the other hand, if there exists some $z \in \partial g(\beta_0)$, where $\partial g(\beta_0)$ is the set of subgradient of $g(\beta)$ at $\beta_0$, such that:

$$\nabla L(\beta_0) + z - \nabla h(\beta_0) = 0, \tag{B.10}$$

- For $i$ such that $\beta_{0i} \neq 0$,

$$z_i = \text{sign}(\beta_{0i}) = 1 \text{ or } -1,$$

- For $i$ such that $\beta_{0i} = 0$,

$$-1 \leq z = -(\nabla L(\beta_0)_i - \nabla h(\beta_0)_i) \leq 1.$$

- For $i$ such that $\beta_{0i} \neq 0$,

$$F'(\beta_0, \beta^{i+} - \beta_0) = 0,$$
$$F'(\beta_0, \beta^{i-} - \beta_0) = 0.$$

- For $i$ such that $\beta_{0i} = 0$,

$$\begin{aligned} F'(\beta_0, \beta^{i+} - \beta_0) &= \nabla L(\beta_0)_i - \nabla h(\beta_0)_i + 1 \\ &= \nabla L(\beta_0)_i - \nabla h(\beta_0)_i + z + 1 - z \\ &= 0 + 1 - z \geq 0 \end{aligned} \tag{B.11}$$

$$\begin{aligned} F'(\beta_0, \beta^{i-} - \beta_0) &= -\nabla L(\beta_0)_i + \nabla h(\beta_0)_i + 1 \\ &= -\nabla L(\beta_0)_i + \nabla h(\beta_0)_i - z + 1 + z \\ &= 0 + 1 + z \geq 0 \end{aligned} \tag{B.12}$$

We thus conclude that the directional derivative of the objective function at $\beta_0$ is always nonnegative in any direction. This complete the proof. □

REMARK B.1. *From the proof Lemma 4.1, we can derive similar conditions for "local maximals" for $\tilde{\beta}$ satisfying the following:*

$$F'(\tilde{\beta}, \beta - \tilde{\beta}) \leq 0. \tag{B.13}$$

- *For $i$ such that $\tilde{\beta}_i \neq 0$,*

$$F'(\tilde{\beta}, \beta^{i+} - \tilde{\beta}) = 0,$$
$$F'(\tilde{\beta}, \beta^{i-} - \tilde{\beta}) = 0.$$

- *For $i$ such that $\tilde{\beta}_i = 0$,*

$$\nabla L(\tilde{\beta})_i - \nabla h(\tilde{\beta})_i + 1 \leq 0,$$
$$-\nabla L(\tilde{\beta})_i + \nabla h(\tilde{\beta})_i + 1 \leq 0.$$

*The above implies that if the stationary solution $\tilde{\beta}$ to the FOC satisfies: $\min_{i=1}^{p}\{\tilde{\beta}_i\} = 0$, it will only satisfy the condition for "local" minimals and thus be a d-stationary solution.*

## B.5 Proof of Lemma 4.2

PROOF. Since $L(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2$ is quadratic and convex, we have

$$L(\beta_2) = L(\beta_1) + \nabla L(\beta_1)^T(\beta_2 - \beta_1) + \frac{1}{2}(\beta_2 - \beta_1)^T \nabla^2 L(\beta_1)(\beta_2 - \beta_1),$$

where $\nabla^2 L(\beta_1)$ is the Hessian matrix of $L(\beta)$ at $\beta_1$. Since $\nu = \beta_1 - \beta_2 \in \mathcal{C}$ and Assumption 3.8 holds on $\mathcal{C}$, we will further have

$$L(\beta_2) \geq L(\beta_1) + \nabla L(\beta_1)^T(\beta_2 - \beta_1) + \frac{\gamma}{2}\|\beta_2 - \beta_1\|_2^2.$$

On the other hand, $h_\lambda(\beta)$ is convex with $0 \leq \eta^+ \leq \frac{h'_\lambda(t_2) - h'_\lambda(t_1)}{t_2 - t_1} \leq \eta^-$, we will have

$$h_\lambda(\beta_2) \leq h_\lambda(\beta_1) + \nabla h_\lambda(\beta_1)^T(\beta_2 - \beta_1) + \frac{\eta^-}{2}\|\beta_2 - \beta_1\|_2^2$$

By combining the above two inequalities, we will be able to get (4.2). □

### B.6 Proof of Theorem 4.2

PROOF. The first part is easy to see since the feasible region is convex and is a subset of $\mathcal{C}$, in which the strong convexity condition holds (Assumption 3.8) for the loss function (in our case, the least square loss function). The minimizer to a strong problem is unique.

For the second conclusion, we first need to show that $X_S^T X_S$ is invertible and $\beta_S^O = (X_S^T X_S)^{-1} X_S^T Y$. This follows easily from the Assumption 3.8, which implies $\gamma_{\min}(X_S^T X_S)$, the minimum eigenvalue of $X_S^T X_S$ is larger than $n\gamma$. We thus have $\beta_S^O = (X_S^T X_S)^{-1} X_S^T Y$.

$$\begin{aligned} \beta_S^O - \beta_S^* &= (X_S^T X_S)^{-1} X_S^T Y - \beta^* \\ &= (X_S^T X_S)^{-1} X_S^T \epsilon \end{aligned} \tag{B.14}$$

Since $e_j (X_S^T X_S)^{-1} X_S^T \epsilon$, where $e_j \in \mathbb{R}^s$with all-zero elements except the $j$-th coordinate. Recall that $\epsilon$ has independent sub-Gaussian coordinates with the same variance parameter $\sigma^2$, we thus will have

$$\mathbb{P}\left(|e_j (X_S^T X_S)^{-1} X_S^T \epsilon| > t\right) \leq 2 \exp - \frac{t^2}{\|e_j (X_S^T X_S)^{-1} X_S^T\|_2^2 \sigma^2}. \tag{B.15}$$

By using Bonferroni bound, the above implies

$$\mathbb{P}\left(\max_{j=1,\cdots,s} |e_j (X_S^T X_S)^{-1} X_S^T \epsilon| > t\right) \leq 2s \exp - \frac{t^2}{\|e_j (X_S^T X_S)^{-1} X_S^T\|_2^2 \sigma^2}. \tag{B.16}$$

Taking $t = C\|e_j (X_S^T X_S)^{-1} X_S^T \epsilon\|_2 \sigma \cdot \sqrt{2 \log s}$ with $C > 0$, we will have

$$\begin{aligned} \|\beta_S^O - \beta_S^*\|_\infty &= \|(X_S^T X_S)^{-1} X_S^T \epsilon\|_\infty \\ &= \max_{j=1,\cdots,s} |e_j (X_S^T X_S)^{-1} X_S^T \epsilon| \\ &\leq C\|e_j (X_S^T X_S)^{-1} X_S^T \epsilon\|_2 \sigma \cdot \sqrt{2 \log s} \end{aligned} \tag{B.17}$$

hold with probability at least $1 - 2 \exp - C^2/s$. Since for any $j \in \{1, \cdots, s\}$,

$$\begin{aligned} \|e_j (X_S^T X_S)^{-1} X_S^T \epsilon\|_2^2 &= e_j (X_S^T X_S)^{-1} X_S^T \epsilon (e_j (X_S^T X_S)^{-1} X_S^T \epsilon^T) \\ &= e_j (X_S^T X_S)^{-1} e_j^T \\ &\leq 1/\gamma_{\min}(X_S^T X_S) \\ &\leq 1/n\gamma. \end{aligned} \tag{B.18}$$

This complete the proof since $\|\beta_{S^c}^O - \beta_{S^c}^*\|_\infty = 0$. □

### B.7 Proof of Lemma 4.3

PROOF. According to Theorem 4.2, we will have for $j \in S$,

$$|\beta_j^O| \geq |\beta_j^*| - \|\beta^O - \beta^*\|_\infty \geq \zeta,$$

which further implies $P_j'(\lambda, \beta_j^O) = 0$.

For $j \notin S$, we will have $h'_\lambda(\beta_j^O) = 0$ and since the errors are sub-Gaussian, there will exist $\xi_{S^c}^O \in \partial\|\beta_{S^c}^O\|_1$ satisfying inequality (4.4) with high probability and $\|\xi_{S^c}^O\|_\infty \leq \frac{1}{10}c$ where $c$ is defined in the Assumption 3.7.

$$\left(\nabla L_n(\beta^O)\right)_{S^c} + \lambda\xi_{S^c}^O = 0.$$

In order to see this, first we notice that $\beta_S^O = (X_S^T X_S)^{-1} X_S^T Y$, $\beta_{S^c}^O = 0$. We thus will have $\left(\nabla L_n(\beta^O)\right)_{S^c} = -\frac{1}{n} X_{S^c}(Y - (X_S^T X_S)^{-1} X_S^T Y)$. Plugging in the true model $Y = X\beta^* + \epsilon$, we will have $\left(\nabla L_n(\beta^O)\right)_{S^c} = -\frac{1}{n} X_{S^c}(I - X_S(X_S^T X_S)^{-1} X_S^T)\epsilon$, where $(I - X_S(X_S^T X_S)^{-1} X_S^T)\epsilon$ is a vector of independent sub-Gaussian random variables. By using the Bonferroni bound, we will have the conclusion. □

### B.8 Proof of Lemma 4.4

Proof. Since both $\hat{\beta}$ and $\beta^O$ are d-stationary, per Lemma 4.3, we have:

$$(\hat{\beta} - \beta^O)^T \left(\nabla f_\lambda(\beta^O) + \lambda\xi^O\right) \geq 0,$$

and $\hat{\beta}$ is d-stationary, there exists a $\hat{\xi} \in \partial\|\hat{\beta}\|_1$ ((where $\partial\|\hat{\beta}\|_1$ stands for the subgradient of function $\|\beta\|_1$ at $\beta = \hat{\beta}$)) such that

$$(\beta^O - \hat{\beta})^T \left(\nabla f_\lambda(\hat{\beta}) + \lambda\hat{\xi}\right) \geq 0.$$

On the one hand,

$$\begin{aligned}
0 &\leq (\beta^O - \hat{\beta})^T \left(\nabla f_\lambda(\hat{\beta}) + \lambda\hat{\xi}\right) \\
&\leq (\beta^O - \hat{\beta})^T \left(\nabla f_\lambda(\hat{\beta})\right) - \lambda\|(\nu)_{S^c}\|_1 + \lambda\|(\nu)_S\|_1 \\
&= -\frac{1}{n}(\beta^O - \hat{\beta})^T X^T X(\beta^* - \beta^O + \beta^O - \hat{\beta}) - \frac{1}{n}(\beta^O - \hat{\beta})^T X^T \epsilon \\
&- (\beta^O - \hat{\beta})^T \nabla h_\lambda(\hat{\beta}) - \lambda\|(\nu)_{S^c}\|_1 + \lambda\|(\nu)_S\|_1 \\
&\leq -\frac{1}{n}(\beta^O - \hat{\beta})^T X^T X(\beta^* - \beta^O + \beta^O - \hat{\beta}) - \frac{1}{n}(\beta^O - \hat{\beta})^T X^T \epsilon \\
&+ \sum_{i \notin S} |\hat{\beta}_i||h'_\lambda(\hat{\beta}_i)| - \lambda\|(\nu)_{S^c}\|_1 + 2\lambda\|(\nu)_S\|_1.
\end{aligned} \tag{B.19}$$

By rearranging the terms, we will have

$$\begin{aligned}
&\frac{1}{n}(\beta^O - \hat{\beta})^T X^T X(\beta^* - \beta^O) + \frac{1}{n}(\beta^O - \hat{\beta})^T X^T X(\beta^O - \hat{\beta}) - \sum_{i \notin S} |\hat{\beta}_i||h'_\lambda(\hat{\beta}_i)| \\
&\leq -\frac{1}{n}(\beta^O - \hat{\beta})^T X^T \epsilon - \lambda\|(\nu)_{S^c}\|_1 + 2\lambda\|(\nu)_S\|_1.
\end{aligned} \tag{B.20}$$

On the other hand, according to the proof of Lemma 4.3, we will have

$$\left(\nabla h_\lambda(\beta^O))\right)_S = \lambda \text{sign}(\beta_S^O),$$

$$\left(\nabla h_\lambda(\beta^O))\right)_{S^c} = 0,$$

$$\lambda\xi_{S^c}^O = -\left(\nabla L_n(\beta^O)\right)_{S^c},$$

$$\|\xi^O_{S^c}\|_\infty \leq \frac{1}{2}.$$

By using the above facts, we will further obtain

$$\begin{aligned} 0 &\leq (\hat{\beta} - \beta^O)^T \left(\nabla f_\lambda(\beta^O) + \lambda \xi^O\right) \\ &= -\frac{1}{n}(\hat{\beta} - \beta^O)^T X^T X(\beta^* - \beta^O) - \frac{1}{n}(\hat{\beta} - \beta^O)^T X^T \epsilon + (\hat{\beta} - \beta^O)^T_{S^c} \lambda \xi^O_{S^c} \\ &\leq -\frac{1}{n}(\hat{\beta} - \beta^O)^T X^T X(\beta^* - \beta^O) - \frac{1}{n}(\hat{\beta} - \beta^O)^T X^T \epsilon + \frac{1}{10} c\lambda \|(\nu)_{S^c}\|_1. \end{aligned} \tag{B.21}$$

By rearranging the terms, we will have

$$\frac{1}{n}(\hat{\beta} - \beta^O)^T X^T \epsilon - \frac{1}{10} c\lambda \|(\nu)_{S^c}\|_1 \leq \frac{1}{n}(\beta^O - \hat{\beta})^T X^T X(\beta^* - \beta^O) \tag{B.22}$$

Plugging inequality (B.22) to inequality (B.20), we will have

$$\frac{1}{n}(\beta^O - \hat{\beta})^T X^T X(\beta^O - \hat{\beta}) \leq \sum_{i \notin S} |\hat{\beta}_i||h'_\lambda(\hat{\beta}_i)| - \lambda \|(\nu)_{S^c}\|_1 + \frac{1}{10} c\lambda \|(\nu)_{S^c}\|_1 + 2\lambda \|(\nu)_S\|_1. \tag{B.23}$$

Under the Assumption 3.7 with use of Bonferroni bound as in Corollary 4.2, we will have

$$0 \leq \frac{1}{n}(\beta^O - \hat{\beta})^T X^T X(\beta^O - \hat{\beta}) \leq -\frac{8}{10} c\lambda \|(\nu)_{S^c}\|_1 + 2\lambda \|(\nu)_S\|_1, \tag{B.24}$$

which implies $\lambda \|(\nu)_{S^c}\|_1 \leq \frac{5}{2c} \lambda \|(\nu)_S\|_1$ and $\nu \in \mathcal{C}$. □

## B.9 Proof of Theorem 4.3

PROOF. According to Lemma 4.2, we will have at $\hat{\beta}$ and $\beta^O$, respectively:

$$f_\lambda(\beta^O) \geq f_\lambda(\hat{\beta}) + \nabla f_\lambda(\hat{\beta})^T(\beta^O - \hat{\beta}) + \frac{\gamma - \eta^-}{2}\|\beta^O - \hat{\beta}\|_2^2,$$

$$f_\lambda(\hat{\beta}) \geq f_\lambda(\beta^O) + \nabla f_\lambda(\beta^O)^T(\hat{\beta} - \beta^O) + \frac{\gamma - \eta^-}{2}\|\hat{\beta} - \beta^O\|_2^2.$$

Since $\ell_1$ norm penalty is convex, we will have

$$\lambda\|\hat{\beta}\|_1 \geq \lambda\|\beta^O\|_1 + \lambda(\hat{\beta} - \beta^O)^T \xi^O,$$

$$\lambda\|\beta^O\|_1 \geq \lambda\|\hat{\beta}\|_1 + \lambda(\beta^O - \hat{\beta})^T \hat{\xi},$$

where $\hat{\xi}$ and $\xi^O$ are the same as in Lemma 4.4. Combine the above together, we will have:

$$0 \geq (\nabla f_\lambda(\hat{\beta}) + \lambda\hat{\xi})^T(\beta^O - \hat{\beta}) + (\nabla f_\lambda(\beta^O) + \lambda\xi^O)^T(\hat{\beta} - \beta^O) + (\gamma - \eta^-)\|\beta^O - \hat{\beta}\|_2^2.$$

Since both $\hat{\beta}$ and $\beta^O$ are d-stationary, we will have

$$(\hat{\beta} - \beta^O)^T \left(\nabla f_\lambda(\beta^O) + \lambda\xi^O\right) \geq 0,$$

and $\hat{\beta}$ is d-stationary, there exists a $\hat{\xi} \in \nabla\{\|\hat{\beta}\|_1\}$ such that

$$(\beta^O - \hat{\beta})^T \left(\nabla f_\lambda(\hat{\beta}) + \lambda\hat{\xi}\right) \geq 0.$$

We thus will have $0 \geq (\gamma - \eta^-)\|\beta^O - \hat{\beta}\|_2^2$, which implies that $\beta^O = \hat{\beta}$. □

## APPENDIX C: PROOFS IN SECTION 5

### C.1 Proof of Theorem 5.1

PROOF. Since $\hat{\beta}$ is a d-stationary solution to Problem (5.1), we have

$$\nabla L(\hat{\beta}) + \lambda z - \nabla h_\lambda(\hat{\beta}) = 0, \tag{C.1}$$

where $z \in \partial g(\hat{\beta})$, where $\partial g(\hat{\beta})$ is the set of subgradient of $g(\beta) = \|\beta\|_1$ at $\hat{\beta}$. We can get the gradient for the loss function

$$\nabla L(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\psi'(X_i^T\hat{\beta})X_i - Y_iX_i).$$

We can further write the above expression as

$$\nabla L(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\big((\psi'(X_i^T\hat{\beta})X_i - \psi'(X_i^T\beta^*)X_i) + (\psi'(X_i^T\beta^*)X_i - Y_iX_i)\big),$$

where the term $(\psi'(X_i^T\beta^*) - Y_iX_i)$ does not depend on the d-stationary solution. Multiply both side of (C.1) by $(\beta^* - \hat{\beta})^T$, we have

$$(\beta^*-\hat{\beta})^T\big\{\frac{1}{n}\sum_{i=1}^{n}\big((\psi'(X_i^T\hat{\beta})X_i-\psi'(X_i^T\beta^*)X_i)+(\psi'(X_i^T\beta^*)X_i-Y_iX_i)\big)+\lambda z-\nabla h_\lambda(\hat{\beta})\big\} = 0. \tag{C.2}$$

By rearranging the terms, we have

$$\begin{aligned}
0 &\le (\beta^* - \hat{\beta})^T\big\{\frac{1}{n}\sum_{i=1}^{n}(\psi'(X_i^T\beta^*)X_i - \psi'(X_i^T\hat{\beta})X_i)\big\} \\
&= (\beta^* - \hat{\beta})^T\big\{\frac{1}{n}\sum_{i=1}^{n}(\psi'(X_i^T\beta^*)X_i - Y_iX_i) + \lambda z - \nabla h_\lambda(\hat{\beta})\big\} \\
&\le (\beta^* - \hat{\beta})^T\big\{\frac{1}{n}\sum_{i=1}^{n}(\psi'(X_i^T\beta^*)X_i - Y_iX_i)\big\} + 2\lambda\|(\beta^* - \hat{\beta})_S\|_1 \\
&- \lambda\|\hat{\beta}_{S^c}\|_1 + \hat{\beta}_{S^c}^T\nabla h_\lambda(\hat{\beta}_{S^c}) \\
&\le (2 + \frac{c}{2})\lambda\|(\beta^* - \hat{\beta})_S\|_1 - \frac{c}{2}\lambda\|\hat{\beta}_{S^c}\|_1,
\end{aligned}$$

where the first "$\le$" is due to the convexity of the cumulant function, and the last one is due to the assumptions. We thus conclude that $\hat{\beta} \in \mathcal{C}$, which is defined in the Assumption 5.3. Given the restricted strong convexity, according to Lemma 4.2, let $f(\beta) = L(\beta) - h_\lambda(\beta)$, we will have

$$f_\lambda(\beta^*) \ge f_\lambda(\hat{\beta}) + \nabla f_\lambda(\hat{\beta})^T(\beta^* - \hat{\beta}) + \frac{\gamma - \eta^-}{2}\|\beta^* - \hat{\beta}\|_2^2$$

and

$$f_\lambda(\hat{\beta}) \ge f_\lambda(\beta^*) + \nabla f_\lambda(\beta^*)^T(\hat{\beta} - \beta^*) + \frac{\gamma - \eta^-}{2}\|\beta^* - \hat{\beta}\|_2^2.$$

Adding the above up, we have

$$\nabla f_\lambda(\hat{\beta})^T(\hat{\beta} - \beta^*) \ge \nabla f_\lambda(\beta^*)^T(\hat{\beta} - \beta^*) + (\gamma - \eta^-)\|\beta^* - \hat{\beta}\|_2^2. \tag{C.3}$$

Adding $\lambda z^T(\hat{\beta}-\beta^*)$ to both side, we will have

(C.4)
$$0=(\nabla f_\lambda(\hat{\beta})^T+\lambda z^T)(\hat{\beta}-\beta^*)\geq \nabla f_\lambda(\beta^*)^T(\hat{\beta}-\beta^*)+\lambda z^T(\hat{\beta}-\beta^*)+(\gamma-\eta^-)\|\beta^*-\hat{\beta}\|_2^2,$$

From inequalities (C.4) and (C.3), we have

$$\begin{aligned}(\gamma-\eta^-)\|\beta^*-\hat{\beta}\|_2^2 &\leq -\nabla f_\lambda(\beta^*)^T(\hat{\beta}-\beta^*)-\lambda z^T(\hat{\beta}-\beta^*)\\ &\leq (2+\frac{c}{2})\lambda\|(\beta^*-\hat{\beta})_S\|_1-\frac{c}{2}\lambda\|\hat{\beta}_{S^c}\|_1\\ &\leq (2+\frac{c}{2})\lambda\sqrt{|S|}\|\hat{\beta}-\beta^*\|_2\end{aligned} \tag{C.5}$$

where the last "$\leq$" is due to the fact that $\|(\hat{\beta}-\beta^*)_S\|_1\leq\sqrt{|S|}\|\hat{\beta}-\beta^*\|_2$. We thus derive the bound that

$$\|\beta^*-\hat{\beta}\|_2\leq\frac{(4+c)\lambda}{2(\gamma-\eta^-)}\sqrt{|S|} \tag{C.6}$$

□

## APPENDIX D: PROOFS IN SECTION 6

### D.1 Proof of Lemma 6.1

PROOF. Given $\beta_k$ as the update in the kth iteration, we adopt the following procedure to update the estimation:

$$\beta_{k+1}=\arg\min L_n(\beta)+\sum_{i=1}^{p}(\lambda-h'(|\beta_{ki}|))|\beta_i|. \tag{D.1}$$

Let $Q(\beta|\beta_k)=L_n(\beta)+\lambda\|\beta_k\|_1-h_\lambda(\beta_k)+\sum_{i=1}^p(\lambda-h'(|\beta_{ki}|))(|\beta_i|-|\beta_{ki}|)$. It can be easily checked that $Q(\beta_k|\beta_k)=F(\beta_k)$ and the following is equivalent to D.1:

$$\beta_{k+1}=\min Q(\beta|\beta_k). \tag{D.2}$$

Since $h_\lambda(\cdot)$ is convex by Assumption 3.4, we have

$$F(\beta)\leq Q(\beta|\beta_k).$$

According to the definition of $\beta_{k+1}$, we have

$$F(\beta_{k+1})\leq Q(\beta_{k+1}|\beta_k)\leq Q(\beta_k|\beta_k)=F(\beta_k).$$

□