Journal of Machine Learning Research 16 (2015) 559-616 Submitted 4/14; Revised 10/14; Published 3/15

Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima

Po-Ling Loh LOH@WHARTON.UPENN.EDU
Department of Statistics
The Wharton School
466 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104, USA

Martin J. Wainwright WAINWRIG@BERKELEY.EDU
Departments of EECS and Statistics
263 Cory Hall
University of California
Berkeley, CA 94720, USA

Editor: Nicolai Meinshausen

Abstract

We provide novel theoretical results regarding local optima of regularized M-estimators, allowing for nonconvexity in both loss and penalty functions. Under restricted strong convexity on the loss and suitable regularity conditions on the penalty, we prove that *any stationary point* of the composite objective function will lie within statistical precision of the underlying parameter vector. Our theory covers many nonconvex objective functions of interest, including the corrected Lasso for errors-in-variables linear models; regression for generalized linear models with nonconvex penalties such as SCAD, MCP, and capped-ℓ_1; and high-dimensional graphical model estimation. We quantify statistical accuracy by providing bounds on the ℓ_1-, ℓ_2-, and prediction error between stationary points and the population-level optimum. We also propose a simple modification of composite gradient descent that may be used to obtain a near-global optimum within statistical precision ϵ_{stat} in $\log(1/\epsilon_{\text{stat}})$ steps, which is the fastest possible rate of any first-order method. We provide simulation studies illustrating the sharpness of our theoretical results.

Keywords: high-dimensional statistics, M-estimation, model selection, nonconvex optimization, nonconvex regularization

1. Introduction

Although recent years have brought about a flurry of work on optimization of convex functions, optimizing nonconvex functions is in general computationally intractable (Nesterov and Nemirovskii, 1987; Vavasis, 1995). Nonconvex functions may possess local optima that are not global optima, and iterative methods such as gradient or coordinate descent may terminate undesirably in local optima. Unfortunately, standard statistical results for nonconvex M-estimators often only provide guarantees for *global* optima. This leads to a significant gap between theory and practice, since computing global optima—or even near-global optima—in an efficient manner may be extremely difficult in practice. Nonetheless, empirical studies have shown that local optima of various nonconvex M-estimators arising in statistical problems appear to be well-behaved (e.g., Breheny and Huang, 2011). This type of observation is the starting point of our work.

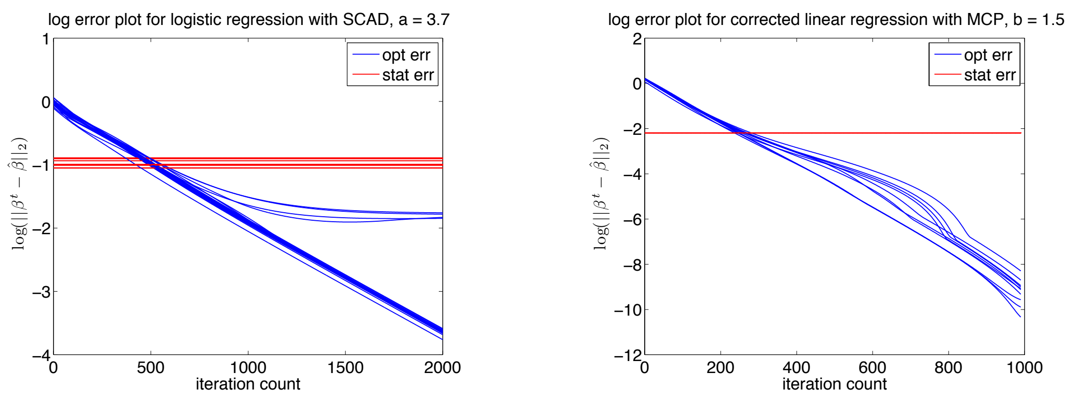
A key insight is that nonconvex functions occurring in statistics are not constructed adversarially, so that "good behavior" might be expected in practice. Our recent work (Loh and Wainwright, 2012) confirmed this intuition for one specific case: a modified version of the Lasso applicable to errors-in-variables regression. Although the Hessian of the modified objective has many negative eigenvalues in the high-dimensional setting, the objective function resembles a strongly convex function when restricted to a cone set that includes the stationary points of the objective. This allows us to establish bounds on the statistical and optimization error.

Our current paper is framed in a more general setting, and we focus on various M-estimators coupled with (nonconvex) regularizers of interest. On the statistical side, we establish bounds on the distance between *any local optimum* of the empirical objective and the unique minimizer of the population risk. Although the nonconvex functions may possess multiple local optima (as demonstrated in simulations), our theoretical results show that all local optima are essentially as good as a global optima from a statistical perspective. The results presented here subsume our previous work (Loh and Wainwright, 2012), and our present proof techniques are much more direct.

Our theory also sheds new light on a recent line of work involving the nonconvex SCAD and MCP regularizers (Fan and Li, 2001; Breheny and Huang, 2011; Zhang, 2010; Zhang and Zhang, 2012). Various methods previously proposed for nonconvex optimization include local quadratic approximation (LQA) (Fan and Li, 2001), minorization-maximization (MM) (Hunter and Li, 2005), local linear approximation (LLA) (Zou and Li, 2008), and coordinate descent (Breheny and Huang, 2011; Mazumder et al., 2011). However, these methods may terminate in local optima, which were not previously known to be well-behaved. In a recent paper, Zhang and Zhang (2012) provided statistical guarantees for global optima of least-squares linear regression with nonconvex penalties and showed that gradient descent starting from a Lasso solution would terminate in specific local minima. Fan et al. (2014) also showed that if the LLA algorithm is initialized at a Lasso optimum satisfying certain properties, the two-stage procedure produces an oracle solution for various nonconvex penalties. Finally, Chen and Gu (2014) showed that specific local optima of nonconvex regularized least-squares problems are stable, so optimization algorithms initialized sufficiently close by will converge to the same optima. See the survey paper (Zhang and Zhang, 2012) for a more complete overview of related work.

In contrast, our paper is the first to establish appropriate regularity conditions under which *all stationary points* (including both local and global optima) lie within a small ball of the population-level minimum. Thus, standard first-order methods such as projected and composite gradient descent (Nesterov, 2007) will converge to stationary points that lie within statistical error of the truth, eliminating the need for specially designed optimization algorithms that converge to specific local optima. Our work provides an important contribution to the growing literature on the tradeoff between statistical accuracy and optimization efficiency in high-dimensional problems, establishing that certain types of nonconvex M-estimators arising in statistical problems possess stationary points that both enjoy strong statistical guarantees and may be located efficiently. For a higher-level description of contemporary problems involving statistical and optimization tradeoffs, see Wainwright (2014) and the references cited therein.

Figure 1 provides an illustration of the type of behavior explained by the theory in this paper. Panel (a) shows the behavior of composite gradient descent for a form of logistic regression with the nonconvex SCAD (Fan and Li, 2001) as a regularizer: the red curve shows the *statistical error*, namely the ℓ_2-norm of the difference between a stationary point and the underlying true regression vector, and the blue curve shows the *optimization error*, meaning the difference between the iterates and a given global optimum. As shown by the blue curves, this problem possesses multiple local optima, since the algorithm converges to different final points depending on the initialization. However, as shown by the red curves, the statistical error of each local optimum is very low, so they are all essentially comparable from a statistical point of view. Panel (b) exhibits the same behavior for a problem in which



(a) (b)

Figure 1: Plots of the optimization error (blue curves) and statistical error (red curves) for a modified form of composite gradient descent, applicable to problems that may involve nonconvex cost functions and regularizers. (a) Plots for logistic regression with the nonconvex SCAD regularizer. (b) Plots for a corrected form of least squares (a nonconvex quadratic program) with the nonconvex MCP regularizer.

both the cost function (a corrected form of least-squares suitable for missing data, described in Loh and Wainwright, 2013a) and the regularizer (the MCP function, described in Zhang, 2010) are nonconvex. Nonetheless, as guaranteed by our theory, we still see the same qualitative behavior of the statistical and optimization error. Moreover, our theory also predicts the geometric convergence rates that are apparent in these plots. More precisely, under the same sufficient conditions for statistical consistency, we show that a modified form of composite gradient descent only requires $\log(1/\epsilon_{\text{stat}})$ steps to achieve a solution that is accurate up to the statistical precision ϵ_{stat}, which is the rate expected for *strongly convex* functions. Furthermore, our techniques are more generally applicable than the methods proposed by previous authors and are not restricted to least-squares or even convex loss functions.

While our paper was under review after its initial arXiv posting (Loh and Wainwright, 2013b), we became aware of an independent line of related work by Wang et al. (2014). Our contributions are substantially different, in that we provide sufficient conditions guaranteeing statistical consistency for *all* local optima, whereas their work is only concerned with establishing good behavior of successive iterates along a certain path-following algorithm. In addition, our techniques are applicable even to regularizers that do not satisfy smoothness constraints on the entire positive axis (such as capped-ℓ_1). Finally, we provide rigorous proofs showing the applicability of our sufficient condition on the loss function to a broad class of generalized linear models, whereas the applicability of their sparse eigenvalue condition to such objectives was not established.

The remainder of the paper is organized as follows. In Section 2, we set up basic notation and provide background on nonconvex regularizers and loss functions of interest. In Section 3, we provide our main theoretical results, including bounds on ℓ_1-, ℓ_2-, and prediction error, and also state corollaries for special cases. Section 4 contains a modification of composite gradient descent that may be used to obtain near-global optima and includes theoretical results establishing the linear convergence of our optimization algorithm. Section 5 supplies the results of various simulations. Proofs are contained in the Appendix. We note that a preliminary form of the results given here, without any proofs or algorithmic details, was presented at the NIPS conference (Loh and Wainwright, 2013c).

Notation: For functions $f(n)$ and $g(n)$, we write $f(n) \precsim g(n)$ to mean that $f(n) \le cg(n)$ for some universal constant $c \in (0, \infty)$, and similarly, $f(n) \succsim g(n)$ when $f(n) \ge c'g(n)$ for some universal 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 $|||M|||_{\max} := \max_{i,j} |m_{ij}|$ 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 $q, r > 0$, let $\mathbb{B}_q(r)$ denote the ℓ_q-ball of radius r centered around 0. We use the term "with high probability" (w.h.p.) to refer to events that occur with probability tending to 1 as $n, p, k \to \infty$. This is a loose requirement, but we will always take care to write out the expression for the probability explicitly (up to constant factors) in the formal statements of our theorems and corollaries below.

2. Problem Formulation

In this section, we develop some general theory for regularized M-estimators. We begin by establishing our notation and basic assumptions, before turning to the class of nonconvex regularizers and nonconvex loss functions to be covered in this paper.

2.1 Background

Given a collection of n samples $Z_1^n = \{Z_1, \dots, Z_n\}$, drawn from a marginal distribution $\mathbb{P}$ over a space $\mathcal{Z}$, consider a loss function $\mathcal{L}_n : \mathbb{R}^p \times (\mathcal{Z})^n \to \mathbb{R}$. The value $\mathcal{L}_n(\beta; Z_1^n)$ serves as a measure of the "fit" between a parameter vector $\beta \in \mathbb{R}^p$ and the observed data. This empirical loss function should be viewed as a surrogate to the *population risk function*

$\mathcal{L} : \mathbb{R}^p \to \mathbb{R}$, given by

$$\mathcal{L}(\beta) := \mathbb{E}_Z\big[\mathcal{L}_n(\beta; Z_1^n)\big].$$

Our goal is to estimate the parameter vector $\beta^* := \arg\min_{\beta \in \mathbb{R}^p} \mathcal{L}(\beta)$ that minimizes the population risk, assumed to be unique.

To this end, we consider a regularized M-estimator of the form

$$\widehat{\beta} \in \arg \min_{g(\beta) \le R,\, \beta \in \Omega} \left\{\mathcal{L}_n(\beta; Z_1^n) + \rho_\lambda(\beta)\right\}, \tag{1}$$

where $\rho_\lambda : \mathbb{R}^p \to \mathbb{R}$ is a *regularizer*, depending on a tuning parameter $\lambda > 0$, which serves to enforce a certain type of structure on the solution. Here, $R > 0$ is another tuning parameter that much be chosen carefully to make β^* a feasible point. In all cases, we consider regularizers that are separable across coordinates, and with a slight abuse of notation, we write

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

Our theory allows for possible nonconvexity in *both* the loss function $\mathcal{L}_n$ and the regularizer ρ_λ. Due to this potential nonconvexity, our M-estimator also includes a side constraint $g : \mathbb{R}^p \to \mathbb{R}_+$, which we require to be a convex function satisfying the lower bound $g(\beta) \ge \|\beta\|_1$ for all $\beta \in \mathbb{R}^p$. Consequently, any feasible point for the optimization problem (1) satisfies the constraint $\|\beta\|_1 \le R$, and as long as the empirical loss and regularizer are continuous, the Weierstrass extreme value theorem guarantees that a global minimum $\widehat{\beta}$ exists. Finally, our theory also allows for an additional side constraint of the form $\beta \in \Omega$, where Ω is some convex set containing β^*. For the graphical Lasso considered in Section 3.4, we take $\Omega = \mathcal{S}_+$ to be the set of positive semidefinite matrices; in settings where such an additional condition is extraneous, we simply set $\Omega = \mathbb{R}^p$.

2.2 Nonconvex Regularizers

We now state and discuss conditions on the regularizer, defined in terms of a univariate function $\rho_\lambda : \mathbb{R} \to \mathbb{R}$.

Assumption 1

(i) The function ρ_λ satisfies $\rho_\lambda(0) = 0$ and is symmetric around zero (i.e., $\rho_\lambda(t) = \rho_\lambda(-t)$ for all $t \in \mathbb{R}$).

(ii) On the nonnegative real line, the function ρ_λ is nondecreasing.

(iii) For $t > 0$, the function $t \mapsto \frac{\rho_\lambda(t)}{t}$ is nonincreasing in t.

(iv) The function ρ_λ is differentiable for all $t \neq 0$ and subdifferentiable at $t = 0$, with $\lim_{t \to 0^+} \rho'_\lambda(t) = \lambda L$.

(v) There exists $\mu > 0$ such that $\rho_{\lambda,\mu}(t) := \rho_\lambda(t) + \frac{\mu}{2} t^2$ is convex.

It is instructive to compare the conditions of Assumption 1 to similar conditions previously proposed in literature. Conditions (i)–(iii) are the same as those proposed in Zhang and Zhang (2012), except we omit the extraneous condition of subadditivity (cf. Lemma 1 of Chen and Gu, 2014). Such conditions are relatively mild and are satisfied for a wide variety of regularizers. Condition (iv) restricts the class of penalties by excluding regularizers such as the bridge (ℓ_q-) penalty, which has infinite derivative at 0; and the capped-ℓ_1 penalty, which has points of non-differentiability on the positive real line. However, one may check that if ρ_λ has an unbounded derivative at zero, then $\widetilde{\beta} = 0$ is *always* a local optimum of the composite objective (1), so there is no hope for $\|\widetilde{\beta} - \beta^*\|_2$ to be vanishingly small. Condition (v), known as *weak convexity* (Vial, 1982), also appears in Chen and Gu (2014) and is a type of curvature constraint that controls the level of nonconvexity of ρ_λ. Although this condition is satisfied by many regularizers of interest, it is again not satisfied by capped-ℓ_1 for any $\mu > 0$. For details on how our arguments may be modified to handle the more tricky capped-ℓ_1 penalty, see Appendix F.

Nonetheless, many regularizers that are commonly used in practice satisfy all the conditions in Assumption 1. It is easy to see that the standard ℓ_1-norm $\rho_\lambda(\beta) = \lambda\|\beta\|_1$ satisfies these conditions. More exotic functions have been studied in a line of past work on nonconvex regularization, and we provide a few examples here:

SCAD penalty: This penalty, due to Fan and Li (2001), takes the form

$$\rho_\lambda(t) := \begin{cases} \lambda|t|, & \text{for } |t| \le \lambda, \\ -(t^2 - 2a\lambda|t| + \lambda^2)/(2(a-1)), & \text{for } \lambda < |t| \le a\lambda, \\ (a+1)\lambda^2/2, & \text{for } |t| > a\lambda, \end{cases} \tag{2}$$

where $a > 2$ is a fixed parameter. As verified in Lemma 6 of Appendix A.2, the SCAD penalty satisfies the conditions of Assumption 1 with $L = 1$ and $\mu = \frac{1}{a-1}$.

MCP regularizer: This penalty, due to Zhang (2010), takes the form

$$\rho_\lambda(t) := \operatorname{sign}(t)\,\lambda \cdot \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz, \tag{3}$$

where $b > 0$ is a fixed parameter. As verified in Lemma 7 in Appendix A.2, the MCP regularizer satisfies the conditions of Assumption 1 with $L = 1$ and $\mu = \frac{1}{b}$.

2.3 Nonconvex Loss Functions and Restricted Strong Convexity

Throughout this paper, we require the loss function $\mathcal{L}_n$ to be differentiable, but we do not require it to be convex. Instead, we impose a weaker condition known as restricted strong convexity (RSC). Such conditions have been discussed in previous literature (Negahban et al., 2012; Agarwal et al., 2012), and involve a lower bound on the remainder in the first-order Taylor expansion of $\mathcal{L}_n$. In particular, our main statistical result is based on the following RSC condition:

$$\langle \nabla \mathcal{L}_n(\beta^* + \Delta) - \nabla \mathcal{L}_n(\beta^*),\, \Delta \rangle \geq \begin{cases} \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log p}{n} \|\Delta\|_1^2, & \forall \|\Delta\|_2 \leq 1, \quad \text{(4a)} \\ \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_1, & \forall \|\Delta\|_2 \geq 1, \quad \text{(4b)} \end{cases}$$

where the α_j's are strictly positive constants and the τ_j's are nonnegative constants.

To understand this condition, note that if $\mathcal{L}_n$ were actually strongly convex, then both these RSC inequalities would hold with $\alpha_1 = \alpha_2 > 0$ and $\tau_1 = \tau_2 = 0$. However, in the high-dimensional setting ($p \gg n$), the empirical loss $\mathcal{L}_n$ will not in general be strongly convex or even convex, but the RSC condition may still hold with strictly positive (α_j, τ_j). In fact, if $\mathcal{L}_n$ is convex (but not strongly convex), the left-hand expression in (4) is always nonnegative, so (4a) and (4b) hold trivially for $\frac{\|\Delta\|_1}{\|\Delta\|_2} \geq \sqrt{\frac{\alpha_1 n}{\tau_1 \log p}}$ and $\frac{\|\Delta\|_1}{\|\Delta\|_2} \geq \frac{\alpha_2}{\tau_2}\sqrt{\frac{n}{\log p}}$, respectively. Hence, the RSC inequalities only enforce a type of strong convexity condition over a cone of the form $\left\{\frac{\|\Delta\|_1}{\|\Delta\|_2} \leq c\sqrt{\frac{n}{\log p}}\right\}$.

It is important to note that the class of functions satisfying RSC conditions of this type is much larger than the class of convex functions; for instance, our own past work (Loh and Wainwright, 2012) exhibits a large family of nonconvex quadratic functions that satisfy the condition (see Section 3.2 below for further discussion). Furthermore, note that we have stated two separate RSC inequalities (4) for different ranges of $\|\Delta\|_2$, unlike in past work (Negahban et al., 2012; Agarwal et al., 2012; Loh and Wainwright, 2012). As illustrated in the corollaries of Sections 3.3 and 3.4 below, an equality of the first type (4a) will only hold locally over Δ when we have more complicated types of loss functions that are only quadratic around a neighborhood of the origin. As proved in Appendix B.1, however, (4b) is implied by (4a) in cases when $\mathcal{L}_n$ is convex, which sustains our theoretical conclusions even under the weaker RSC conditions (4). Further note that by the inequality

$$\mathcal{L}_n(\beta^* + \Delta) - \mathcal{L}_n(\beta^*) \leq \langle \nabla \mathcal{L}_n(\beta^* + \Delta),\, \Delta \rangle,$$

which holds whenever $\mathcal{L}_n$ is convex, the RSC condition appearing in past work (e.g., Agarwal et al., 2012) implies that (4a) holds, so (4b) also holds by Lemma 8 in Appendix B.1. In cases where $\mathcal{L}_n$ is quadratic but not necessarily convex (cf. Section 3.2), our RSC condition (4) is again no stronger than the conditions appearing in past work, since those RSC conditions enforce (4a) globally over $\Delta \in \mathbb{R}^p$, which by Lemma 9 in Appendix B.1 implies that (4b) holds, as well. To allow for more general situations where $\mathcal{L}_n$ may be non-quadratic and/or nonconvex, we prefer to use the RSC formulation (4) in this paper.

Finally, we clarify that whereas Negahban et al. (2012) define an RSC condition with respect to a fixed subset $S \subseteq \{1, \ldots, p\}$, we follow the setup of Agarwal et al. (2012) and Loh and Wainwright (2012) and essentially require an RSC condition of the type defined in Negahban et al. (2012) to hold uniformly over *all* subsets S of size k. Although the results on statistical consistency may be established under the weaker RSC assumption with $S := \text{supp}(\beta^*)$, a uniform RSC condition is preferred because the true support set is not known a priori. The uniform RSC condition may be shown to hold w.h.p. in the sub-Gaussian settings we consider here (cf. Sections 3.2—3.4 below); in fact, the proofs contained in Negahban et al. (2012) establish a uniform RSC condition, as well.

3. Statistical Guarantees and Consequences

With this setup, we now turn to the statements and proofs of our main statistical guarantees, as well as some consequences for various statistical models. Our theory applies to any vector $\widetilde{\beta} \in \mathbb{R}^p$ that satisfies the *first-order necessary conditions* to be a local minimum of the program (1):

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) + \nabla \rho_\lambda(\widetilde{\beta}),\, \beta - \widetilde{\beta} \rangle \geq 0, \qquad \text{for all feasible } \beta \in \mathbb{R}^p. \tag{5}$$

When $\widetilde{\beta}$ lies in the interior of the constraint set, this condition reduces to the usual zero-subgradient condition:

$$\nabla \mathcal{L}_n(\widetilde{\beta}) + \nabla \rho_\lambda(\widetilde{\beta}) = 0.$$

Such vectors $\widetilde{\beta}$ satisfying the condition (5) are also known as *stationary points* (Bertsekas, 1999); note that the set of stationary points also includes interior local maxima. Hence, although some of the discussion below is stated in terms of "local minima," the results hold for interior local maxima, as well.

3.1 Main Statistical Results

Our main theorems are deterministic in nature and specify conditions on the regularizer, loss function, and parameters that guarantee that any local optimum $\widetilde{\beta}$ lies close to the target vector $\beta^* = \arg\min_{\beta \in \mathbb{R}^p} \mathcal{L}(\beta)$. Corresponding probabilistic results will be derived in subsequent sections, where we establish that for appropriate choices of parameters (λ, R), the required conditions hold with high probability. Applying the theorems to particular models requires bounding the random quantity $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty$ and verifying the RSC conditions (4). We begin with a theorem that provides guarantees on the error $\widetilde{\beta} - \beta^*$ as measured in the ℓ_1- and ℓ_2-norms:

Theorem 1 *Suppose the regularizer ρ_λ satisfies Assumption 1, the empirical loss $\mathcal{L}_n$ satisfies the RSC conditions* (4) *with $\frac{3}{4}\mu < \alpha_1$, and β^* is feasible for the objective. Consider any choice of λ such that*

$$\frac{4}{L} \cdot \max\left\{ \|\nabla \mathcal{L}_n(\beta^*)\|_\infty,\ \alpha_2 \sqrt{\frac{\log p}{n}} \right\} \;\leq\; \lambda \;\leq\; \frac{\alpha_2}{6RL}, \tag{6}$$

and suppose $n \geq \frac{16R^2 \max(\tau_1^2, \tau_2^2)}{\alpha_2^2} \log p$. Then any vector $\widetilde{\beta}$ satisfying the first-order necessary conditions (5) *satisfies the error bounds*

$$\|\widetilde{\beta} - \beta^*\|_2 \leq \frac{6\lambda L\sqrt{k}}{4\alpha_1 - 3\mu}, \qquad \text{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \leq \frac{24\lambda L k}{4\alpha_1 - 3\mu}, \tag{7}$$

where $k = \|\beta^*\|_0$.

From the bound (7), note that the squared ℓ_2-error grows proportionally with k, the number of nonzeros in the target parameter, and with λ^2. As will be clarified in the following sections, choosing λ proportional to $\sqrt{\frac{\log p}{n}}$ and R proportional to $\frac{1}{\lambda}$ will satisfy the requirements of Theorem 1 w.h.p. for many statistical models, in which case we have a squared-ℓ_2 error that scales as $\frac{k \log p}{n}$, as expected.

Our next theorem provides a bound on a measure of the prediction error, as defined by the quantity

$$D(\widetilde{\beta}; \beta^*) := \langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \, \widetilde{\beta} - \beta^* \rangle. \tag{8}$$

When the empirical loss $\mathcal{L}_n$ is a convex function, this measure is always nonnegative, and in various special cases, it has a form that is readily interpretable. For instance, in the case of the least-squares objective function $\mathcal{L}_n(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2$, we have

$$D(\widetilde{\beta}; \beta^*) = \frac{1}{n}\|X(\widetilde{\beta} - \beta^*)\|_2^2 = \frac{1}{n}\sum_{i=1}^n \left(\langle x_i, \, \widetilde{\beta} - \beta^* \rangle\right)^2,$$

corresponding to the usual measure of (fixed design) prediction error for a linear regression problem (cf. Corollary 1 below). More generally, when the loss function is the negative log likelihood for a generalized linear model with cumulant function ψ, the error measure (8) is equivalent to the symmetrized Bregman divergence defined by ψ. (See Section 3.3 for further details.)

Theorem 2 *Under the same conditions as Theorem 1, the error measure* (8) *is bounded as*

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \, \widetilde{\beta} - \beta^* \rangle \le \lambda^2 L^2 k \left(\frac{9}{4\alpha_1 - 3\mu} + \frac{27\mu}{(4\alpha_1 - 3\mu)^2} \right). \tag{9}$$

This result shows that the prediction error (8) behaves similarly to the squared Euclidean norm between $\widetilde{\beta}$ and β^*.

Remark on (α_1, μ)**:** It is worthwhile to discuss the quantity $4\alpha_1 - 3\mu$ appearing in the denominator of the bounds in Theorems 1 and 2. Recall that α_1 measures the level of curvature of the loss function $\mathcal{L}_n$, while μ measures the level of nonconvexity of the penalty ρ_λ. Intuitively, the two quantities should play opposing roles in our result: larger values of μ correspond to more severe nonconvexity of the penalty, resulting in worse behavior of the overall objective (1), whereas larger values of α_1 correspond to more (restricted) curvature of the loss, leading to better behavior. However, while the condition $\frac{3}{4}\mu < \alpha_1$ is needed for the proof technique employed in Theorem 1, it does not seem to be strictly necessary in order to guarantee good behavior of local optima. As a careful examination of the proof reveals, the condition may be replaced by the alternate condition $c\mu < \alpha_1$, for any constant $c > \frac{1}{2}$. However, note that the capped-ℓ_1 penalty may be viewed as a limiting version of SCAD when $a \to 1$, or equivalently, $\mu \to \infty$. Viewed in this light, Theorem 4, to be stated and proved in Appendix F, reveals that a condition of the form $c\mu < \alpha_1$ is not necessary, at least in general, for good behavior of local optima. Moreover, Section 5 contains empirical studies using linear regression and the SCAD penalty showing that local optima may be well-behaved when $\alpha_1 < \frac{3}{4}\mu$. Nonetheless, our simulations (see Figure 5) also convey a cautionary message: In extreme cases, where α_1 is significantly smaller than μ, the good behavior of local optima (and the optimization algorithms used to find them) appear to degenerate.

Finally, we note that Negahban et al. (2012) have shown that for convex M-estimators, the arguments used to analyze ℓ_1-regularizers may be generalized to other types of "decomposable" regularizers, such as norms for group sparsity or the nuclear norm for low-rank matrices. In our present setting, where we allow for nonconvexity in the loss and regularizer, our theorems have straightforward and analogous generalizations.

We return to the proofs of Theorems 1 and 2 in Section 3.5. First, we develop various consequences of these theorems for various nonconvex loss functions and regularizers of interest. The main technical challenge is to establish that the RSC conditions (4) hold with high probability for appropriate choices of positive constants $\{(\alpha_j, \tau_j)\}_{j=1}^2$.

3.2 Corrected Linear Regression

We begin by considering the case of high-dimensional linear regression with systematically corrupted observations. Recall that in the framework of ordinary linear regression, we have the linear model

$$y_i = \underbrace{\langle \beta^*, x_i \rangle}_{\sum_{j=1}^p \beta_j^* x_{ij}} + \epsilon_i, \qquad \text{for } i = 1, \dots, n, \tag{10}$$

where $\beta^* \in \mathbb{R}^p$ is the unknown parameter vector and $\{(x_i, y_i)\}_{i=1}^n$ are observations. Following a line of past work (e.g., Rosenbaum and Tsybakov, 2010; Loh and Wainwright, 2012), assume we instead observe pairs $\{(z_i, y_i)\}_{i=1}^n$, where the z_i's are systematically corrupted versions of the corresponding x_i's. Some examples of corruption mechanisms include the following:

(a) *Additive noise:* We observe $z_i = x_i + w_i$, where $w_i \in \mathbb{R}^p$ is a random vector independent of x_i, say zero-mean with known covariance matrix Σ_w.

(b) *Missing data:* For some fraction $\vartheta \in [0, 1)$, we observe a random vector $z_i \in \mathbb{R}^p$ such that for each component j, we independently observe $z_{ij} = x_{ij}$ with probability $1 - \vartheta$, and $z_{ij} = *$ with probability ϑ.

We use the population and empirical loss functions

$$\mathcal{L}(\beta) = \frac{1}{2}\beta^T \Sigma_x \beta - \beta^{*T} \Sigma_x \beta, \qquad \text{and} \qquad \mathcal{L}_n(\beta) = \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta, \tag{11}$$

where $(\widehat{\Gamma}, \widehat{\gamma})$ are estimators for $(\Sigma_x, \Sigma_x \beta^*)$ that depend only on $\{(z_i, y_i)\}_{i=1}^n$. It is easy to see that $\beta^* = \arg\min_\beta \mathcal{L}(\beta)$. From the formulation (1), the corrected linear regression estimator is given by

$$\widehat{\beta} \in \arg \min_{g(\beta) \le R} \left\{ \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta + \rho_\lambda(\beta) \right\}. \tag{12}$$

We now state a concrete corollary in the case of additive noise (model (a) above). In this case, as discussed in Loh and Wainwright (2012), an appropriate choice of the pair $(\widehat{\Gamma}, \widehat{\gamma})$ is given by

$$\widehat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w, \qquad \text{and} \qquad \widehat{\gamma} = \frac{Z^T y}{n}. \tag{13}$$

Here, we assume the noise covariance Σ_w is known or may be estimated from replicates of the data. Such an assumption also appears in canonical errors-in-variables literature (Carroll et al., 1995), but it is an open question how to devise a corrected estimator when an estimate of Σ_w is not readily available. If we assume a sub-Gaussian model on the covariates and errors (i.e., x_i, w_i, and ϵ_i are sub-Gaussian with parameters σ_x^2, σ_w^2, and σ_ϵ^2, respectively), the contribution of the error covariances may be summarized in the error term

$$\varphi = (\sigma_x + \sigma_w)\big(\sigma_\epsilon + \sigma_w \|\beta^*\|_2\big), \tag{14}$$

which appears as a prefactor in the deviation bounds and estimation/prediction error bounds for the subsequent estimators (cf. Lemma 2 in Loh and Wainwright, 2012). We make this dependence explicit in the statement of the corollary for high-dimensional errors-in-variables regression below. Note in particular that φ scales up with both σ_ϵ and σ_w. Hence, even when $\sigma_\epsilon = 0$, corresponding to no additive error, we will have $\varphi \neq 0$ due to errors in the covariates; whereas when $\sigma_w = 0$, corresponding to cleanly observed covariates, we will still have $\varphi \neq 0$ due to the additional additive error introduced by the ϵ_i's, agreeing with canonical results for the Lasso (Bickel et al., 2009).

In the high-dimensional setting ($p \gg n$), the matrix $\widehat{\Gamma}$ in (13) is always negative definite: the matrix $\frac{Z^T Z}{n}$ has rank at most n, and the positive definite matrix Σ_w is then subtracted to obtain $\widehat{\Gamma}$. Consequently, the empirical loss function $\mathcal{L}_n$ previously defined (11) is nonconvex. Other choices of $\widehat{\Gamma}$ are applicable to missing data (model (b)), and also lead to nonconvex programs (see Loh and Wainwright, 2012 for further details).

Corollary 1 *Suppose we have i.i.d. observations $\{(z_i, y_i)\}_{i=1}^n$ from a corrupted linear model with additive noise, where the covariates and error terms are sub-Gaussian. Let φ be defined as in* (14) *with respect to the sub-Gaussian parameters. Suppose (λ, R) are chosen such that β^* is feasible and*

$$c\varphi\sqrt{\frac{\log p}{n}} \le \lambda \le \frac{c'}{R}.$$

Also suppose $\frac{3}{4}\mu < \frac{1}{2}\lambda_{\min}(\Sigma_x)$. Then given a sample size $n \ge C \max\{R^2, k\}\log p$, any stationary point $\widetilde{\beta}$ of the nonconvex program (12) *satisfies the estimation error bounds*

$$\|\widetilde{\beta} - \beta^*\|_2 \le \frac{c_0 \lambda \sqrt{k}}{2\lambda_{\min}(\Sigma_x) - 3\mu}, \qquad \textit{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \le \frac{c_0' \lambda k}{2\lambda_{\min}(\Sigma_x) - 3\mu},$$

and the prediction error bound

$$\widetilde{\nu}^T \widehat{\Gamma} \widetilde{\nu} \le \lambda^2 k \left(\frac{\widetilde{c_0}}{2\lambda_{\min}(\Sigma_x) - 3\mu} + \frac{\widetilde{c_0}' \mu}{(2\lambda_{\min}(\Sigma_x) - 3\mu)^2} \right),$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$, where $\|\beta^*\|_0 = k$.

When $\rho_\lambda(\beta) = \lambda\|\beta\|_1$ and $g(\beta) = \|\beta\|_1$, taking $\lambda \asymp \varphi\sqrt{\frac{\log p}{n}}$ and $R = b_0\sqrt{k}$ for some constant $b_0 \ge \|\beta^*\|_2$ yields the required scaling $n \succsim k \log p$. Hence, the bounds of Corollary 1 agree with bounds previously established in Theorem 1 of Loh and Wainwright (2012). Note, however, that those results are stated only for a *global minimum* $\widehat{\beta}$ of the program (12), whereas Corollary 1 is a much stronger result holding for *any stationary point* $\widetilde{\beta}$. Theorem 2 of our earlier paper (Loh and Wainwright, 2012) provides a rather indirect (algorithmic) route for establishing similar bounds on $\|\widetilde{\beta} - \beta^*\|_1$ and $\|\widetilde{\beta} - \beta^*\|_2$, since the proposed projected gradient descent algorithm may become stuck at a stationary point. In contrast, our argument here is much more direct and does not rely on an algorithmic proof. Furthermore, our result is applicable to a more general class of (possibly nonconvex) penalties beyond the usual ℓ_1-norm.

Corollary 1 also has important consequences in the case where pairs $\{(x_i, y_i)\}_{i=1}^n$ from the linear model (10) are observed cleanly without corruption and ρ_λ is a nonconvex penalty. In that case, the empirical loss $\mathcal{L}_n$ previously defined (11) is equivalent to the least-squares loss, modulo a constant factor. Much existing work, including that of Fan and Li (2001) and Zhang and Zhang (2012), first establishes statistical consistency results concerning *global* minima of the program (12), then provides specialized algorithms such as a local linear approximation (LLA) for obtaining specific local optima that are provably close to the global optima. However, our results show that *any* optimization algorithm guaranteed to converge to a stationary point of the program suffices. See Section 4 for a more detailed discussion of optimization procedures and fast convergence guarantees for obtaining stationary points. In the fully-observed case, we also have $\widehat{\Gamma} = \frac{X^T X}{n}$, so the prediction error bound in Corollary 1 agrees with the familiar scaling $\frac{1}{n}\|X(\widetilde{\beta} - \beta^*)\|_2^2 \precsim \frac{k \log p}{n}$ appearing in ℓ_1-theory.

Furthermore, our theory provides a theoretical motivation for why the usual choice of $a = 3.7$ for linear regression with the SCAD penalty (Fan and Li, 2001) is reasonable. Indeed, as discussed in Section 2.2, we have

$$\mu = \frac{1}{a-1} \approx 0.37$$

in that case. Since $x_i \sim N(0, I)$ in the SCAD simulations, we have $\frac{3}{4}\mu < \frac{1}{2}\lambda_{\min}(\Sigma_x)$ for the choice $a = 3.7$. For further comments regarding the parameter a in the SCAD penalty, see the discussion concerning Figure 3 in Section 5.

3.3 Generalized Linear Models

Moving beyond linear regression, we now consider the case where observations are drawn from a generalized linear model (GLM). Recall that a GLM is characterized by the conditional distribution

$$\mathbb{P}(y_i \mid x_i, \beta, \sigma) = \exp\left\{\frac{y_i\langle\beta,\, x_i\rangle - \psi(x_i^T\beta)}{c(\sigma)}\right\},$$

where $\sigma > 0$ is a scale parameter and ψ is the cumulant function, By standard properties of exponential families (McCullagh and Nelder, 1989; Lehmann and Casella, 1998), we have

$$\psi'(x_i^T\beta) = \mathbb{E}[y_i \mid x_i, \beta, \sigma].$$

In our analysis, we assume that there exists $\alpha_u > 0$ such that $\psi''(t) \leq \alpha_u$, for all $t \in \mathbb{R}$. Note that this boundedness assumption holds in various settings, including linear regression, logistic regression, and multinomial regression, but does not hold for Poisson regression.

The bound will be necessary to establish both statistical consistency results in the present section and fast global convergence guarantees for our optimization algorithms in Section 4.

The population loss corresponding to the negative log likelihood is then given by

$$\mathcal{L}(\beta) = -\mathbb{E}[\log \mathbb{P}(x_i, y_i)] = -\mathbb{E}[\log \mathbb{P}(x_i)] - \frac{1}{c(\sigma)} \cdot \mathbb{E}[y_i \langle \beta,\, x_i \rangle - \psi(x_i^T \beta)],$$

giving rise to the population-level and empirical gradients

$$\nabla \mathcal{L}(\beta) = \frac{1}{c(\sigma)} \cdot \mathbb{E}[(\psi'(x_i^T \beta) - y_i) x_i], \quad \text{and}$$
$$\nabla \mathcal{L}_n(\beta) = \frac{1}{c(\sigma)} \cdot \frac{1}{n} \sum_{i=1}^n \big(\psi'(x_i^T \beta) - y_i\big) x_i.$$

Since we are optimizing over β, we will rescale the loss functions and assume $c(\sigma) = 1$. We may check that if β^* is the true parameter of the GLM, then $\nabla \mathcal{L}(\beta^*) = 0$; furthermore,

$$\nabla^2 \mathcal{L}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \psi''(x_i^T \beta) x_i x_i^T \succeq 0,$$

so $\mathcal{L}_n$ is convex.

We will assume that β^* is sparse and optimize the penalized maximum likelihood program

$$\widehat{\beta} \in \arg \min_{g(\beta) \le R} \left\{ \frac{1}{n} \sum_{i=1}^n \big(\psi(x_i^T \beta) - y_i x_i^T \beta\big) + \rho_\lambda(\beta) \right\}. \tag{15}$$

We then have the following corollary, proved in Appendix B.3:

Corollary 2 *Suppose we have i.i.d. observations $\{(x_i, y_i)\}_{i=1}^n$ from a GLM, where the x_i's are sub-Gaussian. Suppose (λ, R) are chosen such that β^* is feasible and*

$$c\sqrt{\frac{\log p}{n}} \le \lambda \le \frac{c'}{R}.$$

Then given a sample size $n \ge CR^2 \log p$, any stationary point $\widetilde{\beta}$ of the nonconvex program (15) satisfies

$$\|\widetilde{\beta} - \beta^*\|_2 \le \frac{c_0 \lambda \sqrt{k}}{4\alpha_1 - 3\mu}, \qquad \text{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \le \frac{c_0' \lambda k}{4\alpha_1 - 3\mu},$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$, where $\|\beta^*\|_0 = k$. Here, α_1 is a constant depending on $\|\beta^*\|_2$, ψ, $\lambda_{\min}(\Sigma_x)$, and the sub-Gaussian parameter of the x_i's, and we assume $\mu < 2\alpha_1$.

Although $\mathcal{L}_n$ is convex in this case, the overall program may *not* be convex if the regularizer ρ_λ is nonconvex, giving rise to multiple local optima. For instance, see the simulations of Figure 4 in Section 5 for a demonstration of such local optima. In past work, Breheny and Huang (2011) studied logistic regression with SCAD and MCP regularizers, but did not provide any theoretical results on the quality of the local optima. In this context, Corollary 2 shows that their coordinate descent algorithms are guaranteed to converge to a stationary point $\widetilde{\beta}$ within close proximity of the true parameter β^*.

In the statement of Corollary 2, we choose not to write out the form of α_1 explicitly as in Corollary 1, since it is rather complicated. As explained in the proof of Corollary 2 in Appendix B.3, the precise form of α_1 may be traced back to Proposition 2 of Negahban et al. (2012).

3.4 Graphical Lasso

Finally, we specialize our results to the case of the graphical Lasso. Given p-dimensional observations $\{x_i\}_{i=1}^n$, the goal is to estimate the structure of the underlying (sparse) graphical model. Recall that the population and empirical losses for the graphical Lasso are given by

$$\mathcal{L}(\Theta) = \text{trace}(\Sigma\Theta) - \log\det(\Theta), \quad \text{and} \quad \mathcal{L}_n(\Theta) = \text{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta),$$

where $\widehat{\Sigma}$ is an empirical estimate for the covariance matrix $\Sigma = \text{Cov}(x_i)$. The objective function for the graphical Lasso is then given by

$$\widehat{\Theta} \in \arg \min_{g(\Theta) \le R,\, \Theta \succeq 0} \left\{ \text{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \sum_{j,k=1}^{p} \rho_\lambda(\Theta_{jk}) \right\}, \tag{16}$$

where we apply the (possibly nonconvex) penalty function ρ_λ to all entries of Θ, and define $\Omega := \left\{\Theta \in \mathbb{R}^{p \times p} \mid \Theta = \Theta^T,\ \Theta \succeq 0\right\}$.

A host of statistical and algorithmic results have been established for the graphical Lasso in the case of Gaussian observations with an ℓ_1-penalty (Banerjee et al., 2008; Friedman et al., 2008; Rothman et al., 2008; Yuan and Lin, 2007), and more recently, for discrete-valued observations, as well (Loh and Wainwright, 2013a). In addition, a version of the graphical Lasso incorporating a nonconvex SCAD penalty has been proposed (Fan et al., 2009). Our results subsume previous Frobenius error bounds for the graphical Lasso and again imply that even in the presence of a nonconvex regularizer, all stationary points of the nonconvex program (16) remain close to the true inverse covariance matrix Θ^*.

As suggested by Loh and Wainwright (2013a), the graphical Lasso easily accommodates systematically corrupted observations, with the only modification being the form of the sample covariance matrix $\widehat{\Sigma}$. Just as in Corollary 1, the magnitude and form of corruption would occur as a prefactor in the deviation condition captured in (17) below; for instance, in the case of $\widehat{\Sigma} = \frac{Z^T Z}{n} - \Sigma_w$, corresponding to additive noise in the x_i's, the bound (17) would involve a prefactor of σ_z^2 rather than σ_x^2, where σ_z^2 and σ_x^2 are the sub-Gaussian parameters of z_i and x_i, respectively.

Further note that the program (16) is always useful for obtaining a consistent estimate of a sparse inverse covariance matrix, regardless of whether the x_i's are drawn from a distribution for which Θ^* is relevant in estimating the edges of the underlying graph. Note that other variants of the graphical Lasso exist in which only off-diagonal entries of Θ are penalized, and similar results for statistical consistency hold in that case. Here, we assume that all entries are penalized equally in order to simplify our arguments. The same framework is considered by Fan et al. (2009).

We have the following result, proved in Appendix B.4. The statement of the corollary is purely deterministic, but in cases of interest (say, sub-Gaussian observations), the deviation condition (17) holds with probability at least $1 - c_1 \exp(-c_2 \log p)$, translating into the Frobenius norm bound (18) holding with the same probability.

Corollary 3 *Suppose we have an estimate $\widehat{\Sigma}$ of the covariance matrix Σ based on (possibly corrupted) observations $\{x_i\}_{i=1}^n$, such that*

$$\left\|\left\| \widehat{\Sigma} - \Sigma \right\|\right\|_{\max} \le c_0 \sqrt{\frac{\log p}{n}}. \tag{17}$$

Also suppose Θ^* has at most s nonzero entries. Suppose (λ, R) are chosen such that Θ^* is feasible and

$$c\sqrt{\frac{\log p}{n}} \le \lambda \le \frac{c'}{R}.$$

Suppose $\frac{3}{4}\mu < \left(\|\Theta^*\|_2 + 1\right)^{-2}$. Then with a sample size $n > Cs \log p$, for a sufficiently large constant $C > 0$, any stationary point $\widetilde{\Theta}$ of the nonconvex program (16) satisfies

$$\left\|\left\| \widetilde{\Theta} - \Theta^* \right\|\right\|_F \le \frac{c_0' \lambda \sqrt{s}}{4\left(\|\Theta^*\|_2 + 1\right)^{-2} - 3\mu}. \tag{18}$$

When ρ is simply the ℓ_1-penalty, the bound (18) from Corollary 3 matches the minimax rates for Frobenius norm estimation of an s-sparse inverse covariance matrix (Rothman et al., 2008; Ravikumar et al., 2011).

3.5 Proof of Theorems 1 and 2

We now turn to the proofs of our two main theorems.

Proof of Theorem 1: Introducing the shorthand $\widetilde{\nu} := \widetilde{\beta} - \beta^*$, we begin by proving that $\|\widetilde{\nu}\|_2 \le 1$. If not, then (4b) gives the lower bound

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \, \widetilde{\nu} \rangle \ge \alpha_2 \|\widetilde{\nu}\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\nu}\|_1. \tag{19}$$

Since β^* is feasible, we may take $\beta = \beta^*$ in (5), and combining with (19) yields

$$\langle -\nabla \rho_\lambda(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \, \widetilde{\nu} \rangle \ge \alpha_2 \|\widetilde{\nu}\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\nu}\|_1. \tag{20}$$

By Hölder's inequality, followed by the triangle inequality, we also have

$$\begin{aligned} \langle -\nabla \rho_\lambda(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \, \widetilde{\nu} \rangle &\le \left\{ \|\nabla \rho_\lambda(\widetilde{\beta})\|_\infty + \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \right\} \|\widetilde{\nu}\|_1 \\ &\overset{(i)}{\le} \left\{ \lambda L + \frac{\lambda L}{2} \right\} \|\widetilde{\nu}\|_1, \end{aligned}$$

where inequality (i) follows since $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \le \frac{\lambda L}{2}$ by the bound (6), and $\|\nabla \rho_\lambda(\widetilde{\beta})\|_\infty \le \lambda L$ by Lemma 4 in Appendix A.1. Combining this upper bound with (20) and rearranging then yields

$$\|\widetilde{\nu}\|_2 \le \frac{\|\widetilde{\nu}\|_1}{\alpha_2}\left(\frac{3\lambda L}{2} + \tau_2\sqrt{\frac{\log p}{n}}\right) \le \frac{2R}{\alpha_2}\left(\frac{3\lambda L}{2} + \tau_2\sqrt{\frac{\log p}{n}}\right).$$

By our choice of λ from (6) and the assumed lower bound on the sample size n, the right hand side is at most 1, so $\|\widetilde{\nu}\|_2 \le 1$, as claimed.

Consequently, we may apply (4a), yielding the lower bound

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*),\, \widetilde{\nu}\rangle \ge \alpha_1\|\widetilde{\nu}\|_2^2 - \tau_1\frac{\log p}{n}\|\widetilde{\nu}\|_1^2. \tag{21}$$

Since the function $\rho_{\lambda,\mu}(\beta) := \rho_\lambda(\beta) + \frac{\mu}{2}\|\beta\|_2^2$ is convex by assumption, we have

$$\rho_{\lambda,\mu}(\beta^*) - \rho_{\lambda,\mu}(\widetilde{\beta}) \ge \langle \nabla \rho_{\lambda,\mu}(\widetilde{\beta}),\, \beta^* - \widetilde{\beta}\rangle \;=\; \langle \nabla \rho_\lambda(\widetilde{\beta}) + \mu\widetilde{\beta},\, \beta^* - \widetilde{\beta}\rangle,$$

implying that

$$\langle \nabla \rho_\lambda(\widetilde{\beta}),\, \beta^* - \widetilde{\beta}\rangle \le \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \frac{\mu}{2}\|\widetilde{\beta} - \beta^*\|_2^2. \tag{22}$$

Combining (21) with (5) and (22), we obtain

$$\alpha_1\|\widetilde{\nu}\|_2^2 - \tau_1\frac{\log p}{n}\|\widetilde{\nu}\|_1^2 \le -\langle \nabla \mathcal{L}_n(\beta^*),\, \widetilde{\nu}\rangle + \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \frac{\mu}{2}\|\widetilde{\beta} - \beta^*\|_2^2.$$

Rearranging and using Hölder's inequality, we then have

$$\begin{aligned}\left(\alpha_1 - \frac{\mu}{2}\right)\|\widetilde{\nu}\|_2^2 &\le \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \cdot \|\widetilde{\nu}\|_1 + \tau_1\frac{\log p}{n}\|\widetilde{\nu}\|_1^2 \\ &\le \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \left(\|\nabla \mathcal{L}_n(\beta^*)\|_\infty + 4R\tau_1\frac{\log p}{n}\right)\|\widetilde{\nu}\|_1. \end{aligned} \tag{23}$$

Note that by our assumptions, we have

$$\|\nabla \mathcal{L}_n(\beta^*)\|_\infty + 4R\tau_1\frac{\log p}{n} \le \frac{\lambda L}{4} + \alpha_2\sqrt{\frac{\log p}{n}} \le \frac{\lambda L}{2}.$$

Combining this with (23) and (53) in Lemma 4 in Appendix A.1, as well as the subadditivity of ρ_λ, we then have

$$\begin{aligned}\left(\alpha_1 - \frac{\mu}{2}\right)\|\widetilde{\nu}\|_2^2 &\le \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \frac{\lambda L}{2}\cdot\left(\frac{\rho_\lambda(\widetilde{\nu})}{\lambda L} + \frac{\mu}{2\lambda L}\|\widetilde{\nu}\|_2^2\right) \\ &\le \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \frac{\rho_\lambda(\beta^*) + \rho_\lambda(\widetilde{\beta})}{2} + \frac{\mu}{4}\|\widetilde{\nu}\|_2^2,\end{aligned}$$

implying that

$$0 \leq \left(\alpha_1 - \frac{3\mu}{4}\right) \|\widetilde{\nu}\|_2^2 \leq \frac{3}{2}\rho_\lambda(\beta^*) - \frac{1}{2}\rho_\lambda(\widetilde{\beta}). \tag{24}$$

In particular, we have $3\rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) \geq 0$, so we may apply Lemma 5 in Appendix A.1 to conclude that

$$3\rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) \leq 3\lambda L\|\widetilde{\nu}_A\|_1 - \lambda L\|\widetilde{\nu}_{A^c}\|_1, \tag{25}$$

where A denotes the index set of the k largest elements of $\widetilde{\beta} - \beta^*$ in magnitude. In particular, we have the cone condition

$$\|\widetilde{\nu}_{A^c}\|_1 \leq 3\|\widetilde{\nu}_A\|_1. \tag{26}$$

Substituting (25) into (24), we then have

$$\left(2\alpha_1 - \frac{3\mu}{2}\right) \|\widetilde{\nu}\|_2^2 \;\leq\; 3\lambda L\|\widetilde{\nu}_A\|_1 - \lambda L\|\widetilde{\nu}_{A^c}\|_1 \;\leq\; 3\lambda L\|\widetilde{\nu}_A\|_1 \;\leq\; 3\lambda L\sqrt{k}\|\widetilde{\nu}\|_2, \tag{27}$$

from which we conclude that

$$\|\widetilde{\nu}\|_2 \leq \frac{6\lambda L\sqrt{k}}{4\alpha_1 - 3\mu},$$

as wanted. The ℓ_1-bound follows from the ℓ_2-bound and the observation that

$$\|\widetilde{\nu}\|_1 \leq \|\widetilde{\nu}_A\|_1 + \|\widetilde{\nu}_{A^c}\|_1 \leq 4\|\widetilde{\nu}_A\|_1 \leq 4\sqrt{k}\|\widetilde{\nu}\|_2,$$

using the cone inequality (26).

Proof of Theorem 2: In order to establish (9), note that combining the first-order condition (5) with the upper bound (22), we have

$$\begin{aligned} \langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \, \widetilde{\nu} \rangle &\leq \langle -\nabla \rho_\lambda(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \, \widetilde{\nu} \rangle \\ &\leq \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \frac{\mu}{2}\|\widetilde{\nu}\|_2^2 + \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \cdot \|\widetilde{\nu}\|_1. \end{aligned} \tag{28}$$

Furthermore, as noted earlier, Lemma 4 in Appendix A.1 implies that

$$\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \cdot \|\widetilde{\nu}\|_1 \leq \frac{\lambda L}{2} \cdot \left(\frac{\rho_\lambda(\beta^*) + \rho_\lambda(\widetilde{\beta})}{\lambda L} + \frac{\mu}{2\lambda L}\|\widetilde{\nu}\|_2^2\right) \leq \frac{\rho_\lambda(\beta^*) + \rho_\lambda(\widetilde{\beta})}{2} + \frac{\mu}{4}\|\widetilde{\nu}\|_2^2.$$

Substituting this into (28) then gives

$$\begin{aligned} \langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \, \widetilde{\nu} \rangle &\leq \frac{3}{2}\rho_\lambda(\beta^*) - \frac{1}{2}\rho_\lambda(\widetilde{\beta}) + \frac{3\mu}{4}\|\widetilde{\nu}\|_2^2 \\ &\leq \frac{3\lambda L}{2}\|\widetilde{\nu}_A\|_1 - \frac{\lambda L}{2}\|\widetilde{\nu}_{A^c}\|_1 + \frac{3\mu}{4}\|\widetilde{\nu}\|_2^2 \\ &\leq \frac{3\lambda L\sqrt{k}}{2}\|\widetilde{\nu}\|_2 + \frac{3\mu}{4}\|\widetilde{\nu}\|_2^2, \end{aligned}$$

so substituting in the ℓ_2-bound (7) yields the desired result.

4. Optimization Algorithms

We now describe how a version of composite gradient descent (Nesterov, 2007) may be applied to efficiently optimize the nonconvex program (1), and show that it enjoys a linear rate of convergence under suitable conditions. In this section, we focus exclusively on a version of the optimization problem with the side function

$$g_{\lambda,\mu}(\beta) := \frac{1}{\lambda}\Big\{\rho_\lambda(\beta) + \frac{\mu}{2}\|\beta\|_2^2\Big\}. \tag{29}$$

Note that this choice of $g_{\lambda,\mu}$ is convex by Assumption 1. We may then write the program (1) as

$$\widehat{\beta} \in \arg\min_{g_{\lambda,\mu}(\beta)\le R,\ \beta\in\Omega} \Big\{ \underbrace{\Big(\mathcal{L}_n(\beta) - \frac{\mu}{2}\|\beta\|_2^2\Big)}_{\bar{\mathcal{L}}_n} + \lambda g_{\lambda,\mu}(\beta)\Big\}. \tag{30}$$

In this way, the objective function decomposes nicely into a sum of a differentiable but nonconvex function and a possibly nonsmooth but convex penalty. Applied to the representation (30) of the objective function, the composite gradient descent procedure of Nesterov (2007) produces a sequence of iterates $\{\beta^t\}_{t=0}^\infty$ via the updates

$$\beta^{t+1} \in \arg\min_{g_{\lambda,\mu}(\beta)\le R,\ \beta\in\Omega} \left\{\frac{1}{2}\left\|\beta - \left(\beta^t - \frac{\nabla\bar{\mathcal{L}}_n(\beta^t)}{\eta}\right)\right\|_2^2 + \frac{\lambda}{\eta} g_{\lambda,\mu}(\beta)\right\}, \tag{31}$$

where $\frac{1}{\eta}$ is the step size. As discussed in Section 4.2, these updates may be computed in a relatively straightforward manner.

4.1 Fast Global Convergence

The main result of this section is to establish that the algorithm defined by the iterates (31) converges very quickly to a δ-neighborhood of any global optimum, for all tolerances δ that are of the same order (or larger) than the statistical error.

We begin by setting up the notation and assumptions underlying our result. The *Taylor error* around the vector β_2 in the direction $\beta_1 - \beta_2$ is given by

$$\mathcal{T}(\beta_1, \beta_2) := \mathcal{L}_n(\beta_1) - \mathcal{L}_n(\beta_2) - \langle \nabla\mathcal{L}_n(\beta_2),\, \beta_1 - \beta_2\rangle. \tag{32}$$

We analogously define the Taylor error $\bar{\mathcal{T}}$ for the modified loss function $\bar{\mathcal{L}}_n$, and note that

$$\bar{\mathcal{T}}(\beta_1, \beta_2) = \mathcal{T}(\beta_1, \beta_2) - \frac{\mu}{2}\|\beta_1 - \beta_2\|_2^2. \tag{33}$$

For all vectors $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$, we require the following form of restricted strong convexity:

$$\mathcal{T}(\beta_1,\beta_2) \ge \begin{cases} \alpha_1\|\beta_1-\beta_2\|_2^2 - \tau_1\frac{\log p}{n}\|\beta_1-\beta_2\|_1^2, & \forall \|\beta_1-\beta_2\|_2 \le 3, \quad \text{(34a)} \\ \alpha_2\|\beta_1-\beta_2\|_2 - \tau_2\sqrt{\frac{\log p}{n}}\|\beta_1-\beta_2\|_1, & \forall \|\beta_1-\beta_2\|_2 \ge 3. \quad \text{(34b)} \end{cases}$$

The conditions (34) are similar but not identical to the earlier RSC conditions (4). The main difference is that we now require the Taylor difference to be bounded below uniformly over $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$, as opposed to for a fixed $\beta_2 = \beta^*$. In addition, we assume an analogous upper bound on the Taylor series error:

$$\mathcal{T}(\beta_1, \beta_2) \le \alpha_3 \|\beta_1 - \beta_2\|_2^2 + \tau_3 \frac{\log p}{n} \|\beta_1 - \beta_2\|_1^2, \qquad \text{for all } \beta_1, \beta_2 \in \Omega, \tag{35}$$

a condition referred to as *restricted smoothness* in past work (Agarwal et al., 2012). Throughout this section, we assume $2\alpha_i > \mu$ for all i, where μ is the coefficient ensuring the convexity of the function $g_{\lambda,\mu}$ from (29). Furthermore, we define $\alpha = \min\{\alpha_1, \alpha_2\}$ and $\tau = \max\{\tau_1, \tau_2, \tau_3\}$.

The following theorem applies to any population loss function $\mathcal{L}$ for which the population minimizer β^* is k-sparse and $\|\beta^*\|_2 \le 1$. Similar results could be derived for general $\|\beta^*\|_2$, with the radius of the RSC condition (34a) replaced by $3\|\beta^*\|_2$ and Lemma 2 in Section 4.3 adjusted appropriately, but we only include the analysis for $\|\beta^*\|_2 \le 1$ in order to simplify our exposition. We also assume the scaling $n > Ck \log p$, for a constant C depending on the α_i's and τ_i's. Note that this scaling is reasonable, since no estimator of a k-sparse vector in p dimensions can have low ℓ_2-error unless the condition holds (see Raskutti et al., 2011 for minimax rates). We show that the composite gradient updates (31) exhibit a type of *globally geometric convergence* in terms of the quantity

$$\kappa := \frac{1 - \frac{2\alpha - \mu}{8\eta} + \varphi(n,p,k)}{1 - \varphi(n,p,k)}, \qquad \text{where} \quad \varphi(n,p,k) := \frac{c\tau k \frac{\log p}{n}}{2\alpha - \mu}. \tag{36}$$

Under the stated scaling on the sample size, we are guaranteed that $\kappa \in (0,1)$, so it is a *contraction factor*. Roughly speaking, we show that the squared optimization error will fall below δ^2 within $T \asymp \frac{\log(1/\delta^2)}{\log(1/\kappa)}$ iterations. More precisely, our theorem guarantees δ-accuracy for all iterations larger than

$$T^*(\delta) := \frac{2 \log\left(\frac{\phi(\beta^0) - \phi(\widehat{\beta})}{\delta^2}\right)}{\log(1/\kappa)} + \left(1 + \frac{\log 2}{\log(1/\kappa)}\right) \log\log\left(\frac{\lambda R L}{\delta^2}\right), \tag{37}$$

where $\phi(\beta) := \mathcal{L}_n(\beta) + \rho_\lambda(\beta)$ denotes the composite objective function. As clarified in the theorem statement, the squared tolerance δ^2 is not allowed to be arbitrarily small, which would contradict the fact that the composite gradient method may converge to a stationary point. However, our theory allows δ^2 to be of the same order as the squared *statistical error* $\epsilon_{\text{stat}}^2 = \|\widehat{\beta} - \beta^*\|_2^2$, the distance between a fixed global optimum and the target parameter β^*. From a statistical perspective, there is no point in optimizing beyond this tolerance.

With this setup, we now turn to a precise statement of our main optimization-theoretic result. As with Theorems 1 and 2, the statement of Theorem 3 is entirely deterministic.

Theorem 3 *Suppose the empirical loss $\mathcal{L}_n$ satisfies the RSC/RSM conditions* (34) *and* (35), *and suppose the regularizer ρ_λ satisfies Assumption 1. Suppose $\widehat{\beta}$ is any global minimum of* *the program* (30), *with regularization parameters chosen such that*

$$\frac{8}{L} \cdot \max\left\{ \|\nabla \mathcal{L}_n(\beta^*)\|_\infty,\; c'\tau\sqrt{\frac{\log p}{n}} \right\} \le \lambda \le \frac{c''\,\alpha}{RL}.$$

Suppose $\mu < 2\alpha$*. Then for any step size parameter* $\eta \ge \max\{2\alpha_3 - \mu,\, \mu\}$ *and tolerance* $\delta^2 \ge \frac{c\epsilon_{stat}^2}{1-\kappa} \cdot \frac{k\log p}{n}$*, we have*

$$\|\beta^t - \widehat{\beta}\|_2^2 \;\le\; \frac{4}{2\alpha - \mu}\left(\delta^2 + \frac{\delta^4}{\tau} + c\tau\frac{k\log p}{n}\epsilon_{stat}^2\right), \qquad \forall t \ge T^*(\delta). \tag{38}$$

Remark: Note that for the optimal choice of tolerance parameter $\delta \asymp \frac{k\log p}{n}\epsilon_{\text{stat}}$, the error bound appearing in (38) takes the form $\frac{c\epsilon_{\text{stat}}^2}{2\alpha-\mu} \cdot \frac{k\log p}{n}$, meaning that successive iterates of the composite gradient descent algorithm are guaranteed to converge to a region within statistical accuracy of the true global optimum $\widehat{\beta}$. Concretely, if the sample size satisfies $n \gtrsim Ck\log p$ and the regularization parameters are chosen appropriately, Theorem 1 guarantees that $\epsilon_{\text{stat}} = \mathcal{O}\left(\sqrt{\frac{k\log p}{n}}\right)$ with high probability. Combined with Theorem 3, we then conclude that

$$\max\left\{\|\beta^t - \widehat{\beta}\|_2,\; \|\beta^t - \beta^*\|_2\right\} = \mathcal{O}\left(\sqrt{\frac{k\log p}{n}}\right),$$

for all iterations $t \ge T(\epsilon_{\text{stat}})$.

As would be expected, the (restricted) curvature α of the loss function and nonconvexity parameter μ of the penalty function enter into the bound via the denominator $2\alpha - \mu$. Indeed, the bound is tighter when the loss function possesses more curvature or the penalty function is closer to being convex, agreeing with intuition. Similar to our discussion in the remark following Theorem 2, the requirement $\mu < 2\alpha$ is certainly necessary for our proof technique, but it is possible that composite gradient descent still produces good results when this condition is violated. See Section 5 for simulations in scenarios involving mild and severe violations of this condition.

Finally, note that the parameter η must be sufficiently large (or equivalently, the step size must be sufficiently small) in order for the composite gradient descent algorithm to be well-behaved. See Nesterov (2007) for a discussion of how the step size may be chosen via an iterative search when the problem parameters are unknown.

In the case of corrected linear regression (Corollary 1), Lemma 13 of Loh and Wainwright (2012) establishes the RSC/RSM conditions for various statistical models. The following proposition shows that the conditions (34) and (35) hold in GLMs when the x_i's are drawn i.i.d. from a zero-mean sub-Gaussian distribution with parameter σ_x^2 and covariance matrix $\Sigma = \text{cov}(x_i)$. As usual, we assume a sample size $n \ge c\,k\log p$, for a sufficiently large constant $c > 0$. Recall the definition of the Taylor error $\mathcal{T}(\beta_1, \beta_2)$ from (32).

Proposition 1 *[RSC/RSM conditions for generalized linear models] There exists a constant* $\alpha_\ell > 0$*, depending only on the GLM and the parameters* (σ_x^2, Σ)*, such that for all* *vectors* $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$, *we have*

$$\mathcal{T}(\beta_1, \beta_2) \geq \begin{cases} \frac{\alpha_\ell}{2}\|\Delta\|_2^2 - \frac{c^2\sigma_x^2}{2\alpha_\ell}\frac{\log p}{n}\|\Delta\|_1^2, & \text{for all } \|\beta_1 - \beta_2\|_2 \leq 3, \quad (39a) \\ \frac{3\alpha_\ell}{2}\|\Delta\|_2 - 3c\sigma_x\sqrt{\frac{\log p}{n}}\|\Delta\|_1, & \text{for all } \|\beta_1 - \beta_2\|_2 \geq 3, \quad (39b) \end{cases}$$

with probability at least $1 - c_1 \exp(-c_2 n)$. *With the bound* $\|\psi''\|_\infty \leq \alpha_u$, *we also have*

$$\mathcal{T}(\beta_1, \beta_2) \leq \alpha_u \lambda_{max}(\Sigma) \left(\frac{3}{2}\|\Delta\|_2^2 + \frac{\log p}{n}\|\Delta\|_1^2\right), \qquad \text{for all } \beta_1, \beta_2 \in \mathbb{R}^p, \tag{40}$$

with probability at least $1 - c_1 \exp(-c_2 n)$.

For the proof of Proposition 1, see Appendix D.

4.2 Form of Updates

In this section, we discuss how the updates (31) are readily computable in many cases. We begin with the case $\Omega = \mathbb{R}^p$, so we have no additional constraints apart from $g_{\lambda,\mu}(\beta) \leq R$. In this case, given iterate β^t, the next iterate β^{t+1} may be obtained via the following three-step procedure:

(1) First optimize the unconstrained program

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{\frac{1}{2}\left\|\beta - \left(\beta^t - \frac{\nabla \bar{\mathcal{L}}_n(\beta^t)}{\eta}\right)\right\|_2^2 + \frac{\lambda}{\eta} \cdot g_{\lambda,\mu}(\beta)\right\}. \tag{41}$$

(2) If $g_{\lambda,\mu}(\widehat{\beta}) \leq R$, define $\beta^{t+1} = \widehat{\beta}$.

(3) Otherwise, if $g_{\lambda,\mu}(\widehat{\beta}) > R$, optimize the constrained program

$$\beta^{t+1} \in \arg\min_{g_{\lambda,\mu}(\beta) \leq R} \left\{\frac{1}{2}\left\|\beta - \left(\beta^t - \frac{\nabla \bar{\mathcal{L}}_n(\beta^t)}{\eta}\right)\right\|_2^2\right\}. \tag{42}$$

We derive the correctness of this procedure in Appendix C.1. For many nonconvex regularizers ρ_λ of interest, the unconstrained program (41) has a convenient closed-form solution: For the SCAD penalty (2), the program (41) has simple closed-form solution given by

$$\widehat{\beta}_{\text{SCAD}} = \begin{cases} 0 & \text{if } 0 \leq |z| \leq \nu\lambda, \\ z - \text{sign}(z) \cdot \nu\lambda & \text{if } \nu\lambda \leq |z| \leq (\nu + 1)\lambda, \\ \frac{z - \text{sign}(z) \cdot \frac{a\nu\lambda}{a-1}}{1 - \frac{\nu}{a-1}} & \text{if } (\nu + 1)\lambda \leq |z| \leq a\lambda, \\ z & \text{if } |z| \geq a\lambda. \end{cases} \tag{43}$$

For the MCP (3), the optimum of the program (41) takes the form

$$\widehat{\beta}_{\text{MCP}} = \begin{cases} 0 & \text{if } 0 \leq |z| \leq \nu\lambda, \\ \frac{z - \text{sign}(z)\cdot \nu\lambda}{1-\nu/b} & \text{if } \nu\lambda \leq |z| \leq b\lambda, \\ z & \text{if } |z| \geq b\lambda. \end{cases} \tag{44}$$

In both (43) and (44), we have

$$z := \frac{1}{1+\mu/\eta}\left(\beta^t - \frac{\nabla\bar{\mathcal{L}}_n(\beta^t)}{\eta}\right), \qquad \text{and} \qquad \nu := \frac{1/\eta}{1+\mu/\eta},$$

and the operations are taken componentwise. See Appendix C.2 for the derivation of these closed-form updates.

More generally, when $\Omega \subsetneq \mathbb{R}^p$ (such as in the case of the graphical Lasso), the minimum in the program (31) must be taken over Ω, as well. Although the updates are not as simply stated, they still involve solving a convex optimization problem. Despite this more complicated form, however, our results from Section 4.1 on fast global convergence under restricted strong convexity and restricted smoothness assumptions carry over without modification, since they only require RSC/RSM conditions holding over a sufficiently small radius together with feasibility of β^*.

4.3 Proof of Theorem 3

We provide the outline of the proof here, with more technical results deferred to Appendix C. In broad terms, our proof is inspired by a result of Agarwal et al. (2012), but requires various modifications in order to be applied to the much larger family of nonconvex regularizers considered here.

Our first lemma shows that the optimization error $\beta^t - \widehat{\beta}$ lies in an approximate cone set:

Lemma 1 *Under the conditions of Theorem 3, suppose there exists a pair $(\bar{\eta}, T)$ such that*

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \bar{\eta}, \qquad \forall t \geq T. \tag{45}$$

Then for any iteration $t \geq T$, we have

$$\|\beta^t - \widehat{\beta}\|_1 \leq 8\sqrt{k}\|\beta^t - \widehat{\beta}\|_2 + 16\sqrt{k}\|\widehat{\beta} - \beta^*\|_2 + 2\cdot\min\left(\frac{2\bar{\eta}}{\lambda L}, R\right).$$

Our second lemma shows that as long as the composite gradient descent algorithm is initialized with a solution β^0 within a constant radius of a global optimum $\widehat{\beta}$, all successive iterates also lie within the same ball:

Lemma 2 *Under the conditions of Theorem 3, and with an initial vector β^0 such that $\|\beta^0 - \widehat{\beta}\|_2 \leq 3$, we have*

$$\|\beta^t - \widehat{\beta}\|_2 \leq 3, \qquad \textit{for all } t \geq 0. \tag{46}$$

In particular, suppose we initialize the composite gradient procedure with a vector β^0 such that $\|\beta^0\|_2 \le \frac{3}{2}$. Then by the triangle inequality,

$$\|\beta^0 - \widehat{\beta}\|_2 \le \|\beta^0\|_2 + \|\widehat{\beta} - \beta^*\|_2 + \|\beta^*\|_2 \le 3,$$

where we have assumed our scaling of n guarantees $\|\widehat{\beta} - \beta^*\|_2 \le 1/2$.

Finally, recalling our earlier definition (36) of κ, the third lemma combines the results of Lemmas 1 and 2 to establish a bound on the value of the objective function that decays exponentially with t:

Lemma 3 *Under the same conditions of Lemma 2, suppose in addition that* (45) *holds and* $\frac{32k\tau \log p}{n} \le \frac{2\alpha-\mu}{4}$. *Then for any* $t \ge T$*, we have*

$$\phi(\beta^t) - \phi(\widehat{\beta}) \le \kappa^{t-T}(\phi(\beta^T) - \phi(\widehat{\beta})) + \frac{\xi}{1-\kappa}(\epsilon^2 + \bar{\epsilon}^2),$$

where $\bar{\epsilon} := 8\sqrt{k}\epsilon_{stat}$, $\epsilon := 2 \cdot \min\left(\frac{2\bar{\eta}}{\lambda L}, R\right)$, *the quantities* κ *and* φ *are defined according to* (36), *and*

$$\xi := \frac{1}{1-\varphi(n,p,k)} \cdot \frac{\tau \log p}{n} \cdot \left(\frac{2\alpha-\mu}{4\eta} + 2\varphi(n,p,k) + 5\right). \tag{47}$$

The remainder of the proof follows an argument used in Agarwal et al. (2012), so we only provide a high-level sketch. We first prove the following inequality:

$$\phi(\beta^t) - \phi(\widehat{\beta}) \le \delta^2, \qquad \text{for all } t \ge T^*(\delta), \tag{48}$$

as follows. We divide the iterations $t \ge 0$ into a series of epochs $[T_\ell, T_{\ell+1})$ and define tolerances $\bar{\eta}_0 > \bar{\eta}_1 > \cdots$ such that

$$\phi(\beta^t) - \phi(\widehat{\beta}) \le \bar{\eta}_\ell, \qquad \forall t \ge T_\ell.$$

In the first iteration, we apply Lemma 3 with $\bar{\eta}_0 = \phi(\beta^0) - \phi(\widehat{\beta})$ to obtain

$$\phi(\beta^t) - \phi(\widehat{\beta}) \le \kappa^t \left(\phi(\beta^0) - \phi(\widehat{\beta})\right) + \frac{\xi}{1-\kappa}(4R^2 + \bar{\epsilon}^2), \qquad \forall t \ge 0.$$

Let $\bar{\eta}_1 := \frac{2\xi}{1-\kappa}(4R^2 + \bar{\epsilon}^2)$, and note that for $T_1 := \left\lceil \frac{\log(2\bar{\eta}_0/\bar{\eta}_1)}{\log(1/\kappa)} \right\rceil$, we have

$$\phi(\beta^t) - \phi(\widehat{\beta}) \le \bar{\eta}_1 \le \frac{4\xi}{1-\kappa} \max\{4R^2, \bar{\epsilon}^2\}, \qquad \text{for all } t \ge T_1.$$

For $\ell \ge 1$, we now define

$$\bar{\eta}_{\ell+1} := \frac{2\xi}{1-\kappa}(\epsilon_\ell^2 + \bar{\epsilon}^2), \quad \text{and} \quad T_{\ell+1} := \left\lceil \frac{\log(2\bar{\eta}_\ell/\bar{\eta}_{\ell+1})}{\log(1/\kappa)} \right\rceil + T_\ell,$$

where $\epsilon_\ell := 2\min\left\{\frac{\bar{\eta}_\ell}{\lambda L}, R\right\}$. From Lemma 3, we have

$$\phi(\beta^t) - \phi(\widehat{\beta}) \le \kappa^{t-T_\ell}\left(\phi(\beta^{T_\ell}) - \phi(\widehat{\beta})\right) + \frac{\xi}{1-\kappa}(\epsilon_\ell^2 + \bar{\epsilon}^2), \qquad \text{for all } t \ge T_\ell,$$

implying by our choice of $\{(\eta_\ell, T_\ell)\}_{\ell \ge 1}$ that

$$\phi(\beta^t) - \phi(\widehat{\beta}) \le \bar{\eta}_{\ell+1} \le \frac{4\xi}{1-\kappa}\max\{\epsilon_\ell^2, \bar{\epsilon}^2\}, \qquad \forall t \ge T_{\ell+1}.$$

Finally, we use the recursion

$$\bar{\eta}_{\ell+1} \le \frac{4\xi}{1-\kappa}\max\{\epsilon_\ell^2, \bar{\epsilon}^2\}, \qquad T_\ell \le \ell + \frac{\log(2^\ell \bar{\eta}_0/\bar{\eta}_\ell)}{\log(1/\kappa)}, \tag{49}$$

to establish the recursion

$$\bar{\eta}_{\ell+1} \le \frac{\bar{\eta}_\ell}{4^{2^{\ell-1}}}, \qquad \frac{\bar{\eta}_{\ell+1}}{\lambda L} \le \frac{R}{4^{2^\ell}}. \tag{50}$$

Inequality (48) then follows from computing the number of epochs and time steps necessary to obtain $\frac{\lambda R L}{4^{2^{\ell-1}}} \le \delta^2$. For the remaining steps used to obtain (50) from (49), we refer the reader to Agarwal et al. (2012).

Finally, by (85b) in the proof of Lemma 3 in Appendix C.5 and the relative scaling of (n, p, k), we have

$$\begin{aligned}\frac{2\alpha - \mu}{4}\|\beta^t - \widehat{\beta}\|_2^2 &\le \phi(\beta^t) - \phi(\widehat{\beta}) + 2\tau\frac{\log p}{n}\left(\frac{2\delta^2}{\lambda L} + \bar{\epsilon}\right)^2 \\ &\le \delta^2 + 2\tau\frac{\log p}{n}\left(\frac{2\delta^2}{\lambda L} + \bar{\epsilon}\right)^2,\end{aligned}$$

where we have set $\epsilon = \frac{2\delta^2}{\lambda L}$. Rearranging and performing some algebra with our choice of λ gives the ℓ_2-bound.

5. Simulations

In this section, we report the results of simulations we performed to validate our theoretical results. In particular, we present results for two versions of the loss function $\mathcal{L}_n$, corresponding to linear and logistic regression, and three penalty functions, namely the ℓ_1-norm (Lasso), the SCAD penalty, and the MCP, as detailed in Section 2.2. In all cases, we chose regularization parameters $R = \frac{1.1}{\lambda} \cdot \rho_\lambda(\beta^*)$, to ensure feasibility of β^*, and $\lambda = \sqrt{\frac{\log p}{n}}$; in practical applications where β^* is unknown, we would need to tune λ and R using a method such as cross-validation.

Linear regression: In the case of linear regression, we simulated covariates corrupted by additive noise according to the mechanism described in Section 3.2, giving the estimator

$$\widehat{\beta} \in \arg\min_{g_{\lambda,\mu}(\beta) \le R}\left\{\frac{1}{2}\beta^T\left(\frac{Z^TZ}{n} - \Sigma_w\right)\beta - \frac{y^TZ}{n}\beta + \rho_\lambda(\beta)\right\}. \tag{51}$$

We generated i.i.d. samples $x_i \sim N(0, I)$ and set $\Sigma_w = (0.2)^2 I$, and generated additive noise $\epsilon_i \sim N(0, (0.1)^2)$.

Logistic regression: In the case of logistic regression, we also generated i.i.d. samples $x_i \sim N(0, I)$. Since $\psi(t) = \log(1 + \exp(t))$, the program (15) becomes

$$\widehat{\beta} \in \arg \min_{g_{\lambda,\mu}(\beta) \le R} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{\log(1 + \exp(\langle \beta, x_i \rangle) - y_i \langle \beta, x_i \rangle\} + \rho_\lambda(\beta) \right\}. \tag{52}$$

We optimized the programs (51) and (52) using the composite gradient updates (31). In order to compute the updates, we used the three-step procedure described in Section 4.2, together with the updates for SCAD and MCP given by (43) and (44). Note that the updates for the Lasso penalty may be generated more simply and efficiently as discussed in Agarwal et al. (2012).

Figure 2 shows the results of corrected linear regression with Lasso, SCAD, and MCP regularizers for three different problem sizes p. In each case, β^* is a k-sparse vector with $k = \lfloor \sqrt{p} \rfloor$, where the nonzero entries were generated from a normal distribution and the vector was then rescaled so that $\|\beta^*\|_2 = 1$. As predicted by Theorem 1, the three curves corresponding to the same penalty function stack up when the estimation error $\|\widehat{\beta} - \beta^*\|_2$ is plotted against the rescaled sample size $\frac{n}{k \log p}$, and the ℓ_2-error decreases to zero as the number of samples increases, showing that the estimators (51) and (52) are statistically consistent. The Lasso, SCAD, and MCP regularizers are depicted by solid, dotted, and dashed lines, respectively. We chose the parameter $a = 3.7$ for the SCAD penalty, suggested by Fan and Li (2001) to be "optimal" based on cross-validated empirical studies, and chose $b = 3.5$ for the MCP. Each point represents an average over 20 trials.

The simulations in Figure 3 depict the optimization-theoretic conclusions of Theorem 3. Each panel shows two different families of curves, depicting the statistical error $\log(\|\widehat{\beta} - \beta^*\|_2)$ in red and the optimization error $\log(\|\beta^t - \widehat{\beta}\|_2)$ in blue. Here, the vertical axis measures the ℓ_2-error on a logarithmic scale, while the horizontal axis tracks the iteration number. Within each panel, the blue curves were obtained by running the composite gradient descent algorithm from 10 different initial starting points chosen at random, and the optimization error is measured with respect to a stationary point obtained from an earlier run of the composite gradient descent algorithm in place of $\widehat{\beta}$, since a global optimum is unknown. The statistical error is similarly displayed as the distance between β^* and the stationary points computed from successive runs of composite gradient descent. In all cases, we used the parameter settings $p = 128$, $k = \lfloor \sqrt{p} \rfloor$, and $n = \lfloor 20k \log p \rfloor$. As predicted by our theory, the optimization error decreases at a linear rate (on the log scale) until it falls to the level of statistical error. Furthermore, it is interesting to compare the plots in panels (c) and (d), which provide simulation results for two different values of the SCAD parameter a. We see that the choice $a = 3.7$ leads to a tighter cluster of optimization trajectories, providing further evidence that this setting suggested by Fan and Li (2001) is in some sense optimal.

Figure 4 provides analogous results to Figure 3 in the case of logistic regression, using $p = 64, k = \lfloor \sqrt{p} \rfloor$, and $n = \lfloor 20k \log p \rfloor$. The plot shows solution trajectories for 20 different

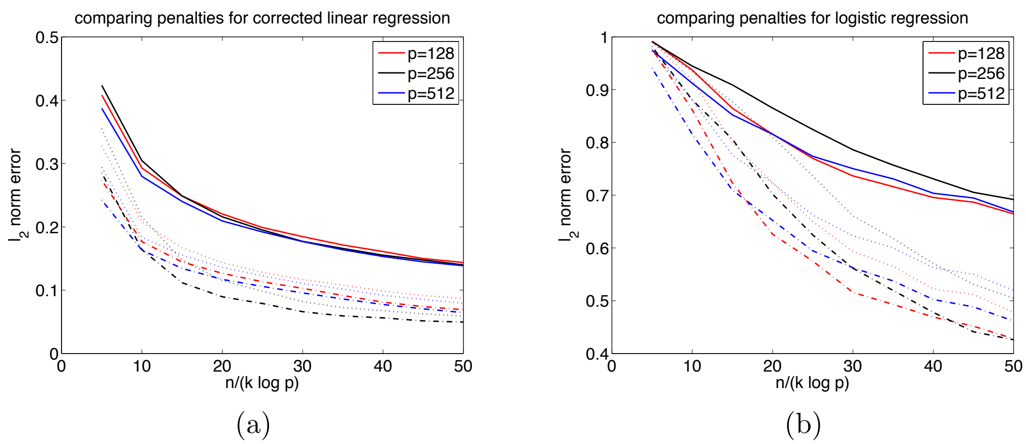


Figure 2: Plots showing statistical consistency of linear and logistic regression with Lasso, SCAD, and MCP regularizers, and with sparsity level $k = \lfloor \sqrt{p} \rfloor$. Panel (a) shows results for corrected linear regression, where covariates are subject to additive noise with $SNR = 5$. Panel (b) shows similar results for logistic regression. Each point represents an average over 20 trials. In both cases, the estimation error $\|\widehat{\beta} - \beta^*\|_2$ is plotted against the rescaled sample size $\frac{n}{k \log p}$. Lasso, SCAD, and MCP results are represented by solid, dotted, and dashed lines, respectively. As predicted by Theorem 1 and Corollaries 1 and 2, the curves for each of the three types stack up for different problem sizes p, and the error decreases to zero as the number of samples increases, showing that our methods are statistically consistent.

initializations of composite gradient descent. Again, we see that the log optimization error decreases at a linear rate up to the level of statistical error, as predicted by Theorem 3. Furthermore, the Lasso penalty yields a unique global optimum $\widehat{\beta}$, since the program (52) is convex, as we observe in panel (a). In contrast, the nonconvex program based on the SCAD penalty produces multiple local optima, whereas the MCP yields a relatively large number of local optima. Note that empirically, all local optima appear to lie within the small ball around β^* defined in Theorem 1. However, if we use $\lambda_{\min}(\nabla^2 \mathcal{L}_n(\beta^*))$ as a surrogate for α_1, we see that $2\alpha_1 < \mu$ in the case of the SCAD or MCP regularizers, which is not covered by our theory.

Finally, Figure 5 explores the behavior of our algorithm when the condition $\mu < 2\alpha_1$ from Theorem 1 is significantly violated. We generated i.i.d. samples $x_i \sim N(0, \Sigma)$, with Σ taken to be a Toeplitz matrix with entries $\Sigma_{ij} = \zeta^{|i-j|}$, for some parameter $\zeta \in [0, 1)$, so that $\lambda_{\min}(\Sigma) \geq (1 - \zeta)^2$. We chose $\zeta \in \{0.5, 0.9\}$, resulting in $\alpha_1 \approx \{0.25, 0.01\}$. The problem parameters were chosen to be $p = 512, k = \lfloor \sqrt{p} \rfloor$, and $n = \lfloor 10k \log p \rfloor$. Panel (a) shows the expected good behavior of ℓ_1-regularization, even for $\alpha_1 = 0.01$; although convergence is slow and the overall statistical error is greater than for $\Sigma = I$ (cf. Figure 3(a)), composite gradient descent still converges at a linear rate. Panel (b) shows that for SCAD parameter $a = 2.5$ (corresponding to $\mu \approx 0.67$), local optima still seem to be well-behaved even for

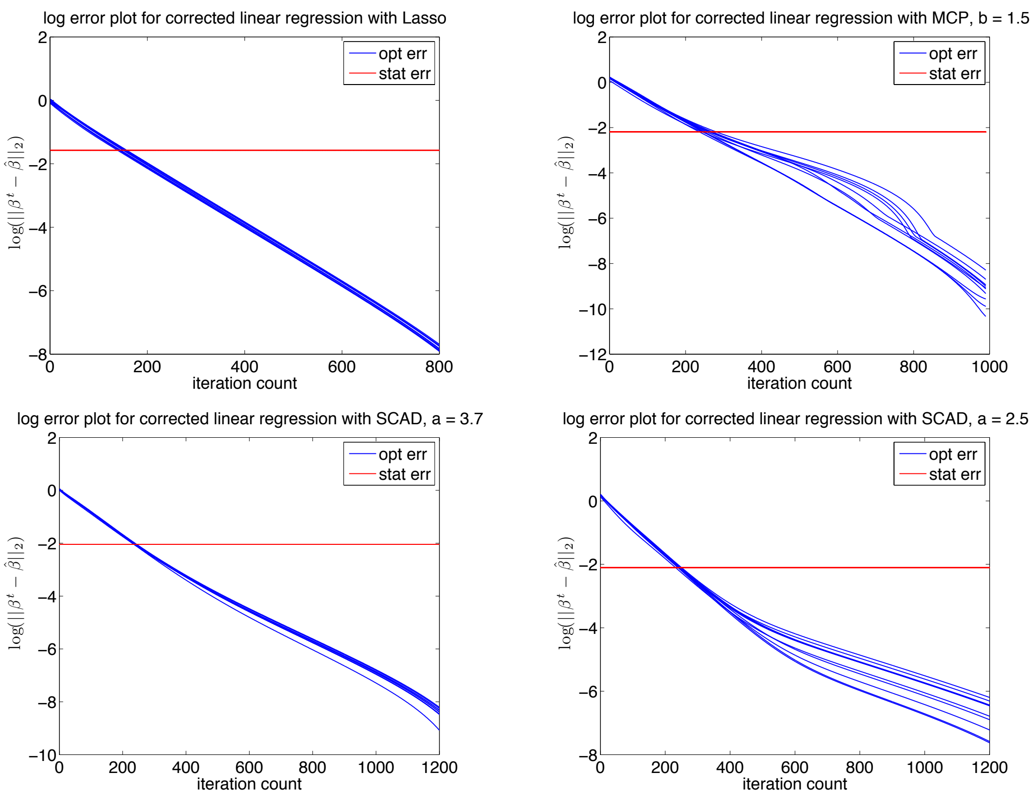
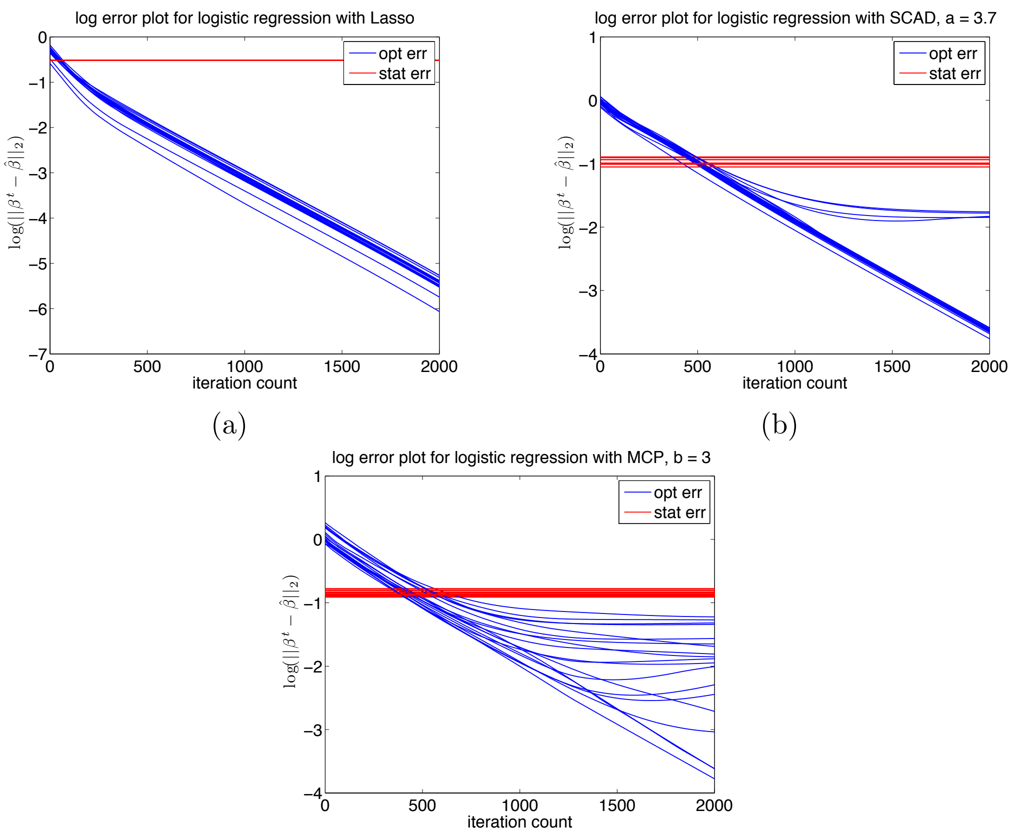


Figure 3: Plots illustrating linear rates of convergence on a log scale for corrected linear regression with Lasso, MCP, and SCAD regularizers, with $p = 128$, $k = \lfloor \sqrt{p} \rfloor$, and $n = \lfloor 20k \log p \rfloor$, where covariates are corrupted by additive noise with $SNR = 5$. Red lines depict statistical error $\log\left(\|\widehat{\beta} - \beta^*\|_2\right)$ and blue lines depict optimization error $\log\left(\|\beta^t - \widehat{\beta}\|_2\right)$. As predicted by Theorem 3, the optimization error decreases linearly when plotted against the iteration number on a log scale, up to statistical accuracy. Each plot shows the solution trajectory for 10 different initializations of the composite gradient descent algorithm. Panels (a) and (b) show the results for Lasso and MCP regularizers, respectively; panels (c) and (d) show results for the SCAD penalty with two different parameter values. Note that the empirically optimal choice $a = 3.7$ proposed by Fan and Li (2001) generates solution paths that exhibit a smaller spread than the solution paths generated for a smaller setting of the parameter a.

$2\alpha_1 = 0.5 < \mu$. However, for much smaller values of α_1, the good behavior breaks down, as seen in panels (c) and (d). Note that in the latter two panels, the composite gradient descent algorithm does not appear to be converging, even as the iteration number increases. Comparing (c) and (d) also illustrates the interplay between the curvature parameter α_1 of $\mathcal{L}_n$ and the nonconvexity parameter μ of ρ_λ. Indeed, the plot in panel (d) is slightly "better" than the plot in panel (c), in the sense that initial iterates at least demonstrate



(a) (b) (c)

Figure 4: Plots that demonstrate linear rates of convergence on a log scale for logistic regression with $p = 64, k = \sqrt{p}$, and $n = \lfloor 20k \log p \rfloor$. Red lines depict statistical error $\log\left(\|\widehat{\beta} - \beta^*\|_2\right)$ and blue lines depict optimization error $\log\left(\|\beta^t - \widehat{\beta}\|_2\right)$. (a) Lasso penalty. (b) SCAD penalty. (c) MCP. As predicted by Theorem 3, the optimization error decreases linearly when plotted against the iteration number on a log scale, up to statistical accuracy. Each plot shows the solution trajectory for 20 different initializations of the composite gradient descent algorithm. Multiple local optima emerge in panels (b) and (c), due to nonconvex regularizers.

some pattern of convergence. This could be attributed to the fact that the SCAD parameter is larger, corresponding to a smaller value of μ.

6. Discussion

We have analyzed theoretical properties of local optima of regularized M-estimators, where both the loss and penalty function are allowed to be nonconvex. Our results are the first to establish that *all stationary points* of such nonconvex problems are close to the truth, implying that any optimization method guaranteed to converge to a stationary point will

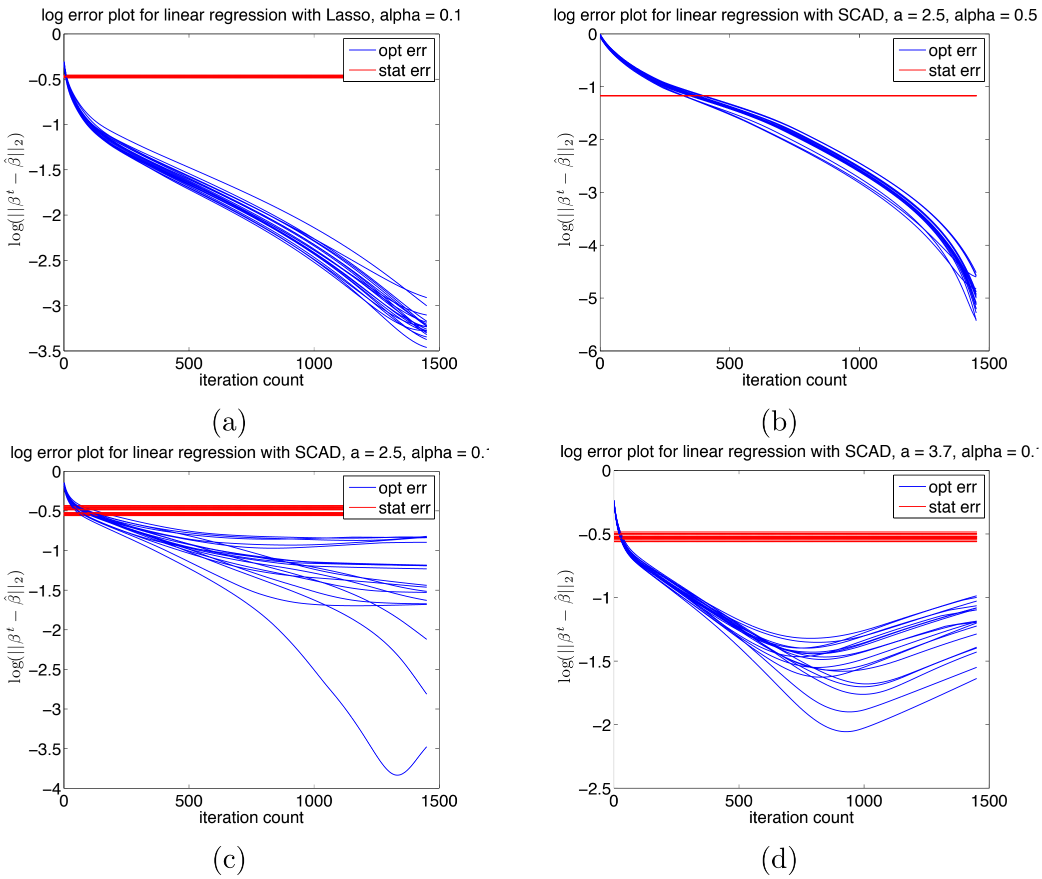


Figure 5: Plots showing breakdown points as a function of the curvature parameter α_1 of the loss function and the nonconvexity parameter μ of the penalty function. The loss comes from ordinary least squares linear regression, where covariates are fully-observed and sampled from a Gaussian distribution with covariance equal to a Toeplitz matrix. Panel (a) depicts the good behavior of Lasso-based linear regression. Panel (b) shows that local optima may still be well-behaved even when $2\alpha_1 < \mu$, although this situation is not covered by our theory. Panels (c) and (d) show that the good behavior nonetheless disintegrates for very small values of α_1 when the regularizer is nonconvex.

provide statistically consistent solutions. We show concretely that a variant of composite gradient descent may be used to obtain near-global optima in linear time, and verify our theoretical results with simulations.

Future directions of research include further generalizing our statistical consistency results to other nonconvex regularizers not covered by our present theory, such as bridge penalties or regularizers that do not decompose across coordinates. In addition, it would be interesting to expand our theory to nonsmooth loss functions such as the hinge loss. For both nonsmooth losses and nonsmooth penalties (including capped-ℓ_1), it remains an open question whether a modified version of composite gradient descent may be used to obtain near-global optima in polynomial time. Finally, it would be useful to develop a general method for establishing RSC and RSM conditions, beyond the specialized methods used for studying GLMs in this paper.

Acknowledgments

The work of PL was supported from a Hertz Foundation Fellowship and an NSF Graduate Research Fellowship. In addition, PL acknowledges support from a PD Award, and MJW from a DY Award. MJW and PL were also partially supported by NSF grant CIF-31712-23800. The authors thank the associate editors and anonymous reviewers for their helpful suggestions that improved the manuscript.

Appendix A. Properties of Regularizers

In this section, we establish properties of some nonconvex regularizers covered by our theory (Appendix A.1) and verify that specific regularizers satisfy Assumption 1 (Appendix A.2). The properties given in Appendix A.1 are used in the proof of Theorem 1.

A.1 General Properties

We begin with some general properties of regularizers that satisfy Assumption 1.

Lemma 4

(a) Under conditions (i)–(ii) of Assumption 1, conditions (iii) and (iv) together imply that ρ_λ is λL-Lipschitz as a function of t. In particular, all subgradients and derivatives of ρ_λ are bounded in magnitude by λL.

(b) Under the conditions of Assumption 1, we have

$$\lambda L \|\beta\|_1 \le \rho_\lambda(\beta) + \frac{\mu}{2}\|\beta\|_2^2, \qquad \forall \beta \in \mathbb{R}^p. \tag{53}$$

Proof (a): Suppose $0 \le t_1 \le t_2$. Then

$$\frac{\rho_\lambda(t_2) - \rho_\lambda(t_1)}{t_2 - t_1} \le \frac{\rho_\lambda(t_1)}{t_1},$$

by condition (iii). Applying (iii) once more, we have

$$\frac{\rho_\lambda(t_1)}{t_1} \le \lim_{t \to 0^+} \frac{\rho_\lambda(t)}{t} = \lambda L,$$

where the last equality comes from condition (iv). Hence,

$$0 \le \rho_\lambda(t_2) - \rho_\lambda(t_1) \le \lambda L(t_2 - t_1).$$

A similar argument applies to the cases when one (or both) of t_1 and t_2 are negative.

(b): Clearly, it suffices to verify the inequality for the scalar case:

$$\lambda L t \le \rho_\lambda(t) + \frac{\mu t^2}{2}, \qquad \forall t \in \mathbb{R}.$$

The inequality is trivial for $t = 0$. For $t > 0$, the convexity of the right-hand expression implies that for any $s \in (0, t)$, we have

$$\left(\rho_\lambda(t) + \frac{\mu t^2}{2}\right) - \left(\rho_\lambda(0) + \frac{\mu \cdot 0^2}{2}\right) \ge (t - 0) \cdot \big(\rho_\lambda'(s) + \mu s\big).$$

Taking a limit as $s \to 0^+$ then yields the desired inequality. The case $t < 0$ follows by symmetry. ∎

Lemma 5 *Suppose ρ_λ satisfies the conditions of Assumption 1. Let $v \in \mathbb{R}^p$, and let A denote the index set of the k largest elements of v in magnitude. Suppose $\xi > 0$ is such that $\xi\rho_\lambda(v_A) - \rho_\lambda(v_{A^c}) \ge 0$. Then*

$$\xi\rho_\lambda(v_A) - \rho_\lambda(v_{A^c}) \le \lambda L(\xi\|v_A\|_1 - \|v_{A^c}\|_1). \tag{54}$$

Moreover, if $\beta^* \in \mathbb{R}^p$ is k-sparse, then for an vector $\beta \in \mathbb{R}^p$ such that $\xi\rho_\lambda(\beta^*) - \rho_\lambda(\beta) > 0$ and $\xi \ge 1$, we have

$$\xi\rho_\lambda(\beta^*) - \rho_\lambda(\beta) \le \lambda L\big(\xi\|\nu_A\|_1 - \|\nu_{A^c}\|_1\big), \tag{55}$$

where $\nu := \beta - \beta^*$ and A is the index set of the k largest elements of ν in magnitude.

Proof We first establish (54). Define $f(t) := \frac{t}{\rho_\lambda(t)}$ for $t > 0$. By our assumptions on ρ_λ, the function f is nondecreasing in $|t|$, so

$$\|v_{A^c}\|_1 = \sum_{j \in A^c} \rho_\lambda(v_j) \cdot f(|v_j|) \le \sum_{j \in A^c} \rho_\lambda(v_j) \cdot f(\|v_{A^c}\|_\infty) = \rho_\lambda(v_{A^c}) \cdot f\left(\|v_{A^c}\|_\infty\right). \tag{56}$$

Again using the nondecreasing property of f, we have

$$\rho_\lambda(v_A) \cdot f(\|v_{A^c}\|_\infty) = \sum_{j \in A} \rho_\lambda(v_j) \cdot f(\|v_{A^c}\|_\infty) \le \sum_{j \in A} \rho_\lambda(v_j) \cdot f(|v_j|) = \|v_A\|_1. \tag{57}$$

Note that for $t > 0$, we have

$$f(t) \ge \lim_{s \to 0^+} f(s) = \lim_{s \to 0^+} \frac{s - 0}{\rho_\lambda(s) - \rho_\lambda(0)} = \frac{1}{\lambda L},$$

where the last equality follows from condition (iv) of Assumption 1. Combining this result with (56) and (57) yields

$$0 \le \xi\rho_\lambda(v_A) - \rho_\lambda(v_{A^c}) \le \frac{1}{f(\|v_{A^c}\|_\infty)} \cdot \big(\xi\|v_A\|_1 - \|v_{A^c}\|_1\big) \le \lambda L\big(\xi\|v_A\|_1 - \|v_{A^c}\|_1\big),$$

as claimed.

We now turn to the proof of the bound (55). Letting $S := \text{supp}(\beta^*)$ denote the support of β^*, the triangle inequality and subadditivity of ρ (see the remark following Assumption 1; cf. Lemma 1 of Chen and Gu, 2014) imply that

$$\begin{aligned} 0 \le \xi\rho_\lambda(\beta^*) - \rho_\lambda(\beta) &= \xi\rho_\lambda(\beta^*_S) - \rho_\lambda(\beta_S) - \rho_\lambda(\beta_{S^c}) \\ &\le \xi\rho_\lambda(\nu_S) - \rho_\lambda(\beta_{S^c}) \\ &= \xi\rho_\lambda(\nu_S) - \rho_\lambda(\nu_{S^c}) \\ &\le \xi\rho_\lambda(\nu_A) - \rho_\lambda(\nu_{A^c}) \\ &\le \lambda L\big(\xi\|\nu_A\|_1 - \|\nu_{A^c}\|_1\big), \end{aligned}$$

thereby completing the proof. ■

A.2 Verification for Specific Regularizers

We now verify that Assumption 1 is satisfied by the SCAD and MCP regularizers. (The properties are trivial to verify for the Lasso penalty.)

Lemma 6 *The SCAD regularizer* (2) *with parameter a satisfies the conditions of Assumption 1 with $L = 1$ and $\mu = \frac{1}{a-1}$.*

Proof Conditions (i)–(iii) were already verified in Zhang and Zhang (2012). Furthermore, we may easily compute the derivative of the SCAD regularizer to be

$$\frac{\partial}{\partial t}\rho_\lambda(t) = \text{sign}(t) \cdot \left(\lambda \cdot \mathbb{I}\{|t| \le \lambda\} + \frac{(a\lambda - |t|)_+}{a-1} \cdot \mathbb{I}\{|t| > \lambda\}\right), \qquad t \neq 0, \tag{58}$$

and any point in the interval $[-\lambda, \lambda]$ is a valid subgradient at $t = 0$, so condition (iv) is satisfied for any $L \ge 1$. Furthermore, we have $\frac{\partial^2}{\partial t^2}\rho_\lambda(t) \ge \frac{-1}{a-1}$, so $\rho_{\lambda,\mu}$ is convex whenever $\mu \ge \frac{1}{a-1}$, giving condition (v). ■

Lemma 7 *The MCP regularizer* (3) *with parameter b satisfies the conditions of Assumption 1 with1 $L = 1$ and $\mu = \frac{1}{b}$.*

Proof Again, the conditions (i)–(iii) are already verified in Zhang and Zhang (2012). We may compute the derivative of the MCP regularizer to be

$$\frac{\partial}{\partial t}\rho_\lambda(t) = \lambda \cdot \text{sign}(t) \cdot \left(1 - \frac{|t|}{\lambda b}\right)_+, \qquad t \neq 0, \tag{59}$$

with subgradient $\lambda[-1, +1]$ at $t = 0$, so condition (iv) is again satisfied for any $L \ge 1$. Taking another derivative, we have $\frac{\partial^2}{\partial t^2}\rho_\lambda(t) \ge \frac{-1}{b}$, so condition (v) of Assumption 1 holds with $\mu = \frac{1}{b}$. ■

Appendix B. Proofs of Corollaries in Section 3

In this section, we provide proofs of the corollaries to Theorem 1 stated in Section 3. Throughout this section, we use the convenient shorthand notation

$$\mathcal{E}_n(\Delta) := \langle \nabla \mathcal{L}_n(\beta^* + \Delta) - \nabla \mathcal{L}_n(\beta^*), \, \Delta \rangle. \tag{60}$$

B.1 General Results for Verifying RSC

We begin with two lemmas that will be useful for establishing the RSC conditions (4) in the special case where $\mathcal{L}_n$ is convex. We assume throughout that $\|\Delta\|_1 \le 2R$, since β^* and $\beta^* + \Delta$ lie in the feasible set.

Lemma 8 *Suppose $\mathcal{L}_n$ is convex. If condition* (4a) *holds and $n \ge 4R^2 \tau_1^2 \log p$, then*

$$\mathcal{E}_n(\Delta) \ge \alpha_1 \|\Delta\|_2 - \sqrt{\frac{\log p}{n}} \|\Delta\|_1, \qquad \textit{for all } \|\Delta\|_2 \ge 1. \tag{61}$$

Proof Fix an arbitrary $\Delta \in \mathbb{R}^p$ with $\|\Delta\|_2 \ge 1$. Since $\mathcal{L}_n$ is convex, the function $f : [0,1] \to \mathbb{R}$ given by $f(t) := \mathcal{L}_n(\beta^* + t\Delta)$ is also convex, so $f'(1) - f'(0) \ge f'(t) - f'(0)$ for all $t \in [0,1]$. Computing the derivatives of f yields the inequality

$$\mathcal{E}_n(\Delta) \; = \; \langle \nabla \mathcal{L}_n(\beta^* + \Delta) - \nabla \mathcal{L}_n(\beta^*), \, \Delta \rangle \ge \frac{1}{t} \, \langle \nabla \mathcal{L}_n(\beta^* + t\Delta) - \nabla \mathcal{L}_n(\beta^*), \, t\Delta \rangle.$$

Taking $t = \frac{1}{\|\Delta\|_2} \in (0,1]$ and applying condition (4a) to the rescaled vector $\frac{\Delta}{\|\Delta\|_2}$ then yields

$$\begin{aligned}
\mathcal{E}_n(\Delta) &\ge \|\Delta\|_2 \left(\alpha_1 - \tau_1 \frac{\log p}{n} \frac{\|\Delta\|_1^2}{\|\Delta\|_2^2} \right) \\
&\ge \|\Delta\|_2 \left(\alpha_1 - \frac{2R\tau_1 \log p}{n} \frac{\|\Delta\|_1}{\|\Delta\|_2^2} \right) \\
&\ge \|\Delta\|_2 \left(\alpha_1 - \sqrt{\frac{\log p}{n}} \frac{\|\Delta\|_1}{\|\Delta\|_2} \right) \\
&= \alpha_1 \|\Delta\|_2 - \sqrt{\frac{\log p}{n}} \|\Delta\|_1,
\end{aligned}$$

where the third inequality uses the assumption on the relative scaling of (n, p) and the fact that $\|\Delta\|_2 \ge 1$. ■

On the other hand, if (4a) holds globally over $\Delta \in \mathbb{R}^p$, we obtain (4b) for free:

Lemma 9 *If inequality* (4a) *holds for all $\Delta \in \mathbb{R}^p$ and $n \ge 4R^2 \tau_1^2 \log p$, then* (4b) *holds, as well.*

Proof Suppose $\|\Delta\|_2 \ge 1$. Then

$$\alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log p}{n} \|\Delta\|_1^2 \ge \alpha_1 \|\Delta\|_2 - 2R\tau_1 \frac{\log p}{n} \|\Delta\|_1 \ge \alpha_1 \|\Delta\|_2 - \sqrt{\frac{\log p}{n}} \|\Delta\|_1,$$

again using the assumption on the scaling of (n, p). ■

B.2 Proof of Corollary 1

Note that $\mathcal{E}_n(\Delta) = \Delta^T \widehat{\Gamma} \Delta$, so in particular,

$$\mathcal{E}_n(\Delta) \geq \Delta^T \Sigma_x \Delta - |\Delta^T (\Sigma_x - \widehat{\Gamma}) \Delta|.$$

Applying Lemma 12 in Loh and Wainwright (2012) with $s = \frac{n}{\log p}$ to bound the second term, we have

$$\begin{aligned}
\mathcal{E}_n(\Delta) &\geq \lambda_{\min}(\Sigma_x) \|\Delta\|_2^2 - \left(\frac{\lambda_{\min}(\Sigma_x)}{2} \|\Delta\|_2^2 + \frac{c \log p}{n} \|\Delta\|_1^2 \right) \\
&= \frac{\lambda_{\min}(\Sigma_x)}{2} \|\Delta\|_2^2 - \frac{c \log p}{n} \|\Delta\|_1^2,
\end{aligned}$$

a bound which holds for all $\Delta \in \mathbb{R}^p$ with probability at least $1 - c_1 \exp(-c_2 n)$ whenever $n \gtrsim k \log p$. Then Lemma 9 in Appendix B.1 implies that the RSC condition (4b) holds. It remains to verify the validity of the specified choice of λ. We have

$$\begin{aligned}
\|\nabla \mathcal{L}_n(\beta^*)\|_\infty = \|\widehat{\Gamma} \beta^* - \widehat{\gamma}\|_\infty &= \|(\widehat{\gamma} - \Sigma_x \beta^*) + (\Sigma_x - \widehat{\Gamma}) \beta^*\|_\infty \\
&\leq \|(\widehat{\gamma} - \Sigma_x \beta^*)\|_\infty + \|(\Sigma_x - \widehat{\Gamma}) \beta^*\|_\infty.
\end{aligned}$$

As shown in previous work (Loh and Wainwright, 2012), both of these terms are upper-bounded by $c' \varphi \sqrt{\frac{\log p}{n}}$ with high probability. Consequently, the claim in the corollary follows by applying Theorem 1.

B.3 Proof of Corollary 2

In the case of GLMs, we have

$$\mathcal{E}_n(\Delta) = \frac{1}{n} \sum_{i=1}^n (\psi'(\langle x_i, \beta^* + \Delta \rangle) - \psi'(\langle x_i, \beta^* \rangle)) \, x_i^T \Delta.$$

Applying the mean value theorem, we find that

$$\mathcal{E}_n(\Delta) = \frac{1}{n} \sum_{i=1}^n \psi''(\langle x_i, \beta^* \rangle + t_i \langle x_i, \Delta \rangle) \left(\langle x_i, \Delta \rangle \right)^2,$$

where $t_i \in [0, 1]$. From (the proof of) Proposition 2 in Negahban et al. (2012), we then have

$$\mathcal{E}_n(\Delta) \geq \alpha_1 \|\Delta\|_2^2 - \tau_1 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 \|\Delta\|_2, \qquad \forall \|\Delta\|_2 \leq 1, \tag{62}$$

with probability at least $1 - c_1 \exp(-c_2 n)$, for an appropriate choice of α_1. Note that by the arithmetic mean-geometric mean inequality,

$$\tau_1 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 \|\Delta\|_2 \leq \frac{\alpha_1}{2} \|\Delta\|_2^2 + \frac{\tau_1^2}{2\alpha_1} \frac{\log p}{n} \|\Delta\|_1^2,$$

and consequently,

$$\mathcal{E}_n(\Delta) \geq \frac{\alpha_1}{2}\|\Delta\|_2^2 - \frac{\tau_1^2}{2\alpha_1}\frac{\log p}{n}\|\Delta\|_1^2,$$

which establishes (4a). Inequality (4b) then follows via Lemma 8 in Appendix B.1.

It remains to show that there are universal constants (c, c_1, c_2) such that

$$\mathbb{P}\left(\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \geq c\sqrt{\frac{\log p}{n}}\right) \leq c_1 \exp(-c_2 \log p). \tag{63}$$

For each $1 \leq i \leq n$ and $1 \leq j \leq p$, define the random variable $V_{ij} := (\psi'(x_i^T\beta^*) - y_i)x_{ij}$. Our goal is to bound $\max_{j=1,\dots,p} |\frac{1}{n}\sum_{i=1}^n V_{ij}|$. Note that

$$\mathbb{P}\left[\max_{j=1,\dots,p} \Big|\frac{1}{n}\sum_{i=1}^n V_{ij}\Big| \geq \delta\right] \leq \mathbb{P}[\mathcal{A}^c] + \mathbb{P}\left[\max_{j=1,\dots,p} \Big|\frac{1}{n}\sum_{i=1}^n V_{ij}\Big| \geq \delta \;\Big|\; \mathcal{A}\right], \tag{64}$$

where

$$\mathcal{A} := \left\{\max_{j=1,\dots,p}\left\{\frac{1}{n}\sum_{i=1}^n x_{ij}^2\right\} \leq 2\mathbb{E}[x_{ij}^2]\right\}.$$

Since the x_{ij}'s are sub-Gaussian and $n \gtrsim \log p$, there exist universal constants (c_1, c_2) such that $\mathbb{P}[\mathcal{A}^c] \leq c_1 \exp(-c_2 n)$. The last step is to bound the second term on the right side of (64). For any $t \in \mathbb{R}$, we have

$$\begin{aligned}\log \mathbb{E}[\exp(tV_{ij}) \mid x_i] &= \log\left[\exp(tx_{ij}\psi'(x_i^T\beta^*)\right] \cdot \mathbb{E}[\exp(-tx_{ij}y_i)] \\ &= tx_{ij}\psi'(x_i^T\beta^*) + \left(\psi(-tx_{ij} + x_i^T\beta^*) - \psi(x_i^T\beta^*)\right),\end{aligned}$$

using the fact that ψ is the cumulant generating function for the underlying exponential family. Thus, by a Taylor series expansion, there is some $v_i \in [0,1]$ such that

$$\log \mathbb{E}[\exp(tV_{ij}) \mid x_i] = \frac{t^2 x_{ij}^2}{2}\,\psi''(x_i^T\beta^* - v_i\, t x_{ij}) \;\leq\; \frac{\alpha_u t^2 x_{ij}^2}{2}, \tag{65}$$

where the inequality uses the boundedness of ψ''. Consequently, conditioned on the event $\mathcal{A}$, the variable $\frac{1}{n}\sum_{i=1}^n V_{ij}$ is sub-Gaussian with parameter at most $\kappa = \alpha_u \cdot \max_{j=1,\dots,p}\mathbb{E}[x_{ij}^2]$, for each $j = 1, \dots, p$. By a union bound, we then have

$$\mathbb{P}\left[\max_{j=1,\dots,p} \Big|\frac{1}{n}\sum_{i=1}^n V_{ij}\Big| \geq \delta \;\Big|\; \mathcal{A}\right] \leq p\,\exp\left(-\frac{n\delta^2}{2\kappa^2}\right).$$

The claimed ℓ_1- and ℓ_2-bounds then follow directly from Theorem 1.

B.4 Proof of Corollary 3

We first verify condition (4a) in the case where $\|\Delta\|_F \leq 1$. A straightforward calculation yields

$$\nabla^2 \mathcal{L}_n(\Theta) = \Theta^{-1} \otimes \Theta^{-1} = (\Theta \otimes \Theta)^{-1}.$$

Moreover, letting $\text{vec}(\Delta) \in \mathbb{R}^{p^2}$ denote the vectorized form of the matrix Δ, applying the mean value theorem yields

$$\mathcal{E}_n(\Delta) = \text{vec}(\Delta)^T \left(\nabla^2 \mathcal{L}_n(\Theta^* + t\Delta)\right) \text{vec}(\Delta) \geq \lambda_{\min}(\nabla^2 \mathcal{L}_n(\Theta^* + t\Delta)) \, |||\Theta|||_F^2, \tag{66}$$

for some $t \in [0, 1]$. By standard properties of the Kronecker product (Horn and Johnson, 1990), we have

$$\begin{aligned}\lambda_{\min}(\nabla^2 \mathcal{L}_n(\Theta^* + t\Delta)) = |||\Theta^* + t\Delta|||_2^{-2} &\geq (|||\Theta^*|||_2 + t\,|||\Delta|||_2)^{-2} \\ &\geq (|||\Theta^*|||_2 + 1)^{-2},\end{aligned}$$

using the fact that $|||\Delta|||_2 \leq |||\Delta|||_F \leq 1$. Plugging back into (66) yields

$$\mathcal{E}_n(\Delta) \geq (|||\Theta^*|||_2 + 1)^{-2} \; |||\Theta|||_F^2,$$

so (4a) holds with $\alpha_1 = (|||\Theta^*|||_2 + 1)^{-2}$ and $\tau_1 = 0$. Lemma 9 then implies (4b) with $\alpha_2 = (|||\Theta^*|||_2 + 1)^{-2}$. Finally, we need to establish that the given choice of λ satisfies the requirement (6) of Theorem 1. By the assumed deviation condition (17), we have

$$|||\nabla \mathcal{L}_n(\Theta^*)|||_{\max} = \left|\left|\left|\widehat{\Sigma} - (\Theta^*)^{-1}\right|\right|\right|_{\max} = \left|\left|\left|\widehat{\Sigma} - \Sigma\right|\right|\right|_{\max} \;\leq\; c_0 \sqrt{\frac{\log p}{n}}.$$

Applying Theorem 1 then implies the desired result.

Appendix C. Auxiliary Optimization-Theoretic Results

In this section, we provide proofs of the supporting lemmas used in Section 4.

C.1 Derivation of Three-Step Procedure

We begin by deriving the correctness of the three-step procedure given in Section 4.2. Let $\widehat{\beta}$ be the unconstrained optimum of the program (41). If $g_{\lambda,\mu}(\widehat{\beta}) \leq R$, we clearly have the update given in step (2). Suppose instead that $g_{\lambda,\mu}(\widehat{\beta}) > R$. Then since the program (31) is convex, the iterate β^{t+1} must lie on the boundary of the feasible set; i.e.,

$$g_{\lambda,\mu}(\beta^{t+1}) = R. \tag{67}$$

By Lagrangian duality, the program (31) is also equivalent to

$$\beta^{t+1} \in \arg \min_{g_{\lambda,\mu}(\beta) \leq R'} \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{\nabla \bar{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 \right\},$$

for some choice of constraint parameter R'. Note that this is projection of $\beta^t - \frac{\nabla \bar{\mathcal{L}}_n(\beta^t)}{\eta}$ onto the set $\{\beta \in \mathbb{R}^p \mid g_{\lambda,\mu}(\beta) \leq R'\}$. Since projection decreases the value of $g_{\lambda,\mu}$, equation (67) implies that

$$g_{\lambda,\mu}\left(\beta^t - \frac{\nabla \bar{\mathcal{L}}_n(\beta^t)}{\eta}\right) \geq R.$$

In fact, since the projection will shrink the vector to the boundary of the constraint set, (67) forces $R' = R$. This yields the update (42) appearing in step (3).

C.2 Derivation of Updates for SCAD and MCP

We now derive the explicit form of the updates (43) and (44) for the SCAD and MCP regularizers, respectively. We may rewrite the unconstrained program (41) as

$$\begin{aligned}\beta^{t+1} &\in \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\left\|\beta-\left(\beta^t-\frac{\nabla\bar{\mathcal{L}}_n(\beta^t)}{\eta}\right)\right\|_2^2+\frac{1}{\eta}\cdot\rho_\lambda(\beta)+\frac{\mu}{2\eta}\|\beta\|_2^2\right\}\\ &= \arg\min_{\beta\in\mathbb{R}^p}\left\{\left(\frac{1}{2}+\frac{\mu}{2\eta}\right)\|\beta\|_2^2-\beta^T\left(\beta^t-\frac{\nabla\bar{\mathcal{L}}_n(\beta^t)}{\eta}\right)+\frac{1}{\eta}\cdot\rho_\lambda(\beta)\right\}\\ &= \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\left\|\beta-\frac{1}{1+\mu/\eta}\left(\beta^t-\frac{\nabla\bar{\mathcal{L}}_n(\beta^t)}{\eta}\right)\right\|_2^2+\frac{1/\eta}{1+\mu/\eta}\cdot\rho_\lambda(\beta)\right\}. \qquad (68)\end{aligned}$$

Since the program in the last line of equation (68) decomposes by coordinate, it suffices to solve the scalar optimization problem

$$\widehat{x}\in\arg\min_x\left\{\frac{1}{2}(x-z)^2+\nu\rho(x;\lambda)\right\}, \qquad (69)$$

for general $z\in\mathbb{R}$ and $\nu>0$.

We first consider the case when ρ is the SCAD penalty. The solution $\widehat{x}$ of the program (69) in the case when $\nu=1$ is given in Fan and Li (2001); the expression (43) for the more general case comes from writing out the subgradient of the objective as

$$(x-z)+\nu\rho'(x;\lambda)=\begin{cases}(x-z)+\nu\lambda[-1,+1] & \text{if } x=0,\\ (x-z)+\nu\lambda & \text{if } 0<x\le\lambda,\\ (x-z)+\frac{\nu(a\lambda-x)}{a-1} & \text{if } \lambda\le x\le a\lambda,\\ x-z & \text{if } x\ge a\lambda,\end{cases}$$

using the equation for the SCAD derivative (58), and setting the subgradient equal to zero.

Similarly, when ρ is the MCP parameterized by (b,λ), the subgradient of the objective takes the form

$$(x-z)+\nu\rho'(x;\lambda)=\begin{cases}(x-z)+\nu\lambda[-1,+1] & \text{if } x=0,\\ (x-z)+\nu\lambda\left(1-\frac{x}{b\lambda}\right) & \text{if } 0<x\le b\lambda,\\ x-z & \text{if } x\ge b\lambda,\end{cases}$$

using the expression for the MCP derivative (59), leading to the closed-form solution given in (44). This agrees with the expression provided in Breheny and Huang (2011) for the special case when $\nu=1$.

C.3 Proof of Lemma 1

We first show that if $\lambda\ge\frac{8}{L}\cdot\|\nabla\mathcal{L}_n(\beta^*)\|_\infty$, then for any feasible β such that

$$\phi(\beta)\le\phi(\beta^*)+\bar{\eta}, \qquad (70)$$

we have

$$\|\beta - \beta^*\|_1 \le 8\sqrt{k}\|\beta - \beta^*\|_2 + 2 \cdot \min\left(\frac{2\bar{\eta}}{\lambda L}, R\right). \tag{71}$$

Defining the error vector $\Delta := \beta - \beta^*$, (70) implies

$$\mathcal{L}_n(\beta^* + \Delta) + \rho_\lambda(\beta^* + \Delta) \le \mathcal{L}_n(\beta^*) + \rho_\lambda(\beta^*) + \bar{\eta},$$

so subtracting $\langle \nabla \mathcal{L}_n(\beta^*), \Delta \rangle$ from both sides gives

$$\mathcal{T}(\beta^* + \Delta, \beta^*) + \rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) \le -\langle \nabla \mathcal{L}_n(\beta^*), \Delta \rangle + \bar{\eta}. \tag{72}$$

We divide the argument into two cases. First suppose $\|\Delta\|_2 \le 3$. Note that if $\bar{\eta} \ge \frac{\lambda L}{4}\|\Delta\|_1$, the claim (71) is trivially true; so assume $\bar{\eta} \le \frac{\lambda L}{4}\|\Delta\|_1$. Then the RSC condition (34a), together with (72), implies that

$$\begin{aligned}\alpha_1\|\Delta\|_2^2 - \tau_1\frac{\log p}{n}\|\Delta\|_1^2 + \rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) &\le \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \cdot \|\Delta\|_1 + \bar{\eta} \\ &\le \frac{\lambda L}{8}\|\Delta\|_1 + \frac{\lambda L}{4}\|\Delta\|_1. \end{aligned} \tag{73}$$

Rearranging and using the assumption $\lambda L \ge 16R\tau_1\frac{\log p}{n}$, along with Lemma 4 in Appendix A.1, we then have

$$\begin{aligned}\alpha_1\|\Delta\|_2^2 &\le \rho_\lambda(\beta^*) - \rho_\lambda(\beta^* + \Delta) + \frac{\lambda L}{2}\|\Delta\|_1 \\ &\le \rho_\lambda(\beta^*) - \rho_\lambda(\beta^* + \Delta) + \frac{\rho_\lambda(\beta^*) + \rho_\lambda(\beta^* + \Delta)}{2} + \frac{\mu}{4}\|\Delta\|_2^2,\end{aligned}$$

implying that

$$0 \le \left(\alpha_1 - \frac{\mu}{4}\right)\|\Delta\|_2^2 \le \frac{3}{2}\rho_\lambda(\beta^*) - \frac{1}{2}\rho_\lambda(\beta^* + \Delta),$$

so

$$\rho_\lambda(\beta^*) - \rho_\lambda(\beta^* + \Delta) \le 3\rho_\lambda(\beta^*) - \rho_\lambda(\beta^* + \Delta) \le 3\lambda L\|\Delta_A\|_1 - \lambda L\|\Delta_{A^c}\|_1, \tag{74}$$

by Lemma 5 in Appendix A.1. Furthermore, note that the bound (73) also implies that

$$\rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) \le \frac{\lambda L}{2}\|\Delta\|_1 + \bar{\eta}. \tag{75}$$

Combining (74) and (75) then gives

$$\|\Delta_{A^c}\|_1 - 3\|\Delta_A\|_1 \le \frac{1}{2}\|\Delta\|_1 + \frac{\bar{\eta}}{\lambda L} \le \frac{1}{2}\|\Delta_A\|_1 + \frac{1}{2}\|\Delta_{A^c}\|_1 + \frac{\bar{\eta}}{\lambda L},$$

so

$$\|\Delta_{A^c}\|_1 \le 7\|\Delta_A\|_1 + \frac{2\bar{\eta}}{\lambda L},$$

implying that

$$\|\Delta\|_1 \le 8\|\Delta_A\|_1 + \frac{2\bar{\eta}}{\lambda L} \le 8\sqrt{k}\|\Delta\|_2 + \frac{2\bar{\eta}}{\lambda L}.$$

In the case when $\|\Delta\|_2 \geq 3$, the RSC condition (34b) gives

$$\alpha_2\|\Delta\|_2 - \tau_2\sqrt{\frac{\log p}{n}}\|\Delta\|_1 + \rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) \leq \|\nabla\mathcal{L}_n(\beta^*)\|_\infty \cdot \|\Delta\|_1 + \bar{\eta} \leq \frac{\lambda L}{8}\|\Delta\|_1 + \frac{\lambda L}{4}\|\Delta\|_1, \tag{76}$$

so

$$\alpha_2\|\Delta\|_2 \leq \rho_\lambda(\beta^*) - \rho_\lambda(\beta^* + \Delta) + \left(\frac{3\lambda L}{8} + \tau_2\sqrt{\frac{\log p}{n}}\right)\|\Delta\|_1.$$

In particular, if $\rho_\lambda(\beta^*) - \rho_\lambda(\beta^* + \Delta) \leq 0$, we have

$$\|\Delta\|_2 \leq \frac{2R}{\alpha_2}\left(\frac{3\lambda L}{8} + \tau_2\sqrt{\frac{\log p}{n}}\right) < 3,$$

a contradiction. Hence, using Lemma 5 in Appendix A.1, we have

$$0 \leq \rho_\lambda(\beta^*) - \rho_\lambda(\beta^* + \Delta) \leq \lambda L\|\Delta_A\|_1 - \lambda L\|\Delta_{A^c}\|_1. \tag{77}$$

Note that under the scaling $\lambda L \geq 4\tau_2\sqrt{\frac{\log p}{n}}$, the bound (76) also implies (75). Combining (75) and (77), we then have

$$\|\Delta_{A^c}\|_1 - \|\Delta_A\|_1 \leq \frac{1}{2}\|\Delta\|_1 + \frac{\bar{\eta}}{\lambda L} = \frac{1}{2}\|\Delta_{A^c}\|_1 + \frac{1}{2}\|\Delta_A\|_1 + \frac{\bar{\eta}}{\lambda L},$$

and consequently,

$$\|\Delta_{A^c}\|_1 \leq 3\|\Delta_A\|_1 + \frac{2\bar{\eta}}{\lambda L},$$

so

$$\|\Delta\|_1 \leq 4\|\Delta_A\|_1 + \frac{2\bar{\eta}}{\lambda L} \leq 4\sqrt{k}\|\Delta\|_2 + \frac{2\bar{\eta}}{\lambda L}.$$

Using the trivial bound $\|\Delta\|_1 \leq 2R$, we obtain the claim (71).

We now apply the implication (70) to the vectors $\widehat{\beta}$ and β^t. Note that by optimality of $\widehat{\beta}$, we have

$$\phi(\widehat{\beta}) \leq \phi(\beta^*),$$

and by the assumption (45), we also have

$$\phi(\beta^t) \leq \phi(\widehat{\beta}) + \bar{\eta} \leq \phi(\beta^*) + \bar{\eta}.$$

Hence,

$$\|\widehat{\beta} - \beta^*\|_1 \leq 8\sqrt{k}\|\widehat{\beta} - \beta^*\|_2, \qquad \text{and}$$
$$\|\beta^t - \beta^*\|_1 \leq 8\sqrt{k}\|\beta^t - \beta^*\|_2 + 2 \cdot \min\left(\frac{2\bar{\eta}}{\lambda L}, R\right).$$

By the triangle inequality, we then have

$$\begin{aligned}\|\beta^t - \widehat{\beta}\|_1 &\le \|\widehat{\beta} - \beta^*\|_1 + \|\beta^t - \beta^*\|_1 \\ &\le 8\sqrt{k} \cdot \Big(\|\widehat{\beta} - \beta^*\|_2 + \|\beta^t - \beta^*\|_2\Big) + 2 \cdot \min\left(\frac{2\bar{\eta}}{\lambda L}, R\right) \\ &\le 8\sqrt{k} \cdot \Big(2\|\widehat{\beta} - \beta^*\|_2 + \|\beta^t - \widehat{\beta}\|_2\Big) + 2 \cdot \min\left(\frac{2\bar{\eta}}{\lambda L}, R\right),\end{aligned}$$

as claimed.

C.4 Proof of Lemma 2

Our proof proceeds via induction on the iteration number t. Note that the base case $t = 0$ holds by assumption. Hence, it remains to show that if $\|\beta^t - \widehat{\beta}\|_2 \le 3$ for some integer $t \ge 1$, then $\|\beta^{t+1} - \widehat{\beta}\|_2 \le 3$, as well.

We assume for the sake of a contradiction that $\|\beta^{t+1} - \widehat{\beta}\|_2 > 3$. By the RSC condition (34b) and the relation (33), we have

$$\overline{\mathcal{T}}(\beta^{t+1}, \widehat{\beta}) \ge \alpha\|\widehat{\beta} - \beta^{t+1}\|_2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 - \frac{\mu}{2}\|\widehat{\beta} - \beta^{t+1}\|_2^2. \tag{78}$$

Furthermore, by convexity of $g := g_{\lambda,\mu}$, we have

$$g(\beta^{t+1}) - g(\widehat{\beta}) - \langle \nabla g(\widehat{\beta}),\, \beta^{t+1} - \widehat{\beta}\rangle \ge 0. \tag{79}$$

Multiplying by λ and summing with (78) then yields

$$\begin{aligned}&\phi(\beta^{t+1}) - \phi(\widehat{\beta}) - \langle \nabla \phi(\widehat{\beta}),\, \beta^{t+1} - \widehat{\beta}\rangle \\ &\qquad \ge \alpha\|\widehat{\beta} - \beta^{t+1}\|_2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 - \frac{\mu}{2}\|\widehat{\beta} - \beta^{t+1}\|_2^2.\end{aligned}$$

Together with the first-order optimality condition $\langle \nabla \phi(\widehat{\beta}),\, \beta^{t+1} - \widehat{\beta}\rangle \ge 0$, we then have

$$\phi(\beta^{t+1}) - \phi(\widehat{\beta}) \ge \alpha\|\widehat{\beta} - \beta^{t+1}\|_2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 - \frac{\mu}{2}\|\widehat{\beta} - \beta^{t+1}\|_2^2. \tag{80}$$

Since $\|\widehat{\beta} - \beta^t\|_2 \le 3$ by the induction hypothesis, applying the RSC condition (34a) to the pair $(\widehat{\beta}, \beta^t)$ also gives

$$\bar{\mathcal{L}}_n(\widehat{\beta}) \ge \bar{\mathcal{L}}_n(\beta^t) + \langle \nabla \bar{\mathcal{L}}_n(\beta^t),\, \widehat{\beta} - \beta^t\rangle + \Big(\alpha - \frac{\mu}{2}\Big) \cdot \|\beta^t - \widehat{\beta}\|_2^2 - \tau\frac{\log p}{n}\|\beta^t - \widehat{\beta}\|_1^2.$$

Combining with the inequality

$$g(\widehat{\beta}) \ge g(\beta^{t+1}) + \langle \nabla g(\beta^{t+1}),\, \widehat{\beta} - \beta^{t+1}\rangle,$$

we then have

$$\begin{aligned}\phi(\widehat{\beta}) &\geq \bar{\mathcal{L}}_n(\beta^t) + \langle \nabla \bar{\mathcal{L}}_n(\beta^t),\, \widehat{\beta} - \beta^t \rangle + \lambda g(\beta^{t+1}) + \lambda \langle \nabla g(\beta^{t+1}),\, \widehat{\beta} - \beta^{t+1} \rangle \\ &\qquad + \left(\alpha - \frac{\mu}{2}\right) \cdot \|\beta^t - \widehat{\beta}\|_2^2 - \tau \frac{\log p}{n} \|\beta^t - \widehat{\beta}\|_1^2 \\ &\geq \bar{\mathcal{L}}_n(\beta^t) + \langle \nabla \bar{\mathcal{L}}_n(\beta^t),\, \widehat{\beta} - \beta^t \rangle + \lambda g(\beta^{t+1}) \\ &\qquad + \lambda \langle \nabla g(\beta^{t+1}),\, \widehat{\beta} - \beta^{t+1} \rangle - \tau \frac{\log p}{n} \|\beta^t - \widehat{\beta}\|_1^2. \end{aligned} \tag{81}$$

Finally, the RSM condition (35) on the pair (β^{t+1}, β^t) gives

$$\begin{aligned}\phi(\beta^{t+1}) &\leq \bar{\mathcal{L}}_n(\beta^t) + \langle \nabla \bar{\mathcal{L}}_n(\beta^t),\, \beta^{t+1} - \beta^t \rangle + \lambda g(\beta^{t+1}) \qquad (82) \\ &\qquad + \left(\alpha_3 - \frac{\mu}{2}\right) \|\beta^{t+1} - \beta^t\|_2^2 + \tau \frac{\log p}{n} \|\beta^{t+1} - \beta^t\|_1^2 \\ &\leq \bar{\mathcal{L}}_n(\beta^t) + \langle \nabla \bar{\mathcal{L}}_n(\beta^t),\, \beta^{t+1} - \beta^t \rangle + \lambda g(\beta^{t+1}) \\ &\qquad + \frac{\eta}{2} \|\beta^{t+1} - \beta^t\|_2^2 + \frac{4R^2 \tau \log p}{n}, \qquad (83)\end{aligned}$$

since $\frac{\eta}{2} \geq \alpha_3 - \frac{\mu}{2}$ by assumption, and $\|\beta^{t+1} - \beta^t\|_1 \leq 2R$. It is easy to check that the update (31) may be written equivalently as

$$\beta^{t+1} \in \arg \min_{g(\beta) \leq R,\, \beta \in \Omega} \left\{ \bar{\mathcal{L}}_n(\beta^t) + \langle \nabla \bar{\mathcal{L}}_n(\beta^t),\, \beta - \beta^t \rangle + \frac{\eta}{2} \|\beta - \beta^t\|_2^2 + \lambda g(\beta) \right\},$$

and the optimality of β^{t+1} then yields

$$\langle \nabla \bar{\mathcal{L}}_n(\beta^t) + \eta(\beta^{t+1} - \beta^t) + \lambda \nabla g(\beta^{t+1}),\, \beta^{t+1} - \widehat{\beta} \rangle \leq 0. \tag{84}$$

Summing up (81), (82), and (84), we then have

$$\begin{aligned}\phi(\beta^{t+1}) - \phi(\widehat{\beta}) &\leq \frac{\eta}{2} \|\beta^{t+1} - \beta^t\|_2^2 + \eta \langle \beta^t - \beta^{t+1},\, \beta^{t+1} - \widehat{\beta} \rangle + \tau \frac{\log p}{n} \|\beta^t - \widehat{\beta}\|_1^2 \\ &\qquad + \frac{4R^2 \tau \log p}{n} \\ &= \frac{\eta}{2} \|\beta^t - \widehat{\beta}\|_2^2 - \frac{\eta}{2} \|\beta^{t+1} - \widehat{\beta}\|_2^2 + \tau \frac{\log p}{n} \|\beta^t - \widehat{\beta}\|_1^2 + \frac{4R^2 \tau \log p}{n}.\end{aligned}$$

Combining this last inequality with (80), we have

$$\begin{aligned}\alpha \|\widehat{\beta} - \beta^{t+1}\|_2 - \tau \sqrt{\frac{\log p}{n}} \|\widehat{\beta} - \beta^{t+1}\|_1 &\leq \frac{\eta}{2} \|\beta^t - \widehat{\beta}\|_2^2 - \frac{\eta - \mu}{2} \|\beta^{t+1} - \widehat{\beta}\|_2^2 + \frac{8R^2 \tau \log p}{n} \\ &\leq \frac{9\eta}{2} - \frac{3(\eta - \mu)}{2} \|\beta^{t+1} - \widehat{\beta}\|_2 + \frac{8R^2 \tau \log p}{n},\end{aligned}$$

since $\|\beta^t - \widehat{\beta}\|_2 \le 3$ by the induction hypothesis and $\|\beta^{t+1} - \widehat{\beta}\|_2 > 3$ by assumption, and using the fact that $\eta \ge \mu$. It follows that

$$\left(\alpha + \frac{3(\eta-\mu)}{2}\right) \cdot \|\widehat{\beta} - \beta^{t+1}\|_2 \le \frac{9\eta}{2} + \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 + \frac{8R^2\tau\log p}{n}$$
$$\le \frac{9\eta}{2} + 2R\tau\sqrt{\frac{\log p}{n}} + \frac{8R^2\tau\log p}{n}$$
$$\le 3\left(\alpha + \frac{3(\eta-\mu)}{2}\right),$$

where the final inequality holds whenever $2R\tau\sqrt{\frac{\log p}{n}} + \frac{8R^2\tau\log p}{n} \le 3\left(\alpha - \frac{3\mu}{2}\right)$. Rearranging gives $\|\beta^{t+1} - \widehat{\beta}\|_2 \le 3$, providing the desired contradiction.

C.5 Proof of Lemma 3

We begin with an auxiliary lemma:

Lemma 10 *Under the conditions of Lemma 3, we have*

$$\bar{\mathcal{T}}(\beta^t, \widehat{\beta}) \ge -2\tau\frac{\log p}{n}(\epsilon + \bar{\epsilon})^2, \quad \textit{and} \tag{85a}$$
$$\phi(\beta^t) - \phi(\widehat{\beta}) \ge \frac{2\alpha - \mu}{4}\|\widehat{\beta} - \beta^t\|_2^2 - \frac{2\tau\log p}{n}(\epsilon + \bar{\epsilon})^2. \tag{85b}$$

We prove this result later, taking it as given for the moment.

Define

$$\phi_t(\beta) := \bar{\mathcal{L}}_n(\beta^t) + \langle \nabla\bar{\mathcal{L}}_n(\beta^t),\, \beta - \beta^t\rangle + \frac{\eta}{2}\|\beta - \beta^t\|_2^2 + \lambda g(\beta),$$

the objective function minimized over the constraint set $\{g(\beta) \le R\}$ at iteration t. For any $\gamma \in [0,1]$, the vector $\beta_\gamma := \gamma\widehat{\beta} + (1-\gamma)\beta^t$ belongs to the constraint set, as well. Consequently, by the optimality of β^{t+1} and feasibility of β_γ, we have

$$\phi_t(\beta^{t+1}) \le \phi_t(\beta_\gamma) = \bar{\mathcal{L}}_n(\beta^t) + \langle \nabla\bar{\mathcal{L}}_n(\beta^t),\, \gamma\widehat{\beta} - \gamma\beta^t\rangle + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2 + \lambda g(\beta_\gamma).$$

Appealing to (85a), we then have

$$\phi_t(\beta^{t+1}) \le (1-\gamma)\bar{\mathcal{L}}_n(\beta^t) + \gamma\bar{\mathcal{L}}_n(\widehat{\beta}) + 2\gamma\tau\frac{\log p}{n}(\epsilon + \bar{\epsilon})^2$$
$$+ \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2 + \lambda g(\beta_\gamma)$$
$$\stackrel{(i)}{\le} \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + 2\gamma\tau\frac{\log p}{n}(\epsilon + \bar{\epsilon})^2 + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2$$
$$\le \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + 2\tau\frac{\log p}{n}(\epsilon + \bar{\epsilon})^2 + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2, \tag{86}$$

where inequality (i) incorporates the fact that

$$g(\beta_\gamma) \le \gamma g(\widehat{\beta}) + (1-\gamma) g(\beta^t),$$

by the convexity of g.

By the RSM condition (35), we also have

$$\bar{\mathcal{T}}(\beta^{t+1}, \beta^t) \le \frac{\eta}{2}\|\beta^{t+1} - \beta^t\|_2^2 + \tau \frac{\log p}{n} \|\beta^{t+1} - \beta^t\|_1^2,$$

since $\alpha_3 - \mu \le \frac{\eta}{2}$ by assumption, and adding $\lambda g(\beta^{t+1})$ to both sides gives

$$\begin{aligned}\phi(\beta^{t+1}) &\le \bar{\mathcal{L}}_n(\beta^t) + \langle \nabla \bar{\mathcal{L}}_n(\beta^t),\, \beta^{t+1} - \beta^t \rangle + \frac{\eta}{2}\|\beta^{t+1} - \beta^t\|_2^2 \\ &\qquad + \tau \frac{\log p}{n}\|\beta^{t+1} - \beta^t\|_1^2 + \lambda g(\beta^{t+1}) \\ &= \phi_t(\beta^{t+1}) + \tau \frac{\log p}{n}\|\beta^{t+1} - \beta^t\|_1^2.\end{aligned}$$

Combining with (86) then yields

$$\begin{aligned}\phi(\beta^{t+1}) &\le \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2 \\ &\qquad + \tau \frac{\log p}{n}\|\beta^{t+1} - \beta^t\|_1^2 + 2\tau \frac{\log p}{n}(\epsilon + \bar{\epsilon})^2. \end{aligned} \tag{87}$$

By the triangle inequality, we have

$$\|\beta^{t+1} - \beta^t\|_1^2 \le \big(\|\Delta^{t+1}\|_1 + \|\Delta^t\|_1\big)^2 \ \le \ 2\|\Delta^{t+1}\|_1^2 + 2\|\Delta^t\|_1^2,$$

where we have defined $\Delta^t := \beta^t - \widehat{\beta}$. Combined with (87), we therefore have

$$\begin{aligned}\phi(\beta^{t+1}) &\le \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \frac{\eta\gamma^2}{2}\|\Delta^t\|_2^2 \\ &\qquad + 2\tau \frac{\log p}{n}(\|\Delta^{t+1}\|_1^2 + \|\Delta^t\|_1^2) + 2\psi(n,p,\epsilon),\end{aligned}$$

where $\psi(n,p,\epsilon) := \tau \frac{\log p}{n}(\epsilon + \bar{\epsilon})^2$. Then applying Lemma 1 to bound the ℓ_1-norms, we have

$$\begin{aligned}\phi(\beta^{t+1}) &\le \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \frac{\eta\gamma^2}{2}\|\Delta^t\|_2^2 \\ &\qquad + ck\tau \frac{\log p}{n}(\|\Delta^{t+1}\|_2^2 + \|\Delta^t\|_2^2) + c'\psi(n,p,\epsilon) \\ &= \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \left(\frac{\eta\gamma^2}{2} + ck\tau\frac{\log p}{n}\right)\|\Delta^t\|_2^2 \\ &\qquad + ck\tau \frac{\log p}{n}\|\Delta^{t+1}\|_2^2 + c'\psi(n,p,\epsilon). \end{aligned} \tag{88}$$

Now introduce the shorthand $\delta_t := \phi(\beta^t) - \phi(\widehat{\beta})$ and $\upsilon(k,p,n) = k\tau\frac{\log p}{n}$. By applying (85b) and subtracting $\phi(\widehat{\beta})$ from both sides of (88), we have

$$\delta_{t+1} \le \big(1-\gamma\big)\delta_t + \frac{\eta\gamma^2 + c\upsilon(k,p,n)}{\alpha - \mu/2}\left(\delta_t + 2\psi(n,p,\epsilon)\right) + \frac{c\upsilon(k,p,n)}{\alpha-\mu/2}\left(\delta_{t+1} + 2\psi(n,p,\epsilon)\right) + c'\psi(n,p,\epsilon).$$

Choosing $\gamma = \frac{2\alpha-\mu}{4\eta} \in (0,1)$ yields

$$\left(1 - \frac{c\upsilon(k,p,n)}{\alpha-\mu/2}\right)\delta_{t+1} \le \left(1 - \frac{2\alpha-\mu}{8\eta} + \frac{c\upsilon(k,p,n)}{\alpha-\mu/2}\right)\delta_t + 2\left(\frac{2\alpha-\mu}{8\eta} + \frac{2c\upsilon(k,p,n)}{\alpha-\mu/2} + c'\right)\psi(n,p,\epsilon),$$

or $\delta_{t+1} \le \kappa\delta_t + \xi(\epsilon+\bar{\epsilon})^2$, where κ and ξ were previously defined in (36) and (47), respectively. Finally, iterating the procedure yields

$$\delta_t \le \kappa^{t-T}\delta_T + \xi(\epsilon+\bar{\epsilon})^2(1+\kappa+\kappa^2+\cdots+\kappa^{t-T-1}) \;\le\; \kappa^{t-T}\delta_T + \frac{\xi(\epsilon+\bar{\epsilon})^2}{1-\kappa}, \tag{89}$$

as claimed.

The only remaining step is to prove the auxiliary lemma.

Proof of Lemma 10: By the RSC condition (34a) and the assumption (46), we have

$$\bar{\mathcal{T}}(\beta^t, \widehat{\beta}) \ge \left(\alpha - \frac{\mu}{2}\right)\|\widehat{\beta} - \beta^t\|_2^2 - \tau\frac{\log p}{n}\|\widehat{\beta} - \beta^t\|_1^2. \tag{90}$$

Furthermore, by convexity of g, we have

$$\lambda\Big(g(\beta^t) - g(\widehat{\beta}) - \langle \nabla g(\widehat{\beta}),\, \beta^t - \widehat{\beta}\rangle\Big) \ge 0, \tag{91}$$

and the first-order optimality condition for $\widehat{\beta}$ gives

$$\langle \nabla\phi(\widehat{\beta}),\, \beta^t - \widehat{\beta}\rangle \ge 0. \tag{92}$$

Summing (90), (91), and (92) then yields

$$\phi(\beta^t) - \phi(\widehat{\beta}) \ge \left(\alpha - \frac{\mu}{2}\right)\|\widehat{\beta} - \beta^t\|_2^2 - \tau\frac{\log p}{n}\|\widehat{\beta} - \beta^t\|_1^2.$$

Applying Lemma 1 to bound the term $\|\widehat{\beta} - \beta^t\|_1^2$ and using the assumption $\frac{ck\tau\log p}{n} \le \frac{2\alpha-\mu}{4}$ yields the bound (85b). On the other hand, applying Lemma 1 directly to (90) with β^t and $\widehat{\beta}$ switched gives

$$\begin{aligned}\bar{\mathcal{T}}(\widehat{\beta}, \beta^t) &\ge \left(\alpha - \frac{\mu}{2}\right)\|\widehat{\beta} - \beta^t\|_2^2 - \tau\frac{\log p}{n}\left(ck\|\widehat{\beta} - \beta^t\|_2^2 + 2(\epsilon+\bar{\epsilon})^2\right) \\ &\ge -2\tau\frac{\log p}{n}(\epsilon+\bar{\epsilon})^2.\end{aligned}$$

This establishes (85a).

Appendix D. Verifying RSC/RSM Conditions

In this Appendix, we provide a proof of Proposition 1, which verifies the RSC (34) and RSM (35) conditions for GLMs.

D.1 Main Argument

Using the notation for GLMs in Section 3.3, we introduce the shorthand $\Delta := \beta_1 - \beta_2$ and observe that, by the mean value theorem, we have

$$\mathcal{T}(\beta_1, \beta_2) = \frac{1}{n} \sum_{i=1}^n \psi''\big(\langle \beta_1, x_i \rangle) + t_i \langle \Delta, x_i \rangle\big) \left(\langle \Delta, x_i \rangle\right)^2, \tag{93}$$

for some $t_i \in [0, 1]$. The t_i's are i.i.d. random variables, with each t_i depending only on the random vector x_i.

Proof of bound (40)**:** The proof of this upper bound is relatively straightforward given earlier results (Loh and Wainwright, 2013a). From the Taylor series expansion (93) and the boundedness assumption $\|\psi''\|_\infty \leq \alpha_u$, we have

$$\mathcal{T}(\beta_1, \beta_2) \leq \alpha_u \cdot \frac{1}{n} \sum_{i=1}^n \big(\langle \Delta, x_i \rangle\big)^2.$$

By known results on restricted eigenvalues for ordinary linear regression (cf. Lemma 13 in Loh and Wainwright (2012)), we also have

$$\frac{1}{n} \sum_{i=1}^n (\langle \Delta, x_i \rangle)^2 \leq \lambda_{\max}(\Sigma) \left(\frac{3}{2} \|\Delta\|_2^2 + \frac{\log p}{n} \|\Delta\|_1^2 \right),$$

with probability at least $1 - c_1 \exp(-c_2 n)$. Combining the two inequalities yields the desired result.

Proof of bounds (39)**:** The proof of the RSC bound is much more involved, and we provide only high-level details here, deferring the bulk of the technical analysis to later in the appendix. We define

$$\alpha_\ell := \left(\inf_{|t| \leq 2T} \psi''(t) \right) \frac{\lambda_{\min}(\Sigma)}{8},$$

where T is a suitably chosen constant depending only on $\lambda_{\min}(\Sigma)$ and the sub-Gaussian parameter σ_x. (In particular, see (99) below, and take $T = 3\tau$.) The core of the proof is based on the following lemma, proved in Section D.2:

Lemma 11 *With probability at least* $1 - c_1 \exp(-c_2 n)$*, we have*

$$\mathcal{T}(\beta_1, \beta_2) \geq \alpha_\ell \|\Delta\|_2^2 - c\sigma_x \|\Delta\|_1 \|\Delta\|_2 \sqrt{\frac{\log p}{n}},$$

uniformly over all pairs (β_1, β_2) *such that* $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$, $\|\beta_1 - \beta_2\|_2 \leq 3$, *and*

$$\frac{\|\Delta\|_1}{\|\Delta\|_2} \leq \frac{\alpha_\ell}{c\sigma_x} \sqrt{\frac{n}{\log p}}. \tag{94}$$

Taking Lemma 11 as given, we now complete the proof of the RSC condition (39). By the arithmetic mean-geometric mean inequality, we have

$$c\sigma_x \|\Delta\|_1 \|\Delta\|_2 \sqrt{\frac{\log p}{n}} \leq \frac{\alpha_\ell}{2} \|\Delta\|_2^2 + \frac{c^2\sigma_x^2}{2\alpha_\ell} \frac{\log p}{n} \|\Delta\|_1^2,$$

so Lemma 11 implies that (39a) holds uniformly over all pairs (β_1, β_2) such that $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$ and $\|\beta_1 - \beta_2\|_2 \leq 3$, whenever the bound (94) holds. On the other hand, if the bound (94) does not hold, then the lower bound in (39a) is negative. By convexity of $\mathcal{L}_n$, we have $\mathcal{T}(\beta_1, \beta_2) \geq 0$, so (39a) holds trivially in that case.

We now show that (39b) holds: in particular, consider a pair (β_1, β_2) with $\beta_2 \in \mathbb{B}_2(3)$ and $\|\beta_1 - \beta_2\|_2 \geq 3$. For any $t \in [0, 1]$, the convexity of $\mathcal{L}_n$ implies that

$$\mathcal{L}_n(\beta_2 + t\Delta) \leq t\mathcal{L}_n(\beta_2 + \Delta) + (1 - t)\mathcal{L}_n(\beta_2),$$

where $\Delta := \beta_1 - \beta_2$. Rearranging yields

$$\mathcal{L}_n(\beta_2 + \Delta) - \mathcal{L}_n(\beta_2) \geq \frac{\mathcal{L}_n(\beta_2 + t\Delta) - \mathcal{L}_n(\beta_2)}{t},$$

so

$$\mathcal{T}(\beta_2 + \Delta, \beta_2) \geq \frac{\mathcal{T}(\beta_2 + t\Delta, \beta_2)}{t}. \tag{95}$$

Now choose $t = \frac{3}{\|\Delta\|_2} \in [0, 1]$ so that $\|t\Delta\|_2 = 1$. Introducing the shorthand $\alpha_1 := \frac{\alpha_\ell}{2}$ and $\tau_1 := \frac{c^2\sigma_x^2}{2\alpha_\ell}$, we may apply (39a) to obtain

$$\begin{aligned} \frac{\mathcal{T}(\beta_2 + t\Delta, \beta_2)}{t} &\geq \frac{\|\Delta\|_2}{3} \left(\alpha_1 \left(\frac{3\|\Delta\|_2}{\|\Delta\|_2} \right)^2 - \tau_1 \frac{\log p}{n} \left(\frac{3\|\Delta\|_1}{\|\Delta\|_2} \right)^2 \right) \\ &= 3\alpha_1 \|\Delta\|_2 - 9\tau_1 \frac{\log p}{n} \frac{\|\Delta\|_1^2}{\|\Delta\|_2}. \end{aligned} \tag{96}$$

Note that (39b) holds trivially unless $\frac{\|\Delta\|_1}{\|\Delta\|_2} \leq \frac{\alpha_\ell}{2c\sigma_x} \sqrt{\frac{n}{\log p}}$, due to the convexity of $\mathcal{L}_n$. In that case, (95) and (96) together imply

$$\mathcal{T}(\beta_2 + \Delta, \beta_2) \geq 3\alpha_1 \|\Delta\|_2 - \frac{9\tau_1 \, \alpha_\ell}{2c\sigma_x} \sqrt{\frac{\log p}{n}} \|\Delta\|_1,$$

which is exactly the bound (39b).

D.2 Proof of Lemma 11

For a truncation level $\tau' > 0$ to be chosen, define the functions

$$\varphi_{\tau'}(u) = \begin{cases} u^2, & \text{if } |u| \le \frac{\tau'}{2}, \\ (\tau' - u)^2, & \text{if } \frac{\tau'}{2} \le |u| \le \tau', \\ 0, & \text{if } |u| \ge \tau'. \end{cases}$$

By construction, $\varphi_{\tau'}$ is τ'-Lipschitz and

$$\varphi_{\tau'}(u) \le u^2 \cdot \mathbb{I}\{|u| \le \tau'\}, \quad \text{for all } u \in \mathbb{R}. \tag{97}$$

In addition, we define the trapezoidal function

$$\gamma_\tau'(u) = \begin{cases} 1, & \text{if } |u| \le \frac{\tau'}{2}, \\ 2 - \frac{2}{\tau'}|u|, & \text{if } \frac{\tau'}{2} \le |u| \le \tau', \\ 0, & \text{if } |u| \ge \tau', \end{cases}$$

and note that γ_τ' is $\frac{2}{\tau'}$-Lipschitz and $\gamma_\tau'(u) \le \mathbb{I}\{|u| \le \tau'\}$.

Taking $T \ge 3\tau'$ so that $T \ge \tau'\|\Delta\|_2$ (since $\|\Delta\|_2 \le 3$ by assumption), and defining

$$L_\psi(T) := \inf_{|u| \le 2T} \psi''(u),$$

we have the following inequality:

$$\begin{aligned} \mathcal{T}(\beta + \Delta, \beta) &= \frac{1}{n}\sum_{i=1}^n \psi''(x_i^T\beta + t_i \cdot x_i^T\Delta) \cdot (x_i^T\Delta)^2 \\ &\ge L_\psi(T) \cdot \sum_{i=1}^n (x_i^T\Delta)^2 \cdot \mathbb{I}\{|x_i^T\Delta| \le \tau'\|\Delta\|_2\} \cdot \mathbb{I}\{|x_i^T\beta| \le T\} \\ &\ge L_\psi(T) \cdot \frac{1}{n}\sum_{i=1}^n \varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta), \end{aligned} \tag{98}$$

where the first equality is the expansion (93) and the second inequality uses the bound (97).

Now define the subset of $\mathbb{R}^p \times \mathbb{R}^p$ via

$$\mathbb{A}_\delta := \left\{(\beta, \Delta) : \beta \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R),\ \Delta \in \mathbb{B}_2(3),\ \frac{\|\Delta\|_1}{\|\Delta\|_2} \le \delta\right\},$$

as well as the random variable

$$Z(\delta) := \sup_{(\beta,\Delta)\in\mathbb{A}_\delta} \frac{1}{\|\Delta\|_2^2}\left|\frac{1}{n}\sum_{i=1}^n \varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta) - \mathbb{E}\left[\varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta)\,\gamma_T(x_i^T\beta)\right]\right|.$$

For any pair $(\beta, \Delta) \in \mathbb{A}_\delta$, we have

$$\begin{aligned}
&\mathbb{E}\big[(x_i^T\Delta)^2 - \varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\big] \\
&\quad\le \mathbb{E}\left[(x_i^T\Delta)^2\mathbb{I}\left\{|x_i^T\Delta| \ge \frac{\tau'\|\Delta\|_2}{2}\right\}\right] + \mathbb{E}\left[(x_i^T\Delta)^2\mathbb{I}\left\{|x_i^T\beta| \ge \frac{T}{2}\right\}\right] \\
&\quad\le \sqrt{\mathbb{E}\left[(x_i^T\Delta)^4\right]}\cdot\left(\sqrt{\mathbb{P}\left(|x_i^T\Delta| \ge \frac{\tau'\|\Delta\|_2}{2}\right)} + \sqrt{\mathbb{P}\left(|x_i^T\beta| \ge \frac{T}{2}\right)}\right) \\
&\quad\le \sigma_x^2\|\Delta\|_2^2\cdot c\exp\left(-\frac{c'\tau'^2}{\sigma_x^2}\right),
\end{aligned}$$

where we have used Cauchy-Schwarz and a tail bound for sub-Gaussians, assuming $\beta \in \mathbb{B}_2(3)$. It follows that for τ' chosen such that

$$c\sigma_x^2\exp\left(-\frac{c'\tau'^2}{\sigma_x^2}\right) = \frac{\lambda_{\min}\left(\mathbb{E}[x_ix_i^T]\right)}{2}, \tag{99}$$

we have the lower bound

$$\mathbb{E}\left[\varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\right] \ge \frac{\lambda_{\min}\left(\mathbb{E}[x_ix_i^T]\right)}{2}\cdot\|\Delta\|_2^2. \tag{100}$$

By construction of φ, each summand in the expression for $Z(\delta)$ is sandwiched as

$$0 \le \frac{1}{\|\Delta\|_2^2}\cdot\varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta) \le \frac{\tau'^2}{4}.$$

Consequently, applying the bounded differences inequality yields

$$\mathbb{P}\left(Z(\delta) \ge \mathbb{E}[Z(\delta)] + \frac{\lambda_{\min}\left(\mathbb{E}[x_ix_i^T]\right)}{4}\right) \le c_1\exp(-c_2 n). \tag{101}$$

Furthermore, by Lemmas 12 and 13 in Appendix E, we have

$$\mathbb{E}[Z(\delta)] \le 2\sqrt{\frac{\pi}{2}}\cdot\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta}\frac{1}{\|\Delta\|_2^2}\left|\frac{1}{n}\sum_{i=1}^n g_i\Big(\varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\Big)\right|\right], \tag{102}$$

where the g_i's are i.i.d. standard Gaussians. Conditioned on $\{x_i\}_{i=1}^n$, define the Gaussian processes

$$Z_{\beta,\Delta} := \frac{1}{\|\Delta\|_2^2}\cdot\frac{1}{n}\sum_{i=1}^n g_i\Big(\varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\Big),$$

and note that for pairs (β, Δ) and $(\widetilde{\beta}, \widetilde{\Delta})$, we have

$$\operatorname{var}\left(Z_{\beta,\Delta} - Z_{\widetilde{\beta},\widetilde{\Delta}}\right) \le 2\operatorname{var}\left(Z_{\beta,\Delta} - Z_{\widetilde{\beta},\Delta}\right) + 2\operatorname{var}\left(Z_{\widetilde{\beta},\Delta} - Z_{\widetilde{\beta},\widetilde{\Delta}}\right),$$

with

$$\operatorname{var}\left(Z_{\beta,\Delta} - Z_{\widetilde{\beta},\Delta}\right) = \frac{1}{\|\Delta\|_2^4} \cdot \frac{1}{n^2} \sum_{i=1}^n \varphi_{\tau'\|\Delta\|_2}^2 (x_i^T \Delta) \cdot \left(\gamma_T(x_i^T \beta) - \gamma_T(x_i^T \widetilde{\beta})\right)^2$$
$$\leq \frac{1}{n^2} \sum_{i=1}^n \frac{{\tau'}^4}{16} \cdot \frac{4}{T^2} \left(x_i^T (\beta - \widetilde{\beta})\right)^2,$$

since $\varphi_{\tau'\|\Delta\|_2} \leq \frac{{\tau'}^2 \|\Delta\|_2^2}{4}$ and γ_T is $\frac{2}{T}$-Lipschitz. Similarly, using the homogeneity property

$$\frac{1}{c^2} \cdot \varphi_{ct}(cu) = \varphi_t(u), \qquad \forall c > 0,$$

and the fact that $\varphi_{\tau'\|\Delta\|_2}$ is $\tau'\|\Delta\|_2$-Lipschitz, we have

$$\operatorname{var}\left(Z_{\widetilde{\beta},\Delta} - Z_{\widetilde{\beta},\widetilde{\Delta}}\right) \leq \frac{1}{n^2} \sum_{i=1}^n \gamma_T^2 (x_i^T \widetilde{\beta}) \left(\frac{\varphi_{\tau'\|\Delta\|_2}(x_i^T \Delta)}{\|\Delta\|_2^2} - \frac{\varphi_{\tau'\|\widetilde{\Delta}\|_2}(x_i^T \widetilde{\Delta})}{\|\widetilde{\Delta}\|_2^2}\right)^2$$
$$= \frac{1}{n^2} \sum_{i=1}^n \frac{\gamma_T^2 (x_i^T \widetilde{\beta})}{\|\Delta\|_2^4} \left(\varphi_{\tau'\|\Delta\|_2}(x_i^T \Delta) - \varphi_{\tau'\|\Delta\|_2}\left(x_i^T \widetilde{\Delta} \cdot \frac{\|\Delta\|_2}{\|\widetilde{\Delta}\|_2}\right)\right)^2$$
$$\leq \frac{1}{n^2} \sum_{i=1}^n \frac{{\tau'}^2}{\|\Delta\|_2^2} \left(x_i^T \Delta - x_i^T \widetilde{\Delta} \cdot \frac{\|\Delta\|_2}{\|\widetilde{\Delta}\|_2}\right)^2$$
$$= \frac{1}{n^2} \sum_{i=1}^n {\tau'}^2 \left(\frac{x_i^T \Delta}{\|\Delta\|_2} - \frac{x_i^T \widetilde{\Delta}}{\|\widetilde{\Delta}\|_2}\right)^2.$$

Defining the centered Gaussian process

$$Y_{\beta,\Delta} := \frac{{\tau'}^2}{\sqrt{2}T} \cdot \frac{1}{n} \sum_{i=1}^n \widehat{g}_i \cdot x_i^T \beta + \frac{\sqrt{2}\tau'}{\|\Delta\|_2} \cdot \frac{1}{n} \sum_{i=1}^n \widetilde{g}_i \cdot x_i^T \Delta,$$

where the $\widehat{g}_i$'s and $\widetilde{g}_i$'s are independent standard Gaussians, it follows that

$$\operatorname{var}\left(Z_{\beta,\Delta} - Z_{\widetilde{\beta},\widetilde{\Delta}}\right) \leq \operatorname{var}\left(Y_{\beta,\Delta} - Y_{\widetilde{\beta},\widetilde{\Delta}}\right).$$

Applying Lemma 14 in Appendix E, we then have

$$\mathbb{E}\left[\sup_{(\beta,\Delta) \in \mathbb{A}_\delta} Z_{\beta,\Delta}\right] \leq 2 \cdot \mathbb{E}\left[\sup_{(\beta,\Delta) \in \mathbb{A}_\delta} Y_{\beta,\Delta}\right]. \tag{103}$$

Note further (cf. p.77 of Ledoux and Talagrand (1991)) that

$$\mathbb{E}\left[\sup_{(\beta,\Delta) \in \mathbb{A}_\delta} |Z_{\beta,\Delta}|\right] \leq \mathbb{E}\left[|Z_{\beta_0,\Delta_0}|\right] + 2\mathbb{E}\left[\sup_{(\beta,\Delta) \in \mathbb{A}_\delta} Z_{\beta,\Delta}\right], \tag{104}$$

for any $(\beta_0, \Delta_0) \in \mathbb{A}_\delta$, and furthermore,

$$\mathbb{E}\left[|Z_{\beta_0,\Delta_0}|\right] \le \sqrt{\frac{2}{\pi}} \cdot \sqrt{\mathrm{var}\left(Z_{\beta_0,\Delta_0}\right)} \le c_0 \cdot \sqrt{\frac{2}{\pi}} \cdot \sqrt{\frac{\tau'^2}{4n}}. \tag{105}$$

Finally,

$$\begin{aligned} \mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} Y_{\beta,\Delta}\right] &\le \frac{\tau'^2 R}{\sqrt{2}T} \cdot \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \widehat{g}_i x_i\right\|_\infty\right] + \sqrt{2}\tau'\delta \cdot \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \widetilde{g}_i x_i\right\|_\infty\right] \\ &\le \frac{c\tau'^2 R\sigma_x}{T}\sqrt{\frac{\log p}{n}} + c\tau'\delta\sigma_x \cdot \sqrt{\frac{\log p}{n}}, \end{aligned} \tag{106}$$

by Lemma 16 in Appendix E. Combining (102), (103), (104), (105), and (106), we then obtain

$$\mathbb{E}[Z(\delta)] \le \frac{c'\tau'^2 R\sigma_x}{T}\sqrt{\frac{\log p}{n}} + c'\tau'\delta\sigma_x \cdot \sqrt{\frac{\log p}{n}}. \tag{107}$$

Finally, combining (100), (101), and (107), we see that under the scaling $R\sqrt{\frac{\log p}{n}} \precsim 1$, we have

$$\begin{aligned} &\frac{1}{\|\Delta\|_2^2} \cdot \frac{1}{n}\sum_{i=1}^n \varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta) \\ &\qquad \ge \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{4} - \left(\frac{c'\tau'^2 R\sigma_x}{T}\sqrt{\frac{\log p}{n}} + c'\tau'\delta\sigma_x\sqrt{\frac{\log p}{n}}\right) \\ &\qquad \ge \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{8} - c'\tau'\delta\sigma_x\sqrt{\frac{\log p}{n}}, \end{aligned} \tag{108}$$

uniformly over all $(\beta, \Delta) \in \mathbb{A}_\delta$, with probability at least $1 - c_1 \exp(-c_2 n)$.

It remains to extend this bound to one that is uniform in the ratio $\frac{\|\Delta\|_1}{\|\Delta\|_2}$, which we do via a peeling argument (Alexander, 1987; van de Geer, 2000). Consider the inequality

$$\frac{1}{\|\Delta\|_2^2} \cdot \frac{1}{n}\sum_{i=1}^n \varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta) \ge \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{8} - 2c'\tau'\sigma_x \frac{\|\Delta\|_1}{\|\Delta\|_2}\sqrt{\frac{\log p}{n}}, \tag{109}$$

as well as the event

$$\mathcal{E} := \left\{\text{Inequality (109) holds } \forall \|\beta\|_2 \le 3 \text{ and } \frac{\|\Delta\|_1}{\|\Delta\|_2} \le \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{16c'\tau\sigma_x}\sqrt{\frac{n}{\log p}}\right\}.$$

Define the function

$$f(\beta, \Delta; X) := \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{8} - \frac{1}{\|\Delta\|_2^2} \cdot \frac{1}{n}\sum_{i=1}^n \varphi_{\tau'\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta), \tag{110}$$

along with

$$g(\delta) := c'\tau'\sigma_x\delta\sqrt{\frac{\log p}{n}}, \qquad \text{and} \qquad h(\beta, \Delta) := \frac{\|\Delta\|_1}{\|\Delta\|_2}.$$

Note that (108) implies

$$\mathbb{P}\left(\sup_{h(\beta,\Delta)\le\delta} f(\beta,\Delta;X) \ge g(\delta)\right) \le c_1\exp(-c_2 n), \quad \text{for any } \delta > 0, \tag{111}$$

where the sup is also restricted to $\{(\beta,\Delta) : \beta \in \mathbb{B}_2(3)\cap\mathbb{B}_1(R),\ \Delta\in\mathbb{B}_2(3)\}$.

Since $\frac{\|\Delta\|_1}{\|\Delta\|_2} \ge 1$, we have

$$1 \le\ h(\beta,\Delta)\ \le \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{16c'\tau'\sigma_x}\sqrt{\frac{n}{\log p}}, \tag{112}$$

over the region of interest. For each integer $m \ge 1$, define the set

$$\mathbb{V}_m := \left\{(\beta,\Delta)\ \mid\ 2^{m-1}\mu \le g(h(\beta,\Delta)) \le 2^m\mu\right\},$$

where $\mu = c'\tau'\sigma_x\sqrt{\frac{\log p}{n}}$. By a union bound, we then have

$$\mathbb{P}(\mathcal{E}^c) \le \sum_{m=1}^{M}\mathbb{P}\left(\exists(\beta,\Delta)\in\mathbb{V}_m \text{ s.t. } f(\beta,\Delta;X) \ge 2g(h(\beta,\Delta))\right),$$

where the index m ranges up to $M := \left\lceil \log\left(c\sqrt{\frac{n}{\log p}}\right)\right\rceil$ over the relevant region (112). By the definition (110) of f, we have

$$\mathbb{P}(\mathcal{E}^c) \le \sum_{m=1}^{M}\mathbb{P}\left(\sup_{h(\beta,\Delta)\le g^{-1}(2^m\mu)} f(\beta,\Delta;X) \ge 2^m\mu\right) \overset{(i)}{\le} M\cdot c_1\exp(-c_2 n),$$

where inequality (i) applies the tail bound (111). It follows that

$$\mathbb{P}(\mathcal{E}^c) \le c_1\exp\left(-c_2 n + \log\log\left(\frac{n}{\log p}\right)\right) \le c_1'\exp\left(-c_2' n\right).$$

Multiplying through by $\|\Delta\|_2^2$ then yields the desired result.

Appendix E. Auxiliary Results

In this section, we provide some auxiliary results that are useful for our proofs. The first lemma concerns symmetrization and desymmetrization of empirical processes via Rademacher random variables:

Lemma 12 (Lemma 2.3.6 in van der Vaart and Wellner, 1996) *Let $\{Z_i\}_{i=1}^n$ be independent zero-mean stochastic processes. Then*

$$\frac{1}{2}\mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n \epsilon_i Z_i(t_i)\right|\right] \le \mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n Z_i(t_i)\right|\right] \le 2\mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n \epsilon_i(Z_i(t_i)-\mu_i)\right|\right],$$

where the ϵ_i's are independent Rademacher variables and the functions $\mu_i : \mathcal{F}\to\mathbb{R}$ are arbitrary.

We also have a useful lemma that bounds the Gaussian complexity in terms of the Rademacher complexity:

Lemma 13 (Lemma 4.5 in Ledoux and Talagrand, 1991) *Let $Z_1, \ldots, Z_n$ be independent stochastic processes. Then*

$$\mathbb{E}\left[\sup_{t \in T}\left|\sum_{i=1}^n \epsilon_i Z_i(t_i)\right|\right] \le \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}\left[\sup_{t \in T}\left|\sum_{i=1}^n g_i Z_i(t_i)\right|\right],$$

where the ϵ_i's are Rademacher variables and the g_i's are standard normal.

We next state a version of the Sudakov-Fernique comparison inequality:

Lemma 14 (Corollary 3.14 in Ledoux and Talagrand, 1991) *Given a countable index set T, let $\{X(t), t \in T\}$ and $\{Y(t), t \in T\}$ be centered Gaussian processes such that*

$$\text{var}\left(Y(s) - Y(t)\right) \le \text{var}\left(X(s) - X(t)\right), \qquad \forall (s,t) \in T \times T.$$

Then

$$\mathbb{E}\left[\sup_{t \in T} Y(t)\right] \le 2 \cdot \mathbb{E}\left[\sup_{t \in T} X(t)\right].$$

A zero-mean random variable Z is sub-Gaussian with parameter σ if $\mathbb{P}(Z > t) \le \exp(-\frac{t^2}{2\sigma^2})$ for all $t \ge 0$. The next lemma provides a standard bound on the expected maximum of N such variables (cf. Equation 3.6 in Ledoux and Talagrand, 1991):

Lemma 15 *Suppose $X_1, \ldots, X_N$ are zero-mean sub-Gaussian random variables such that $\max_{j=1,\ldots,N} \|X_j\|_{\psi_2} \le \sigma$. Then $\mathbb{E}\left[\max_{j=1,\ldots,p} |X_j|\right] \le c_0\, \sigma \sqrt{\log N}$, where $c_0 > 0$ is a universal constant.*

We also have a lemma about maxima of products of sub-Gaussian variables:

Lemma 16 *Suppose $\{g_i\}_{i=1}^n$ are i.i.d. standard Gaussians and $\{X_i\}_{i=1}^n \subseteq \mathbb{R}^p$ are i.i.d. sub-Gaussian vectors with parameter bounded by σ_x. Then as long as $n \ge c\sqrt{\log p}$ for some constant $c > 0$, we have*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i X_i\right\|_\infty\right] \le c' \sigma_x \sqrt{\frac{\log p}{n}}.$$

Proof Conditioned on $\{X_i\}_{i=1}^n$, for each $j = 1, \ldots, p$, the variable $\left|\frac{1}{n}\sum_{i=1}^n g_i X_{ij}\right|$ is zero-mean and sub-Gaussian with parameter bounded by $\frac{\sigma_x}{n}\sqrt{\sum_{i=1}^n X_{ij}^2}$. Hence, by Lemma 15, we have

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i X_i\right\|_\infty \middle| X\right] \le \frac{c_0 \sigma_x}{n} \cdot \max_{j=1,\ldots,p} \sqrt{\sum_{i=1}^n X_{ij}^2} \cdot \sqrt{\log p},$$

implying that

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i X_i\right\|_\infty\right] \le c_0\sigma_x\sqrt{\frac{\log p}{n}} \cdot \mathbb{E}\left[\max_j \sqrt{\frac{\sum_{i=1}^n X_{ij}^2}{n}}\right]. \tag{113}$$

Furthermore, $Z_j := \frac{\sum_{i=1}^n X_{ij}^2}{n}$ is an i.i.d. average of subexponential variables, each with parameter bounded by $c\sigma_x$. Since $\mathbb{E}[Z_j] \le 2\sigma_x^2$, we have

$$\mathbb{P}\left(Z_j - \mathbb{E}[Z_j] \ge u + 2\sigma_x^2\right) \le c_1 \exp\left(-\frac{c_2 nu}{\sigma_x}\right), \qquad \forall u \ge 0 \text{ and } 1 \le j \le p. \tag{114}$$

Now fix some $t \ge \sqrt{2\sigma_x^2}$. Since the $\{Z_j\}_{j=1}^p$ are all nonnegative, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{j=1,\dots,p}\sqrt{Z_j}\right] &\le t + \int_t^\infty \mathbb{P}\left(\max_{j=1,\dots,p}\sqrt{Z_j} > s\right) ds \\
&\le t + \sum_{j=1}^p \int_t^\infty \mathbb{P}\left(\sqrt{Z_j} > s\right) ds \\
&\le t + c_1 p \int_t^\infty \exp\left(-\frac{c_2 n(s^2 - 2\sigma_x^2)}{\sigma_x}\right) ds
\end{aligned}$$

where the final inequality follows from the bound (114) with $u = s^2 - 2\sigma_x^2$, valid as long as $s^2 \ge t^2 \ge 2\sigma_x^2$. Integrating, we have the bound

$$\mathbb{E}\left[\max_{j=1,\dots,p}\sqrt{Z_j}\right] \le t + c_1' p\sigma_x \exp\left(-\frac{c_2' n(t^2 - 2\sigma_x^2)}{\sigma_x^2}\right).$$

Since $n \gtrsim \sqrt{\log p}$ by assumption, setting t equal to a constant implies $\mathbb{E}\left[\max_j \sqrt{Z_j}\right] = \mathcal{O}(1)$, which combined with (113) gives the desired result. ■

Appendix F. Capped-ℓ_1 Penalty

In this section, we show how our results on nonconvex but subdifferentiable regularizers may be extended to include certain types of more complicated regularizers that do not possess (sub)gradients everywhere, such as the capped-ℓ_1 penalty.

In order to handle the case when ρ_λ has points where neither a gradient nor subderivative exists, we assume the existence of a function $\widetilde{\rho}_\lambda$ (possibly defined according to the particular local optimum $\widetilde{\beta}$ of interest), such that the following conditions hold:

Assumption 2

(i) The function $\widetilde{\rho}_\lambda$ is differentiable/subdifferentiable everywhere, and $\|\nabla\widetilde{\rho}_\lambda(\widetilde{\beta})\|_\infty \le \lambda L$.

(ii) For all $\beta \in \mathbb{R}^p$, we have $\widetilde{\rho}_\lambda(\beta) \ge \rho_\lambda(\beta)$.

(iii) The equality $\widetilde{\rho}_\lambda(\widetilde{\beta}) = \rho_\lambda(\widetilde{\beta})$ holds.

(iv) There exists $\mu_1 \geq 0$ such that $\widetilde{\rho}_\lambda(\beta) + \frac{\mu_1}{2}\|\beta\|_2^2$ is convex.

(v) For some index set A with $|A| \leq k$ and some parameter $\mu_2 \geq 0$, we have

$$\widetilde{\rho}_\lambda(\beta^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}) \leq \lambda L\|\widetilde{\beta}_A - \beta^*_A\|_1 - \lambda L\|\widetilde{\beta}_{A^c} - \beta^*_{A^c}\|_1 + \frac{\mu_2}{2}\|\widetilde{\beta} - \beta^*\|_2^2.$$

In addition, we assume conditions (i)–(iii) of Assumption 1 in Section 2.2 above.

When $\rho_\lambda(\beta) + \frac{\mu_1}{2}\|\beta\|_2^2$ is convex for some $\mu_1 \geq 0$ (as in the case of SCAD or MCP), we may take $\widetilde{\rho}_\lambda = \rho_\lambda$ and $\mu_2 = 0$ (cf. Lemma 5 in Appendix A.1). When no such convexification of ρ_λ exists (as in the case of the capped-ℓ_1 penalty), we instead construct a separate convex function $\widetilde{\rho}_\lambda$ to upper-bound ρ_λ and take $\mu_1 = 0$.

Under the conditions of Assumption 2, we have the following variant of Theorems 1 and 2:

Theorem 4 *Suppose $\mathcal{L}_n$ satisfies the RSC conditions* (4)*, and the functions ρ_λ and $\widetilde{\rho}_\lambda$ satisfy Assumption 1 and Assumption 2, respectively. Suppose λ is chosen according to the bound* (6) *and $n \geq \frac{16R^2 \max(\tau_1^2, \tau_2^2)}{\alpha_2^2} \log p$. Then for any stationary point $\widetilde{\beta}$ of the program* (1)*, we have*

$$\|\widetilde{\beta} - \beta^*\|_2 \leq \frac{7\lambda L\sqrt{k}}{4\alpha_1 - 2\mu_1 - 2\mu_2}, \qquad \text{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \leq \frac{28\lambda Lk}{2\alpha_1 - \mu_1 - \mu_2},$$

along with the prediction error bound

$$\langle \nabla\mathcal{L}_n(\widetilde{\beta}) - \nabla\mathcal{L}_n(\beta^*),\, \widetilde{\nu}\rangle \leq \lambda^2 L^2 k \left(\frac{21}{8\alpha_1 - 4\mu_1 - 4\mu_2)} + \frac{49(\mu_1 + \mu_2)}{8(2\alpha_1 - \mu_1 - \mu_2)^2}\right).$$

Proof

The proof is essentially the same as the proofs of Theorems 1 and 2, so we only mention a few key modifications here. First note that any local minimum $\widetilde{\beta}$ of the program (1) is a local minimum of $\mathcal{L}_n + \widetilde{\rho}_\lambda$, since

$$\mathcal{L}_n(\widetilde{\beta}) + \widetilde{\rho}_\lambda(\widetilde{\beta}) = \mathcal{L}_n(\widetilde{\beta}) + \rho_\lambda(\widetilde{\beta}) \leq \mathcal{L}_n(\beta) + \rho_\lambda(\beta) \leq \mathcal{L}_n(\beta) + \widetilde{\rho}_\lambda(\beta),$$

locally for all β in the constraint set, where the first inequality comes from the fact that $\widetilde{\beta}$ is a local minimum of $\mathcal{L}_n + \rho_\lambda$, and the second inequality holds because $\widetilde{\rho}_\lambda$ upper-bounds ρ_λ. Hence, the first-order condition (5) still holds with ρ_λ replaced by $\widetilde{\rho}_\lambda$. Consequently, (20) holds, as well.

Next, note that (22) holds as before, with ρ_λ replaced by $\widetilde{\rho}_\lambda$ and μ replaced by μ_1. By condition (v) on $\widetilde{\rho}_\lambda$, we then have (27) with μ replaced by $\mu_1 + \mu_2$. The remainder of the proof is essentially the same as before. Note that condition (v) does not include the extra factor of ξ appearing in Lemma 5, but this condition is actually strong enough for the proof arguments to hold, since we do not impose a positivity condition on the difference between $\|\widetilde{\nu}_A\|_1$ and $\|\widetilde{\nu}_{A^c}\|_1$. ■

Specializing now to the case of the capped-ℓ_1 penalty, we have the following lemma. For a fixed parameter $c \geq 1$, the capped-ℓ_1 penalty (Zhang and Zhang, 2012) is given by

$$\rho_\lambda(t) := \min\left\{\frac{\lambda^2 c}{2},\ \lambda|t|\right\}. \tag{115}$$

Lemma 17 *The capped-ℓ_1 regularizer* (115) *with parameter c satisfies the conditions of Assumption 2, with $\mu_1 = 0$, $\mu_2 = \frac{4}{c}$, and $L = 1$.*

Proof We will show how to construct an appropriate choice of $\widetilde{\rho}_\lambda$. Note that ρ_λ is piecewise linear and locally equal to $|t|$ in the range $\left[-\frac{\lambda c}{2}, \frac{\lambda c}{2}\right]$, and takes on a constant value outside that region. However, ρ_λ does not have either a gradient or subgradient at $t = \pm\frac{\lambda c}{2}$, hence is not "convexifiable" by adding a squared-ℓ_2 term.

For a fixed local optimum $\widetilde{\beta}$, we define the functions $\widetilde{\rho}_\lambda^j : \mathbb{R} \to \mathbb{R}$ via

$$\widetilde{\rho}_\lambda^j(t) = \begin{cases} \lambda|t|, & \text{if } |\widetilde{\beta}_j| \le \frac{\lambda c}{2}, \\ \frac{\lambda^2 c}{2}, & \text{otherwise,} \end{cases}$$

and let $\widetilde{\rho}_\lambda(\beta) = \sum_{j=1}^p \widetilde{\rho}_\lambda^j(\beta_j)$, for $\beta \in \mathbb{R}^p$. Then

$$\widetilde{\rho}_\lambda(\beta) = \sum_{j \in T} \lambda|\beta_j| + \sum_{j \in T^c} \frac{\lambda^2 c}{2},$$

where $T := \left\{ j \mid |\widetilde{\beta}_j| \le \frac{\lambda c}{2} \right\}$. It is easy to see that $\widetilde{\rho}_\lambda$ is a convex upper bound on ρ_λ, with $\widetilde{\rho}_\lambda(\widetilde{\beta}) = \rho_\lambda(\widetilde{\beta})$, since $\widetilde{\rho}_\lambda^j(\widetilde{\beta}_j) = \rho_\lambda(\widetilde{\beta}_j)$ for all j. Then

$$\widetilde{\rho}_\lambda(\beta^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}) \;=\; \sum_{j \in S \cap T} \left(\widetilde{\rho}_\lambda^j(\beta_j^*) - \widetilde{\rho}_\lambda^j(\widetilde{\beta}_j) \right) + \sum_{j \in S^c \cap T} \left(\widetilde{\rho}_\lambda^j(\beta_j^*) - \widetilde{\rho}_\lambda^j(\widetilde{\beta}_j) \right), \tag{116}$$

using decomposability of $\widetilde{\rho}_\lambda$. Furthermore, $\widetilde{\rho}_\lambda^j(\beta_j^*) = \widetilde{\rho}_\lambda^j(0) = 0$ for $j \in S^c \cap T$, and for $j \in T$, we have

$$\widetilde{\rho}_\lambda^j(\beta_j^*) - \widetilde{\rho}_\lambda^j(\widetilde{\beta}_j) \le \lambda|\beta_j^*| - \lambda|\widetilde{\beta}_j| \le \lambda|\widetilde{\nu}_j|,$$

whereas for $j \notin T$, we have $\widetilde{\rho}_\lambda^j(\beta_j^*) - \widetilde{\rho}_\lambda^j(\widetilde{\beta}_j) = 0 \le \lambda|\widetilde{\nu}_j|$. Combined with (116), we obtain

$$\begin{aligned} \widetilde{\rho}_\lambda(\beta^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}) &\le \sum_{j \in S \cap T} \lambda|\widetilde{\nu}_j| - \sum_{j \in S^c \cap T} \widetilde{\rho}_\lambda^j(\widetilde{\beta}_j) \\ &= \lambda\|\widetilde{\nu}_{S \cap T}\|_1 - \sum_{j \in S^c \cap T} \rho_\lambda(\widetilde{\beta}_j) \\ &= \lambda\|\widetilde{\nu}_{S \cap T}\|_1 - \lambda\|\widetilde{\nu}_{S^c \cap T}\|_1 + \sum_{j \in S^c \cap T} \left(\lambda|\widetilde{\beta}_j| - \rho_\lambda(\widetilde{\beta}_j) \right). \end{aligned} \tag{117}$$

Now observe that

$$\lambda|t| - \rho_\lambda(t) = \begin{cases} 0, & \text{if } \; |t| \le \frac{\lambda c}{2}, \\ \lambda|t| - \frac{\lambda^2 c}{2}, & \text{if } \; |t| > \frac{\lambda c}{2}, \end{cases}$$

and moreover, the derivative of $\frac{t^2}{c}$ always exceeds λ for $|t| > \frac{\lambda c}{2}$. Consequently, we have $\lambda|t| - \rho_\lambda(t) \le \frac{t^2}{c}$ for all $t \in \mathbb{R}$. Substituting this bound into (117) yields

$$\widetilde{\rho}_\lambda(\beta^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}) \le \lambda\|\widetilde{\nu}_{S \cap T}\|_1 - \lambda\|\widetilde{\nu}_{S^c \cap T}\|_1 + \frac{1}{c}\|\widetilde{\nu}_{S^c \cap T}\|_2^2 \le \lambda\|\widetilde{\nu}_S\|_1 - \lambda\|\widetilde{\nu}_{S^c \cap T}\|_1 + \frac{1}{c}\|\widetilde{\nu}_{S^c \cap T}\|_2^2. \tag{118}$$

Finally, note that

$$\lambda \|\widetilde{\nu}_{S^c \cap T^c}\|_1 = \lambda \|\widetilde{\beta}_{S^c \cap T^c}\|_1 \le \frac{2}{c} \|\widetilde{\beta}_{S^c \cap T^c}\|_2^2 = \frac{2}{c} \|\widetilde{\nu}_{S^c \cap T^c}\|_2^2, \tag{119}$$

since $\lambda |t| \le \frac{2t^2}{c}$ when $|t| \ge \frac{\lambda c}{2}$. Combining (118) and (119) then yields

$$\widetilde{\rho}_\lambda(\beta^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}) \le \lambda \|\widetilde{\nu}_S\|_1 - \lambda \|\widetilde{\nu}_{S^c}\|_1 + \frac{2}{c} \|\widetilde{\nu}_{S^c}\|_2^2 \le \lambda \|\widetilde{\nu}_S\|_1 - \lambda \|\widetilde{\nu}_{S^c}\|_1 + \frac{2}{c} \|\widetilde{\nu}\|_2^2,$$

which is condition (v) of Assumption 2 on $\widetilde{\rho}_\lambda$ with $L = 1$, $A = S$, and $\mu_2 = \frac{4}{c}$. The remaining conditions are easy to verify (see also Zhang and Zhang, 2012). ■

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40(5):2452–2482, 2012.

K. S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1999.

P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1):232–253, 2011.

R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement Error in Nonlinear Models.* Chapman and Hall, 1995.

L. Chen and Y. Gu. The convergence guarantees of a non-convex approach for sparse recovery. *IEEE Transactions on Signal Processing*, 62(15):3754–3767, 2014.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics*, pages 521–541, 2009.

J. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849, 06 2014.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, July 2008.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

D. R. Hunter and R. Li. Variable selection using MM algorithms. *Annals of Statistics*, 33 (4):1617–1642, 2005.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Verlag, 1998.

P. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40(3):1637–1664, 2012.

P. Loh and M. J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6):3022–3049, 12 2013a.

P. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *arXiv e-prints*, May 2013b. Available at `http://arxiv.org/abs/1305.2436`.

P. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *NIPS*, pages 476–484, 2013c.

R. Mazumder, J. H. Friedman, and T. Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

P. McCullagh and J. A. Nelder. *Generalized Linear Models (Second Edition)*. London: Chapman & Hall, 1989.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, December 2012. See arXiv version for lemma/propositions cited here.

Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Universit Catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007. URL `http://EconPapers.repec.org/RePEc:cor:louvco:2007076`.

Y. Nesterov and A. Nemirovskii. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM studies in applied and numerical mathematics. Society for Industrial and Applied Mathematics, 1987.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.

M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *Annals of Statistics*, 38:2620–2651, 2010.

A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

S. A. Vavasis. Complexity issues in global optimization: A survey. In *Handbook of Global Optimization*, pages 27–41. Kluwer, 1995.

J.-P. Vial. Strong convexity of sets and functions. *Journal of Mathematical Economics*, 9 (1-2):187–205, January 1982.

M. J. Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1(1):233–253, 2014.

Z. Wang, H. Liu, and T. Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, 42(6):2164–2201, 12 2014.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010.

C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.