The Annals of Statistics
2017, Vol. 45, No. 6, 2455–2482
https://doi.org/10.1214/16-AOS1530

SUPPORT RECOVERY WITHOUT INCOHERENCE: A CASE FOR NONCONVEX REGULARIZATION

BY PO-LING LOH[1,2] AND MARTIN J. WAINWRIGHT[2]

University of Wisconsin—Madison and University of California, Berkeley

We develop a new primal-dual witness proof framework that may be used to establish variable selection consistency and ℓ_∞-bounds for sparse regression problems, even when the loss function and regularizer are nonconvex. We use this method to prove two theorems concerning support recovery and ℓ_∞-guarantees for a regression estimator in a general setting. Notably, our theory applies to all potential stationary points of the objective and certifies that the stationary point is unique under mild conditions. Our results provide a strong theoretical justification for the use of nonconvex regularization: For certain nonconvex regularizers with vanishing derivative away from the origin, any stationary point can be used to recover the support without requiring the typical incoherence conditions present in ℓ_1-based methods. We also derive corollaries illustrating the implications of our theorems for composite objective functions involving losses such as least squares, nonconvex modified least squares for errors-in-variables linear regression, the negative log likelihood for generalized linear models and the graphical Lasso. We conclude with empirical studies that corroborate our theoretical predictions.

1. Introduction. The last two decades have generated a significant body of work involving convex relaxations of nonconvex problems arising in high-dimensional sparse regression (e.g., see the papers [2, 5, 7, 9, 28, 31]). In broad terms, the goal is to identify a relatively sparse solution from among a larger set of candidates that yield good fits to the data. A hard sparsity constraint is most directly encoded in terms of the ℓ_0-"norm," which counts the number of nonzero entries in a vector. However, this results in a nonconvex optimization problem that may be NP-hard to solve or even approximate [23, 30]. As a result, much work has focused on a slightly different problem, where the ℓ_0-constraint is replaced by the *convex* ℓ_1-norm (e.g., see the papers [2, 5, 22, 29] and references therein).

Although the ℓ_1-norm encourages sparsity, however, it differs from the ℓ_0-norm in a crucial aspect: whereas the ℓ_0-norm is equal to a constant value for

Received May 2015; revised July 2016.

[1] Supported in part from a Hertz Foundation Fellowship, an NSF Graduate Research Fellowship while studying at Berkeley.

[2] Supported in part by NSF Grants DMS-11-07000 and CIF-31712-23800, Air Force Office of Scientific Research Grant AFOSR-FA9550-14-1-0016, and DY and PD Fellowships, respectively.

MSC2010 subject classifications. 62F12.

Key words and phrases. M-estimator, Lasso, sparsity, nonconvex regularizer, high-dimensional statistics, variable selection.

any nonzero argument, the ℓ_1-norm increases linearly with the absolute value of the argument. This linear increase biases the resulting ℓ_1-regularized solution and noticeably affects the performance of the estimator in finite-sample settings [3, 10, 21]. Accordingly, several authors have proposed alternative forms of nonconvex regularization, including the smoothly clipped absolute deviation (SCAD) penalty [10], the minimax concave penalty (MCP) [36] and the log-sum penalty (LSP) [6]. Such regularizers may be viewed as a hybrid of ℓ_0- and ℓ_1-regularizers: they resemble the ℓ_1-norm within a neighborhood of the origin, but become (asymptotically) constant at larger values. Furthermore, although the nonconvexity of the regularizer causes the overall optimization problem to be nonconvex, numerous empirical studies have shown that gradient-based optimization methods—while only guaranteed to find local optima—often produce estimators with consistently smaller estimation error than the estimators produced by the convex ℓ_1-penalty [3, 10, 15, 21, 40].

In recent years, significant progress has been made in the theory of nonconvex regularizers. Zhang and Zhang [35] showed that *global* optima of nonconvex regularized least squares problems are statistically consistent for the true regression vector, leaving open the question of how to find such optima efficiently. Fan et al. [13] showed that applying one step of a local linear approximation (LLA) algorithm, initialized at a Lasso solution with low ℓ_∞-error, yields a local optimum of the nonconvex regularized least squares problem that satisfies oracle properties; Wang et al. [33] established similar guarantees for the output of a particular path-following algorithm. Our own past work [17] supplies a general set of sufficient conditions under which all stationary points of the nonconvex regularized problem are guaranteed to lie within the statistical precision, as measured in terms of the ℓ_2-error, of the true parameter; this result substantially simplifies the optimization problem to one of finding stationary points, rather than global optima.

Despite these advances, however, an important question has remained open. To wit, are stationary points of such nonconvex problems also consistent for variable selection? Existing results in nonconvex regularization guarantee that the global optimum, or certain local optima, are statistically consistent (e.g., [10, 13, 36, 40]). However, whenever the underlying problem is nonconvex, such results do not preclude the unpleasant possibility of having multiple stationary points within close proximity of the global optimum, only one of which has support equal to the support of the true regression vector. Indeed, recent work by Zhang et al. [37] exhibits ensembles of “hard” sparse regression problems possessing a huge number of local optima that trap local algorithms. Our paper can be viewed as making a complementary contribution, in particular by establishing conditions under which a nonconvex optimization problem *cannot* exhibit such behavior.

For convex objectives, various standard proof techniques for variable selection consistency now exist, including approaches that leverage the Karush–Kuhn–Tucker (KKT) optimality conditions, as well as the primal-dual witness argument (e.g., [16, 31, 38]), which combines the KKT conditions with a strict dual feasibility condition to establish uniqueness. However, the validity of such arguments has relied heavily upon the convexity of both the loss and the regularizer. The first main contribution of our paper is to show that the primal-dual witness proof technique may be modified and extended to handle a certain class of *nonconvex* problems. Our proof involves checking sufficient conditions for local minimality of a properly constructed oracle solution, and the key technical step hinges on the notion of generalized gradients from nonsmooth analysis [8], as well as classical optimization-theoretic results for norm-regularized, smooth, but possibly nonconvex functions [14]. Our main result thereby establishes sufficient conditions for variable selection consistency when both the loss and regularizer are allowed to be nonconvex, provided the loss function satisfies a form of restricted strong convexity and the regularizer satisfies suitable mild conditions. Remarkably, our results demonstrate that for a certain class of regularizers—including the SCAD and MCP regularizers—we may dispense with the usual incoherence conditions required by ℓ_1-based methods, and still guarantee support recovery consistency for all stationary points of the resulting nonconvex program.

We also establish that for the same class of nonconvex regularizers, the unique stationary point is in fact equal to the oracle solution—this is striking, given the long line of work focusing on providing theoretical guarantees for specific local optima. (Note that some authors have established ℓ_∞-bounds for convex penalties under different ℓ_∞-curvature assumptions, without need for incoherence, but the imposed curvature conditions are generally stronger than the restricted strong convexity assumptions used to derive ℓ_1- and ℓ_2-bounds [4, 20, 29]. Our work implies that if a nonconvex penalty is used, one may obtain ℓ_∞-bounds without recourse to the stronger curvature assumptions.) This provides a strong theoretical reason for why certain nonconvex regularizers should be favored over their convex counterparts.

Relation to previous work. Let us now briefly compare the main results of this paper to other related work in the literature, paying attention to key aspects that are crucial to appreciating the novelty of our paper. Several authors have investigated the potential for nonconvex regularizers to deliver estimation and support recovery guarantees under weaker assumptions than those required by the Lasso penalty [10–13, 34, 39, 40]; this line of work demonstrates that in the absence of incoherence conditions, nonconvex regularized problems possess certain local optima that are statistically consistent and satisfy an oracle property. However, the arguments in this line of work are based on applying the KKT stationarity conditions *at a specific point*; thus, they are only able to derive good behavior of *certain* stationary points, as opposed to all stationary points. Since nonconvex programs may possess multiple local optima, it is entirely possible that stationary points with the incorrect support could exist. In contrast, the PDW proof technique developed in our paper provides a rather different guarantee: via a strict dual feasibility condition and additional second-order conditions, the proof certifies that *all* stationary points are consistent for variable selection.

Another line of recent work has focused on establishing theoretical guarantees for specific stationary points that correspond to the output of particular optimization algorithms. Wang et al. [33] propose a path-following homotopy algorithm for obtaining solutions to nonconvex regularized M-estimators, and show that iterates of the homotopy algorithm converge at a linear rate to the oracle solution of the M-estimation problem. In contrast to theory of this type—applicable only to a particular algorithm—the theory in our paper is purely statistical and does *not* concern iterates of a specific optimization algorithm. Again, the novelty of our theoretical results lies in the fact that we establish support recovery consistency for *all* stationary points, showing that *any* optimization algorithm that is guaranteed to converge to a stationary point is suitable for optimization. Pan and Zhang [24] also provide related but weaker guarantees showing that under restricted eigenvalue assumptions on the design matrix that are less stringent than the standard restricted eigenvalue conditions, a certain class of nonconvex regularizers yields estimates that are consistent in ℓ_2-norm. They provide bounds on the sparsity of approximate global and approximate sparse (AGAS) solutions, a notion also studied in earlier work [35]. However, their theoretical development stops short of providing conditions for recovering the exact support of the underlying regression vector.

Finally, as pointed out by a reviewer, the result of Fan et al. [13] implies variable selection consistency of stationary points that are close enough to β^*, since stationary points are fixed points for the second stage of their two-step optimization algorithm. However, a careful examination of the paper [13] reveals that only stationary points that are close in ℓ_∞-norm are guaranteed to be consistent for variable selection. The ℓ_1- and ℓ_2-error bounds derived in our earlier work [17] are *not* strong enough to imply that all stationary points are close enough in ℓ_∞-norm; in fact, our Theorem 2 below, which we derive via our novel PDW machinery, is the first result we are aware of that gives $\mathcal{O}(\sqrt{\frac{\log p}{n}})$ rates in ℓ_∞-norm for *all* stationary points of this class of nonconvex problems.

The remainder of our paper is organized as follows. In Section 2, we provide background on regularized M-estimators and set up the assumptions on loss functions and regularizers to be analyzed in the paper. We also outline the primal-dual witness proof method. Section 3 is devoted to the statements of our main results concerning support recovery and ℓ_∞-bounds, followed by corollaries that specialize our results to particular objective functions. In each case, we contrast our conditions for nonconvex regularizers to those required by convex regularizers and discuss the implications of our significantly weaker assumptions. We provide proof sketches outlining the key components in the proofs of our main results in Section 4, with proofs of more technical lemmas contained in the supplemental appendix [19]. Section 5 contains illustrative simulations that confirm our theoretical results.

Notation. Unless specifically noted, we use the standard notational convention that constants c_1, c_2, etc., refer to universal positive constants, with values that may differ from line to line. For functions $f(n)$ and $g(n)$, we write $f(n) \precsim g(n)$ to mean $f(n) \le cg(n)$ for some constant $c \in (0, \infty)$, and similarly, $f(n) \succsim g(n)$ when $f(n) \ge c'g(n)$ for some constant $c' \in (0, \infty)$. We write $f(n) \asymp g(n)$ when $f(n) \precsim g(n)$ and $f(n) \succsim g(n)$ hold simultaneously. For a vector $v \in \mathbb{R}^p$ and a subset $S \subseteq \{1, \dots, p\}$, we write $v_S \in \mathbb{R}^S$ to denote the vector v restricted to S. For a matrix M, we write $|||M|||_2$ and $|||M|||_F$ to denote the spectral and Frobenius norms, respectively, and write $|||M|||_\infty$ to denote the ℓ_∞-operator norm. We write $\|M\|_{\max} := \max_{i,j} |m_{ij}|$ to denote the elementwise ℓ_∞-norm of M. For a function $h : \mathbb{R}^p \to \mathbb{R}$, we write ∇h to denote a gradient or subgradient, if it exists. Finally, for $q, r > 0$, we write $\mathbb{B}_q(r)$ to denote the ℓ_q-ball of radius r centered around 0.

2. Problem formulation.

In this section, we briefly review the theory of regularized M-estimators. We also outline the primal-dual witness method that underlies our proofs of variable selection consistency.

2.1. *Regularized M-estimators.*

The analysis in this paper applies to regularized M-estimators of the form

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \le R, \beta \in \Omega} \{\mathcal{L}_n(\beta) + \rho_\lambda(\beta)\}, \tag{2.1}$$

where $\mathcal{L}_n$ denotes the empirical loss function and ρ_λ denotes the penalty function, both assumed to be continuous. In our framework, both of these functions are allowed to be nonconvex. The prototypical example of a loss function is the least squares objective $\mathcal{L}_n(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2$. We include the side constraint $\|\beta\|_1 \le R$ in order to ensure that a global minimum $\widehat{\beta}$ exists.[3] For modeling purposes, we have also allowed for an additional constraint, $\beta \in \Omega$, where Ω is an open convex set; note that we may take $\Omega = \mathbb{R}^p$ when this extra constraint is not needed.

The analysis of this paper is restricted to the class of *coordinate-separable regularizers*, meaning that ρ_λ is expressible as the sum:

$$\rho_\lambda(\beta) = \sum_{j=1}^{p} \rho_\lambda(\beta_j). \tag{2.2}$$

We have engaged in a minor abuse of notation; the functions $\rho_\lambda : \mathbb{R} \to \mathbb{R}$ appearing on the right-hand side of equation (2.2) are univariate functions acting upon each

[3] In the sequel, we will give examples of nonconvex loss functions for which the global minimum fails to exist without such a side constraint (cf. Section 2.3 below).

coordinate. Our results readily extend to the inhomogenous case, where coordinate j has regularizer ρ_λ^j.

From a statistical perspective, the purpose of solving the program (2.1) is to estimate the vector $\beta^* \in \mathbb{R}^p$ that minimizes the expected loss:

$$\beta^* := \arg\min_{\beta \in \Omega} \mathbb{E}[\mathcal{L}_n(\beta)], \tag{2.3}$$

where we assume that β^* is unique and independent of the sample size. Our goal is to develop conditions under which a minimizer $\widehat{\beta}$ of the composite objective (2.1) is consistent for β^*. Consequently, we will always choose $R \geq \|\beta^*\|_1$, which ensures that β^* is a feasible point.

2.2. *Assumptions on regularizers.*

We will study the class of regularizers $\rho_\lambda : \mathbb{R} \to \mathbb{R}$ that are amenable in the following sense.

Amenable regularizers. For some with $\mu \geq 0$, we say that ρ_λ is μ-amenable if:

(i) The function $t \mapsto \rho_\lambda(t)$ is symmetric around zero [i.e., $\rho_\lambda(t) = \rho_\lambda(-t)$ for all t], and $\rho_\lambda(0) = 0$.
(ii) The function $t \mapsto \rho_\lambda(t)$ is nondecreasing on $\mathbb{R}^+$.
(iii) The function $t \mapsto \frac{\rho_\lambda(t)}{t}$ is nonincreasing on $\mathbb{R}^+$.
(iv) The function $t \mapsto \rho_\lambda(t)$ is differentiable, for $t \neq 0$.
(v) The function $t \mapsto \rho_\lambda(t) + \frac{\mu}{2}t^2$ is convex, for some $\mu > 0$.
(vi) $\lim_{t \to 0^+} \rho_\lambda'(t) = \lambda$.

We say that ρ_λ is (μ, γ)-amenable if, in addition:

(vii) There exists a scalar $\gamma \in (0, \infty)$ such that $\rho_\lambda'(t) = 0$, for all $t \geq \gamma\lambda$.

Conditions (vi) and (vii) are also known as the *selection* and *unbiasedness* properties, respectively.

Note that the usual ℓ_1-penalty $\rho_\lambda(t) = \lambda|t|$ is 0-amenable, but it is *not* $(0, \gamma)$-amenable, for any $\gamma < \infty$. The notion of μ-amenability was also used in our past work on ℓ_2-bounds for nonconvex regularizers [17], without the selection property (vi). Since the goal of the current paper is to obtain *stronger* conclusions, in terms of variable selection and ℓ_∞-bounds, we will also require ρ_λ to satisfy conditions (vi)–(vii).

Note that if we define $q_\lambda(t) := \lambda|t| - \rho_\lambda(t)$, the conditions (iv) and (vi) together imply that q_λ is everywhere differentiable. Furthermore, if ρ_λ is (μ, γ)-amenable, we have $q_\lambda'(t) = \lambda \cdot \text{sign}(t)$, for all $|t| \geq \gamma\lambda$. Many popular regularizers are either μ-amenable or (μ, γ)-amenable. Appendix A.1 contains definitions of common regularizers discussed in the paper, and Appendix A.2 supplies additional useful results concerning amenable regularizers.

2.3. *Assumptions on loss functions.* We now describe the types of nonconvex loss functions we will discuss in this paper. We will consider loss functions that are twice-differentiable and satisfy a form of *restricted strong convexity*, as used in large body of past work on high-dimensional sparse M-estimators (e.g., [2, 17, 22, 29]). In order to provide intuition before stating the formal definition, note that for any convex and differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ that is globally convex and locally strongly convex around a point $\beta \in \mathbb{R}^p$, there exists a constant $\alpha > 0$ such that

$$\langle \nabla f(\beta + \Delta) - \nabla f(\beta), \Delta \rangle \ge \alpha \cdot \min\{\|\Delta\|_2, \|\Delta\|_2^2\}, \tag{2.4}$$

for all $\Delta \in \mathbb{R}^p$. The notion of restricted strong convexity (with respect to the ℓ_1-norm) weakens this requirement by adding a tolerance term:

Restricted strong convexity (RSC). Given any pair of vectors $\beta, \Delta \in \mathbb{R}^p$, the loss function $\mathcal{L}_n$ satisfies an (α, τ)-RSC condition if:

$$\begin{aligned} &\text{(2.5a)} \\ &\langle \nabla \mathcal{L}_n(\beta + \Delta) - \nabla \mathcal{L}_n(\beta), \Delta \rangle \ge \\ &\text{(2.5b)} \end{aligned} \begin{cases} \alpha_1 \|\Delta\|_2^2 - \tau_1 \dfrac{\log p}{n} \|\Delta\|_1^2, & \forall \|\Delta\|_2 \le 1, \\ \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\dfrac{\log p}{n}} \|\Delta\|_1, & \forall \|\Delta\|_2 \ge 1, \end{cases}$$

where $\alpha_1, \alpha_2 > 0$ and $\tau_1, \tau_2 \ge 0$.

As noted in inequality (2.4), any locally strongly convex function that is also globally convex satisfies the RSC condition with $\tau_1 = \tau_2 = 0$. For $\tau_1, \tau_2 > 0$, the RSC condition imposes strong curvature only in certain directions of p-dimensional space—namely, those nonzero directions $\Delta \in \mathbb{R}^p$ for which the ratio $\frac{\|\Delta\|_1}{\|\Delta\|_2}$ is relatively small. Note that for any k-sparse vector Δ, we have $\frac{\|\Delta\|_1}{\|\Delta\|_2} \le \sqrt{k}$, so that the RSC definition guarantees a form of strong convexity for all k-sparse vectors when $n \gtrsim k \log p$.

A line of past work (e.g., [17, 22, 25, 27]) shows that the RSC condition holds, with high probability, for many types of convex and nonconvex objectives arising in statistical estimation problems. We will elaborate on some specific examples of interest in Section 3 below.

2.4. *Primal-dual witness proof technique.* Finally, we outline the main steps of our new primal-dual witness (PDW) proof construction, which will yield support recovery and ℓ_∞-bounds for the program (2.1). We emphasize that the steps described below provide a significant generalization of the previous proof construction proposed in literature to establish support recovery in high-dimensional estimation. In particular, whereas such a technique was previously applicable only under assumptions of convexity on $\mathcal{L}_n$ and ρ_λ, we show that this machinery may be extended via a careful analysis of local optima of norm-regularized functions based on generalized gradients.

As stated in Theorem 1 below, the success of the PDW construction guarantees that stationary points of the nonconvex objective are consistent for variable selection consistency—in fact, they are unique. Recall that $\widetilde{\beta} \in \mathbb{R}^p$ is a *stationary point* of the program (2.1) if $\langle \nabla \mathcal{L}_n(\widetilde{\beta}) + \nabla \rho_\lambda(\widetilde{\beta}), \beta - \widetilde{\beta} \rangle \geq 0$, for all β in the feasible region [1]. Due to the possible nondifferentiability of ρ_λ at 0, we abuse notation slightly and denote

$$\langle \nabla \rho_\lambda(\widetilde{\beta}), \beta - \widetilde{\beta} \rangle = \lim_{t \to 0^+} \langle \nabla \rho_\lambda(\widetilde{\beta} + t(\beta - \widetilde{\beta})), \beta - \widetilde{\beta} \rangle$$

(see, e.g., Clarke [8] for more details on generalized gradients). The set of stationary points includes all local/global minima of the program (2.1), as well as any interior local maxima.

The key steps of the PDW argument are as follows. Note that the PDW construction is merely a *proof technique* for establishing properties of stationary points, rather than a construction to be performed on the data.

Steps of PDW construction.

(i) Optimize the *restricted program*:

$$\widehat{\beta}_S \in \arg \min_{\beta \in \mathbb{R}^S : \|\beta\|_1 \leq R, \beta \in \Omega} \{\mathcal{L}_n(\beta) + \rho_\lambda(\beta)\}, \tag{2.6}$$

where we enforce the additional constraint $\operatorname{supp}(\widehat{\beta}_S) \subseteq \operatorname{supp}(\beta^*) := S$. Establish that $\|\widehat{\beta}_S\|_1 < R$; that is, $\widehat{\beta}_S$ is in the interior of the feasible set.

(ii) Define $\widehat{z}_S \in \partial \|\widehat{\beta}_S\|_1$, and choose $\widehat{z}_{S^c}$ to satisfy the zero-subgradient condition:

$$\nabla \mathcal{L}_n(\widehat{\beta}) - \nabla q_\lambda(\widehat{\beta}) + \lambda \widehat{z} = 0, \tag{2.7}$$

where $\widehat{z} = (\widehat{z}_S, \widehat{z}_{S^c})$, $\widehat{\beta} := (\widehat{\beta}_S, 0_{S^c})$, and $q_\lambda(t) := \lambda |t| - \rho_\lambda(t)$. Establish *strict dual feasibility* of $\widehat{z}_{S^c}$; that is, $\|\widehat{z}_{S^c}\|_\infty < 1$.

(iii) Show that $\widehat{\beta}$ is a *local minimum* of the full program (2.1), and moreover, *all* stationary points of the program (2.1) are supported on S.

Note that the output $(\widehat{\beta}, \widehat{z})$ of the PDW construction depends on λ and R.

Under the RSC condition, the restricted problem (2.6) minimized in step (i) is actually a convex program. Hence, if $\|\widehat{\beta}_S\|_1 < R$, the zero-subgradient condition (2.7) must hold at $\widehat{\beta}_S$ for the restricted problem (2.6). Note that when $\mathcal{L}_n$ is convex and ρ_λ is the ℓ_1-penalty as in the conventional setting, the additional ℓ_1-constraint in the programs (2.1) and (2.6) is omitted. If also $\Omega = \mathbb{R}^p$, the vector $\widehat{\beta}_S$ is automatically a zero-subgradient point if it is a global minimum of the restricted program (2.6), which greatly simplifies the analysis. Our refined analysis shows that under suitable restrictions, global optimality still holds for $\widehat{\beta}_S$ and $\widehat{\beta}$, and the convexity of the restricted program therefore implies uniqueness.

To further highlight the differences between our PDW construction and the construction mentioned in previous literature [31], we remark that due to the nonconvexity of the objective function, a zero-subgradient condition alone cannot guarantee that the zero-padded vector $\widehat{\beta}$ constructed from the optimum $\widehat{\beta}_S$ of the restricted program (2.6) is a local (global) minimum of the full program (2.1). However, as we will argue in more detail in Section 4 below, the RSC condition, combined with certain sufficient conditions for local optima of norm-regularized functions (cf. Lemma 10 in Appendix G.2), allows us to establish that $\widehat{\beta}$ is indeed a local optimum in step (iii) of the construction. An additional technical argument is employed to show that all stationary points of the program (2.1) are also supported on S.

3. Main results and consequences.

In the sections to follow, we use the primal-dual witness proof technique to establish support recovery results for general nonconvex regularized M-estimators and derive conditions under which stationary points of the program (2.1) are unique. We then specialize our results to specific problems of interest.

3.1. *Main results*.

Our main statistical results concern stationary points of the regularized M-estimator (2.1), where the loss function satisfies the RSC condition (2.5) with parameters $\{(\alpha_j, \tau_j)\}_{j=1}^2$, and the regularizer is μ-amenable with $\frac{3}{4}\mu < \alpha_1$. Our first theorem concerns the success of the PDW construction described in Section 2.4. The theorem guarantees that the support of the vector $\widehat{\beta}$ obtained from step (ii) of the PDW construction is the unique stationary point of the regularized program (2.1), provided two conditions are met, the first involving an appropriate choice of λ and R and the second involving strict dual feasibility of the vector $\widehat{z}$. In particular, we may conclude that $\operatorname{supp}(\widehat{\beta}) \subseteq \operatorname{supp}(\beta^*)$. Note that it is through validating the second condition that the incoherence assumption arises in the usual ℓ_1-analysis, but we demonstrate in our corollaries to follow that strict dual feasibility may be guaranteed under *weaker* conditions when a (μ, γ)-amenable regularizer is used. (See Appendix C for a technical discussion.)

THEOREM 1 (PDW construction for nonconvex functions). *Suppose $\mathcal{L}_n$ is a twice-differentiable, (α, τ)-RSC function and ρ_λ is μ-amenable, for some $\frac{3}{4}\mu < \alpha_1$. Further suppose that*:

(a) *The parameters (λ, R) satisfy the bounds*

$$4\max\left\{\|\nabla\mathcal{L}_n(\beta^*)\|_\infty, \alpha_2\sqrt{\frac{\log k}{n}}\right\} \le \lambda \le \sqrt{\frac{(4\alpha_1 - 3\mu)\alpha_2}{384k}}, \quad \textit{and} \tag{3.1a}$$

$$\max\left\{2\|\beta^*\|_1, \frac{48k\lambda}{4\alpha_1 - 3\mu}\right\} \le R \le \min\left\{\frac{\alpha_2}{8\lambda}, \frac{\alpha_2}{\tau_2}\sqrt{\frac{n}{\log p}}\right\}. \tag{3.1b}$$

(b) *For some* $\delta \in [\frac{4R\tau_1 \log p}{n\lambda}, 1]$, *the dual vector* $\widehat{z}$ *from the PDW construction satisfies the strict dual feasibility condition*:

$$\|\widehat{z}_{S^c}\|_\infty \leq 1 - \delta. \tag{3.2}$$

Then if $n \geq \frac{2\tau_1}{2\alpha_1 - \mu} k \log p$ *and* β^* *is* k*-sparse*, *the program* (2.1) *has a unique stationary point*, *given by the primal output* $\widehat{\beta}$ *of the PDW construction.*

Of course, Theorem 1 is vacuous unless proper choices of λ, R and δ exist. In the corollaries to follow, we show that $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \leq c\sqrt{\frac{\log p}{n}}$, with high probability, in many settings of interest. In particular, we may choose $\lambda \asymp \sqrt{\frac{\log p}{n}}$ to satisfy inequality (3.1a) when $n \gtrsim k \log p$. Note that $R \asymp \frac{1}{\lambda}$ then causes inequality (3.1b) to be satisfied under the same sample size scaling. Finally, note that the inequality $\frac{4R\tau_1 \log p}{n\lambda} \leq 1$ is satisfied as long as $R \leq \frac{n\lambda}{4\tau_1 \log p}$, which is guaranteed by the preceding choice of (λ, R) and the scaling $n \gtrsim k \log p$. This ensures the existence of an appropriate δ.[4]

We now remark briefly on the proof of Theorem 1; more details are provided in Section 4. As outlined in Section 2.4, the proof proceeds by constructing a vector $\widehat{\beta}$ supported on S that we will show is the unique optimum. Clearly, the appropriate vector is a zero-filled version of the $|S|$-dimensional vector obtained by minimizing the program (2.6). We first use results on ℓ_1-consistency to argue that the constructed vector lies in the interior of the feasible region. In order to establish that it is a local optimum of the full program (2.1), we construct a dual witness vector $\widehat{z}$ satisfying first-order necessary conditions. By verifying the appropriate second-order sufficient conditions, which follow from strict dual feasibility and restricted strong convexity in a neighborhood around β^*, we establish that $\widehat{\beta}$ is indeed a local optimum. Finally, further algebraic manipulations with respect to $\widehat{\beta}$ show that other stationary points must be supported on S, as well.

We also note that our results require the assumption $3\mu < 4\alpha_1$, where a smaller gap of $(4\alpha_1 - 3\mu)$ translates into a larger sample size requirement. This consideration may motivate an advantage of using the LSP regularizer over a regularizer such as SCAD or MCP; as discussed in Appendix A.1, the SCAD and MCP regularizers have μ equal to a constant value, whereas $\mu = \lambda^2 \to 0$ for the LSP. On the other hand, the LSP is *not* (μ, γ)-amenable, which as discussed later, allows us to remove the incoherence condition for SCAD and MCP when establishing strict dual feasibility (3.2). This suggests that for more incoherent designs, the LSP may be preferred for variable selection, whereas for less incoherent designs, SCAD or MCP may be better. (In simulations, however, the LSP regularizer only performs negligibly better than the ℓ_1-penalty in situations where the incoherence

[4] Note that the parameter δ does not appear in the statistical estimation procedure and is simply a byproduct of the PDW analysis. Hence, it is not necessary to know or estimate a valid value of δ.

condition holds and the same regularization parameter λ is chosen.) Finally, note that although the conditions of Theorem 1 are only *sufficient* conditions. Indeed, as confirmed experimentally, many situations exist where the condition $3\mu < 4\alpha_1$ does *not* hold, yet the stationary points of the program (2.1) still appear to be supported on S and/or unique.

Our second theorem provides control on the ℓ_∞-error between any stationary point and β^*, and shows that if the regularizer is (μ, γ)-amenable, the unique local/global optimum is the *oracle estimator*, which is the unpenalized estimator obtained from minimizing $\mathcal{L}_n$ over the true support set S. Let $\widehat{\beta}_S^{\mathcal{O}} := \arg\min_{\beta_S \in \mathbb{R}^S}\{\mathcal{L}_n(\beta_S, 0_{S^c})\}$, and let $\widehat{\beta}^{\mathcal{O}} := (\widehat{\beta}_S^{\mathcal{O}}, 0_{S^c})$ be the oracle estimator. Note that under the assumed RSC conditions, the restricted function $\mathcal{L}_n|_S$ is strictly convex and $\widehat{\beta}_S^{\mathcal{O}}$ is uniquely defined. With this notation, we have the following result.

THEOREM 2. *Suppose the assumptions of Theorem 1 are satisfied. The unique stationary point $\widehat{\beta}$ of the program (2.1) has the following properties*:

(a) *Let $\widehat{Q} := \int_0^1 \nabla^2 \mathcal{L}_n(\beta^* + t(\widehat{\beta} - \beta^*))\,dt$, and suppose $\widehat{Q}_{SS}$ is invertible. Then*

$$\|\widehat{\beta} - \beta^*\|_\infty \le \|(\widehat{Q}_{SS})^{-1}\nabla\mathcal{L}_n(\beta^*)_S\|_\infty + \lambda |||(\widehat{Q}_{SS})^{-1}|||_\infty. \tag{3.3}$$

(b) *Moreover, if ρ_λ is (μ, γ)-amenable and $\beta^*_{\min} := \min_{j \in S}|\beta^*_j|$ is lower-bounded as*

$$\beta^*_{\min} \ge \lambda(\gamma + |||(\widehat{Q}_{SS})^{-1}|||_\infty) + \|(\widehat{Q}_{SS})^{-1}\nabla\mathcal{L}_n(\beta^*)_S\|_\infty, \tag{3.4a}$$

then $\widehat{\beta}$ agrees with the oracle estimator $\widehat{\beta}^{\mathcal{O}}$, and we have

$$\|\widehat{\beta} - \beta^*\|_\infty \le \|(\widehat{Q}_{SS})^{-1}\nabla\mathcal{L}_n(\beta^*)_S\|_\infty. \tag{3.4b}$$

The proof of Theorem 2 is provided in Section 4.3. Note that in part (a), the integral appearing in the definition of $\widehat{Q}$ is taken componentwise.

Theorem 2 underscores the strength of (μ, γ)-amenable regularizers: with the addition of a beta-min condition (3.4a), the unbiasedness property allows us to remove the second term in inequality (3.3) and obtain a faster oracle rate (3.4b). In the corollaries below, we demonstrate typical scenarios where the right-hand expression in inequality (3.4b) is $\mathcal{O}(\sqrt{\frac{\log p}{n}})$, with high probability, when the spectrum of $\nabla^2\mathcal{L}_n(\beta^*)$ is bounded appropriately.

It is worth noting the relationship between the ℓ_∞-bounds guaranteed by Theorem 2(b) and the results of Fan et al. [13], who prove certain results about one step of the LLA algorithm when initialized at a Lasso solution with low ℓ_∞-error. Since stationary points of the nonconvex regularized program (2.1) are fixed points of the LLA algorithm, the results of that paper guarantee that stationary points with ℓ_∞-norm error on the order of $\mathcal{O}(\sqrt{\frac{\log p}{n}})$ are equal to the oracle estimator. This conclusion is consistent with Theorem 2; however, our theorem also shows

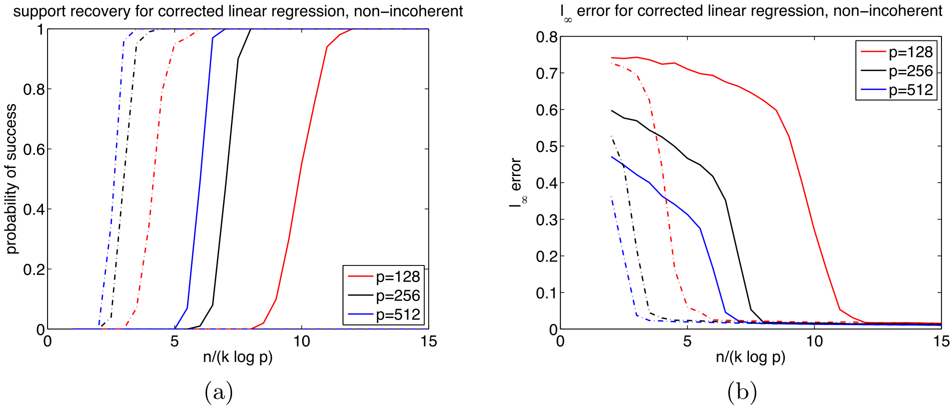


FIG. 1. *Simulation results for least squares linear regression with covariates corrupted by additive noise.* (a) *Plot showing variable selection consistency with the SCAD* (*solid*) *and MCP* (*dash-dotted*) *regularizers. The probability of success in recovering the correct signed support transitions sharply from* 0 *to* 1 *as a function of the sample size, agreeing with the predictions of Theorem* 1; *such behavior is* not *observed when using the* ℓ_1*-penalty or LSP.* (b) *Plot showing* ℓ_∞*-error* $\|\widehat{\beta} - \beta^*\|_\infty$ *with the SCAD* (*solid*) *and MCP* (*dash-dotted*) *regularizers. As predicted by Theorem* 2, *both regularizers demonstrate consistency in* ℓ_∞*-error.*

that when the PDW construction succeeds, the program (2.1) does not possess any stationary points outside the designated ℓ_∞-norm ball, either. Thus, we have the same guarantees but without assuming the initial Lasso estimate has low ℓ_∞-error.

Figure 1 illustrates the guarantees delivered by Theorems 1 and 2 in the case of errors-in-variables linear regression. As detailed in Section 3.3 below, the PDW construction is shown to succeed with high probability when using the SCAD or MCP regularizer, implying that the estimated vector $\hat{\beta}$ has the correct signed support when $n \gtrsim k \log p$ and the ℓ_∞-error is $\mathcal{O}(\sqrt{\frac{\log p}{n}})$. In contrast, the PDW construction does *not* succeed when only a μ-amenable regularizer is used, due to non-incoherence of the design matrix.

We now unpack Theorems 1 and 2 in several concrete settings. Theory for the graphical Lasso is developed in Appendix E.

3.2. *Linear regression.*

Our first application focuses on the setting of ordinary least squares, together with the nonconvex SCAD, MCP and LSP regularizers. We compare the consequences of Theorems 1 and 2 for each of these regularizers with the corresponding results for the convex ℓ_1-penalty. Our theory demonstrates a clear advantage of using nonconvex regularizers such as SCAD and MCP that are (μ, γ)-amenable; whereas support recovery using ℓ_1-based methods requires fairly stringent incoherence conditions, our corollaries show that methods based on nonconvex regularizers guarantee support recovery even without incoherence conditions.

Consider the standard linear regression model, in which we observe i.i.d. pairs $\{(x_i, y_i)\}_{i=1}^n$, linked by the linear model:

$$y_i = x_i^T \beta^* + \varepsilon_i, \qquad \text{for } i = 1, \dots, n,$$

and the goal is to estimate $\beta^* \in \mathbb{R}^p$. A standard loss function in this case is the least squares function $\mathcal{L}_n(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2$, where $y \in \mathbb{R}^n$ is the vector of responses and $X \in \mathbb{R}^{n\times p}$ is the design matrix with $x_i^T \in \mathbb{R}^p$ as its ith row. For any $\beta, \Delta \in \mathbb{R}^p$, we have

$$\langle \nabla\mathcal{L}_n(\beta + \Delta) - \nabla\mathcal{L}_n(\beta), \Delta\rangle = \Delta^T\left(\frac{X^T X}{n}\right)\Delta = \frac{\|X\Delta\|_2^2}{n}.$$

Consequently, for the least squares loss function, the RSC condition is essentially equivalent to lower-bounding sparse restricted eigenvalues [2, 29].

Setting $\Omega = \mathbb{R}^p$, the ρ_λ-regularized least squares estimator takes the form

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \le R}\left\{\frac{1}{2}\beta^T \frac{X^T X}{n}\beta - \frac{y^T X}{n}\beta + \rho_\lambda(\beta)\right\}. \tag{3.5}$$

Note that the Hessian of the loss function is $\nabla^2\mathcal{L}_n(\beta) = \frac{X^T X}{n}$. Although the sample covariance matrix is always positive semidefinite, it has rank at most n. Hence, in high-dimensional settings where $n < p$, the Hessian of the loss function has at least $p - n$ zero eigenvalues, implying that *any* nonconvex regularizer ρ_λ makes the overall program (3.5) nonconvex.

In analyzing the family of estimators (3.5), we assume that $n \ge c_0 k \log p$, for some constant c_0. (By known information-theoretic results [32], such a lower bound is required for any method to recover the support of a k-sparse signal.) The following result is proved in Appendix D.

COROLLARY 1. *Suppose X and ε are sub-Gaussian, and the parameters (λ, R) are chosen such that $\|\beta^*\|_1 \le \frac{R}{2}$ and $c_\ell\sqrt{\frac{\log p}{n}} \le \lambda \le \frac{c_u}{R}$, for some constants c_ℓ and c_u. Also suppose the sample covariance matrix $\widehat{\Gamma} = \frac{X^T X}{n}$ satisfies the condition:*

$$|||\widehat{\Gamma}_{SS}^{-1}|||_\infty \le c_\infty. \tag{3.6}$$

(a) *Suppose ρ_λ is μ-amenable, with $3\mu < 2\lambda_{\min}(\Sigma_x)$, and $\widehat{\Gamma}$ also satisfies the incoherence condition:*

$$|||\widehat{\Gamma}_{S^c S}\widehat{\Gamma}_{SS}^{-1}|||_\infty \le \eta < 1. \tag{3.7}$$

Then with probability at least $1 - c_1\exp(-c_2\min\{k, \log p\})$, the nonconvex objective (3.5) has a unique stationary point $\widehat{\beta}$, which corresponds to the global optimum. Furthermore, $\operatorname{supp}(\widehat{\beta}) \subseteq \operatorname{supp}(\beta^*)$, and

$$\|\widehat{\beta} - \beta^*\|_\infty \le C\sqrt{\frac{\log p}{n}} + c_\infty\lambda. \tag{3.8}$$

(b) *Suppose the regularizer* ρ_λ *is* (μ, γ)*-amenable, with* $3\mu < 2\lambda_{\min}(\Sigma_x)$. *Also suppose*

$$\beta^*_{\min} \geq \lambda(\gamma + c_\infty) + C\sqrt{\frac{\log p}{n}}.$$

Then with probability at least $1 - c_1 \exp(-c_2 \min\{k, \log p\})$, *the nonconvex objective* (3.5) *has a unique stationary point* $\widehat{\beta}$ *given by the oracle estimator* $\widehat{\beta}^{\mathcal{O}}$, *and*

$$\|\widehat{\beta} - \beta^*\|_\infty \leq C\sqrt{\frac{\log p}{n}}. \tag{3.9}$$

Note that if we also have the beta-min condition $\beta^*_{\min} \geq 2(C\sqrt{\frac{\log p}{n}} + c_\infty \lambda)$ in part (a), then $\widehat{\beta}$ is still a sign-consistent estimate of β^*; however, the bound (3.8) is looser than the oracle bound (3.9) derived in part (b).

Remark. The constants c_u, c_ℓ and C appearing in the statement of the corollary above depend on the sub-Gaussian parameters σ_x and σ_ε of X and ε, respectively, but we have chosen to suppress this dependence in the statement of the corollary in order to simplify our presentation. Indeed, the required values of c_u and c_ℓ may be derived from equation (3.1) in Theorem 1: Since $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty = \|\frac{X^T \varepsilon}{n}\|_\infty$ and $\alpha_2 \asymp \lambda_{\min}(\Sigma_x)$, we would need $c_\ell \asymp \sigma_x \sigma_\varepsilon$ and $c_u \asymp \lambda_{\min}(\Sigma_x)$. Consequently, the optimal choice of λ would scale explicitly as $\lambda \asymp \sigma_x \sigma_\varepsilon \sqrt{\frac{\log p}{n}}$, as well. As shown in the proof of the corollary, we may take $C = \lambda_{\max}^{1/2}(\Sigma_x)\sigma_\varepsilon \sqrt{2}$. Finally, note that a term of the form c_∞ in the $\beta^*_{\min}$ condition is also *necessary* for support recovery in the case of the μ-amenable regularizers studied here (cf. Theorem 2 in Wainwright [31]).

The distinguishing point between parts (a) and (b) in Corollary 1 is that using (μ, γ)-amenable regularizers allows us to dispense with an incoherence assumption (3.7) and guarantees that the unique stationary point is in fact equal to the oracle estimator. Regularizers satisfying the conditions of part (b) include the SCAD and MCP. Recall that for the SCAD penalty, we have $\mu = \frac{1}{a-1}$; and for the MCP, we have $\mu = \frac{1}{b}$ (cf. Appendix A.1). Hence, the lower-eigenvalue condition translates into $\frac{3}{a-1} < 2\lambda_{\min}(\Sigma_x)$ and $\frac{3}{b} < 2\lambda_{\min}(\Sigma_x)$, respectively. The LSP penalty is an example of a regularizer that satisfies the conditions of part (a), but not part (b): with this choice, we have $\mu = \lambda^2$, so the condition $3\mu < 2\lambda_{\min}(\Sigma_x)$ is satisfied asymptotically whenever $\lambda_{\min}(\Sigma_x)$ is bounded below by a constant. A version of part (a) also holds for the ℓ_1-penalty, as shown in past work [31].

Perhaps the closest existing work on nonconvex regularization of ordinary least squares is in the characterization of “approximate local” and “approximate global” solutions due to Zhang and Zhang [35], defined according to a bound on the value of the subgradient and the value of the objective function, respectively. Their paper establishes approximate sparsity of such solutions, which is a relaxed version of support recovery. They also prove that the objective has a unique approximate local solution that is also sparse, thereby coinciding with the oracle solution, and that any local solution which is also approximately global also coincides with this oracle estimator. However, our Corollary 1 addresses all local solutions simultaneously, showing that these local solutions all have the correct support and are unique—without recourse to approximate global properties or prior knowledge of sparsity. In addition, when the regularizer is (μ, γ)-amenable, corresponding to the regularizers studied in Zhang and Zhang [35], we again achieve the oracle estimator.

Regarding removing incoherence conditions under nonconvex penalization, Zhang [34] shows that the two-step MC+ estimator [beginning with a global optimum of the program (3.5) with the MCP regularizer] is guaranteed to be consistent for variable selection, under only a sparse eigenvalue assumption on the design matrix. Our result shows that the global optimum obtained in the MCP step is actually already guaranteed to be consistent for variable selection, under only slightly stronger assumptions involving lower- and upper-eigenvalue bounds on the design matrix. In another related paper, Wainwright [32] establishes necessary conditions for support recovery in a linear regression setting when the covariates are Gaussian. As remarked in that paper, the necessary conditions only require eigenvalue bounds on the design matrix, in contrast to the more stringent incoherence conditions appearing in analysis of the Lasso [31, 38].

3.3. *Linear regression with corrupted covariates.* We now shift our focus to a setting where the loss function is nonconvex. Consider a simple extension of the linear model: The pairs $\{(x_i, y_i)\}_{i=1}^n$ are again drawn according to the standard linear model, $y_i = x_i^T \beta^* + \varepsilon_i$. However, instead of observing the covariates $x_i \in \mathbb{R}^p$ directly, we observe the corrupted vectors $z_i = x_i + w_i$, where $w_i \in \mathbb{R}^p$ is a noise vector. This setup is a particular instantiation of a more general errors-in-variables model for linear regression; note that the standard Lasso estimate [applied to the observed pairs $\{(z_i, y_i)\}_{i=1}^n$] is *inconsistent* in this setting.

As studied in our previous work [18], it is natural to consider a corrected version of the Lasso, which we state in terms of the quadratic objective,

$$\mathcal{L}_n(\beta) = \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta. \tag{3.10}$$

Our past work shows that for specific choices of $(\widehat{\Gamma}, \widehat{\gamma})$, any global minimizer $\widehat{\beta}$ of the appropriately regularized problem (2.1) is a consistent estimate for β^* [18]. In the additive corruption model described in the previous paragraph, a natural choice is

$$(\widehat{\Gamma}, \widehat{\gamma}) := \left(\frac{Z^T Z}{n} - \Sigma_w, \frac{Z^T y}{n}\right), \tag{3.11}$$

where the covariance matrix $\Sigma_w = \operatorname{Cov}(w_i)$ is assumed to be known. However, in the high-dimensional setting ($n \ll p$), the matrix $\widehat{\Gamma}$ is *not* positive semidefinite, so the quadratic objective function (3.10) is nonconvex. [This is also a concrete instance where the objective function (2.1) requires the constraint $\|\beta\|_1 \le R$ in order to be bounded below.] Nonetheless, our past work [17, 18] shows that under certain tail conditions on the covariates and noise vectors, the loss function (3.10) does satisfy restricted strong convexity.

To simplify our discussion, we only state an explicit corollary for the case when ρ_λ is the convex ℓ_1-penalty; the most general case, involving a nonconvex quadratic form and a nonconvex regularizer, is simply a hybrid of the analysis below and the arguments of the previous section. Our goal is to illustrate the applicability of the primal-dual witness technique for nonconvex loss functions. This setup leads to the following estimator for β^*:

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \le R} \left\{ \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta + \lambda \|\beta\|_1 \right\}. \tag{3.12}$$

In the following corollary, we assume that $n \ge k^2$ and $n \ge c_0 k \log p$, for a sufficiently large constant c_0. As in the statement of Corollary 2, we do not include an explicit choice of c_ℓ, c_u, and C in the statement of the corollary to avoid clutter, but it is easy to check that $c_\ell \asymp (\sigma_x + \sigma_w)(\sigma_\varepsilon + \sigma_w \|\beta^*\|_2)$ and $c_u \asymp \lambda_{\min}(\Sigma_x)$ suffices. The expression for C is more complicated, but an inspection of the proof shows that it should scale with $|||(\Sigma_x)_{SS}^{-1}|||_2$, in addition to the sub-Gaussian parameters of X, W and ε.

COROLLARY 2. *Suppose* (X, w, ε) *are sub-Gaussian*, $\lambda_{\min}(\Sigma_x) > 0$, *and* (λ, R) *are chosen such that* $\|\beta^*\|_1 \le \frac{R}{2}$ *and* $c_\ell \sqrt{\frac{\log p}{n}} \le \lambda \le \frac{c_u}{R}$. *If in addition*,

$$|||\widehat{\Gamma}_{SS}^{-1}|||_\infty \le c_\infty, \quad \text{and} \quad |||\widehat{\Gamma}_{S^cS}\widehat{\Gamma}_{SS}^{-1}|||_\infty \le \eta < 1, \tag{3.13}$$

then with probability at least $1 - c_1 \exp(-c_2 \min\{k, \log p\})$, *the objective* (3.12) *has a unique stationary point* $\widehat{\beta}$ (*corresponding to the global optimum*) *such that* $\operatorname{supp}(\widehat{\beta}) \subseteq \operatorname{supp}(\beta^*)$, *and*

$$\|\widehat{\beta} - \beta^*\|_\infty \le C\sqrt{\frac{\log p}{n}} + c_\infty \lambda. \tag{3.14}$$

We prove Corollary 2 in Appendix D.2. Note that similar results regarding the uniqueness of stationary points and ℓ_∞-error bounds hold without assuming the incoherence condition $|||\widehat{\Gamma}_{S^cS}\widehat{\Gamma}_{SS}^{-1}|||_\infty \le \eta < 1$ if a (μ, γ)-amenable regularizer is used instead [cf. Corollary 1(b)]. Also note that if in addition, we have the bound $\beta^*_{\min} \ge 2(C\sqrt{\frac{\log p}{n}} + c_\infty \lambda)$, we are guaranteed that $\widehat{\beta}$ is sign-consistent for β^*.

Corollary 2 shows that the primal-dual witness technique may be used even in a setting where the loss function is nonconvex. Under the same incoherence assumption (3.13) and the sample size scaling $n \ge c_0 k \log p$, stationary points of the modified (nonconvex) Lasso program (3.12) are also consistent for support recovery. Surprisingly, although the objective (3.12) is indeed nonconvex whenever $n < p$ and $\Sigma_w \succ 0$, it nonetheless has a *unique* stationary point that is in fact equal to the global optimum. This further clarifies the simulation results appearing in Loh and Wainwright [18]: those simulations are performed with $\Gamma = I_p$, so the incoherence condition (3.13) holds, with high probability, with η close to 0.

In order to produce the plots shown in Figure 1 above, we generated i.i.d. covariates $x_i \sim N(0, \Sigma_x)$, where $\Sigma_x = M_1(\theta)$ was obtained from the family of non-incoherent matrices (F.10), with $\theta = \frac{2.5}{k}$. We chose $k \approx \sqrt{p}$ and $\beta^* = (\frac{1}{\sqrt{k}}, \dots, \frac{1}{\sqrt{k}}, 0, \dots, 0)$, and generated response variables according to the linear model $y_i = x_i^T \beta^* + \varepsilon_i$, where $\varepsilon_i \sim N(0, (0.1)^2)$. In addition, we generated corrupted covariates $z_i = x_i + w_i$, where $w_i \sim N(0, (0.2)^2)$, and $w_i \perp\!\!\!\perp x_i$. We ran the composite gradient descent algorithm (cf. Appendix F.1) on the objective function given by equations (3.10) and (3.11), and with regularization parameters $R = 1.1\|\beta^*\|_1$ and $\lambda = \sqrt{\frac{\log p}{n}}$. In panel (a), we see that the probability of correct support recovery transitions sharply from 0 to 1 as the sample size increases and ρ_λ is the SCAD or MCP regularizer. In contrast, the probability of recovering the correct support remains at 0 when ρ_λ is the ℓ_1-penalty or LSP—by the structure of Σ_x, regularization with the ℓ_1-penalty or LSP results in an estimator $\widehat{\beta}$ that puts nonzero weight on the $(k+1)$st coordinate, as well. Note that we have rescaled the horizontal axis according to $\frac{n}{k \log p}$ in order to match the scaling prescribed by our theory; the three sets of curves for each regularizer roughly align, as predicted by Theorem 1. Panel (b) confirms that the ℓ_∞-error $\|\widehat{\beta} - \beta^*\|_\infty$ decreases to 0 when using the SCAD and MCP, as predicted by Theorem 2. Further note that the ℓ_1-penalty, LSP, SCAD, and MCP regularizers are all ℓ_2-consistent; however, since a lower-eigenvalue bound on the covariance matrix of the design is sufficient in that case [17].

3.4. *Generalized linear models.* We now move to the case where the loss function is the negative log likelihood of a generalized linear model, and show that the incoherence condition may again be removed if the regularizer ρ_λ is (μ, γ)-amenable.

Suppose the pairs $\{(x_i, y_i)\}_{i=1}^n$ are drawn from a generalized linear model (GLM). Recall that the conditional distribution for a GLM takes the form

$$\mathbb{P}(y_i | x_i, \beta, \sigma) = \exp\left(\frac{y_i x_i^T \beta - \psi(x_i^T \beta)}{c(\sigma)}\right), \tag{3.15}$$

where $\sigma > 0$ is a scale parameter and ψ is the cumulant function. The loss function corresponding to the negative log likelihood is given by

$$\mathcal{L}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \bigl(\psi(x_i^T \beta) - y_i x_i^T \beta\bigr), \tag{3.16}$$

and it is easy to see that equation (3.16) reduces to equation (3.10) when $\psi(t) = \frac{t^2}{2}$. Using properties of exponential families, we may also verify that equation (2.3) holds. Negahban et al. [22] showed that restricted strong convexity holds for a broad class of generalized linear models.

Taking $\Omega = \mathbb{R}^p$, we then construct the composite objective:

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \le R} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\psi(x_i^T \beta) - y_i x_i^T \beta) + \rho_\lambda(\beta) \right\}. \tag{3.17}$$

We further impose the following technical conditions.

ASSUMPTION 1.

(i) The covariates are uniformly bounded: $\|x_i\|_\infty \le M$, for all $1 \le i \le n$.
(ii) There are positive constants κ_2 and κ_3, such that $\|\psi''\|_\infty \le \kappa_2$ and $\|\psi'''\|_\infty \le \kappa_3$.

The conditions of Assumption 1, although somewhat stringent, are satisfied in various settings of interest. In the case of logistic regression, we have $\psi(t) = \log(1 + \exp(t))$, and we may easily verify that the boundedness conditions in Assumption 1(ii) are satisfied with $\kappa_2 = 0.25$ and $\kappa_3 = 0.1$. Also note that the uniform bound on ψ''' is used implicitly in the proof for support recovery consistency in the logistic regression analysis of Ravikumar et al. [26], whereas the uniform bound on ψ'' also appears in the conditions for ℓ_1- and ℓ_2-consistency in other past work [17, 22]. The uniform boundedness condition in Assumption 1(i) is somewhat less desirable: although it always holds for categorical data, it does not hold for Gaussian covariates. We suspect that is possible to relax this constraint.

In what follows, let $Q^* := \mathbb{E}[\frac{1}{n}\sum_{i=1}^n \psi''(x_i^T \beta^*) x_i x_i^T]$ denote the Fisher information matrix.

COROLLARY 3. *Suppose Assumption 1 holds, and suppose ρ_λ is (μ, γ)-amenable with $\mu < c_\psi \lambda_{\min}(\Sigma_x)$, where c_ψ is a constant depending only on ψ. Also suppose $n \ge c_0 k^3 \log p$, and (λ, R) are chosen such that $\|\beta^*\|_1 \le \frac{R}{2}$ and $c_\ell \sqrt{\frac{\log p}{n}} \le \lambda \le \frac{c_u}{R}$. Further suppose $|||(Q_{SS}^*)^{-1}|||_\infty \le c_\infty$, and $\beta_{\min}^* \ge \lambda(\gamma + 2c_\infty) + C\sqrt{\frac{\log p}{n}}$. Then with probability at least $1 - c_1 \exp(-c_2 \min\{k, \log p\})$, the program (3.17) has a unique stationary point $\widehat{\beta}$ given by the oracle estimator $\widehat{\beta}^{\mathcal{O}}$, and*

$$\|\widehat{\beta} - \beta^*\|_\infty \le C\sqrt{\frac{\log p}{n}}. \tag{3.18}$$

Corollary 3 may be compared with the result for ℓ_1-regularized logistic regression given in Ravikumar et al. [26] (see Theorem 1 in their paper). Both results require that the sample size is lower-bounded as $n \geq c_0 k^3 \log p$, but Ravikumar et al. [26] also require Q^* to satisfy the incoherence condition:

$$\||Q^*_{S^cS}(Q^*_{SS})^{-1}\||_\infty \leq \eta < 1. \tag{3.19}$$

As noted in their paper and by other authors, the incoherence condition (3.19) is difficult to interpret and verify for general GLMs. In contrast, Corollary 3 shows that by using a properly chosen nonconvex regularizer, this incoherence requirement may be removed entirely. Furthermore, Corollary 3 applies to more than just the logistic case with an ℓ_1-penalty and extends to various nonconvex problems where the uniqueness of stationary points is not evident a priori. The proof of Corollary 3 is contained in Appendix D.3. Due to increasing technicality, we do not state explicit choices of c_ℓ, c_u and C, but we refer the reader to the supplement for more details.

4. Proofs of main theorems. We now outline the proofs of the theorems stated in Section 3. In the proof of Theorem 1, we show that the PDW construction in Section 2.4 may be applied to establish support recovery even for nonconvex objectives. In the proof of Theorem 2, we highlight how the support recovery results achieved in Theorem 1 may be used to derive novel ℓ_∞-error bounds on the unique stationary point $\widehat{\beta}$. Furthermore, we show how (μ, γ)-amenability implies an even tighter bound for the oracle estimator.

4.1. *Proof of Theorem* 1. We follow the outline of the primal-dual witness construction described in Section 2.4. For step (i) of the construction, we use Lemma 9 in Appendix G, where we simply replace p by k and $\mathcal{L}_n$ by $(\mathcal{L}_n)|_S$, which is the function $\mathcal{L}_n$ restricted to $\mathbb{R}^S$. It follows that as long as $n \geq \frac{16R^2 \max(\tau_1^2, \tau_2^2)}{\alpha_2^2} \log k$, we are guaranteed that $\|\widehat{\beta}_S - \beta^*_S\|_1 \leq \frac{24\lambda k}{4\alpha_1 - 3\mu}$, whence $\|\widehat{\beta}_S\|_1 \leq \|\beta^*\|_1 + \|\widehat{\beta}_S - \beta^*_S\|_1 \leq \frac{R}{2} + \frac{24\lambda k}{4\alpha_1 - 3\mu} < R$. Here, the final inequality follows by the lower bound in inequality (3.1b). We conclude that $\widehat{\beta}_S$ must be in the interior of the feasible region.

Moving to step (ii) of the PDW construction, we define the shifted objective function $\overline{\mathcal{L}}_n(\beta) := \mathcal{L}_n(\beta) - q_\lambda(\beta)$. Since $\widehat{\beta}_S$ is an interior point, it must be a zero-subgradient point for the restricted program (2.6), so $\nabla(\overline{\mathcal{L}}_n)|_S(\widehat{\beta}_S) + \lambda \widehat{z}_S = 0$, where $\widehat{z}_S \in \partial\|\widehat{\beta}_S\|_1$ is the dual vector. By the chain rule, this implies that $(\nabla\overline{\mathcal{L}}_n(\widehat{\beta}))_S + \lambda\widehat{z}_S = 0$, where $\widehat{\beta} := (\widehat{\beta}_S, 0_{S^c})$. Accordingly, we may define the subvector $\widehat{z}_{S^c} \in \mathbb{R}^{S^c}$ such that

$$\nabla\overline{\mathcal{L}}_n(\widehat{\beta}) + \lambda\widehat{z} = 0, \tag{4.1}$$

where $\widehat{z} := (\widehat{z}_S, \widehat{z}_{S^c})$ is the extended subgradient. Under the assumption (3.2), this completes step (ii) of the construction.

For step (iii), we first establish that $\widehat{\beta}$ is a local minimum for the program (2.1) by verifying the sufficient conditions of Lemma 10 in Appendix G, with functions

$f = \overline{\mathcal{L}}_n$ and $g = q_\lambda$, and $(x^*, v^*, w^*, \mu^*) = (\widehat{\beta}, \widehat{z}, \widehat{z}, 0)$. Note that Lemma 5(b) from Appendix A.2 ensures the concavity and differentiability of $g(x) - \frac{\mu}{2}\|x\|_2^2$. Since $\mu^* = 0$, condition (G.2a) is trivially satisfied. Furthermore, condition (G.2b) holds by equation (4.1). Hence, it remains to verify the condition (G.2c).

We first show that $G^* \subseteq \mathbb{R}^S$. Supposing the contrary, consider a vector $\nu \in G^*$ such that $\operatorname{supp}(\nu) \subsetneq S$. Fixing some index $j \in S^c$ such that $\nu_j \neq 0$, and using the definition of G^*, we have

$$\sup_{v \in \partial \|\widehat{\beta}\|_1} \nu^T \big(\nabla \overline{\mathcal{L}}_n(\widehat{\beta}) + \lambda v\big) = 0. \tag{4.2}$$

However, if $\widetilde{z}$ denotes the vector $\widehat{z}$ with entry j replaced by $\operatorname{sign}(s_j) \in \{-1, 1\}$, we clearly still have $\widetilde{z} \in \partial \|\widehat{\beta}\|_1$. On the other hand,

$$\nu^T \big(\nabla \overline{\mathcal{L}}_n(\widehat{\beta}) + \lambda \widetilde{z}\big) > \nu^T \big(\nabla \overline{\mathcal{L}}_n(\widehat{\beta}) + \lambda \widehat{z}\big) = 0,$$

where the strict inequality holds because $\|\widehat{z}_{S^c}\|_\infty < 1$, by our assumption. We have thus obtained a contradiction to equation (4.2); consequently, our initial assumption was false, and we may conclude that $G^* \subseteq \mathbb{R}^S$.

The following lemma, proved Appendix B.1, guarantees that a shifted form of the loss function is strictly convex over a k-dimensional subspace.

LEMMA 1. *Consider any twice-differentiable, (α, τ)-RSC loss function $\mathcal{L}_n$ and any μ-amenable regularizer ρ_λ, with $\mu < \alpha_1$. Suppose $|S| = k$. If $n \geq \frac{2\tau_1}{\alpha_1 - \mu} k \log p$, the function $\mathcal{L}_n(\beta) - \frac{\mu}{2}\|\beta\|_2^2$ is strictly convex on $\beta \in \mathbb{R}^S$, and the restricted program (2.6) is also strictly convex.*

In particular, since $G^* \subseteq S$ and $\operatorname{supp}(\widehat{\beta}) \subseteq S$, Lemma 1 implies condition (G.2c) of Lemma 10, so $\widehat{\beta}$ is a local minimum of the program (2.1).

The following lemma, proved in Section 4.2, shows that all stationary points of the program (2.1) are supported on S. It involves a fairly technical argument, using the strict dual feasibility of $\widehat{z}_{S^c}$ to deduce that *any* stationary point $\widetilde{\beta}$ also has the correct support.

LEMMA 2. *Suppose $\widetilde{\beta}$ is a stationary point of the program (2.1) and the conditions of Theorem 1 hold. Then* $\operatorname{supp}(\widetilde{\beta}) \subseteq S$.

The uniqueness assertion now follows fairly easily. Note that since all stationary points are supported in S by Lemma 2, any stationary point $\widetilde{\beta}$ of the program (2.1) must satisfy $\widetilde{\beta} = (\widetilde{\beta}_S, 0_{S^c})$, where $\widetilde{\beta}_S$ is a stationary point of the restricted program (2.6). By Lemma 1, the restricted program is strictly convex. Hence, the vector $\widetilde{\beta}_S$, and consequently also $\widetilde{\beta}$, is unique.

4.2. *Proof of Lemma* 2. Let $\widetilde{\nu} := \widetilde{\beta} - \beta^*$. We first show that $\|\widetilde{\nu}\|_2 \le 1$. Suppose on the contrary that $\|\widetilde{\nu}\|_2 > 1$. By inequality (2.5b), we have $\langle \nabla\mathcal{L}_n(\widetilde{\beta}) - \nabla\mathcal{L}_n(\widehat{\beta}), \widetilde{\nu}\rangle \ge \alpha_2\|\widetilde{\nu}\|_2 - \tau_2\sqrt{\frac{\log p}{n}}\|\widetilde{\nu}\|_1$. Moreover, since $\widehat{\beta}$ is feasible, the first-order optimality condition gives

$$0 \le \langle \nabla\mathcal{L}_n(\widetilde{\beta}) + \nabla\rho_\lambda(\widetilde{\beta}), \widehat{\beta} - \widetilde{\beta}\rangle. \tag{4.3}$$

Summing the two preceding inequalities yields

$$\alpha_2\|\widetilde{\nu}\|_2 - \tau_2\sqrt{\frac{\log p}{n}}\|\widetilde{\nu}\|_1 \le \langle -\nabla\mathcal{L}_n(\widehat{\beta}) - \nabla\rho_\lambda(\widetilde{\beta}), \widetilde{\nu}\rangle. \tag{4.4}$$

Since $\widehat{\beta}$ is an interior local minimum, we have $\nabla\mathcal{L}_n(\widehat{\beta}) + \nabla\rho_\lambda(\widehat{\beta}) = 0$. Inequality (4.4) implies

$$\begin{aligned}\alpha_2\|\widetilde{\nu}\|_2 - \tau_2\sqrt{\frac{\log p}{n}}\|\widetilde{\nu}\|_1 &\le \langle \nabla\rho_\lambda(\widehat{\beta}) - \nabla\rho_\lambda(\widetilde{\beta}), \widetilde{\nu}\rangle \\ &\le (\|\nabla\rho_\lambda(\widehat{\beta})\|_\infty + \|\nabla\rho_\lambda(\widetilde{\beta})\|_\infty)\|\widetilde{\nu}\|_1 \\ &\le 2\lambda\|\widetilde{\nu}\|_1,\end{aligned}$$

where the bound $\|\nabla\rho_\lambda(\beta)\|_\infty \le \lambda$ holds by Lemma 5 in Appendix A.2. Rearranging, we have $\|\widetilde{\nu}\|_2 \le \frac{\|\widetilde{\nu}\|_1}{\alpha_2}(2\lambda + \tau_2\sqrt{\frac{\log p}{n}}) \le \frac{2R}{\alpha_2}(2\lambda + \tau_2\sqrt{\frac{\log p}{n}})$. Since $\lambda \le \frac{\alpha_2}{8R}$ and $n \ge \frac{16R^2\tau_2^2}{\alpha_2^2}\log p$, this implies that $\|\widetilde{\nu}\|_2 \le 1$, as claimed.

Now, applying the RSC condition (2.5b), we have

$$\langle \nabla\overline{\mathcal{L}}_n(\widetilde{\beta}) - \nabla\overline{\mathcal{L}}_n(\widehat{\beta}), \widetilde{\nu}\rangle \ge (\alpha_1 - \mu)\|\widetilde{\nu}\|_2^2 - \tau_1\frac{\log p}{n}\|\widetilde{\nu}\|_1^2. \tag{4.5}$$

By inequality (4.3), we also have

$$0 \le \langle \nabla\overline{\mathcal{L}}_n(\widetilde{\beta}), \widehat{\beta} - \widetilde{\beta}\rangle + \lambda \cdot \langle \widetilde{z}, \widehat{\beta} - \widetilde{\beta}\rangle, \tag{4.6}$$

where $\widetilde{z} \in \partial\|\widetilde{\beta}\|_1$. From the zero-subgradient condition (2.7), we further have $\langle \nabla\overline{\mathcal{L}}_n(\widehat{\beta}) + \lambda\widehat{z}, \widetilde{\beta} - \widehat{\beta}\rangle = 0$. Combining with inequality (4.6) then yields

$$0 \le \langle \nabla\overline{\mathcal{L}}_n(\widehat{\beta}) - \nabla\overline{\mathcal{L}}_n(\widetilde{\beta}), \widetilde{\beta} - \widehat{\beta}\rangle + \lambda \cdot \langle \widehat{z}, \widetilde{\beta}\rangle - \lambda\|\widehat{\beta}\|_1 + \lambda \cdot \langle \widetilde{z}, \widehat{\beta}\rangle - \lambda\|\widetilde{\beta}\|_1. \tag{4.7}$$

Rearranging, we have

$$\begin{aligned}\lambda\|\widetilde{\beta}\|_1 - \lambda \cdot \langle \widehat{z}, \widetilde{\beta}\rangle &\le \langle \nabla\overline{\mathcal{L}}_n(\widehat{\beta}) - \nabla\overline{\mathcal{L}}_n(\widetilde{\beta}), \widetilde{\beta} - \widehat{\beta}\rangle + \lambda \cdot \langle \widetilde{z}, \widehat{\beta}\rangle - \lambda\|\widehat{\beta}\|_1 \\ &\le \langle \nabla\overline{\mathcal{L}}_n(\widehat{\beta}) - \nabla\overline{\mathcal{L}}_n(\widetilde{\beta}), \widetilde{\beta} - \widehat{\beta}\rangle \\ &\le \tau_1\frac{\log p}{n}\|\widetilde{\nu}\|_1^2 - (\alpha_1 - \mu)\|\widetilde{\nu}\|_2^2,\end{aligned} \tag{4.8}$$

where the second inequality follows because $\langle \widetilde{z}, \widehat{\beta}\rangle \le \|\widetilde{z}\|_\infty \cdot \|\widehat{\beta}\|_1 \le \|\widehat{\beta}\|_1$, and the third inequality comes from the bound (4.5). Finally, we show that $\widetilde{\nu}$ lies in a cone set.

LEMMA 3. *If* $\|\widehat{z}_{S^c}\|_\infty \le 1-\delta$ *for some* $\delta \in (0,1]$ *and* $\lambda \ge \frac{4R\tau_1 \log p}{\delta n}$, *then*

$$\|\widetilde{\nu}\|_1 \le \left(\frac{4}{\delta}+2\right)\sqrt{k}\|\widetilde{\nu}\|_2.$$

The proof of Lemma 3 is provided in Appendix B.2. Combining Lemma 3 with inequality (4.8) then gives

$$\lambda\|\widetilde{\beta}\|_1 - \lambda\cdot\langle\widehat{z},\widetilde{\beta}\rangle \le \tau_1 \frac{k\log p}{n}\left(\frac{4}{\delta}+2\right)^2\|\widetilde{\nu}\|_2^2 - (\alpha_1-\mu)\|\widetilde{\nu}\|_2^2.$$

Hence, if $n \ge \frac{2\tau_1}{\alpha_1-\mu}(\frac{4}{\delta}+2)^2 k\log p$, then $\lambda\|\widetilde{\beta}\|_1 - \lambda\langle\widehat{z},\widetilde{\beta}\rangle \le -\frac{\alpha_1-\mu}{2}\|\widetilde{\nu}\|_2^2 \le 0$. On the other hand, Hölder's inequality gives $\lambda\langle\widehat{z},\widetilde{\beta}\rangle \le \lambda\|\widehat{z}\|_\infty\|\widetilde{\beta}\|_1 \le \lambda\|\widetilde{\beta}\|_1$. It follows that we must have $\langle\widehat{z},\widetilde{\beta}\rangle = \|\widetilde{\beta}\|_1$. Since $\|\widehat{z}_{S^c}\|_\infty < 1$, we conclude that $\widetilde{\beta}_j = 0$, for all $j \notin S$. Hence, $\operatorname{supp}(\widetilde{\beta}) \subseteq S$, as claimed.

4.3. *Proof of Theorem* 2. Note that by the fundamental theorem of calculus for vector-valued functions, we have $\widehat{Q}(\widehat{\beta}-\beta^*) = \nabla\mathcal{L}_n(\widehat{\beta}) - \nabla\mathcal{L}_n(\beta^*)$. Furthermore, we have $\widehat{\beta}_{S^c} = \beta^*_{S^c} = 0$, by construction. Using the zero-subgradient condition (2.7) and the assumed invertibility of $\widehat{Q}_{SS}$, we have

$$\widehat{\beta}_S - \beta^*_S = (\widehat{Q}_{SS})^{-1}(-\nabla\mathcal{L}_n(\beta^*)_S + \nabla q_\lambda(\widehat{\beta}_S) - \lambda\widehat{z}_S),$$

so combined with the support recovery result of Theorem 1, we have

$$\|\widehat{\beta}-\beta^*\|_\infty \le \|(\widehat{Q}_{SS})^{-1}(\nabla\mathcal{L}_n(\beta^*)_S - \nabla q_\lambda(\widehat{\beta}_S) + \lambda\widehat{z}_S)\|_\infty. \tag{4.9}$$

Lemma 5 in Appendix A.2 guarantees that $\|(\nabla q_\lambda(\widehat{\beta}_S) - \lambda\widehat{z}_S)\|_\infty \le \lambda$, so

$$\begin{aligned}\|\widehat{\beta}-\beta^*\|_\infty &\le \|(\widehat{Q}_{SS})^{-1}\nabla\mathcal{L}_n(\beta^*)_S\|_\infty + \|(\widehat{Q}_{SS})^{-1}(\nabla q_\lambda(\widehat{\beta}_S) - \lambda\widehat{z}_S)\|_\infty \\ &\le \|(\widehat{Q}_{SS})^{-1}\nabla\mathcal{L}_n(\beta^*)_S\|_\infty + \lambda|\!|\!|(\widehat{Q}_{SS})^{-1}|\!|\!|_\infty,\end{aligned}$$

which is inequality (3.3).

For inequality (3.4b), we use the following lemma, proved in Appendix B.3:

LEMMA 4. *Suppose* ρ_λ *is* (μ,γ)-*amenable, and the bound* (3.4a) *holds. Then* $|\widehat{\beta}_j| \ge \gamma\lambda$ *for all* $j \in S$, *and in particular,* $q'_\lambda(\widehat{\beta}_j) = \lambda\cdot\operatorname{sign}(\widehat{\beta}_j)$.

Lemma 4 implies that $\nabla q_\lambda(\widehat{\beta}_S) = \lambda\widehat{z}_S$. Hence, the zero-subgradient condition (2.7) reduces to $(\nabla\mathcal{L}_n(\widehat{\beta}_S))|_S = 0$. Since $\mathcal{L}_n$ is strictly convex on $\mathbb{R}^S$ by Lemma 1, this zero-gradient condition implies that $\widehat{\beta}_S$ is the unique global minimum of $(\mathcal{L}_n)|_S$, so $\widehat{\beta}_S = \widehat{\beta}^{\mathcal{O}}_S$ and $\widehat{\beta} = \widehat{\beta}^{\mathcal{O}}$, as claimed. Finally, inequality (4.9) simplifies to inequality (3.4b), since $\nabla q_\lambda(\widehat{\beta}_S) = \lambda\widehat{z}_S$.

5. Simulations. In this section, we report the results of additional illustrative simulations. We ran experiments with the loss function coming from (a) ordinary least squares regression and (b) logistic regression, together with the ℓ_1-penalty, LSP, SCAD ($a = 2.5$) and MCP ($b = 1.5$) regularizers. In each setting, we tuned the regularization parameters (R, λ) via 10-fold cross-validation over a grid of values with $R = c_1\|\beta^*\|_1$ and $\lambda = c_2\sqrt{\frac{\log p}{n}}$, using the squared error loss for linear regression and deviance measure for logistic regression. To locate stationary points, we ran a version of the composite gradient descent algorithm (for more details on the optimization algorithm, see Appendix F.1).

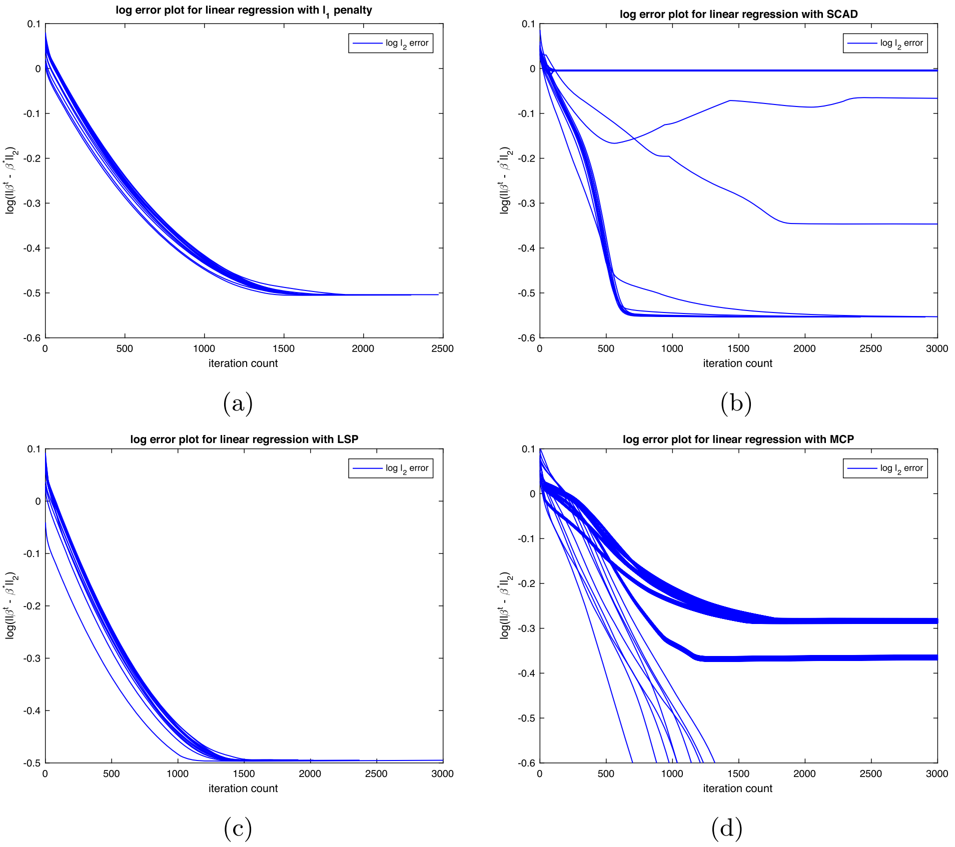
In our first set of simulations, we explore the uniqueness of stationary points in the linear regression setting. We generated i.i.d. observations $x_i \sim N(0, \Sigma_x)$ with $\Sigma_x = M_2(\theta)$ coming from the family of spiked identity models (F.11), for $\theta = 0.9$, and independent additive noise, $\varepsilon_i \sim N(0, (0.1)^2)$. We set the problem dimensions to be $p = 256$, $k = 4$ and $n \approx 25k \log p$, and generated β^* to have k nonzero values $\pm\frac{1}{\sqrt{k}}$ with equal probability for each sign. In particular, the design matrix satisfies the incoherence property with high probability, so the theory of Section 3.2 guarantees that stationary points are unique for all regularizers whenever $3\mu < 4\alpha_1$. When $4\alpha_1 \leq 3\mu$, convergence of the composite gradient descent algorithm and consistent support recovery are no longer guaranteed; in practice, we observe that multiple initializations of the composite gradient descent algorithm still appear to converge to a single stationary point with the correct support when 3μ is slightly larger than $4\alpha_1$. However, when the condition is violated more severely, the composite gradient descent algorithm indeed terminates at several distinct stationary points.

Figure 2 displays the results of our simulations. When using the SCAD or MCP regularizers [panels (b) and (d)], distinct stationary points emerge, the recovered support is incorrect, and the optimization algorithm sometimes has difficulty converging, since $4\alpha_1 < 3\mu$. In contrast, the ℓ_1-penalty and LSP still continue to produce unique stationary points with the correct support [panels (a) and (c)]. [For smaller values of θ, the solution trajectories for the SCAD and MCP regularizers exhibit the same nice behavior as the trajectories in panels (a) and (c).] Also observe that the error $\|\beta^t - \beta^*\|_2$ decreases at a rate that is linear on a log scale, as predicted by Theorem 3 of Loh and Wainwright [17], until it reaches the threshold of statistical accuracy.

In our second set of simulations, we used the maximum likelihood loss function for logistic regression:

$$\mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\{\log(1 + \exp(x_i^T\beta)) - y_i x_i^T\beta\}.$$

We generated $x_i \sim N(0, \sigma_x^2 I)$, with $\sigma_x \in \{0.8, 2\}$, and set the problem dimensions to be $p = 256$, $k = 3$, and $n \approx 3k^3 \log p$. We generated β^* to have k nonzero



(a)

(b)

(c)

(d)

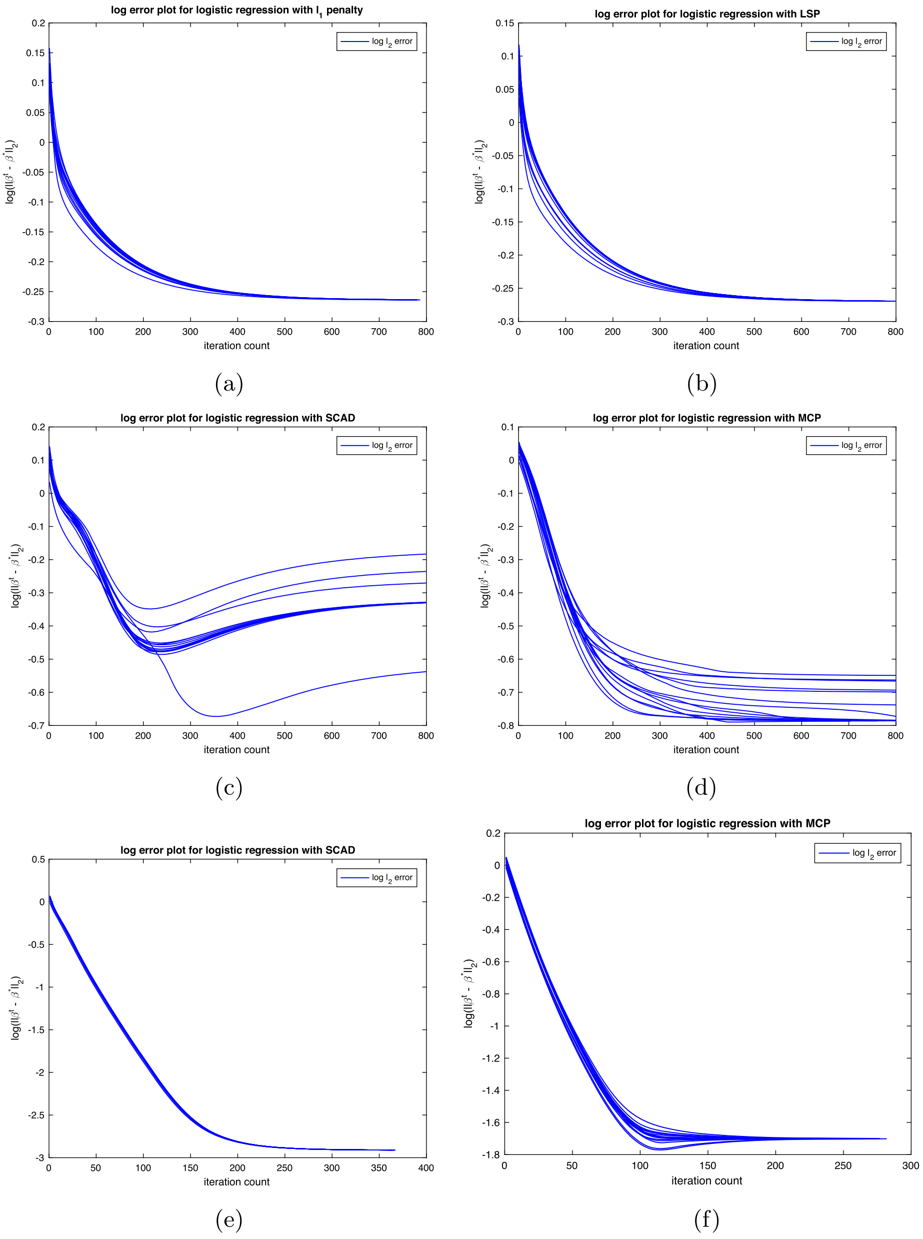
FIG. 2. *Plots showing log ℓ_2-error* $\log(\|\beta^t - \beta^*\|_2)$ *as a function of iteration number t for OLS linear regression with a variety of regularizers and* 15 *random initializations of composite gradient descent. As seen in panels* (b) *and* (d), *the SCAD and MCP regularizers give rise to multiple distinct stationary points.*

values $\pm\frac{1}{\sqrt{k}}$, with equal probability for each sign, and generated response variables $y_i \in \{0, 1\}$ according to

$$\mathbb{P}(y_i = 1 | x_i, \beta^*) = \frac{\exp(x_i^T \beta^*)}{1 + \exp(x_i^T \beta^*)}.$$

[Note that although the covariates in our simulations for logistic regression do not satisfy the boundedness Assumption 1(i) imposed in our corollary, the generated plots still agree qualitatively with our predicted theoretical results.]

Figure 3 displays the results of our simulations. Panels (a)–(d) plot the log ℓ_2-error as a function of iteration number, when $\sigma_x = 0.8$. Note that in this case, an empirical evaluation shows that $\lambda_{\min}(\nabla^2 \mathcal{L}(\beta^*)) \approx 0.10$, so we expect $\alpha_1 \approx 0.10$ and $3\mu \not< 4\alpha_1$. As in the plots of Figure 2, multiple stationary points emerge in



(a)

(b)

(c)

(d)

(e)

(f)

FIG. 3. *Plots showing log ℓ_2-error* $\log(\|\beta^t - \beta^*\|_2)$ *as a function of iteration number t for logistic regression, with a variety of regularizers and* 15 *random initializations of composite gradient descent. The covariates are normally distributed according to* $x_i \sim N(0, \sigma_x^2 I)$, *with* $\sigma_x = 0.8$ *in plots* (a)–(d), *and* $\sigma_x = 2$ *in plots* (e)–(f). *In panels* (c) *and* (d), *the composite gradient descent algorithm settles into multiple distinct stationary points, which exist because* $4\alpha_1 < 3\mu$ *for the SCAD and MCP. However, when the covariates have a larger covariance, the SCAD and MCP regularizers produce unique stationary points, as observed in panels* (e) *and* (f).

panels (c) and (d) when ρ_λ is the SCAD or MCP regularizer; in contrast, we see from panels (a) and (b) that all 15 runs of composite gradient descent converge to the same stationary point when ρ_λ is the ℓ_1-penalty or LSP. In panels (e) and (f), we repeat the simulations with $\sigma_x = 2$. In this case, $\lambda_{\min}(\nabla^2 \mathcal{L}(\beta^*)) \approx 0.24$, and we see from our plots that although the condition $3\mu < 4\alpha_1$ is still violated, the larger value of α_1 is enough to make the stationary points under SCAD or MCP regularization unique. We may again observe the geometric rate of convergence of the ℓ_2-error $\|\beta^t - \beta^*\|_2$ in each plot, up to a certain small threshold. The improved performance from using the SCAD and MCP regularizers may also be observed empirically by comparing the vertical axes in the panels of Figure 3.

6. Discussion.

We have developed a novel framework for analyzing a variety of nonconvex problems via the primal-dual witness proof technique. Our results apply to composite optimization programs where both the loss and regularizer function are allowed to be nonconvex, and our analysis significantly generalizes the machinery previously proposed to establish variable selection consistency for convex functions. As a consequence, we have provided a powerful reason for using nonconvex regularizers such as the SCAD and MCP rather than the convex ℓ_1-penalty: In addition to being consistent in ℓ_2-error, the nonconvex regularizers actually produce an overall estimator that is consistent for support recovery when the design matrix is non-incoherent and the usual ℓ_1-regularized program fails in recovering the correct support. We have also established guarantees concerning the uniqueness of stationary points of certain nonconvex regularized problems that subsume several recent results in the high-dimensional regression literature.

Future research directions include devising theoretical guarantees when the condition $3\mu < 4\alpha_1$ is only mildly violated, since the condition does not appear to be strictly necessary based on our simulations; and justifying why the SCAD and MCP regularizers perform appreciably better than the ℓ_1-penalty even in terms of ℓ_2-error, in situations where the assumptions are not strong enough for an oracle result to apply. It would be useful to be able to compute the RSC constants (α_1, α_2) empirically from data, so as to assign a nonconvex regularizer with the proper amount of curvature.

Acknowledgments. The authors thank the Associate Editor and anonymous reviewers for helpful comments that substantially improved the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Support recovery without incoherence: A case for nonconvex regularization" (DOI: 10.1214/16-AOS1530SUPP; .pdf). We provide additional technical details for the results provided in the main body of the paper.

REFERENCES

[1] BERTSEKAS, D. P. (1999). *Nonlinear Programming*. Athena Scientific, Belmont, MA.

[2] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[3] BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253. MR2810396

[4] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

[5] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[6] CANDÈS, E. J., WAKIN, M. B. and BOYD, S. P. (2008). Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.* **14** 877–905. MR2461611

[7] CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

[8] CLARKE, F. H. (1975). Generalized gradients and applications. *Trans. Amer. Math. Soc.* **205** 247–262. MR0367131

[9] DONOHO, D. L. and STARK, P. B. (1989). Uncertainty principles and signal recovery. *SIAM J. Appl. Math.* **49** 906–931.

[10] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

[11] FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. MR2849368

[12] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961.

[13] FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988

[14] FLETCHER, R. and WATSON, G. A. (1980). First- and second-order conditions for a class of nondifferentiable optimization problems. *Math. Program.* **18** 291–307. MR0571992

[15] HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642.

[16] LEE, J. D., SUN, Y. and TAYLOR, J. E. (2015). On model selection consistency of regularized M-estimators. *Electron. J. Stat.* **9** 608–642. MR3331852

[17] LOH, P. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616.

[18] LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038

[19] LOH, P.-L. and WAINWRIGHT, M. J. (2017). Supplement to "Support recovery without incoherence: A case for nonconvex regularization." DOI:10.1214/16-AOS1530SUPP.

[20] LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087

[21] MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). *SparseNet*: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. MR2894769

[22] NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

[23] NESTEROV, Y. and NEMIROVSKII, A. (1987). *Interior Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, PA.

[24] PAN, Z. and ZHANG, C. (2015). Relaxed sparse eigenvalue conditions for sparse estimation via non-convex regularized regression. *Pattern Recognit.* **48** 231–243.

[25] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259.

[26] RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

[27] RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory* **59** 3434–3447. MR3061256

[28] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[29] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

[30] VAVASIS, S. A. (1995). Complexity issues in global optimization: A survey. In *Handbook of Global Optimization. Nonconvex Optim. Appl.* **2** 27–41. Kluwer Academic, Dordrecht. MR1377083

[31] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

[32] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55** 5728–5741.

[33] WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201.

[34] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[35] ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135

[36] ZHANG, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

[37] ZHANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2017). Optimal prediction for sparse linear models? Lower bounds for coordinate-separable M-estimators. *Electron. J. Stat.* **11** 752–799. MR3622646

[38] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2567.

[39] ZHENG, Z., FAN, Y. and LV, J. (2014). High dimensional thresholded regression and shrinkage effect. *J. Roy. Statist. Soc. Ser. B* **76** 627–649.

[40] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

DEPARTMENT OF ECE
AND
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN—MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: loh@ece.wisc.edu

DEPARTMENT OF STATISTICS
AND
DEPARTMENT OF EECS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: wainwrig@stat.berkeley.edu