Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima

Po-Ling Loh[1] Martin J. Wainwright[1,2]
ploh@berkeley.edu wainwrig@stat.berkeley.edu
Department of Statistics[1] Department of EECS[2] UC Berkeley, Berkeley, CA 94720

May 2013

Abstract

We establish theoretical results concerning all local optima of various regularized M-estimators, where both loss and penalty functions are allowed to be nonconvex. Our results show that as long as the loss function satisfies restricted strong convexity and the penalty function satisfies suitable regularity conditions, *any local optimum* of the composite objective function lies within statistical precision of the true parameter vector. Our theory covers a broad class of nonconvex objective functions, including corrected versions of the Lasso for error-in-variables linear models; regression in generalized linear models using nonconvex regularizers such as SCAD and MCP; and graph and inverse covariance matrix estimation. On the optimization side, we show that a simple adaptation of composite gradient descent may be used to compute a global optimum up to the statistical precision ϵ_{stat} in $\log(1/\epsilon_{\text{stat}})$ iterations, which is the fastest possible rate of any first-order method. We provide a variety of simulations to illustrate the sharpness of our theoretical predictions.

1 Introduction

The problem of optimizing a nonconvex function is known to be computationally intractable in general [18, 24]. Unlike convex functions, nonconvex functions may possess local optima that are not global optima, and standard iterative methods such as gradient descent and coordinate descent are only guaranteed to converge to a *local* optimum. Unfortunately, statistical results regarding nonconvex M-estimation often provide guarantees about the accuracy of *global* optima. Computing such global optima—or even a local optimum that is suitably "close" to a global optimum—may be extremely difficult in practice, which leaves a significant gap in the theory.

However, nonconvex functions arising from statistical estimation problems are often not constructed in an adversarial manner, leading to the natural intuition that the behavior of such functions might be "better" than predicted by worst-case theory. Recent work [13] has confirmed this intuition in one very specific case: a modified version of the Lasso designed for error-in-variables regression. Although the Hessian of this modified Lasso objective always has a large number of negative eigenvalues in the high-dimensional setting (hence is nonconvex), it nonetheless resembles a strongly convex function when restricted to a cone set, leading to provable bounds on statistical and optimization error.

In this paper, we study the question of whether it is possible to certify "good behavior," in both a statistical and computational sense, for nonconvex M-estimators. On the statistical level, we provide an abstract result, applicable to a broad class of (potentially nonconvex) M-estimators, which bounds the distance between and *any local optimum* and the unique minimum of the population risk. Although local optima of nonconvex objectives may not

coincide with global optima, our theory shows that any local optimum is essentially as good as a global optimum from a statistical perspective. The class of M-estimators covered by our theory includes the modified Lasso as a special case, but our results concerning local optima are based on a much simpler argument than the arguments used to establish similar results in previous work [13].

In addition to nonconvex loss functions, our theory also applies to nonconvex regularizers, thereby shedding new light on a long line of recent work involving the nonconvex SCAD and MCP regularizers [6, 4, 26, 27]. Various methods have been proposed for optimizing convex loss functions with nonconvex penalties, including local quadratic approximation (LQA) [6], minorization-maximization (MM) [10], and local linear approximation (LLA) [28]. However, these methods are only guaranteed to generate local optima of the composite objective, which have not been proven to be well-behaved. More recently, Zhang and Zhang [27] provided statistical guarantees concerning global optima of least-squares linear regression with various nonconvex penalties, and proposed that gradient descent initialized at a Lasso optimum could be used to obtain specific local minima. In the same spirit, Fan et al. [7] showed that if the LLA algorithm is initialized at a Lasso optimum that satisfies certain properties, then the two-stage procedure produces an oracle solution for various nonconvex penalties. For a more complete overview of existing work, we refer the reader to the survey paper by Zhang and Zhang [27] and the references cited therein.

In contrast to these previous results, our work provides a set of regularity conditions under which *all local/global optima* are guaranteed to lie within a small ball of the population-level minimum, which ensures that standard methods such as projected and composite gradient descent [17] are sufficient for obtaining estimators that lie within statistical error of the truth. This eliminates the need to design specialized optimization algorithms that will locate specific local optima, as prescribed by previous authors. In fact, we establish that under suitable conditions, a modified form of composite gradient descent only requires $\log(1/\epsilon_{\text{stat}})$ iterations to obtain a solution that is accurate up to statistical precision ϵ_{stat}. Furthermore, our methods are not restricted to least-squares or even convex loss functions, and cover various nonconvex loss functions, as well.

Notation. For functions $f(n)$ and $g(n)$, we write $f(n) \precsim g(n)$ to mean that $f(n) \leq cg(n)$ for some universal constant $c \in (0, \infty)$, and similarly, $f(n) \succsim g(n)$ when $f(n) \geq c'g(n)$ 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 $S \subseteq \{1, \ldots, p\}$, 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. Finally, for a function $h : \mathbb{R}^p \to \mathbb{R}$, we write ∇h to denote a gradient or subgradient, if it exists.

2 Problem formulation

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

2.1 Background

Given a collection of n samples $Z_1^n = \{Z_1, \ldots, Z_n\}$, 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

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

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. 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 $g(\beta) \geq \|\beta\|_1$, for all $\beta \in \mathbb{R}^p$. Consequently, any feasible point for the optimization problem (1) satisfies the constraint $\|\beta\|_1 \leq R$, 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, we allow for an additional side constraint $\beta \in \Omega$, 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 the conditions imposed 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 positive real line, the function ρ_λ is nondecreasing and subadditive, meaning $\rho_\lambda(s+t) \leq \rho_\lambda(s) + \rho_\lambda(t)$ for all $s, t \geq 0$.

(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 subgradients at $t = 0$ bounded by λL.

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

Conditions (i)–(iii) were previously proposed in Zhang and Zhang [27] and are satisfied for a variety of regularizers, including the usual ℓ_1-norm and nonconvex regularizers such as SCAD, MCP, and capped-ℓ_1. However, conditions (iv)–(v) exclude the capped-ℓ_1 penalty; for details on how a modified version of our arguments may be used to analyze capped-ℓ_1, see Appendix F. Note that condition (v) is a type of curvature constraint that controls the level of nonconvexity of ρ_λ.

Many types of regularizers that are relevant in practice satisfy Assumption 1. For instance, the usual ℓ_1-norm, $\rho_\lambda(\beta) = \|\beta\|_1$, satisfies the 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 [6], 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 7 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 [26], takes the form

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

where $b > 0$ is a fixed parameter. As verified in Lemma 8 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 [16, 1], 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:

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

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 $\alpha_1 = \alpha_2 = \alpha > 0$ and $\tau_1 = \tau_2 = 0$. However, in the high-dimensional setting ($p \gg n$), the empirical loss $\mathcal{L}_n$ can never be strongly convex, but the RSC condition may still hold with strictly positive (α_j, τ_j). On the other hand, if $\mathcal{L}_n$ is convex (but not strongly convex), the left-hand expression in inequality (4) is always nonnegative, so

inequalities (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 set 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; past work [13] exhibits a large family of nonconvex quadratic functions that satisfy this condition (see Section 3.2 below for further discussion). Finally, note that we have stated two separate RSC inequalities (4), unlike in past work [16, 1, 13], which only imposes the first condition (4a) over the entire range of Δ. As illustrated in the corollaries of Sections 3.3 and 3.4 below, the first inequality (4a) can only hold locally over Δ for more complicated types of functions; in contrast, as proven in Appendix B.1, inequality (4b) is implied by inequality (4a) in cases where $\mathcal{L}_n$ is convex.

3 Statistical guarantee and consequences

With this setup, we now turn to the statement and proof of our main statistical guarantee, 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:

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

3.1 Main statistical result

Our main theorem is deterministic in nature, specifying conditions on the regularizer, loss function, and parameters, which guarantee that any local optimum $\widetilde{\beta}$ lies close to the target vector $\beta^* = \arg\min_{\beta \in \in \mathbb{R}^p} \mathcal{L}(\beta)$.

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

$$\frac{2}{L} \cdot \max\left\{\|\nabla \mathcal{L}_n(\beta^*)\|_\infty,\, \alpha_2\sqrt{\frac{\log p}{n}}\right\} \;\leq\; \lambda \;\leq\; \frac{\alpha_2}{6RL}, \tag{6}$$

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

$$\|\widetilde{\beta} - \beta^*\|_2 \leq \frac{7\lambda\sqrt{k}}{4(\alpha_1 - \mu)}, \qquad \text{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \leq \frac{63\lambda k}{4(\alpha_1 - \mu)}, \tag{7}$$

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

From the bound (7), note that the squared ℓ_2-error grows proportionally with k, the number of non-zeros in the target parameter, and with λ^2. As will be clarified momentarily, 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.

We stress that the statement Theorem 1 is entirely deterministic. Corresponding probabilistic results will be derived in subsequent sections, where we establish that, with appropriate choices of (λ, R), the required conditions hold w.h.p. In particular, applying the theorem to a particular model requires bounding the random quantity $\|\mathcal{L}_n(\beta^*)\|_\infty$ and verifying the RSC condition (4).

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

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle \geq \alpha_2 \|\widetilde{\nu}\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\nu}\|_1. \tag{8}$$

Since β^* is feasible, we may take $\beta = \beta^*$ in inequality (5), and combining with inequality (8) yields

$$\langle \nabla \rho_\lambda(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle \geq \alpha_2 \|\widetilde{\nu}\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\nu}\|_1. \tag{9}$$

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

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

where inequality (i) follows since $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \leq \frac{\lambda L}{2}$ by the bound (6), and $\|\nabla \rho_\lambda(\widetilde{\beta})\|_\infty \leq \lambda L$ by Lemma 5 in Appendix A.1. Combining this upper bound with inequality (9) and rearranging then yields

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

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

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

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle \geq \alpha_1 \|\widetilde{\nu}\|_2^2 - \tau_1 \frac{\log p}{n} \|\widetilde{\nu}\|_1^2. \tag{10}$$

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

$$\rho_{\lambda,\mu}(\beta^*) - \rho_{\lambda,\mu}(\widetilde{\beta}) \geq \langle \nabla \rho_{\lambda,\mu}(\widetilde{\beta}), \beta^* - \widetilde{\beta} \rangle \;=\; \langle \nabla \rho_\lambda(\widetilde{\beta}) + 2\mu \widetilde{\beta}, \beta^* - \widetilde{\beta} \rangle,$$

implying that

$$\langle \nabla \rho_\lambda(\widetilde{\beta}), \beta^* - \widetilde{\beta} \rangle \leq \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \mu \|\widetilde{\beta} - \beta^*\|_2^2. \tag{11}$$

Combining inequality (10) with inequalities (5) and (11), we obtain

$$\begin{aligned}\alpha_1\|\widetilde{\nu}\|_2^2 - \tau_1\frac{\log p}{n}\|\widetilde{\nu}\|_1^2 &\leq -\langle\nabla\mathcal{L}_n(\beta^*),\,\widetilde{\nu}\rangle + \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \mu\|\widetilde{\beta}-\beta^*\|_2^2 \\ &\overset{(i)}{\leq} \|\nabla\mathcal{L}_n(\beta^*)\|_\infty\cdot\|\widetilde{\nu}\|_1 + \lambda L\left(\|\widetilde{\nu}_A\|_1 - \|\widetilde{\nu}_{A^c}\|_1\right) + \mu\|\widetilde{\nu}\|_2^2 \\ &\overset{(ii)}{\leq} \frac{3\lambda L}{2}\|\widetilde{\nu}_A\|_1 - \frac{\lambda L}{2}\|\widetilde{\nu}_{A^c}\|_1 + \mu\|\widetilde{\nu}\|_2^2, \end{aligned} \tag{12}$$

where inequality (i) is obtained by applying Hölder's inequality to the first term and Lemma 6 in Appendix A.1 to the middle two terms, and inequality (ii) uses the bound

$$\|\widetilde{\nu}\|_1 \leq \|\widetilde{\nu}_A\|_1 + \|\widetilde{\nu}_{A^c}\|_1.$$

Rearranging inequality (12), we find that

$$\begin{aligned}0 \leq 2(\alpha_1-\mu)\|\widetilde{\nu}\|_2^2 &\leq 3\lambda L\|\widetilde{\nu}_A\|_1 - \lambda L\|\widetilde{\nu}_{A^c}\|_1 + 4R\tau_1\frac{\log p}{n}\|\widetilde{\nu}\|_1 \\ &\leq 3\lambda L\|\widetilde{\nu}_A\|_1 - \lambda L\|\widetilde{\nu}_{A^c}\|_1 + \alpha_2\sqrt{\frac{\log p}{n}}\|\widetilde{\nu}\|_1 \\ &\leq \frac{7\lambda L}{2}\|\widetilde{\nu}_A\|_1 - \frac{\lambda L}{2}\|\widetilde{\nu}_{A^c}\|_1, \end{aligned} \tag{13}$$

implying that $\|\widetilde{\nu}_{A^c}\|_1 \leq 8\|\widetilde{\nu}_A\|_1$. Consequently,

$$\|\widetilde{\nu}\|_1 = \|\widetilde{\nu}_A\|_1 + \|\widetilde{\nu}_{A^c}\|_1 \leq 9\|\widetilde{\nu}_A\|_1 \leq 9\sqrt{k}\|\widetilde{\nu}_A\|_2 \leq 9\sqrt{k}\|\widetilde{\nu}\|_2. \tag{14}$$

Furthermore, inequality (13) implies that

$$2(\alpha_1-\mu)\|\widetilde{\nu}\|_2^2 \leq \frac{7\lambda}{2}\|\widetilde{\nu}_A\|_1 \leq \frac{7\lambda\sqrt{k}}{2}\|\widetilde{\nu}\|_2.$$

Rearranging yields the ℓ_2-bound, whereas the ℓ_1-bound follows from by combining the ℓ_2-bound with the cone inequality (14). □

Remark 1. For convex M-estimators, Negahban et al. [16] have shown that arguments applied to ℓ_1-regularizers may be generalized in a straightforward manner to other types of "decomposable" regularizers, including various types of norms for group sparsity, the nuclear norm for low-rank matrices, etc. In our present setting, where we allow for nonconvexity in the loss and regularizer, Theorem 1 has an analogous generalization.

We now turn to various consequences of Theorem 1 for nonconvex loss functions and regularizers of interest. The main challenge in moving from Theorem 1 to these consequences is to establish that the RSC conditions (4) hold w.h.p. 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

$$y_i = \underbrace{\langle\beta^*,\,x_i\rangle}_{\sum_{j=1}^p\beta_j^*x_{ij}} + \epsilon_i, \qquad \text{for } i=1,\ldots,n, \tag{15}$$

where $\beta^* \in \mathbb{R}^p$ is the unknown parameter vector and $\{(x_i, y_i)\}_{i=1}^n$ are observations. Following Loh and Wainwright [13], 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

$$\mathcal{L}(\beta) = \frac{1}{2}\beta^T \Sigma_x \beta - \beta^{*T} \Sigma_x \beta, \qquad \text{and} \qquad \mathcal{L}_n(\beta) = \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta, \tag{16}$$

where $(\widehat{\Gamma}, \widehat{\gamma})$ are estimators for $(\Sigma_x, \Sigma_x \beta^*)$ depending 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

$$\widehat{\beta} \in \arg \min_{g(\beta) \leq R} \left\{ \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta + \rho_\lambda(\beta) \right\}. \tag{17}$$

We now state a concrete corollary in the case of the additive noise (model (a) above). In this case, as discussed in Loh and Wainwright [13], an appropriate choice of the pair $(\widehat{\Gamma}, \widehat{\gamma})$ is given by

$$\widehat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w, \qquad \text{and} \qquad \widehat{\gamma} = \frac{Z^T y}{n}. \tag{18}$$

In the high-dimensional setting ($p \gg n$), the matrix $\widehat{\Gamma}$ is always negative-definite: the matrix $\frac{Z^T Z}{n}$ has rank at most n, and then the positive definite matrix Σ_w is subtracted to obtain $\widehat{\Gamma}$. Consequently, the empirical loss function $\mathcal{L}_n$ previously defined (16) is nonconvex. Other choices of $\widehat{\Gamma}$ are applicable to the missing data (model (b)), and also lead to nonconvex programs (see the paper [13] 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 x_i's are sub-Gaussian. Suppose (λ, R) are chosen such that β^* is feasible and

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

Then given a sample size $n \geq C \max\{R^2, k\} \log p$, any local optimum $\widetilde{\beta}$ of the nonconvex program (17) satisfies the bounds

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

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

Remark 2. When $\rho_\lambda(\beta) = \lambda\|\beta\|_1$ and $g(\beta) = \|\beta\|_1$, then taking $\lambda \asymp \sqrt{\frac{\log p}{n}}$ and $R = b_0\sqrt{k}$ for some constant $b_0 \geq \|\beta^*\|_2$ yields the required scaling $n \gtrsim k \log p$. Hence, the bounds of Corollary 1 agree with bounds previously established in Theorem 1 of Loh and Wainwright [13].

Note, however, that those results are stated only for a *global minimum* $\widehat{\beta}$ of the program (17), whereas Corollary 1 is a much stronger result holding for *any local minimum* $\widetilde{\beta}$.

Theorem 2 of Loh and Wainwright [13] indirectly implies similar bounds on $\|\widetilde{\beta} - \beta^*\|_1$ and $\|\widetilde{\beta} - \beta^*\|_2$, since the proposed projected gradient descent algorithm may become stuck in a local minimum. 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 (15) are observed cleanly without corruption and ρ_λ is a nonconvex penalty. In that case, the empirical loss $\mathcal{L}_n$ previously defined (16) is equivalent to the least-squares loss, modulo a constant factor. Much existing work, including that of Fan and Li [6] and Zhang and Zhang [27], first establishes statistical consistency results concerning *global* minima of the program (17), then provides specialized algorithms such as a local linear approximation (LLA) for obtaining specific local optima that are provably close to global optima. However, our results show that *any* optimization algorithm guaranteed to converge to a local optimum of the program suffices. See Section 4 for a more detailed discussion of optimization procedures and fast convergence guarantees for obtaining local minima.

Our theory also provides a theoretical justification for why the usual choice of $a = 3.7$ for linear regression with the SCAD penalty [6] 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 $\lambda_{\min}(\Sigma_x) > 2\mu$ for the choice $a = 3.7$. For further comments regarding the parameter a in the SCAD penalty, see the discussion concerning Figure 2 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 [15, 12], 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

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

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

$$\widehat{\beta} \in \arg\min_{g(\beta)\le R} \left\{ \frac{1}{n}\sum_{i=1}^n \big(\psi(x_i^T\beta) - y_i x_i^T\beta\big) + \rho_\lambda(\beta) \right\}. \tag{19}$$

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 local optimum $\widetilde{\beta}$ of the nonconvex program (19) satisfies

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

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

Remark 3. 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 3 in Section 5 for a demonstration of such local optima. In past work, Breheny and Huang [4] 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 local optimum $\widetilde{\beta}$ within close proximity of the true parameter β^*.

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

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

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

$$\widehat{\Theta} \in \arg\min_{g(\Theta) \leq R,\ \Theta \succeq 0} \left\{ \mathrm{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \sum_{j,k=1}^{p} \rho_\lambda(\Theta_{jk}) \right\}, \tag{20}$$

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

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 [3, 8, 21, 25], and more recently for discrete-valued observations, as well [14]. In addition, a version of the graphical Lasso incorporating a nonconvex SCAD penalty has been proposed [5]. Our results subsume previous Frobenius error bounds for the graphical Lasso, and again imply that even in the presence of a nonconvex regularizer, all local optima of the nonconvex program (20) remain close to the true inverse covariance matrix Θ^*.

As suggested by Loh and Wainwright [14], the graphical Lasso easily accommodates systematically corrupted observations, with the only modification being the form of the sample covariance matrix $\widehat{\Sigma}$. Furthermore, the program (20) 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 all entries are penalized equally in order to simplify our arguments. The same framework is considered by Fan et al. [5].

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 (21) holds with probability at least $1 - c_1 \exp(-c_2 \log p)$, translating into the Frobenius norm bound (22) 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

$$\left\| \widehat{\Sigma} - \Sigma \right\|_{\max} \leq c_0 \sqrt{\frac{\log p}{n}}. \tag{21}$$

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

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

Then with a sample size $n > Cs \log p$, for sufficiently large constant $C > 0$, any local optimum $\widetilde{\Theta}$ of the nonconvex program (20) satisfies

$$\left\| \widetilde{\Theta} - \Theta^* \right\|_F \leq \frac{c_0' \lambda \sqrt{s}}{(\|\Theta^*\|_2 + 1)^{-2} - \mu}. \tag{22}$$

When ρ is simply the ℓ_1-penalty, the bound (22) from Corollary 3 matches the minimax rates for Frobenius norm estimation of an s-sparse inverse covariance matrix [21, 20].

4 Optimization algorithms

We now describe how a version of composite gradient descent 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

$$g_{\lambda,\mu}(\beta) := \frac{1}{\lambda}\Big\{\rho_\lambda(\beta) + \mu\|\beta\|_2^2\Big\}, \tag{23}$$

which is convex by Assumption 1. We may then write the program (1) as

$$\widehat{\beta} \in \arg\min_{g_{\lambda,\mu}(\beta)\leq R,\ \beta\in\Omega} \Big\{ \underbrace{\big(\mathcal{L}_n(\beta) - \mu\|\beta\|_2^2\big)}_{\bar{\mathcal{L}}_n} + \lambda g_{\lambda,\mu}(\beta)\Big\}. \tag{24}$$

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 (24) of the objective function, the composite gradient descent procedure of Nesterov [17] produces a sequence of iterates $\{\beta^t\}_{t=0}^{\infty}$ via the updates

$$\beta^{t+1} \in \arg\min_{g_{\lambda,\mu}(\beta)\leq R} \left\{ \frac{1}{2}\left\|\beta - \left(\beta^t - \frac{\nabla\bar{\mathcal{L}}_n(\beta^t)}{\eta}\right)\right\|_2^2 + \frac{\lambda}{\eta} g_{\lambda,\mu}(\beta)\right\}, \tag{25}$$

where $\frac{1}{\eta}$ is the stepsize. 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 (25) 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

$$\mathcal{T}(\beta_1, \beta_2) := \mathcal{L}_n(\beta_1) - \mathcal{L}_n(\beta_2) - \langle \nabla\mathcal{L}_n(\beta_2),\, \beta_1 - \beta_2\rangle. \tag{26}$$

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

$$\overline{\mathcal{T}}(\beta_1, \beta_2) = \mathcal{T}(\beta_1, \beta_2) - \mu\|\beta_1 - \beta_2\|_2^2. \tag{27}$$

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

$$\mathcal{T}(\beta_1, \beta_2) \geq \begin{cases} \alpha_1\|\beta_1 - \beta_2\|_2^2 - \tau_1\dfrac{\log p}{n}\|\beta_1 - \beta_2\|_1^2, & \text{for all } \|\beta_1 - \beta_2\|_2 \leq 3, \quad (28a) \\ \alpha_2\|\beta_1 - \beta_2\|_2 - \tau_2\sqrt{\dfrac{\log p}{n}}\|\beta_1 - \beta_2\|_1, & \text{for all } \|\beta_1 - \beta_2\|_2 \geq 3. \quad (28b) \end{cases}$$

The conditions (28) 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:

$$\mathcal{T}(\beta_1, \beta_2) \leq \alpha_3 \|\beta_1 - \beta_2\|_2^2 + \tau_3 \frac{\log p}{n} \|\beta_1 - \beta_2\|_1^2, \qquad \text{for all } \beta_1, \beta_2 \in \Omega, \tag{29}$$

a condition referred to as *restricted smoothness* in past work [1]. Throughout this section, we assume $\alpha_i > \mu$ for all i, where μ is the coefficient ensuring the convexity of the function $g_{\lambda,\mu}$ from equation (23). 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 $\|\beta^*\|_2 \leq 1$, and under 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 the paper [19] for minimax rates). We show that the composite gradient updates (25) exhibit a type of *globally geometric convergence* in terms of the quantity

$$\kappa := \frac{1 - \frac{\alpha - \mu}{4\eta} + \varphi(n, p, k)}{1 - \varphi(n, p, k)}, \qquad \text{where} \quad \varphi(n, p, k) := \frac{128 \tau k \frac{\log p}{n}}{\alpha - \mu}. \tag{30}$$

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

$$T^*(\delta) := \frac{2 \log\left(\frac{\phi(\beta^0) - \phi(\widehat{\beta})}{\delta^2}\right)}{\log(1/\kappa)} + \left(1 + \frac{\log 2}{\log(1/\kappa)}\right) \log\log\left(\frac{\lambda R L}{\delta^2}\right), \tag{31}$$

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 local optimum. 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:

Theorem 2. Suppose the empirical loss $\mathcal{L}_n$ satisfies the RSC/RSM conditions (28) and (29), and suppose the regularizer ρ_λ satisfies Assumption 1. Suppose $\widehat{\beta}$ is any global minimum of the program (24), with regularization parameters chosen such that

$$R\sqrt{\frac{\log p}{n}} \leq c, \qquad \text{and} \qquad \lambda \geq \frac{4}{L} \cdot \max\left\{ \|\nabla \mathcal{L}_n(\beta^*)\|_\infty,\ \tau\sqrt{\frac{\log p}{n}} \right\}.$$

Then for any stepsize parameter $\eta \geq 2 \cdot \max\{\alpha_3 - \mu, \mu\}$ and tolerance parameter $\delta^2 \geq \frac{c\epsilon_{\text{stat}}^2}{1-\kappa}$, we have

$$\|\beta^t - \widehat{\beta}\|_2^2 \ \leq \ \frac{2}{\alpha - \mu} \left(\delta^2 + \frac{\delta^4}{\tau} + 128 \tau \frac{k \log p}{n} \epsilon_{\text{stat}}^2 \right), \qquad \text{for all iterations } t \geq T^*(\delta). \tag{32}$$

Remark 4. As with Theorem 1, the statement of Theorem 2 is entirely deterministic. In Section 4.4 below, we will establish that the required RSC and RSM conditions hold w.h.p. for various GLMs.

Also note that for the optimal choice of tolerance parameter $\delta \asymp \epsilon_{\text{stat}}$, the error bound appearing in inequality (32) takes the form $\frac{c\epsilon_{\text{stat}}^2}{\alpha-\mu}$, 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 that $\epsilon_{\text{stat}} = \mathcal{O}\left(\sqrt{\frac{k \log p}{n}}\right)$ with high probability. Combined with Theorem 2, 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 $t \geq T(\epsilon_{\text{stat}})$.

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 $\alpha - \mu$. Indeed, the bound is tighter when the loss function possesses more curvature or the penalty function is closer to being convex, agreeing with intuition.

Finally, the parameter η needs to be sufficiently large (or equivalently, the stepsize must be sufficiently small) in order for the composite gradient descent algorithm to be well-behaved. See Nesterov [17] for a discussion of how the stepsize may be chosen via an iterative search when the problem parameters are unknown.

4.2 Form of updates

In this section, we discuss how the updates (25) 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, the next iterate β^{t+1} may be obtained via the following three-step procedure:

(1) First optimize the unconstrained program

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 + \frac{\lambda}{\eta} \cdot g_{\lambda,\mu}(\beta) \right\}. \tag{33}$$

(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

$$\beta^{t+1} \in \arg\min_{g_{\lambda,\mu}(\beta) \leq R} \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{\nabla \mathcal{L}_n(\beta^t)}{\eta} \right) \right\|_2^2 \right\}. \tag{34}$$

We derive the correctness of this procedure in Appendix C.1. For many nonconvex regularizers ρ_λ of interest, the unconstrained program (33) has a convenient closed-form solution:

For the SCAD penalty (2), the program (33) has simple closed-form solution given by

$$\widehat{x}_{\text{SCAD}} = \begin{cases} 0 & \text{if } 0 \leq |z| \leq \nu\lambda, \\ z - \text{sign}(z) \cdot \nu\lambda & \text{if } \nu\lambda \leq |z| \leq (\nu+1)\lambda, \\ \frac{z - \text{sign}(z) \cdot \frac{a\nu\lambda}{a-1}}{1 - \frac{\nu}{a-1}} & \text{if } (\nu+1)\lambda \leq |z| \leq a\lambda, \\ z & \text{if } |z| \geq a\lambda. \end{cases} \tag{35}$$

For the MCP penalty (3), the optimum of the program (33) takes the form

$$\widehat{x}_{\text{MCP}} = \begin{cases} 0 & \text{if } 0 \leq |z| \leq \nu\lambda, \\ \frac{z - \text{sign}(z) \cdot \nu\lambda}{1 - \nu/b} & \text{if } \nu\lambda \leq |z| \leq b\lambda, \\ z & \text{if } |z| \geq b\lambda. \end{cases} \tag{36}$$

In both equations (35) and (36), we have

$$z := \frac{1}{1 + 2\mu/\eta} \left(\beta^t - \frac{\nabla \bar{\mathcal{L}}_n(\beta^t)}{\eta} \right), \qquad \text{and} \qquad \nu := \frac{1/\eta}{1 + 2\mu/\eta}.$$

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 (25) 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 2

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. [1], 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 2, suppose that there exists a pair $(\bar{\eta}, T)$ such that

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \bar{\eta}, \qquad \forall t \geq T. \tag{37}$$

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

$$\|\beta^t - \widehat{\beta}\|_1 \leq 4\sqrt{k}\|\beta^t - \widehat{\beta}\|_2 + 8\sqrt{k}\|\widehat{\beta} - \beta^*\|_2 + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right).$$

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 2, and with an initial vector β^0 such that $\|\beta^0 - \widehat{\beta}\|_2 \leq 3$, we have

$$\|\beta^t - \widehat{\beta}\|_2 \leq 3, \qquad \text{for all } t \geq 0. \tag{38}$$

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

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

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

Finally, recalling our earlier definition (30) 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 inequality (37) holds and $\frac{32k\tau \log p}{n} \leq \frac{\alpha-\mu}{2}$. Then for any $t \geq T$, we have

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \kappa^{t-T}(\phi(\beta^T) - \phi(\widehat{\beta})) + \frac{\xi}{1-\kappa}(\epsilon^2 + \epsilon_{\text{stat}}^2),$$

where $\bar{\epsilon} := 8\sqrt{k}\epsilon_{\text{stat}}$, $\epsilon := 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right)$, the quantities κ and φ are defined according to equations (30), and

$$\xi := \frac{1}{1 - \varphi(n,p,k)} \cdot \frac{2\tau \log p}{n} \cdot \left(\frac{\alpha - \mu}{4\eta} + 2\varphi(n,p,k) + 5\right). \tag{39}$$

The remainder of the proof follows an argument used in Agarwal et al. [1], so we only provide a high-level sketch. We first prove the following inequality:

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \delta^2, \qquad \text{for all } t \geq T^*(\delta), \tag{40}$$

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

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

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

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

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

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

For $\ell \geq 1$, 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

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

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

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

Finally, we use the recursion

$$\bar{\eta}_{\ell+1} \leq \frac{4\xi}{1-\kappa} \max\{\epsilon_\ell^2, \bar{\epsilon}^2\}, \qquad T_\ell \leq \ell + \frac{\log(2^\ell \bar{\eta}_0/\bar{\eta}_\ell)}{\log(1/\kappa)}, \tag{41}$$

to establish the recursion

$$\bar{\eta}_{\ell+1} \leq \frac{\bar{\eta}_\ell}{4^{2^{\ell-1}}}, \qquad \frac{\bar{\eta}_{\ell+1}}{\lambda L} \leq \frac{R}{4^{2^\ell}}. \tag{42}$$

Inequality (40) then follows from computing the number of epochs and timesteps necessary to obtain $\frac{\lambda R L}{4^{2^{\ell-1}}} \leq \delta^2$. For the remaining steps used to obtain inequalities (42) from inequalities (41), we refer the reader to Agarwal et al. [1].

Finally, by inequality (79b) in the proof of Lemma 3 in Appendix C.5 and the relative scaling of (n, p, k), we have

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

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.

4.4 Verifying RSC/RSM conditions

We now address how to establish versions of the RSC conditions (28) and RSM condition (29). In the case of corrected linear regression (Corollary 1), Lemma 13 of Loh and Wainwright [13] establish these conditions w.h.p. for various statistical models. Here, we focus on establishing the conditions for GLMs when the covariates x_i are drawn i.i.d. from a zero-mean sub-Gaussian distribution with parameter σ_x and covariance matrix $\Sigma = \mathrm{cov}(x_i)$. As usual, we assume a sample size $n \geq c\, k \log p$, for a sufficiently large constant $c > 0$. Recall the definition of the Taylor error $\mathcal{T}(\beta_1, \beta_2)$ from equation (26).

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

$$\mathcal{T}(\beta_1, \beta_2) \geq \begin{cases} \frac{\alpha_\ell}{2} \|\Delta\|_2^2 - \frac{c^2 \sigma_x^2}{2\alpha_\ell} \frac{\log p}{n} \|\Delta\|_1^2, & \text{for all } \|\beta_1 - \beta_2\|_2 \leq 3, \qquad (43a) \\ \frac{3\alpha_\ell}{2} \|\Delta\|_2 - 3c\sigma_x \sqrt{\frac{\log p}{n}} \|\Delta\|_1 & \text{for all } \|\beta_1 - \beta_2\|_2 \geq 3, \qquad (43b) \end{cases}$$

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

$$\mathcal{T}(\beta_1, \beta_2) \leq \alpha_u \lambda_{max}(\Sigma) \left(\frac{3}{2} \|\Delta\|_2^2 + \frac{\log p}{n} \|\Delta\|_1^2 \right), \qquad \text{for all } \beta_1, \beta_2 \in \mathbb{R}^p, \tag{44}$$

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

Proof. 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

$$\mathcal{T}(\beta_1, \beta_2) = \frac{1}{n} \sum_{i=1}^{n} \psi''\big(\langle \beta_1, x_i \rangle) + t_i \langle \Delta, x_i \rangle\big) \, (\langle \Delta, x_i \rangle)^2, \tag{45}$$

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 (44): The proof of this upper bound is relatively straightforward given earlier results [14]. From the Taylor series expansion (45) 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 [13]), 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 (43): 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 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 equation (89) below, and take $T = 3\tau$). The core of the proof is based on the following lemma, proved in Appendix D:

Lemma 4. 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

$$\frac{\|\Delta\|_1}{\|\Delta\|_2} \leq \frac{\alpha_\ell}{c\sigma_x} \sqrt{\frac{n}{\log p}}. \tag{46}$$

Taking Lemma 4 as given, we now complete the proof of the RSC condition (43). 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 4 implies that inequality (43a) 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 (46) holds. On the other hand, if the bound (46) does not hold, then the lower bound in inequality (43a) is negative. By convexity of $\mathcal{L}_n$, we have $\mathcal{T}(\beta_1, \beta_2) \geq 0$, so inequality (43a) holds trivially in that case.

We now show that inequality (43b) 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

$$\mathcal{T}(\beta_2 + \Delta, \beta_2) \geq \frac{\mathcal{T}(\beta_2 + t\Delta, \beta_2)}{t}. \tag{47}$$

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 inequality (43a) to obtain

$$\frac{\mathcal{T}(\beta_2 + t\Delta, \beta_2)}{t} \geq \frac{\|\Delta\|_2}{3}\left(\alpha_1\left(\frac{3\|\Delta\|_2}{\|\Delta\|_2}\right)^2 - \tau_1\frac{\log p}{n}\left(\frac{3\|\Delta\|_1}{\|\Delta\|_2}\right)^2\right) = 3\alpha_1\|\Delta\|_2 - 9\tau_1\frac{\log p}{n}\frac{\|\Delta\|_1^2}{\|\Delta\|_2}. \tag{48}$$

Note that inequality (43b) 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, inequalities (47) and (48) 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 (43b). □

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 version 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.

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

$$\widehat{\beta} \in \arg\min_{g_{\lambda,\mu}(\beta) \leq R}\left\{\frac{1}{2}\beta^T\left(\frac{X^TX}{n} - \Sigma_w\right)\beta - \frac{y^TZ}{n}\beta + \rho_\lambda(\beta)\right\}. \tag{49}$$

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 (19) becomes

$$\widehat{\beta} \in \arg \min_{g_{\lambda,\mu}(\beta) \leq R} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{\log(1 + \exp(\langle \beta, x_i \rangle) - y_i \langle \beta, x_i \rangle\} + \rho_\lambda(\beta) \right\}. \tag{50}$$

We optimized the programs (49) and (50) using the composite gradient updates (25). 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 equations (35) and (36). Note that the updates for the Lasso penalty may be generated more simply and efficiently as discussed in Agarwal et al. [1].

Figure 1 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 $\|\beta^*\|_2 = 1$. As predicted by Theorem 1, the three curves corresponding to the same penalty function stack up nicely 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 (49) and (50) are statistically consistent. The Lasso, SCAD, and MCP regularizers are depicted by solid, dotted, and dashed lines, respectively. In the case of linear regression, we set the parameter $a = 3.7$ for the SCAD penalty, suggested by Fan and Li [6] to be "optimal" based on cross-validated empirical studies. We chose $b = 1.5$ for the MCP. The remaining parameters were set as $\lambda = \sqrt{\frac{\log p}{n}}$ and $R = \frac{1.1}{\lambda} \cdot \rho_\lambda(\beta^*)$. Each point represents an average over 100 trials. In the case of logistic regression, we set $a = 3.7$ for SCAD and $b = 3$ for MCP, and took $\lambda = 0.5\sqrt{\frac{\log p}{n}}$ and $R = \frac{2}{\lambda} \cdot \rho_\lambda(\beta^*)$. Each point represents an average over 50 trials.

In Figure 2, we provide the results of simulations to illustrate the optimization-theoretic conclusions of Theorem 2. Each panel shows two different families of curves, corresponding to statistical error (red) and optimization error (blue). Here, the vertical axis measures the ℓ_2-error on a logarithmic scale, while the horizontal axis tracks the iteration number. Within each block, the curves were obtained by running the composite gradient descent algorithm from 10 different initial starting points chosen at random. In all cases, we used the parameter settings $p = 128$, $k = \lfloor \sqrt{p} \rfloor$, and $n = \lfloor 20k \log p \rfloor$, and took $\lambda = \sqrt{\frac{\log p}{n}}$ and $R = \frac{1.1}{\lambda} \rho_\lambda(\beta^*)$. 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 local optima, providing further evidence that this setting suggested by Fan and Li [6] is in some sense optimal.

Finally, Figure 3 provides analogous results to Figure 2 in the case of logistic regression, using $p = 64, k = \lfloor \sqrt{p} \rfloor, n = \lfloor 20k \log p \rfloor$, and regularization parameters $\lambda = 0.5\sqrt{\frac{\log p}{n}}$ and $R = \frac{1.1}{\lambda} \cdot \rho_\lambda(\beta^*)$. The plot shows solution trajectories for 20 different 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 2. Furthermore, the Lasso penalty yields a unique local/optimum $\widehat{\beta}$, since the program (50) is convex, as we observe in panel (a). In contrast, the nonconvex program based on the SCAD penalty produces multiple local

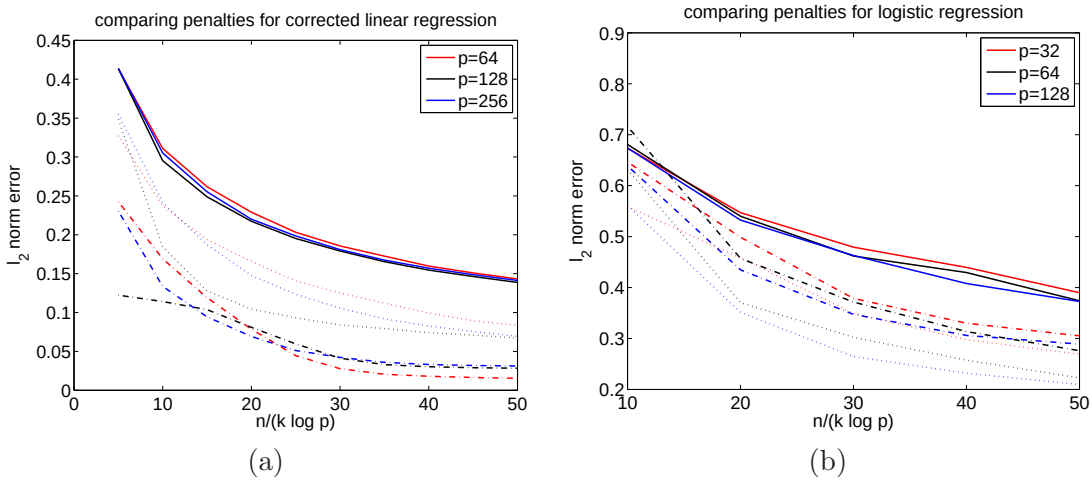

Figure 1. Plots showing statistical consistency of linear and logistic regression with Lasso, SCAD, and MCP regularizers, at sparsity level $k = \lfloor \sqrt{p} \rfloor$. Panel (a) shows results for corrected linear regression, where covariates are subject to additive noise with $SNR = 5$. Each point represents an average over 100 trials. Panel (b) shows similar results for logistic regression, where each point represents an average over 50 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.

optima, whereas the MCP penalty yields a relatively large number of local optima, albeit all guaranteed to lie within a small ball of β^* by Theorem 1.

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 local optima* of such nonconvex problems are close to the truth, implying that any optimization method guaranteed to converge to a local optimum will 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 interesting to develop a general method for establishing RSC and RSM conditions, beyond the specialized methods used for studying GLMs in this paper.

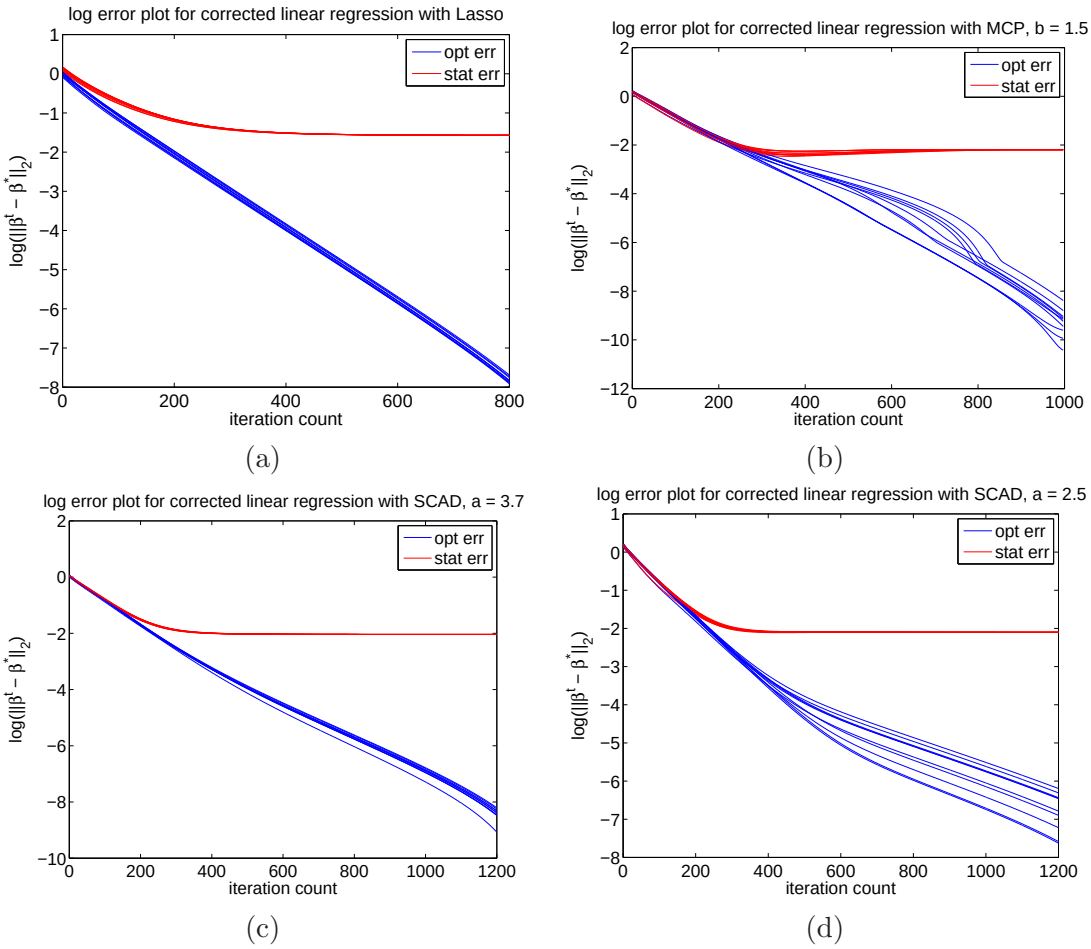

Figure 2. Plots illustrating linear rates of convergence on a log scale for corrected linear regression with Lasso, MCP, and SCAD regularizers, with $p = 128$, $k = \lfloor \sqrt{p} \rfloor$, and $n = \lfloor 20k \log p \rfloor$, where covariates are corrupted by additive noise with $SNR = 5$. Red lines 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 2, 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 [6] generates local optima that exhibit a smaller spread than the local optima generated for a smaller setting of the parameter a.

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 grants NSF-DMS-0907632 and AFOSR-09NL184.

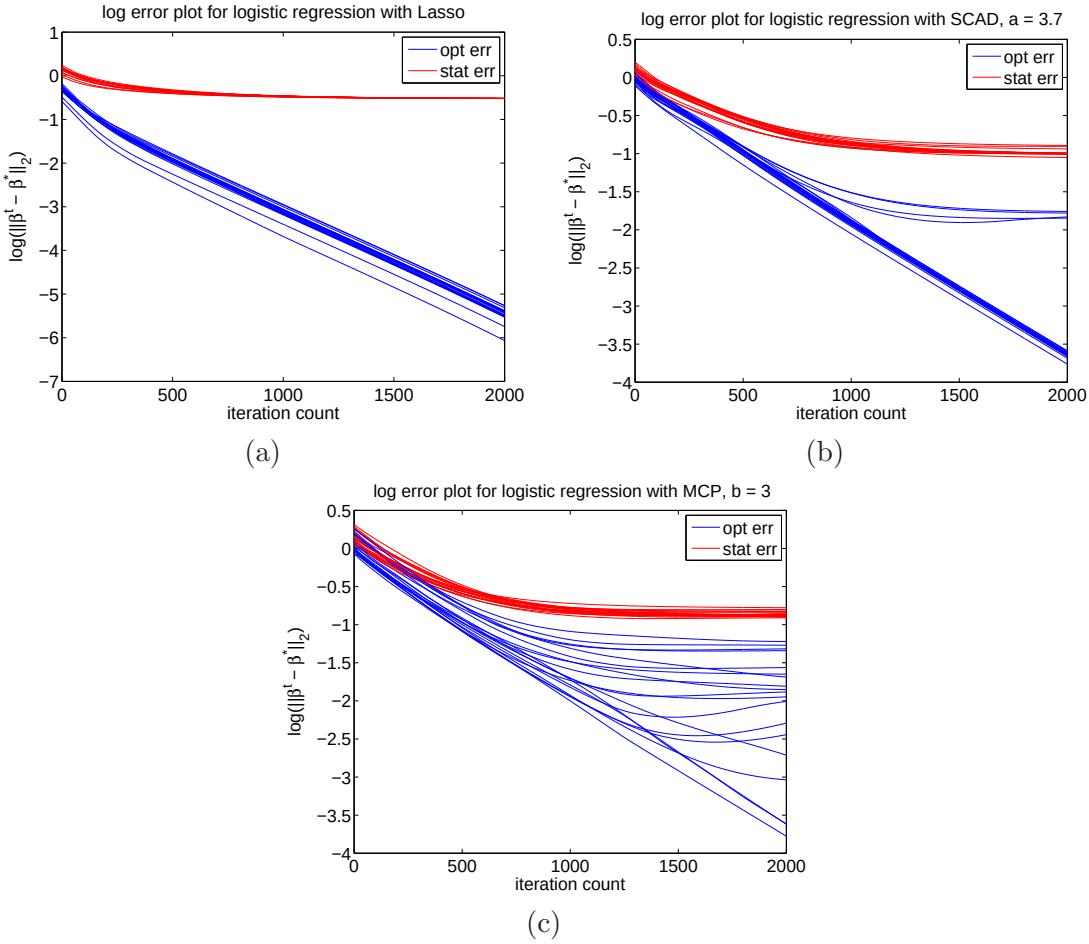

Figure 3. Plots that demonstrate linear rates of convergence on a log scale for logistic regression with $p = 64, k = \sqrt{p}$, and $n = \lfloor 20k \log p \rfloor$. Red lines depict statistical error $\log\left(\|\widehat{\beta} - \beta^*\|_2\right)$ and blue lines depict optimization error $\log\left(\|\beta^t - \widehat{\beta}\|_2\right)$. (a) Lasso penalty. (b) SCAD penalty. (c) MCP. As predicted by Theorem 2, 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.

A Properties of regularizers

In this section, we establish properties of some nonconvex regularizers covered by our theory (Section A.1) and verify that specific regularizers satisfy Assumption 1 (Section A.2). The properties given in Section 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 5. 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.

Proof. Suppose $0 \leq t_1 \leq t_2$. Then

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

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

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

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

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

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

Lemma 6. For any vector $v \in \mathbb{R}^p$, let A denote the index set of its k largest elements in magnitude. Under Assumption 1, we have

$$\rho_\lambda(v_A) - \rho_\lambda(v_{A^c}) \leq \lambda L(\|v_A\|_1 - \|v_{A^c}\|_1). \tag{51}$$

Moreover, for an arbitrary vector $\beta \in \mathbb{R}^p$, we have

$$\rho_\lambda(\beta^*) - \rho_\lambda(\beta) \leq \lambda L(\|\nu_A\|_1 - \|\nu_{A^c}\|_1), \tag{52}$$

where $\nu := \beta - \beta^*$ and β^* is k-sparse.

Proof. We first establish inequality (51). Define the function $f(t) := \frac{t}{\rho_\lambda(t)}$ for $t > 0$. By our assumptions on ρ_λ, the function f is nondecreasing in $|t|$, so

$$\|v_{A^c}\|_1 = \sum_{j \in A^c} \rho_\lambda(v_j) \cdot f(|v_j|) \leq \sum_{j \in A^c} \rho_\lambda(v_j) \cdot f(\|v_{A^c}\|_\infty) = \rho_\lambda(v_{A^c}) \cdot f\left(\|v_{A^c}\|_\infty\right). \tag{53}$$

Again using the nondecreasing property of f, we have

$$\rho_\lambda(v_A) \cdot f(\|v_{A^c}\|_\infty) = \sum_{j \in A} \rho_\lambda(v_j) \cdot f(\|v_{A^c}\|_\infty) \leq \sum_{j \in A} \rho_\lambda(v_j) \cdot f(|v_j|) = \|v_A\|_1. \tag{54}$$

Note that for $t > 0$, we have

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

where the last inequality follows from the bounds on the subgradients of ρ_λ from Lemma 5. Combining this result with inequalities (53) and (54) yields

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

as claimed.

We now turn to the proof of the bound (52). Letting $S := \mathrm{supp}(\beta^*)$ denote the support of β^*, the triangle inequality and subadditivity of ρ imply that

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

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 7. 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 [27]. Furthermore, we may easily compute the derivative of the SCAD regularizer to be

$$\frac{\partial}{\partial t}\rho_\lambda(t) = \operatorname{sign}(t) \cdot \left(\lambda \cdot \mathbb{I}\left\{|t| \le \lambda\right\} + \frac{(a\lambda - |t|)_+}{a-1} \cdot \mathbb{I}\left\{|t| > \lambda\right\}\right), \qquad t \neq 0, \tag{55}$$

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 8. 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 [27]. We may compute the derivative of the MCP regularizer to be

$$\frac{\partial}{\partial t}\rho_\lambda(t) = \lambda \cdot \operatorname{sign}(t) \cdot \left(1 - \frac{|t|}{\lambda b}\right)_+, \qquad t \neq 0, \tag{56}$$

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 $\mu = \frac{1}{b}$. □

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

$$\mathcal{E}_n(\Delta) := \langle \nabla \mathcal{L}_n(\beta^* + \Delta) - \nabla \mathcal{L}_n(\beta^*),\, \Delta \rangle. \tag{57}$$

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

$$\mathcal{E}_n(\Delta) \ge \alpha_1 \|\Delta\|_2 - \sqrt{\frac{\log p}{n}}\|\Delta\|_1, \qquad \text{for all } |\Delta\|_2 \ge 1. \tag{58}$$

Proof. Fix an arbitrary $\Delta \in \mathbb{R}^p$ with $\|\Delta\|_2 \geq 1$. 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 $f'(1) - f'(0) \geq f'(t) - f'(0)$ for all $t \in [0,1]$. Computing the derivatives of f yields the inequality

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

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

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

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

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

Lemma 10. If inequality (4a) holds for all $\Delta \in \mathbb{R}^p$ and $n \geq 4R^2\tau_1^2 \log p$, then inequality (4b) holds, as well.

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

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

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 [13] 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 10 in Appendix B.1 implies that the RSC condition (4a) 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 [13], both of these terms are upper-bounded by $c' \sqrt{\frac{\log p}{n}}$ 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 of Negahban et al. [16], we then have

$$\mathcal{E}_n(\Delta) \geq \alpha_1 \|\Delta\|_2^2 - \tau_1 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 \|\Delta\|_2, \qquad \forall \|\Delta\|_2 \leq 1, \tag{59}$$

with probability at least $1 - c_1 \exp(-c_2 n)$, where $\alpha_1 \asymp \lambda_{\min}(\Sigma_x)$. 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 inequality (4a). Inequality (4b) then follows via Lemma 9 in Appendix B.1.

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

$$\mathbb{P}\left(\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \geq c \sqrt{\frac{\log p}{n}} \right) \leq c_1 \exp(-c_2 \log p). \tag{60}$$

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

$$\mathbb{P}\left[\max_{j=1,\ldots,p} |\frac{1}{n} \sum_{i=1}^{n} V_{ij}| \geq \delta \right] \leq \mathbb{P}[\mathcal{A}^c] + \mathbb{P}\left[\max_{j=1,\ldots,p} |\frac{1}{n} \sum_{i=1}^{n} V_{ij}| \geq \delta \;\middle|\; \mathcal{A} \right], \tag{61}$$

where $\mathcal{A} := \left\{ \max_{j=1,\ldots,p} \left\{ \frac{1}{n} \sum_{i=1}^n x_{ij}^2 \right\} \leq 2\mathbb{E}[x_{ij}^2] \right\}$. Since the x_{ij}'s are sub-Gaussian and $n \gtrsim \log p$, there are 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 inequality (61). For any $t \in \mathbb{R}$, we have

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

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

$$\log \mathbb{E}[\exp(tV_{ij}) \mid x_i] = \frac{t^2 x_{ij}^2}{2} \psi''(x_i^T \beta^* - v_i\, t x_{ij}) \;\leq\; \frac{\alpha_u t^2 x_{ij}^2}{2}, \tag{62}$$

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 $\kappa = \alpha_u \cdot \max_{j=1,\ldots,p} \mathbb{E}[x_{ij}^2]$, for each $j = 1, \ldots, p$. By a union bound, we then have

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

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

$$\mathcal{E}_n(\Delta) = \mathrm{vec}(\Delta)^T \left(\nabla^2 \mathcal{L}_n(\Theta^* + t\Delta)\right) \mathrm{vec}(\Delta) \geq \lambda_{\min}(\nabla^2 \mathcal{L}_n(\Theta^* + t\Delta))\, \|\Theta\|_F^2, \tag{63}$$

for some $t \in [0,1]$. By standard properties of the Kronecker product [9], we have

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

using the fact that $\|\Delta\|_2 \leq \|\Delta\|_F \leq 1$. Plugging back into inequality (63) yields

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

which shows that inequality (4a) holds with $\alpha_1 = (\|\Theta^*\|_2 + 1)^{-2}$ and $\tau_1 = 0$. Lemma 10 then implies inequality (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 (21), we have

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

Applying Theorem 1 then implies the desired result.

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 (33). 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 (25) is convex, the iterate β^{t+1} must lie on the boundary of the feasible set; i.e.,

$$g_{\lambda,\mu}(\beta^{t+1}) = R. \tag{64}$$

By Lagrangian duality, the program (25) is also equivalent to

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

for some choice of constraint parameter R'. Note that this is projection of $\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta}$ 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 (64) implies that

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

In fact, since the projection will shrink the vector to the boundary of the constraint set, equation (64) forces $R' = R$. This yields the update (34) appearing in step (3).

C.2 Derivation of updates for SCAD and MCP

We now derive the explicit form of the updates (35) and (36) for the SCAD and MCP regularizers, respectively. We may rewrite the unconstrained program (33) as

$$\begin{aligned}
\beta^{t+1} &\in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 + \frac{1}{\eta} \cdot \rho_\lambda(\beta) + \frac{\mu}{\eta} \|\beta\|_2^2 \right\} \\
&= \arg \min_{\beta \in \mathbb{R}^p} \left\{ \left(\frac{1}{2} + \frac{\mu}{\eta} \right) \|\beta\|_2^2 - \beta^T \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) + \frac{1}{\eta} \cdot \rho_\lambda(\beta) \right\} \\
&= \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \left\| \beta - \frac{1}{1 + 2\mu/\eta} \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 + \frac{1/\eta}{1 + 2\mu/\eta} \cdot \rho_\lambda(\beta) \right\}.
\end{aligned} \tag{65}$$

Since the program in the last line of equation (65) decomposes by coordinate, it suffices to solve the scalar optimization problem

$$\widehat{x} \in \arg \min_x \left\{ \frac{1}{2}(x - z)^2 + \nu \rho(x; \lambda) \right\}, \tag{66}$$

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 (66) in the case when $\nu = 1$ is given in Fan and Li [6]; the expression (35) for the more general case comes from writing out the subgradient of the objective as

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

using the equation for the SCAD derivative (55), and setting the subgradient equal to zero.

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

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

using the expression for the MCP derivative (56), leading to the closed-form solution given in equation (36). This agrees with the expression provided in Breheny and Huang [4] for the special case when $\nu = 1$.

C.3 Proof of Lemma 1

We first show that if $\lambda \geq \frac{4}{L} \cdot \|\nabla \mathcal{L}_n(\beta^*)\|_\infty$, then for any feasible β such that

$$\phi(\beta) \leq \phi(\beta^*) + \bar{\eta}, \tag{67}$$

we have

$$\|\beta - \beta^*\|_1 \leq 4\sqrt{k}\|\beta - \beta^*\|_2 + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right). \tag{68}$$

Defining the error vector $\Delta := \beta - \beta^*$, inequality (67) implies

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

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

$$\mathcal{T}(\beta^* + \Delta, \beta^*) + \rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) \leq -\langle \nabla \mathcal{L}_n(\beta^*), \Delta \rangle + \bar{\eta}. \tag{69}$$

We claim that

$$\rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) \leq \frac{\lambda L}{2}\|\Delta\|_1 + \bar{\eta}. \tag{70}$$

We divide the argument into two cases. First suppose $\|\Delta\|_2 \leq 3$. Since $\mathcal{L}_n$ satisfies the RSC condition (28a), we may lower-bound the left side of inequality (69) and apply Hölder's inequality to obtain

$$\alpha_1\|\Delta\|_2^2 - \tau_1\frac{\log p}{n}\|\Delta\|_1^2 + \rho(\beta^* + \Delta) - \rho(\beta^*) \leq \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \cdot \|\Delta\|_1 + \bar{\eta} \leq \frac{\lambda L}{4}\|\Delta\|_1 + \bar{\eta}. \tag{71}$$

Since $\|\Delta\|_1 \leq 2R$ by the feasibility of β^* and $\beta^* + \Delta$, we see that inequality (71) together with the condition $\lambda L \geq \frac{4R\tau_1 \log p}{n}$ gives inequality (70). On the other hand, when $\|\Delta\|_2 \geq 3$, the RSC condition (28b) gives

$$\alpha_2\|\Delta\|_2 - \tau_2\sqrt{\frac{\log p}{n}}\|\Delta\|_1 + \rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) \leq \frac{\lambda L}{4}\|\Delta\|_1 + \bar{\eta},$$

so for $\lambda L \geq 4\tau_2\sqrt{\frac{\log p}{n}}$, we also arrive at inequality (70).

By Lemma 6 in Appendix A.1, we have

$$\rho_\lambda(\beta^*) - \rho_\lambda(\beta) \leq \lambda L(\|\Delta_A\|_1 - \|\Delta_{A^c}\|_1),$$

where A indexes the top k components of Δ in magnitude. Combining this bound with inequality (70) then implies that

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

and consequently,

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

Putting together the pieces, we have

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

Using the bound $\|\Delta\|_1 \leq 2R$ once more, we obtain inequality (68).

We now apply the implication (67) 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 (37), we also have

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

Hence,

$$\|\widehat{\beta} - \beta^*\|_1 \leq 4\sqrt{k}\|\widehat{\beta} - \beta^*\|_2, \qquad \text{and} \qquad \|\beta^t - \beta^*\|_1 \leq 4\sqrt{k}\|\beta^t - \beta^*\|_2 + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right).$$

By the triangle inequality, we then have

$$\begin{aligned}
\|\beta^t - \widehat{\beta}\|_1 &\leq \|\widehat{\beta} - \beta^*\|_1 + \|\beta^t - \beta^*\|_1 \\
&\leq 4\sqrt{k} \cdot \left(\|\widehat{\beta} - \beta^*\|_2 + \|\beta^t - \beta^*\|_2\right) + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right) \\
&\leq 4\sqrt{k} \cdot \left(2\|\widehat{\beta} - \beta^*\|_2 + \|\beta^t - \widehat{\beta}\|_2\right) + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right),
\end{aligned}$$

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 $\|\beta^t - \widehat{\beta}\|_2 \leq 3$ for some integer $t \geq 1$, then $\|\beta^{t+1} - \widehat{\beta}\|_2 \leq 3$, as well.

We assume for the sake of a contradiction that $\|\beta^{t+1} - \widehat{\beta}\|_2 > 3$. By the RSC condition (28b) and the relation (27), we have

$$\overline{\mathcal{T}}(\beta^{t+1}, \widehat{\beta}) \geq \alpha\|\widehat{\beta} - \beta^{t+1}\|_2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 - \mu\|\widehat{\beta} - \beta^{t+1}\|_2^2. \tag{72}$$

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

$$g(\beta^{t+1}) - g(\widehat{\beta}) - \langle \nabla g(\widehat{\beta}),\, \beta^{t+1} - \widehat{\beta} \rangle \geq 0. \tag{73}$$

Multiplying by λ and summing with inequality (72) then yields

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

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

$$\phi(\beta^{t+1}) - \phi(\widehat{\beta}) \geq \alpha\|\widehat{\beta} - \beta^{t+1}\|_2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 - \mu\|\widehat{\beta} - \beta^{t+1}\|_2^2. \tag{74}$$

Since $\|\widehat{\beta} - \beta^t\|_2 \leq 3$ by the induction hypothesis, applying the RSC condition (28a) to the pair $(\widehat{\beta}, \beta^t)$ also gives

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

Combining with the inequality

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

we then have

$$\begin{aligned}
\phi(\widehat{\beta}) &\geq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla\overline{\mathcal{L}}_n(\beta^t),\, \widehat{\beta} - \beta^t\rangle + \lambda g(\beta^{t+1}) + \lambda\langle \nabla g(\beta^t),\, \widehat{\beta} - \beta^{t+1}\rangle \\
&\quad + (\alpha - \mu)\cdot\|\beta^t - \widehat{\beta}\|_2^2 - \tau\frac{\log p}{n}\|\beta^t - \widehat{\beta}\|_1^2 \\
&\geq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla\overline{\mathcal{L}}_n(\beta^t),\, \widehat{\beta} - \beta^t\rangle + \lambda g(\beta^{t+1}) + \lambda\langle \nabla g(\beta^{t+1}),\, \widehat{\beta} - \beta^{t+1}\rangle - \tau\frac{\log p}{n}\|\beta^t - \widehat{\beta}\|_1^2.
\end{aligned} \tag{75}$$

Finally, the RSM condition (29) on the pair (β^{t+1}, β^t) gives

$$\begin{aligned}
\phi(\beta^{t+1}) &\leq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla\overline{\mathcal{L}}_n(\beta^t),\, \beta^{t+1} - \beta^t\rangle + \lambda g(\beta^{t+1}) \qquad (76) \\
&\quad + (\alpha_3 - \mu)\|\beta^{t+1} - \beta^t\|_2^2 + \tau\frac{\log p}{n}\|\beta^{t+1} - \beta^t\|_1^2 \\
&\leq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla\overline{\mathcal{L}}_n(\beta^t),\, \beta^{t+1} - \beta^t\rangle + \lambda g(\beta^{t+1}) + \frac{\eta}{2}\|\beta^{t+1} - \beta^t\|_2^2 + \frac{4R^2\tau\log p}{n}, \qquad (77)
\end{aligned}$$

since $\frac{\eta}{2} \geq \alpha_3 - \mu$ by assumption, and $\|\beta^{t+1} - \beta^t\|_1 \leq 2R$. It is easy to check that the update (25) may be written equivalently as

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

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

$$\langle \nabla\overline{\mathcal{L}}_n(\beta^t) + \eta(\beta^{t+1} - \beta^t) + \lambda\nabla g(\beta^{t+1}),\, \beta^{t+1} - \widehat{\beta}\rangle \leq 0. \tag{78}$$

Summing up inequalities (75), (76), and (78), we then have

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

Combining this last inequality with inequality (74), we have

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

since $\|\beta^t-\widehat{\beta}\|_2\leq 3$ by the induction hypothesis and $\|\beta^{t+1}-\widehat{\beta}\|_2>3$ by assumption, and using the fact that $\eta\geq 2\mu$. It follows that

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

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

C.5 Proof of Lemma 3

We begin with an auxiliary lemma:

Lemma 11. Under the conditions of Lemma 3, we have

$$\overline{\mathcal{T}}(\beta^t,\widehat{\beta})\geq-2\tau\frac{\log p}{n}(\epsilon+\bar{\epsilon})^2,\quad\text{and}\tag{79a}$$
$$\phi(\beta^t)-\phi(\widehat{\beta})\geq\frac{\alpha-\mu}{2}\|\widehat{\beta}-\beta^t\|_2^2-\frac{2\tau\log p}{n}(\epsilon+\bar{\epsilon})^2.\tag{79b}$$

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

Define

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

the objective function minimized over the constraint set $\{g(\beta)\leq R\}$ 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

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

Appealing to inequality (79a), we then have

$$\phi_t(\beta^{t+1})\leq(1-\gamma)\overline{\mathcal{L}}_n(\beta^t)+\gamma\overline{\mathcal{L}}_n(\widehat{\beta})+2\gamma\tau\frac{\log p}{n}(\epsilon+\epsilon_{\text{stat}})^2+\frac{\eta\gamma^2}{2}\|\widehat{\beta}-\beta^t\|_2^2+\lambda g(\beta_\gamma)$$
$$\overset{(i)}{\leq}\phi(\beta^t)-\gamma(\phi(\beta^t)-\phi(\widehat{\beta}))+2\gamma\tau\frac{\log p}{n}(\epsilon+\epsilon_{\text{stat}})^2+\frac{\eta\gamma^2}{2}\|\widehat{\beta}-\beta^t\|_2^2$$
$$\leq\phi(\beta^t)-\gamma(\phi(\beta^t)-\phi(\widehat{\beta}))+2\tau\frac{\log p}{n}(\epsilon+\epsilon_{\text{stat}})^2+\frac{\eta\gamma^2}{2}\|\widehat{\beta}-\beta^t\|_2^2,\tag{80}$$

where inequality (i) incorporates the fact that

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

by the convexity of g.

By the RSM condition (29), we also have

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

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

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

Combining with inequality (80) then yields

$$\phi(\beta^{t+1}) \leq \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2 + \tau\frac{\log p}{n}\|\beta^{t+1} - \beta^t\|_1^2 + 2\tau\frac{\log p}{n}(\epsilon + \bar{\epsilon})^2. \tag{81}$$

By the triangle inequality, we have

$$\|\beta^{t+1} - \beta^t\|_1^2 \leq \left(\|\Delta^{t+1}\|_1 + \|\Delta^t\|_1\right)^2 \;\leq\; 2\|\Delta^{t+1}\|_1^2 + 2\|\Delta^t\|_1^2,$$

where we have defined $\Delta^t := \beta^t - \widehat{\beta}$. Combined with inequality (81), we therefore have

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

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

$$\begin{aligned}\phi(\beta^{t+1}) &\leq \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \frac{\eta\gamma^2}{2}\|\Delta^t\|_2^2 + 64k\tau\frac{\log p}{n}(\|\Delta^{t+1}\|_2^2 + \|\Delta^t\|_2^2) + 10\psi(n,p,\epsilon) \\ &= \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \left(\frac{\eta\gamma^2}{2} + 64k\tau\frac{\log p}{n}\right)\|\Delta^t\|_2^2 + 64k\tau\frac{\log p}{n}\|\Delta^{t+1}\|_2^2 \\ &\qquad + 10\psi(n,p,\epsilon). \end{aligned} \tag{82}$$

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 inequality (79b) and subtracting $\phi(\widehat{\beta})$ from both sides of inequality (82), we have

$$\begin{aligned}\delta_{t+1} \leq \big(1-\gamma\big)\delta_t + \frac{\eta\gamma^2 + 128\upsilon(k,p,n)}{\alpha - \mu}\left(\delta_t + 2\psi(n,p,\epsilon)\right) + \frac{128\upsilon(k,p,n)}{\alpha-\mu}\left(\delta_{t+1} + 2\psi(n,p,\epsilon)\right) \\ + 10\psi(n,p,\epsilon).\end{aligned}$$

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

$$\begin{aligned}\left(1 - \frac{128\upsilon(k,p,n)}{\alpha-\mu}\right)\delta_{t+1} \leq \left(1 - \frac{\alpha-\mu}{4\eta} + \frac{128\upsilon(k,p,n)}{\alpha-\mu}\right)\delta_t \\ + 2\left(\frac{\alpha-\mu}{4\eta} + \frac{256\upsilon(k,p,n)}{\alpha-\mu} + 5\right)\psi(n,p,\epsilon),\end{aligned}$$

or $\delta_{t+1} \leq \kappa\delta_t + \xi(\epsilon + \bar{\epsilon})^2$, where κ and ξ were previously defined in equations (30) and (39), respectively. Finally, iterating the procedure yields

$$\delta_t \leq \kappa^{t-T}\delta_T + \xi(\epsilon+\bar{\epsilon})^2(1+\kappa+\kappa^2+\cdots+\kappa^{t-T-1}) \;\leq\; \kappa^{t-T}\delta_T + \frac{\xi(\epsilon+\bar{\epsilon})^2}{1-\kappa}, \tag{83}$$

as claimed.

The only remaining step is to prove the auxiliary lemma.

Proof of Lemma 11: By the RSC condition (28a) and the assumption (38), we have

$$\overline{\mathcal{T}}(\beta^t, \widehat{\beta}) \geq (\alpha - \mu)\ \|\widehat{\beta} - \beta^t\|_2^2 - \tau\frac{\log p}{n}\|\widehat{\beta} - \beta^t\|_1^2. \tag{84}$$

Furthermore, by convexity of g, we have

$$\lambda\Big(g(\beta^t) - g(\widehat{\beta}) - \langle \nabla g(\widehat{\beta}),\, \beta^t - \widehat{\beta}\rangle\Big) \geq 0, \tag{85}$$

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

$$\langle \nabla\phi(\widehat{\beta}),\, \beta^t - \widehat{\beta}\rangle \geq 0. \tag{86}$$

Summing inequalities (84), (85), and (86) then yields

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

Applying Lemma 1 to bound the term $\|\widehat{\beta} - \beta^t\|_1^2$ and using the assumption $\frac{32k\tau\log p}{n} \leq \frac{\alpha-\mu}{2}$ yields the bound (79b). On the other hand, applying Lemma 1 directly to inequality (84) with β^t and $\widehat{\beta}$ switched gives

$$\begin{aligned}\overline{\mathcal{T}}(\widehat{\beta}, \beta^t) &\geq (\alpha-\mu)\|\widehat{\beta} - \beta^t\|_2^2 - \tau\frac{\log p}{n}\Big(32k\|\widehat{\beta} - \beta^t\|_2^2 + 2(\epsilon+\bar{\epsilon})^2\Big) \\ &\geq -2\tau\frac{\log p}{n}(\epsilon+\bar{\epsilon})^2.\end{aligned}$$

This establishes inequality (79a).

D Proof of Lemma 4

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

$$\varphi_\tau(u) = \begin{cases} u^2, & \text{if } |u| \leq \frac{\tau}{2}, \\ (\tau - u)^2, & \text{if } \frac{\tau}{2} \leq |u| \leq \tau, \\ 0, & \text{if } |u| \geq \tau, \end{cases} \qquad \text{and} \qquad \alpha_\tau(u) = \begin{cases} u, & \text{if } |u| \leq \tau, \\ 0, & \text{if } |u| \geq \tau. \end{cases}$$

By construction, φ_τ is τ-Lipschitz and

$$\varphi_\tau(u) \leq u^2 \cdot \mathbb{I}\{|u| \leq \tau\}, \quad \text{for all } u \in \mathbb{R}. \tag{87}$$

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:

$$\begin{aligned} \mathcal{T}(\beta + \Delta, \beta) &= \frac{1}{n}\sum_{i=1}^n \psi''(x_i^T\beta + t_i \cdot x_i^T\Delta) \cdot (x_i^T\Delta)^2 \\ &\ge L_\psi(T) \cdot \sum_{i=1}^n (x_i^T\Delta)^2 \cdot \mathbb{I}\{|x_i^T\Delta| \le \tau\|\Delta\|_2\} \cdot \mathbb{I}\{|x_i^T\beta| \le T\} \\ &\ge L_\psi(T) \cdot \frac{1}{n}\sum_{i=1}^n \varphi_{\tau\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta), \end{aligned} \tag{88}$$

where the first equality is the expansion (45) and the second inequality uses the bound (87).

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

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

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

$$c\sigma_x^2 \exp\left(-\frac{c'\tau^2}{\sigma_x^2}\right) = \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{2}, \tag{89}$$

we have the lower bound

$$\mathbb{E}\left[\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\right] \geq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{2}\cdot\|\Delta\|_2^2. \tag{90}$$

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

$$0 \leq \frac{1}{\|\Delta\|_2}\cdot\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta) \leq \frac{\tau^2}{4}.$$

Consequently, applying the bounded differences inequality yields

$$\mathbb{P}\left(Z(\delta) \geq \mathbb{E}[Z(\delta)] + \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{4}\right) \leq c_1\exp(-c_2 n). \tag{91}$$

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

$$\mathbb{E}[Z(\delta)] \leq 2\sqrt{\frac{\pi}{2}}\cdot\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta}\frac{1}{\|\Delta\|_2^2}\left|\frac{1}{n}\sum_{i=1}^n g_i\Big(\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\Big)\right|\right], \tag{92}$$

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

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

with

$$\begin{aligned}\operatorname{var}\left(Z_{\beta,\Delta} - Z_{\widetilde{\beta},\Delta}\right) &= \frac{1}{\|\Delta\|_2^4}\cdot\frac{1}{n^2}\sum_{i=1}^n\varphi_{\tau\|\Delta\|_2}^2(x_i^T\Delta)\cdot\Big(\gamma_T(x_i^T\beta) - \gamma_T(x_i^T\widetilde{\beta})\Big)^2\\ &\leq \frac{1}{n^2}\sum_{i=1}^n\frac{\tau^4}{16}\cdot\frac{4}{T^2}\left(x_i^T(\beta-\widetilde{\beta})\right)^2,\end{aligned}$$

since $\varphi_{\tau\|\Delta\|_2} \leq \frac{\tau^2\|\Delta\|_2^2}{4}$ and γ_T is $\frac{2}{T}$-Lipschitz. Similarly,

$$\begin{aligned}\operatorname{var}\left(Z_{\widetilde{\beta},\Delta} - Z_{\widetilde{\beta},\widetilde{\Delta}}\right) &\leq \frac{1}{\|\Delta\|_2^4}\cdot\frac{1}{n^2}\sum_{i=1}^n\gamma_T^2(x_i^T\widetilde{\beta})\cdot\Big(\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta) - \varphi_{\tau\|\Delta\|_2}(x_i^T\widetilde{\Delta})\Big)^2\\ &\leq \frac{1}{\|\Delta\|_2^2}\cdot\frac{1}{n^2}\sum_{i=1}^n\tau^2\left(x_i^T(\Delta-\widetilde{\Delta})\right)^2.\end{aligned}$$

Defining the centered Gaussian process

$$Y_{\beta,\Delta} := \frac{\tau^2}{2T}\cdot\frac{1}{n}\sum_{i=1}^n\widehat{g}_i\cdot x_i^T\beta + \frac{\tau}{\|\Delta\|_2}\cdot\frac{1}{n}\sum_{i=1}^n\widetilde{g}_i\cdot x_i^T\Delta,$$

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

$$\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} Z_{\beta,\Delta}\right] \leq 2 \cdot \mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} Y_{\beta,\Delta}\right]. \tag{93}$$

Note further (cf. p.77 of Ledoux and Talagrand [11]) that

$$\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} |Z_{\beta,\Delta}|\right] \leq \mathbb{E}\left[|Z_{\beta_0,\Delta_0}|\right] + 2\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} Z_{\beta,\Delta}\right], \tag{94}$$

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

$$\mathbb{E}\left[|Z_{\beta_0,\Delta_0}|\right] \leq \sqrt{\frac{2}{\pi}} \cdot \sqrt{\operatorname{var}\left(Z_{\beta_0,\Delta_0}\right)} \leq \frac{1}{\|\Delta\|_2} \cdot \sqrt{\frac{2}{\pi}} \cdot \sqrt{\frac{\tau^2}{4n}}. \tag{95}$$

Finally,

$$\begin{aligned}\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} Y_{\beta,\Delta}\right] &\leq \frac{\tau^2 R}{2T} \cdot \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \widehat{g}_i x_i\right\|_\infty\right] + \tau\delta \cdot \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \widetilde{g}_i x_i\right\|_\infty\right] \\ &\leq \frac{c\tau^2 R\sigma_x}{2T}\sqrt{\frac{\log p}{n}} + c\tau\delta\sigma_x \cdot \sqrt{\frac{\log p}{n}},\end{aligned} \tag{96}$$

by Lemma 16 in Appendix E. Combining inequalities (92), (93), (94), (95), and (96), we then obtain

$$\mathbb{E}[Z(\delta)] \leq \frac{c'\tau^2 R\sigma_x}{2T}\sqrt{\frac{\log p}{n}} + c'\tau\delta\sigma_x \cdot \sqrt{\frac{\log p}{n}}. \tag{97}$$

Finally, combining inequalities (90), (91), and (97), we see that under the scaling $R\sqrt{\frac{\log p}{n}} \precsim 1$, we have

$$\begin{aligned}\frac{1}{\|\Delta\|_2^2} \cdot \frac{1}{n}\sum_{i=1}^n \varphi_{\tau\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta) &\geq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{4} - \left(\frac{c'\tau^2 R\sigma_x}{2T}\sqrt{\frac{\log p}{n}} + c'\tau\delta\sigma_x\sqrt{\frac{\log p}{n}}\right) \\ &\geq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{8} - c'\tau\delta\sigma_x\sqrt{\frac{\log p}{n}},\end{aligned} \tag{98}$$

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 [2, 22]. Consider the inequality

$$\frac{1}{\|\Delta\|_2^2} \cdot \frac{1}{n}\sum_{i=1}^n \varphi_{\tau\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta) \geq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{8} - 2c'\tau\sigma_x\frac{\|\Delta\|_1}{\|\Delta\|_2}\sqrt{\frac{\log p}{n}}, \tag{99}$$

as well as the event

$$\mathcal{E} := \left\{\text{inequality (99) holds for all } \|\beta\|_2 \leq 3 \text{ and } \frac{\|\Delta\|_1}{\|\Delta\|_2} \leq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{16c'\tau\sigma_x}\sqrt{\frac{n}{\log p}}\right\}.$$

Define the function

$$f(\beta, \Delta; X) := \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{8} - \frac{1}{\|\Delta\|_2^2} \cdot \frac{1}{n} \sum_{i=1}^{n} \varphi_\tau(x_i^T \Delta) \cdot \gamma_T(x_i^T \beta), \tag{100}$$

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 inequality (98) implies

$$\mathbb{P}\left(\sup_{h(\beta,\Delta) \leq \delta} f(\beta, \Delta; X) \geq g(\delta) \right) \leq c_1 \exp(-c_2 n), \quad \text{for any } \delta > 0, \tag{101}$$

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

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

$$1 \leq\ h(\beta, \Delta)\ \leq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{16 c' \tau \sigma_x} \sqrt{\frac{n}{\log p}}, \tag{102}$$

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

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

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

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

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 (102). By the definition (100) of f, we have

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

where inequality (i) applies the tail bound (101). It follows that

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

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

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 [23])**.** Let $Z_1, \ldots, Z_n$ 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 [11])**.** 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 [11])**.** Given a countable index set T, let $X(t)$ and $Y(t)$ be centered Gaussian processes such that

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

Then

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

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 [11]):

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

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} g_i X_i\right\|_\infty\right] \leq c_0\sigma_x\sqrt{\frac{\log p}{n}} \cdot \mathbb{E}\left[\max_j \sqrt{\frac{\sum_{i=1}^n X_{ij}^2}{n}}\right]. \tag{103}$$

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 $\mathbb{E}[Z_j] \leq 2\sigma_x^2$, we have

$$\mathbb{P}\left(Z_j - \mathbb{E}[Z_j] \geq u + 2\sigma_x^2\right) \leq c_1 \exp\left(-\frac{c_2 n u}{\sigma_x}\right), \qquad \text{for all } u \geq 0 \text{ and } 1 \leq j \leq p. \tag{104}$$

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

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

where the final inequality follows from the bound (104) with $u = s^2 - 2\sigma_x^2$, valid as long as $s^2 \geq t^2 \geq 2\sigma_x^2$. Integrating, we have the bound

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

Since $n \gtrsim \sqrt{\log p}$ by assumption, taking t to be a constant implies $\mathbb{E}\left[\max_j \sqrt{Z_j}\right] = \mathcal{O}(1)$, which combined with inequality (103) gives the desired result. □

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 $\|\nabla\widetilde{\rho}_\lambda(\widetilde{\beta})\|_\infty \leq \lambda L$.

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

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

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

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

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

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

Remark 5. When $\rho_\lambda(\beta) + \mu_1 \|\beta\|_2^2$ 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 $\mu_2 = 0$. (See Lemma 6 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 variation of Theorem 1:

Theorem 3. 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. With λ is chosen according to the bound (6) and $n \geq \frac{16R^2\tau_2^2}{\alpha_2^2} \log p$, then

$$\|\widetilde{\beta} - \beta^*\|_2 \leq \frac{7\lambda\sqrt{k}}{4(\alpha_1 - \mu_1 - \mu_2)}, \qquad \text{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \leq \frac{63\lambda k}{4(\alpha_1 - \mu_1 - \mu_2)}.$$

Proof. The proof is essentially the same as the proof of Theorem 1, 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, inequality (9) holds, as well.

Next, note that inequality (11) 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 inequality (12) with μ replaced by $\mu_1 + \mu_2$. The remainder of the proof is exactly as before. □

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 [27] is given by

$$\rho_\lambda(t) := \min\left\{\frac{\lambda^2 c}{2},\ \lambda|t|\right\}. \tag{105}$$

Lemma 17. The capped-ℓ_1 regularizer (105) with parameter c satisfies the conditions of Assumption 2, with $\mu_1 = 0$, $\mu_2 = \frac{1}{c}$, 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.

We begin by defining the function $\widetilde{\rho} : \mathbb{R} \to \mathbb{R}$ via

$$\widetilde{\rho}_\lambda(t) = \begin{cases} \lambda|t|, & \text{if} \quad |t| \leq \frac{\lambda c}{2}, \\ \frac{\lambda^2 c}{2}, & \text{if} \quad |t| > \frac{\lambda c}{2}. \end{cases}$$

For a fixed local optimum $\widetilde{\beta}$, note that we have $\widetilde{\rho}_\lambda(\beta) = \sum_{j \in T} \lambda |\widetilde{\beta}_j| + \sum_{j \in T^c} \frac{\lambda^2 c}{2}$, where $T := \left\{ j \mid |\widetilde{\beta}_j| \leq \frac{\lambda c}{2} \right\}$. Clearly, $\widetilde{\rho}_\lambda$ is a convex upper bound on ρ_λ, with $\widetilde{\rho}_\lambda(\widetilde{\beta}) = \rho_\lambda(\widetilde{\beta})$. Furthermore, by the convexity of $\widetilde{\rho}_\lambda$, we have

$$\langle \nabla \widetilde{\rho}_\lambda(\widetilde{\beta}), \beta^* - \widetilde{\beta} \rangle \leq \widetilde{\rho}_\lambda(\beta^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}) = \sum_{j \in T} \Big(\widetilde{\rho}_\lambda(\beta_j^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}_j) \Big) - \sum_{j \notin T} \widetilde{\rho}_\lambda(\widetilde{\beta}_j), \tag{106}$$

using decomposability of $\widetilde{\rho}$. For $j \in T$, we have $\widetilde{\rho}_\lambda(\beta_j^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}_j) = \lambda |\beta_j^*| - \lambda |\widetilde{\beta}_j| \leq \lambda |\widetilde{\nu}_j|$, whereas for $j \notin T$, we have $\widetilde{\rho}_\lambda(\beta_j^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}_j) = 0 \leq \lambda |\widetilde{\nu}_j|$. Combined with the bound (106), we obtain

$$\begin{aligned} \langle \nabla \widetilde{\rho}_\lambda(\widetilde{\beta}), \beta^* - \widetilde{\beta} \rangle &\leq \sum_{j \in T} \lambda |\widetilde{\nu}_j| - \sum_{j \notin T} \widetilde{\rho}_\lambda(\widetilde{\beta}_j) \\ &= \lambda \|\widetilde{\nu}_S\|_1 - \sum_{j \notin S} \rho_\lambda(\widetilde{\beta}_j) \\ &= \lambda \|\widetilde{\nu}_T\|_1 - \lambda \|\widetilde{\nu}_{T^c}\|_1 + \sum_{j \notin T} \Big(\lambda |\widetilde{\beta}_j| - \rho_\lambda(\widetilde{\beta}_j) \Big). \end{aligned} \tag{107}$$

Now observe that

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

and moreover, the derivative of $\frac{t^2}{c}$ always exceeds λ for $|t| > \frac{\lambda c}{2}$. Consequently, we have $\lambda |t| - \rho_\lambda(t) \leq \frac{t^2}{c}$ for all $t \in \mathbb{R}$. Substituting this bound into inequality (107) yields

$$\langle \nabla \widetilde{\rho}_\lambda(\widetilde{\beta}), \beta^* - \widetilde{\beta} \rangle \leq \lambda \|\widetilde{\nu}_S\|_1 - \lambda \|\widetilde{\nu}_{S^c}\|_1 + \frac{1}{c} \|\widetilde{\nu}_{S^c}\|_2^2,$$

which is condition (v) of Assumption 2 on $\widetilde{\rho}_\lambda$ with $L = 1$, $A = S$, and $\mu_2 = \frac{1}{c}$. The remaining conditions are easy to verify (see also Zhang and Zhang [27]). □

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] J. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. *arXiv e-prints*, October 2013. Available at http://arxiv.org/abs/1210.5992.

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

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

[10] D. R. Hunter and R. Li. Variable selection using MM algorithms. *Ann. Statist.*, 33(4):1617–1642, 2005.

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

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

[13] 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.

[14] P. Loh and M.J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *arXiv e-prints*, December 2012. Available at http://arxiv.org/abs/1212.0478.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

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

[23] 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.

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

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

[26] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2012.

[27] 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.

[28] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, 36(4):1509–1533, 2008.