The Annals of Statistics
2017, Vol. 45, No. 2, 866–896
DOI: 10.1214/16-AOS1471

STATISTICAL CONSISTENCY AND ASYMPTOTIC NORMALITY FOR HIGH-DIMENSIONAL ROBUST M-ESTIMATORS

BY PO-LING LOH

University of Wisconsin-Madison

We study theoretical properties of regularized robust M-estimators, applicable when data are drawn from a sparse high-dimensional linear model and contaminated by heavy-tailed distributions and/or outliers in the additive errors and covariates. We first establish a form of local statistical consistency for the penalized regression estimators under fairly mild conditions on the error distribution: When the derivative of the loss function is bounded and satisfies a local restricted curvature condition, all stationary points within a constant radius of the true regression vector converge at the minimax rate enjoyed by the Lasso with sub-Gaussian errors. When an appropriate nonconvex regularizer is used in place of an ℓ_1-penalty, we show that such stationary points are in fact unique and equal to the local oracle solution with the correct support; hence, results on asymptotic normality in the low-dimensional case carry over immediately to the high-dimensional setting. This has important implications for the efficiency of regularized nonconvex M-estimators when the errors are heavy-tailed. Our analysis of the local curvature of the loss function also has useful consequences for optimization when the robust regression function and/or regularizer is nonconvex and the objective function possesses stationary points outside the local region. We show that as long as a composite gradient descent algorithm is initialized within a constant radius of the true regression vector, successive iterates will converge at a linear rate to a stationary point within the local region. Furthermore, the global optimum of a convex regularized robust regression function may be used to obtain a suitable initialization. The result is a novel two-step procedure that uses a convex M-estimator to achieve consistency and a nonconvex M-estimator to increase efficiency. We conclude with simulation results that corroborate our theoretical findings.

1. Introduction. Ever since robustness entered the statistical scene in Box's classical paper of 1953 [Box (1953)], many significant steps have been taken toward analyzing and quantifying robust statistical procedures—notably the work of Tukey (1960), Huber (1964), and Hampel (1968), among others. Huber's seminal work on M-estimators [Huber (1964)] established asymptotic properties of a class of statistical estimators containing the maximum likelihood estimator, and provided initial theory for constructing regression functions that are robust

Received January 2015; revised April 2016.
MSC2010 subject classifications. 62F12.
Key words and phrases. Robust regression, high-dimensional statistics, statistical consistency, support recovery, nonconvex optimization.

to deviations from normality. Despite the substantial body of work on robust M-estimators, however, research on high-dimensional regression estimators has mostly been limited to penalized likelihood-based approaches [e.g., Fan and Li (2001), Friedman, Hastie and Tibshirani (2008), Ravikumar, Wainwright and Lafferty (2010), Tibshirani (1996)]. Several recent papers [Negahban et al. (2012), Loh and Wainwright (2014, 2015)] have shed new light on high-dimensional M-estimators, presenting a fairly unified framework for analyzing statistical and optimization properties of such estimators. However, whereas the M-estimators studied in those papers are finite-sample versions of globally convex functions, many important M-estimators, such as those arising in classical robust regression, only possess convex curvature over local regions—*even at the population level*. In this paper, we present new theoretical results, based only on *local* curvature assumptions, which may be used to establish statistical and optimization properties of regularized M-estimators with highly nonconvex loss functions.

Broadly, we are interested in regression estimators that are robust in the following senses:

(a) *Model misspecification.* The ordinary least squares objective function may be viewed as a maximum likelihood estimator for linear regression when the additive errors ε_i are normally distributed. It is well known that the ℓ_1-penalized ordinary least squares estimator is still consistent when the ε_i's are sub-Gaussian [Bickel, Ritov and Tsybakov (2009), Wainwright (2009)]; however, if the distribution of the ε_i's deviates more wildly from the normal distribution (e.g., the ε_i's are heavy-tailed), the regression estimator based on the least squares loss no longer converges at optimal rates. In addition, whereas the usual regularity assumptions on the design matrix such as the restricted eigenvalue condition have been shown to hold with high probability when the covariates are sub-Gaussian [Raskutti, Wainwright and Yu (2010), Rudelson and Zhou (2013)], we wish to devise estimators that are also consistent under weaker assumptions on the distribution of the covariates.

(b) *Outliers.* Even when the covariates and error terms are normally distributed, the regression estimator may be inconsistent when observations are contaminated by outliers in the predictors and/or response variables [Rousseeuw and Leroy (2005)]. Whereas the standard ordinary least squares loss function is nonrobust to outliers, alternative estimators exist in a low-dimensional setting that are robust to a certain degree of contamination. We wish to extend this theory to high dimensions.

Inspired by the classical theory on robust estimators for linear regression [Huber (1981), Maronna, Martin and Yohai (2006), Hampel et al. (1986)], we study regularized versions of low-dimensional robust regression estimators and establish statistical guarantees in a high-dimensional setting. As we will see, the regularized robust regression functions continue to enjoy good behavior in high dimensions,

and we can quantify the degree to which the high-dimensional estimators are robust to the types of deviations described above.

Our first main contribution is to provide a general set of sufficient conditions under which optima of regularized robust M-estimators are statistically consistent, even in the presence of heavy-tailed errors and outlier contamination. The conditions involve a bound on the derivative of the regression function, as well as restricted strong convexity of the loss function in a neighborhood of constant radius about the true parameter vector, and the conclusions are given in terms of the tails of the error distribution. The notion of restricted strong convexity, as used previously in the literature [Negahban et al. (2012), Agarwal, Negahban and Wainwright (2012), Loh and Wainwright (2014, 2015)], traditionally involves a global condition on the behavior of the loss function. However, due to the highly nonconvex behavior of the robust regression functions of interest, we assume only a *local* condition of restricted strong convexity in the development of our statistical results. Consequently, our main theorem provides guarantees only for stationary points within the local region of strong curvature. We show that all such local stationary points are statistically consistent estimators for the true regression vector; when the covariates are sub-Gaussian, the rate of convergence agrees (up to a constant factor) with the rate of convergence for ℓ_1-penalized ordinary least squares regression with sub-Gaussian errors. We also use the same framework to study generalized M-estimators and provide results for statistical consistency of local stationary points under weaker distributional assumptions on the covariates.

The wide applicability of our theorem on statistical consistency of high-dimensional robust M-estimators opens the door to an important question regarding the design of robust regression estimators, which is the topic of our second contribution: If all regression estimators with bounded derivative are statistically consistent with rates agreeing up to a constant factor, why use a complicated nonconvex regression function instead of a simple convex function such as the Huber loss? In the low-dimensional setting, several independent lines of work provide reasons for using nonconvex M-estimators over their convex alternatives [Huber (1981), Shevlyakov, Morgenthaler and Shurygin (2008)]. One compelling justification is from the viewpoint of statistical efficiency. Indeed, the log likelihood function of the heavy-tailed t-distribution with one degree of freedom gives rise to the nonconvex Cauchy loss, which is consequently asymptotically efficient [Lehmann and Casella (1998)]. In our second main theorem, we prove that by using a suitable nonconvex regularizer [Fan and Li (2001), Zhang (2010a)], we may guarantee that local stationary points of the regularized robust M-estimator agree with a local oracle solution defined on the correct support. Thus, provided the sample size scales sufficiently quickly with the level of sparsity, results on asymptotic normality of low-dimensional M-estimators with a diverging number of parameters [Huber (1973), Yohai and Maronna (1979), Portnoy (1985), Mammen (1989), He and Shao (2000)] may be used to establish asymptotic normality of their high-dimensional counterparts. In particular, when the loss function equals

the negative log-likelihood of the error distribution, stationary points of the high-dimensional M-estimator will also be efficient in an asymptotic sense. Our oracle result and subsequent conclusions regarding asymptotic normality resemble a variety of other results in the literature on nonconvex regularization [Fan and Peng (2004), Bradic, Fan and Wang (2011), Li, Peng and Zhu (2011)], but our result is stronger, because it provides guarantees for *all* stationary points in the local region. Our proof technique leverages the primal-dual witness construction recently proposed in Loh and Wainwright (2014); however, we require a more refined analysis here in order to extend the result to one involving only local properties of the loss function.

Our third and final contribution addresses algorithms used to optimize our proposed M-estimators. Since our statistical consistency and oracle results only provide guarantees for the behavior of *local* solutions, we need to devise an optimization algorithm that always converges to a stationary point inside the local region. Indeed, local optima that are statistically inconsistent are the bane of nonconvex M-estimators, even in low-dimensional settings [Freedman and Diaconis (1982)]. To remedy this issue, we propose a novel two-step algorithm that is *guaranteed* to converge to a stationary point within the local region of restricted strong convexity. Our algorithm consists of optimizing two separate regularized M-estimators in succession and may be applied to situations where both the loss and regularizer are nonconvex. In the first step, we optimize a convex regularized M-estimator to obtain a sufficiently close point that is then used to initialize an optimization algorithm for the original (nonconvex) M-estimator in the second step. We use the composite gradient descent algorithm [Nesterov (2007)] in both steps of the algorithm, and prove rigorously that if the initial point in the second step lies within the local region of restricted curvature, all successive iterates will continue to lie in the region and converge at a linear rate to an appropriate stationary point. Any convex, statistically consistent M-estimator suffices for the first step; we use the ℓ_1-penalized Huber loss in our simulations involving sub-Gaussian covariates with heavy-tailed errors, since global optima are statistically consistent by our earlier theory. Our resulting two-step estimator, which first optimizes a convex Huber loss to obtain a consistent estimator and then optimizes a (possibly nonconvex) robust M-estimator to obtain a more efficient estimator, is reminiscent of the one-step estimators common in the robust regression literature [Bickel (1975)]; however, here we require full runs of composite gradient descent in each step of the algorithm, rather than a single Newton–Raphson step. Note that if the goal is to optimize an M-estimator involving a convex loss and nonconvex regularizer, such as the SCAD-penalized Huber loss, our two-step algorithm is also applicable, where we optimize the ℓ_1-penalized loss in the first step.

Related work. We close this section by highlighting three recent papers on related topics. The analysis in this paper most closely resembles the work of Lozano and Meinshausen (2013), in that we study stationary points of nonconvex functions

used for robust high-dimensional linear regression within a local neighborhood of the true regression vector. Although the technical tools we use here are similar, we focus on regression functions that are expressible as M-estimators; the minimum distance loss function proposed in that paper does not fall into this category. In addition, we formalize the notion of basins of attraction for optima of nonconvex M-estimators and develop a two-step optimization algorithm that consists of optimizing successive regularized M-estimators, which goes beyond their results about local convergence of a composite gradient descent algorithm.

Another related work is that of Fan, Li and Wang (2014). While that paper focuses exclusively on developing estimation bounds for penalized robust regression with the Huber loss function, the results presented in our paper are strictly more general, since they also hold for *nonconvex* M-estimators, giving rise to solutions that are efficient as well as consistent. The analysis of the ℓ_1-penalized Huber loss is still relevant to our analysis, because as shown below, its global convergence guarantees provide us with a good initialization for the composite gradient algorithm that we use in the first step of our two-step algorithm.

Finally, we draw attention to the recent work by Mendelson (2014). In that paper, careful derivations based on empirical process theory demonstrate the advantage of using differently parametrized convex loss functions tuned according to distributional properties of the additive noise in the model. Our analysis also reveals the impact of different parameter choices for the regression function on the resulting estimator, but the rates of Mendelson (2014) are much sharper than ours (albeit agreeing up to a constant factor). However, our analysis is not limited to convex loss functions, and covers nonconvex loss functions possessing local curvature, as well. Finally, whereas Mendelson (2014) is primarily concerned with optimizing the estimator with respect to ℓ_1- and ℓ_2-error, our oracle results suggest that it is also instructive to consider second-order properties such as the variance and asymptotic efficiency. Indeed, such considerations may lead to a different parameter choice for a robust loss than if the primary goal is to minimize the bias alone.

The remainder of our paper is organized as follows: In Section 2, we provide basic background concerning (generalized) M-estimators, and introduce robust loss functions and regularizers to be discussed in the sequel. In Section 3, we present our main theorem concerning statistical consistency of robust high-dimensional M-estimators and unpack the distributional conditions required for the assumptions of the theorem to hold for specific estimators via a series of propositions. We also present our main theorem concerning oracle properties of nonconvex regularized M-estimators, with a corollary illustrating the types of asymptotic normality conclusions that may be derived from the oracle result. Section 4 provides our two-step optimization algorithm and corresponding theoretical guarantees. We conclude in Section 5 with a variety of simulation results. A brief review of robustness measures is provided in Appendix A, and proofs of the main theorems and all supporting lemmas and propositions are contained in the remaining Supplementary Material [Loh (2016)].

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 define $f(n) \succsim g(n)$ analogously. We write $f(n) \asymp g(n)$ when $f(n) \precsim g(n)$ and $f(n) \succsim g(n)$ hold simultaneously. For a vector $v \in \mathbb{R}^p$, we write $\operatorname{supp}(v) \subseteq \{1, \ldots, p\}$ to denote the support of v, and for an arbitrary 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 $|||M|||_2$ to denote the spectral norm. For a function $h : \mathbb{R}^p \to \mathbb{R}$, we write ∇h to denote a gradient or subgradient of the function. We use the relation $\perp\!\!\!\perp$ to denote independence. Finally, $\mathbb{B}_r(v)$ denotes the ℓ_2-ball of radius r centered around v.

2. Background and problem setup. In this section, we provide some background on M-estimators for robust regression and the nonconvex regularizers covered by our theory.

Throughout, we assume that $\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. observations from the linear model

$$y_i = x_i^T \beta^* + \varepsilon_i, \qquad \forall 1 \leq i \leq n, \tag{1}$$

where $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, and $\beta^* \in \mathbb{R}^p$ is a k-sparse vector; that is, $|\operatorname{supp}(\beta^*)| \leq k$. We also assume that $x_i \perp\!\!\!\perp \varepsilon_i$ and both are zero-mean random variables. We are interested in high-dimensional regression estimators of the form

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \{\mathcal{L}_n(\beta) + \rho_\lambda(\beta)\}, \tag{2}$$

where $\mathcal{L}_n$ is the empirical loss function and ρ_λ is a penalty function. For instance, the Lasso program is given by the loss $\mathcal{L}_n(\beta) = \frac{1}{n} \sum_{i=1}^n (x_i^T \beta - y_i)^2$ and penalty $\rho_\lambda(\beta) = \lambda \|\beta\|_1$, but this framework allows for much more general settings. Since we are interested in cases where the loss and regularizer may be nonconvex, we include the side condition $\|\beta\|_1 \leq R$ in the program (2) in order to guarantee the existence of local/global optima. We will require $R \geq \|\beta^*\|_1$, so that the true regression vector β^* is feasible for the program.

In the scenarios below, we will consider loss functions $\mathcal{L}_n$ that satisfy

$$\mathbb{E}\big[\nabla \mathcal{L}_n(\beta^*)\big] = 0. \tag{3}$$

When the population-level loss $\mathcal{L}(\beta) := \mathbb{E}[\mathcal{L}_n(\beta)]$ is a convex function, equation (3) implies that β^* is a global optimum of $\mathcal{L}(\beta)$. When $\mathcal{L}$ is nonconvex, the condition (3) ensures that β^* is at least a stationary point of the function. Our goal is to develop conditions under which certain stationary points of the program (2) are statistically consistent for β^*.

2.1. *Robust M-estimators.* We wish to study functions $\mathcal{L}_n$ that are robust to outliers and/or model misspecification. Consequently, we borrow our loss functions from the classical theory of robust regression; the additional regularizer ρ_λ

appearing in the program (2) encourages sparsity and endows it with appealing behavior in high dimensions. Here, we provide a brief review of M-estimators used for robust linear regression. For a more detailed treatment of the basic concepts of robust regression, see the books [Huber (1981), Maronna, Martin and Yohai (2006), Hampel et al. (1986)] and the many references cited therein.

Let ℓ denote the regression function defined on an individual observation pair (x_i, y_i). The corresponding M-estimator is then

$$\mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i^T\beta - y_i). \tag{4}$$

Note that

$$\mathbb{E}\big[\nabla \mathcal{L}_n(\beta^*)\big] = \mathbb{E}\big[\ell'(x_i^T\beta^* - y_i)x_i\big] = \mathbb{E}\big[\ell'(\varepsilon_i)x_i\big] = \mathbb{E}\big[\ell'(\varepsilon_i)\big]\cdot\mathbb{E}[x_i] = 0,$$

so the condition (3) is always satisfied. In particular, the maximum likelihood estimator corresponds to the choice $\ell(u) = -\log p_\varepsilon(u)$, where p_ε is the probability density function of the additive errors ε_i. Note that when $\varepsilon_i \sim N(0,1)$, the MLE corresponds to the choice $\ell(u) = \frac{u^2}{2}$, and the resulting loss function is convex.

Several popular robust loss functions that we will study in this paper include the Huber, Tukey and Cauchy losses, which are reviewed in Appendix B.1. Although second and third derivatives do not exist for all these loss functions, a unifying property is that the derivative ℓ' is *bounded* in each case. This turns out to be an important property for robustness of the resulting estimator. Intuitively, we may view a solution $\widehat{\beta}$ of the program (2) as an approximate sparse solution to the estimating equation $\nabla\mathcal{L}_n(\beta) = 0$, or equivalently,

$$\frac{1}{n}\sum_{i=1}^{n} \ell'(x_i^T\beta - y_i)x_i = 0. \tag{5}$$

When $\beta = \beta^*$, equation (5) becomes

$$\frac{1}{n}\sum_{i=1}^{n} \ell'(\varepsilon_i)x_i = 0. \tag{6}$$

In particular, if ε_i is an outlier, its contribution to the sum in equation (6) is bounded when ℓ' is bounded, lessening the contamination effect of gross outliers.

In the robust regression literature, a *redescending M-estimator* has the additional property that there exists $\xi_0 > 0$ such that $|\ell'(u)| = 0$, for all $|u| \geq \xi_0$. Then ξ_0 is known as a *finite rejection point*, since outliers (x_i, y_i) with $|\varepsilon_i| \geq \xi_0$ will be completely eliminated from the summand in equation (6). For instance, the Tukey loss gives rise to a redescending M-estimator.[1] Note that redescending M-estimators will always be nonconvex, so computational efficiency will be sacrificed at the expense of finite rejection properties. For an in-depth discussion of

[1]The Cauchy loss has the property that $\lim_{u\to\infty}|\ell'(u)| = 0$, but it is not redescending for any finite ξ_0.

redescending M-estimators, namely different measures of robustness, see the article by Shevlyakov, Morgenthaler and Shurygin (2008).

2.2. *Generalized M-estimators.* Whereas the M-estimators described in Section 2.1 are robust with respect to outliers in the additive noise terms ε_i, they are nonrobust to outliers in the covariates x_i. This may be quantified using the concept of influence functions (see Appendix A). Intuitively, an outlier in x_i may cause the corresponding term in equation (6) to behave arbitrarily badly. This motivates the use of *generalized M-estimators* that downweight large values of x_i (also known as leverage points). The resulting estimating equation is then defined as follows:

$$\sum_{i=1}^{n} \eta(x_i, x_i^T \beta - y_i) x_i = 0, \tag{7}$$

where $\eta : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ is defined appropriately. As will be discussed in the sequel, generalized M-estimators may allow us to relax the distributional assumptions on the covariates, for example, from sub-Gaussian to sub-exponential. Recall the following definitions [Vershynin (2012)].

DEFINITION 1 (Sub-Gaussian and sub-exponential random variables). (i) A random variable X is *sub-Gaussian with parameter* σ_x if $(\mathbb{E}(|X|^s))^{1/s} \le \sigma_x \sqrt{s}$, for all $s = 1, 2, \ldots$.

(ii) A random variable X is *sub-exponential with parameter* σ_x if $(\mathbb{E}(|X|^s))^{1/s} \le \sigma_x s$, for all $s = 1, 2, \ldots$.

A matrix $X \in \mathbb{R}^{n \times p}$ with i.i.d. rows $\{x_i^T\}$ is sub-Gaussian (sub-exponential) with parameter σ_x if for any unit vector $u \in \mathbb{R}^p$, the random variable $u^T x_i$ is sub-Gaussian (sub-exponential) with parameter σ_x.

We will focus on functions η that take the form

$$\eta(x_i, r_i) = w(x_i) \ell'\big(r_i \cdot v(x_i)\big), \tag{8}$$

where $w, v > 0$ are weighting functions. Note that the M-estimators considered in Section 2.1 may also be written in this form, where $w \equiv v \equiv 1$. The Mallows, Hill–Ryan and Schweppe functions, which are popular weight functions in robust statistics and are of the form presented in equation (8), are defined in Appendix B.2.

Note that when η takes the form in equation (8), the estimating equation (7) may again be seen as a zero-gradient condition $\nabla \mathcal{L}_n(\beta) = 0$, where

$$\mathcal{L}_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \frac{w(x_i)}{v(x_i)} \ell\big((x_i^T \beta - y_i) v(x_i)\big). \tag{9}$$

Under reasonable conditions, such as oddness of ℓ' and symmetry of the error distribution, the condition (3) may be seen to hold [cf. condition (2) of Proposition 1 below and the following remark]. The overall program for a generalized

M-estimator then takes the form

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

2.3. *Nonconvex regularizers*. Finally, we provide some background on the types of regularizers we will use in our analysis of the composite objective function (2). Following Loh and Wainwright (2014, 2015), we require the regularizer ρ_λ to satisfy the following properties.

ASSUMPTION 1 (Amenable regularizers). The regularizer is coordinate-separable:

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

for some scalar function $\rho_\lambda : \mathbb{R} \mapsto \mathbb{R}$. In addition:

(i) The function $t \mapsto \rho_\lambda(t)$ is symmetric around zero and $\rho_\lambda(0) = 0$.
(ii) The function $t \mapsto \rho_\lambda(t)$ is nondecreasing on $\mathbb{R}^+$.
(iii) The function $t \mapsto \frac{\rho_\lambda(t)}{t}$ is nonincreasing on $\mathbb{R}^+$.
(iv) The function $t \mapsto \rho_\lambda(t)$ is differentiable for $t \neq 0$.
(v) $\lim_{t \to 0^+} \rho'_\lambda(t) = \lambda$.
(vi) There exists $\mu > 0$ such that the function $t \mapsto \rho_\lambda(t) + \frac{\mu}{2} t^2$ is convex.
(vii) There exists $\gamma \in (0, \infty)$ such that $\rho'_\lambda(t) = 0$ for all $t \ge \gamma\lambda$.

If ρ_λ satisfies conditions (i)–(vi) of Assumption 1, we say that ρ_λ is μ-*amenable*. If ρ_λ also satisfies condition (vii), we say that ρ_λ is (μ, γ)-*amenable*. In particular, if ρ_λ is μ-amenable, then $q_\lambda(t) := \lambda|t| - \rho_\lambda(t)$ is everywhere differentiable. Defining the vector version $q_\lambda : \mathbb{R}^p \to \mathbb{R}$ accordingly, it is easy to see that $\frac{\mu}{2}\|\beta\|_2^2 - q_\lambda(\beta)$ is convex.

Some popular examples of nonconvex regularizers satisfying the above properties, particularly the SCAD and MCP, are provided in Appendix B.3. As studied in detail in Loh and Wainwright (2014) and leveraged in the results of Section 3.3 below, using (μ, γ)-amenable regularizers allows us to derive a powerful oracle result concerning local stationary points, which will be useful for our discussion of asymptotic normality.

3. Main statistical results. We now present our core results concerning stationary points of the high-dimensional robust M-estimators described in Section 2. We begin with a general deterministic result that ensures statistical consistency of stationary points of the program (2) when the loss function satisfies restricted strong convexity and the regularizer is μ-amenable. Next, we interpret the consequences of our theorem for specific M-estimators and generalized M-estimators

through a series of propositions, and provide conditions on the distributions of the covariates and error terms under which the assumptions hold with high probability. Lastly, we provide a theorem establishing that stationary points are equal to a local oracle estimator when the regularizer is nonconvex and (μ, γ)-amenable.

Recall that $\widetilde{\beta}$ is a *stationary point* of the program (2) if

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) + \nabla \rho_\lambda(\widetilde{\beta}), \beta - \widetilde{\beta} \rangle \geq 0,$$

for all feasible β, where we abuse notation slightly and write $\nabla \rho_\lambda(\widetilde{\beta}) = \lambda \operatorname{sign}(\widetilde{\beta}) - \nabla q_\lambda(\widetilde{\beta})$ (recall that q_λ is differentiable by our assumptions). The set of stationary points includes all local and global minima, as well as interior local maxima [Bertsekas (1999), Clarke (1983)].

3.1. *General statistical theory.* We require the loss function $\mathcal{L}_n$ to be differentiable and satisfy the following local restricted strong convexity (RSC) condition.

ASSUMPTION 2 (Local RSC condition). There exist $\alpha > 0$ and $\tau \geq 0$, and a radius $r > 0$, such that for all $\beta_1, \beta_2 \in \mathbb{B}_r(\beta^*)$,

$$\langle \nabla \mathcal{L}_n(\beta_1) - \nabla \mathcal{L}_n(\beta_2), \beta_1 - \beta_2 \rangle \geq \alpha \|\beta_1 - \beta_2\|_2^2 - \tau \frac{\log p}{n} \|\beta_1 - \beta_2\|_1^2. \tag{10}$$

The notion of restricted strong convexity was introduced by Negahban et al. (2012) and Agarwal, Negahban and Wainwright (2012) to analyze statistical and optimization properties of convex regularized M-estimators, and extended by Loh and Wainwright (2015) to the case of nonconvex functions. However, unlike the analogous definition in Loh and Wainwright (2015), Assumption 2 imposes *no* conditions on the behavior of $\mathcal{L}_n$ outside the ball of radius r centered at β^*. The intuition behind the RSC condition is that it is a weaker version of strong convexity, which is exactly the condition when $\alpha > 0$ and $\tau = 0$. However, for $\tau > 0$, we allow the inner product on the left-hand side of inequality (10) to be negative for some values of β_1 and β_2. Note that when $p > n$, the loss function $\mathcal{L}_n$ is not strongly convex even when ℓ is convex, since directions exist in which $\mathcal{L}_n$ has zero curvature. Condition (10) therefore only imposes positive curvature of $\mathcal{L}_n$ for vectors in the set where $\frac{\|\beta_1 - \beta_2\|_1}{\|\beta_1 - \beta_2\|_2} \leq c\sqrt{\frac{n}{\log p}}$, which turns out to be sufficient for stationary points to be consistent. (This set includes all k-sparse vectors $\beta_1 - \beta_2$, provided $n \geq c^2 k \log p$.)

Since the loss functions used in robust regression are often nonconvex further away from the origin, Assumption 2 only requires the RSC condition to hold *locally*. Accordingly, the theoretical guarantees derived in this paper only concern the behavior of local behavior of stationary points around β^*. As discussed in more detail below, we will take r to scale as a constant independent of n, p, and k. The ball of radius r essentially cuts out a local basin of attraction around β^* in which

stationary points of the M-estimator are well behaved. Furthermore, our optimization results in Section 4 guarantee that we may efficiently locate stationary points within this constant-radius region via a two-step M-estimator.

We have the following main result, which requires the regularizer and loss functions to satisfy Assumptions 1 and 2, respectively. The theorem guarantees that stationary points within the local region of restricted strong convexity are statistically consistent. Recall that k refers to the sparsity level of the true parameter vector β^*.

THEOREM 1. *Suppose $\mathcal{L}_n$ satisfies the local RSC condition (10) with $\beta_2 = \beta^*$, for all $\beta_1 \in \mathbb{B}_r(\beta^*)$. Also suppose ρ_λ is μ-amenable with $\frac{3}{4}\mu < \alpha$. Further suppose $n \geq \frac{C}{r^2} \cdot k \log p$ and $R \geq \|\beta^*\|_1$, and suppose*

$$\lambda \geq \max\left\{4\|\nabla \mathcal{L}_n(\beta^*)\|_\infty, 8\tau R \frac{\log p}{n}\right\}. \tag{11}$$

Then there exists a stationary point $\widetilde{\beta}$ of the program (2) such that $\|\widetilde{\beta} - \beta^*\|_2 \leq r$. Furthermore,

$$\|\widetilde{\beta} - \beta^*\|_2 \leq \frac{24\lambda\sqrt{k}}{4\alpha - 3\mu} \quad \text{and} \quad \|\widetilde{\beta} - \beta^*\|_1 \leq \frac{96\lambda k}{4\alpha - 3\mu}. \tag{12}$$

In particular, for the choice $\lambda \asymp \sqrt{\frac{\log p}{n}}$, the inequalities (12) imply that

$$\|\widetilde{\beta} - \beta^*\|_2 \leq \frac{24}{4\alpha - 3\mu}\sqrt{\frac{k \log p}{n}} \quad \text{and} \quad \|\widetilde{\beta} - \beta^*\|_2 \leq \frac{96}{4\alpha - 3\mu} k\sqrt{\frac{\log p}{n}}.$$

The proof of Theorem 1 is contained in Section C.1. Note that the statement of the theorem is entirely deterministic, and the distributional properties of the covariates and error terms come into play in verifying that inequality (11) and the local RSC condition (10) hold with high probability. As will be shown below (cf. Proposition 1), we have $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \precsim \sqrt{\frac{\log p}{n}}$ under fairly mild distributional assumptions, provided ℓ' is bounded. Furthermore, for $R \leq c\sqrt{k}$, we also have $\frac{R \log p}{n} \precsim \sqrt{\frac{\log p}{n}}$ under the prescribed sample size scaling, thus justifying the choice $\lambda \asymp \sqrt{\frac{\log p}{n}}$ suggested in the statement of Theorem 1. Finally, note that the theorem does not require the condition (10) to hold uniformly over all pairs in the ball $\mathbb{B}_r(\beta^*)$, as in Assumption 2, but only for $\beta_2 = \beta^*$ and $\beta_1 \in \mathbb{B}_r(\beta^*)$.

REMARK. Although Theorem 1 only guarantees the statistical consistency of stationary points within the local region of radius r, it is essentially the strongest conclusion one can draw based on a local RSC assumption (10) alone. The power of Theorem 1 lies in the fact that when r is chosen to be a constant and

$\frac{k \log p}{n} = o(1)$, as is the case in our robust regression settings of interest, all stationary points within the constant-radius region are guaranteed to fall within a shrinking ball of radius $\mathcal{O}(\sqrt{\frac{k \log p}{n}})$ centered around β^*. Hence, the stationary points in the local region are statistically consistent at the usual minimax rate expected for ℓ_1-penalized ordinary least squares regression with sub-Gaussian data. Furthermore, the dimensions n, p and k are all allowed to grow; in particular, we do not assume that k is bounded by a constant. As we will illustrate in more detail in the next section, if robust loss functions with bounded derivatives are used in place of the ordinary least squares loss, the statistical consistency conclusion of Theorem 1 still holds even when the additive errors follow a heavy-tailed distribution or are contaminated by outliers.

As a corollary to the proof of Theorem 1, we have the following result, which holds when $\mathcal{L}_n$ is convex. It applies to loss functions such as the Huber loss, and will be relevant to our discussion of two-step estimators in Section 4.2, where we will use a convex robust loss in the first step of an optimization procedure designed to find stationary points of the program (2), even when ℓ is nonconvex. The proof of Corollary 1 is provided in Appendix E.1.

COROLLARY 1. *Suppose, in addition to the assumptions stated in Theorem 1, that the loss function $\mathcal{L}_n$ is convex. Also suppose $n \geq \frac{2\tau}{\alpha} \cdot k \log p$. Then the program (2) possesses a unique stationary point $\widetilde{\beta}$, and $\widetilde{\beta}$ is contained in the ball $\mathbb{B}_r(\beta^*)$.*

3.2. *Establishing sufficient conditions.* From Theorem 1, we see that the key ingredients for statistical consistency of local stationary points are (i) the boundedness of $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty$ in inequality (11), which ultimately dictates the ℓ_2-rate of convergence of $\widetilde{\beta}$ to β^* up to a factor of $\sqrt{k}$, and (ii) the local RSC condition (10) in Assumption 2. We provide more interpretable sufficient conditions in this section via a series of propositions.

For the results of this section, we will require some boundedness conditions on the derivatives of the loss function ℓ, which we state in the following assumption.

ASSUMPTION 3 (Derivative assumptions). Suppose there exist $\kappa_1, \kappa_2 \geq 0$ such that

$$|\ell'(u)| \leq \kappa_1, \qquad \forall u, \tag{13}$$

$$\ell''(u) \geq -\kappa_2, \qquad \forall u. \tag{14}$$

Note that the bounded derivative assumption (13) holds for all the robust loss functions highlighted in Appendix B.1 (but *not* for the ordinary least squares loss), and $\kappa_1 \asymp \xi$ in each of those cases. Indeed, it is well known from classical robust

statistics (cf. Appendix A) that loss functions with a bounded derivative exactly correspond to bounded influence functions in the fixed-covariate setting. Furthermore, inequality (14) holds with $\kappa_2 = 0$ when ℓ is convex and twice-differentiable, but the inequality also holds for nonconvex losses such as the Tukey and Cauchy loss with $\kappa_2 > 0$. By a more careful argument, we may eschew the condition (14) if ℓ is a convex function that is in C^1 but not C^2, as in the case of the Huber loss, since Theorem 1 only requires first-order differentiability of $\mathcal{L}_n$ and q_λ; however, we state the propositions with Assumption 3 for the sake of simplicity.

We have the following proposition, which establishes the gradient bound (11) with high probability under fairly mild assumptions.

PROPOSITION 1. *Suppose ℓ satisfies the bounded derivative condition* (13) *and the following conditions also hold*:

(1) *$w(x_i)x_i$ is sub-Gaussian with parameter σ_w.*
(2) *Either*

$$\text{(a)} \quad v(x_i) = 1 \quad \textit{and} \quad \mathbb{E}\big[w(x_i)x_i\big] = 0, \quad \textit{or} \quad \text{(b)} \quad \mathbb{E}\big[\ell'\big(\varepsilon_i \cdot v(x_i)\big) \mid x_i\big] = 0.$$

With probability at least $1 - c_1 \exp(-c_2 \log p)$, the loss function defined by equation (9) *satisfies the bound*

$$\big\|\nabla \mathcal{L}_n(\beta^*)\big\|_\infty \le c\kappa_1 \sigma_w \sqrt{\frac{\log p}{n}}.$$

The proof of Proposition 1 is a simple but important application of sub-Gaussian tail bounds and is provided in Appendix D.1.

REMARK. Note that for the unweighted M-estimator (4), conditions (1) and (2a) of Proposition 1 hold when x_i is sub-Gaussian and $\mathbb{E}[x_i] = 0$. If the x_i's are not sub-Gaussian, condition (1) nonetheless holds whenever $w(x_i)x_i$ is bounded. Furthermore, condition (2b) holds whenever ε_i has a symmetric distribution and ℓ' is an odd function. We further highlight the fact that aside from a possible mild requirement of symmetry, the concentration result given in Proposition 1 is *independent of the distribution of* ε_i, and holds equally well for heavy-tailed error distributions. The distributional effect of the x_i's is captured in the sub-Gaussian parameter σ_w; in settings where the contaminated data still follow a sub-Gaussian distribution, but the sub-Gaussian parameter is inflated due to large leverage points, using a weight function as defined in equation (4) may lead to a significant decrease in the value of σ_w. This decreases the finite-sample bias of the overall estimator.

Establishing the local RSC condition in Assumption 2 is more subtle, and the propositions described below depend in a more complex fashion on the distribution of the ε_i's. As noted above, the statistical consistency result in Theorem 1 only

requires inequality (10) to hold with $\beta_2 = \beta^*$. However, for the stronger oracle result of Theorem 2, we will require the full form of Assumption 2 to hold over all pairs (β_1, β_2) in the local region. We will quantify the parameters of the RSC condition in terms of an additional parameter $T > 0$, which is treated as a fixed constant. Define the tail probability

$$\varepsilon_T := \mathbb{P}\left(|\varepsilon_i| \geq \frac{T}{2}\right), \tag{15}$$

and the lower-curvature bound

$$\alpha_T := \min_{|u| \leq T} \ell''(u) > 0, \tag{16}$$

where ℓ'' is assumed to exist on the interval $[-T, T]$. We assume that T is chosen small enough so that $\alpha_T > 0$.

We first consider the case where the loss function takes the usual form of an unweighted M-estimator (4). We have the following proposition, proved in Appendix D.2.

PROPOSITION 2. *Suppose the x_i's are drawn from a sub-Gaussian distribution with parameter σ_x and the loss function is defined by equation* (4). *Also suppose*

$$c\sigma_x^2\left(\varepsilon_T^{1/2} + \exp\left(-\frac{c'T^2}{\sigma_x^2 r^2}\right)\right) \leq \frac{\alpha_T}{\alpha_T + \kappa_2} \cdot \frac{\lambda_{\min}(\Sigma_x)}{2}. \tag{17}$$

Suppose ℓ satisfies Assumption 3 and the sample size satisfies $n \geq c_0 k \log p$. With probability at least $1 - c\exp(-c' \log p)$, the loss function $\mathcal{L}_n$ satisfies Assumption 2 with

$$\alpha = \alpha_T \cdot \frac{\lambda_{\min}(\Sigma_x)}{16} \quad \textit{and} \quad \tau = \frac{C(\alpha_T + \kappa_2)^2 \sigma_x^2 T^2}{r^2}.$$

REMARK. Note that for a fixed value of T, inequality (17) places a tail condition on the distribution of ε_i via the term ε_T. This may be interpreted as a bound on the variance of the error distribution when ε_i is sub-Gaussian, or a bound on the fraction of outliers when ε_i has a contaminated distribution. Furthermore, the exponential term decreases as a function of the ratio $\frac{T}{r}$. Hence, for a larger value of ε_T, the radius r will need to be smaller in order to satisfy the bound (17). This agrees with the intuition that the local basin of good behavior for the M-estimator is smaller for larger levels of contamination. Finally, note that although α_T and κ_2 are deterministic functions of the known regression function ℓ and could be computed, the values of $\lambda_{\min}(\Sigma_x)$ and σ_x^2 are usually unknown a priori. Hence, Proposition 2 should be viewed as more of a qualitative result describing the behavior of the RSC parameters as the amount of contamination of the error distribution increases, rather than a bound that can be used to select a suitable robust loss function.

The situation where $\mathcal{L}_n$ takes the form of a generalized M-estimator (9) is more difficult to analyze in its most general form, so we will instead focus on verifying the local RSC condition (10) for the Mallows and Hill–Ryan estimators described in Section 2.2. We will show that the RSC condition holds under weaker conditions on the distribution of the x_i's. We have the following lemmas, proved in Appendices D.3 and D.4.

PROPOSITION 3 (Mallows estimator). *Suppose the x_i's are drawn from a subexponential distribution with parameter σ_x and the loss function is defined by*

$$\mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} w(x_i)\ell\big(x_i^T\beta - y_i\big),$$

and $w(x_i) = \min\{1, \frac{b}{\|Bx_i\|_2}\}$. Also suppose

$$cb\|\!|B^{-1}|\!\|_2\sigma_x^2\left(\varepsilon_T^{1/2} + \exp\left(-\frac{c'T}{\sigma_x r}\right)\right) \le \frac{\alpha_T}{2(\alpha_T+\kappa_2)}\cdot\lambda_{\min}\big(\mathbb{E}\big[w(x_i)x_ix_i^T\big]\big).$$

Suppose ℓ satisfies Assumption 3, and suppose the sample size satisfies $n \ge c_0 k\log p$. With probability at least $1 - c\exp(-c'\log p)$, the loss function $\mathcal{L}_n$ satisfies Assumption 2 with

$$\alpha = \alpha_T\cdot\frac{\lambda_{\min}(\mathbb{E}[w(x_i)x_ix_i^T])}{16} \quad \text{and} \quad \tau = \frac{C(\alpha_T+\kappa_2)^2\sigma_x^2T^2}{r^2}.$$

PROPOSITION 4 (Hill–Ryan estimator). *Suppose the loss function is defined by*

$$\mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} w(x_i)\ell\big(\big(x_i^T\beta - y_i\big)w(x_i)\big),$$

where $w(x_i) = \min\{1, \frac{b}{\|Bx_i\|_2}\}$. Also suppose

$$\begin{aligned} &cb^2|\!|\!|B^{-1}|\!|\!|_2^2\left(\varepsilon_T^{1/2} + \exp\left(-\frac{c'T^2}{b^2|\!|\!|B^{-1}|\!|\!|_2^2\sigma_x^2r^2}\right)\right) \\ &\quad\le \frac{\alpha_T}{2(\alpha_T+\kappa_2)}\cdot\lambda_{\min}\big(\mathbb{E}\big[w(x_i)x_ix_i^T\big]\big). \end{aligned} \tag{18}$$

Suppose ℓ satisfies Assumption 3, and suppose the sample size satisfies $n \ge c_0 k\log p$. With probability at least $1 - c\exp(-c'\log p)$, the loss function $\mathcal{L}_n$ satisfies Assumption 2 with

$$\alpha = \alpha_T\cdot\frac{\lambda_{\min}(w(x_i)x_ix_i^T)}{16} \quad \text{and} \quad \tau = \frac{C(\alpha_T+\kappa_2)b^2|\!|\!|B^{-1}|\!|\!|_2^2T^2}{r^2}.$$

REMARK. Due to the presence of the weighting function $w(x_i)$, Proposition 3 imposes weaker distributional requirements on the x_i's than Proposition 2, and the requirements imposed in Proposition 4 are still weaker. In fact, a version of Proposition 3 could be derived with $w(x_i) = \min\{1, \frac{b^2}{\|Bx_i\|_2^2}\}$, which would not require the x_i's to be sub-exponential. The tradeoff in comparing Proposition 4 to Propositions 2 and 3 is that although the RSC condition holds under weaker distributional assumptions on the covariates, the absolute bound $b^2|||B^{-1}|||_2^2$ used in place of the sub-Gaussian/exponential parameter σ_x may be much larger. Hence, the relative size of ε_T and the radius r will need to be smaller in order for inequality (18) to be satisfied, relative to the requirement for inequality (17).

In Section 5 below, we explore the consequences of Propositions 1–4 for heavy-tailed, outlier and sub-exponential distributions.

3.3. *Oracle results and asymptotic normality.* As discussed in the preceding two subsections, penalized robust M-estimators produce local stationary points that enjoy ℓ_1- and ℓ_2-consistency whenever ℓ' is bounded and the errors and covariates satisfy suitable mild assumptions. In fact, a distinguishing aspect of different robust loss functions lies not in first-order comparisons, but in second-order considerations concerning the variance of the estimator. This is a well-known concept in classical robust regression analysis, where p is fixed, $n \to \infty$ and the objective function does not contain a penalty term. By the Cramér–Rao bound and under fairly general regularity conditions [Lehmann and Casella (1998)], the optimal choice of ℓ that minimizes the asymptotic variance in the low-dimensional setting is the MLE function, $\ell(u) = -\log p_\varepsilon(u)$, where p_ε is the probability density function of ε_i. When the class of regression functions is constrained to those with bounded influence functions, however, a more complex analysis reveals that choices of ℓ corresponding, for example, to the losses introduced in Section 2.2 produce better performance [Huber (1981)].

In this section, we establish oracle properties of penalized robust M-estimators. Our main result shows that under many of the assumptions stated earlier, local stationary points of the regularized M-estimators coincide with the local oracle result, defined by

$$\widehat{\beta}_S^{\mathcal{O}} \in \arg \min_{\beta \in \mathbb{R}^S : \|\beta - \beta^*\|_2 \le r} \{\mathcal{L}_n(\beta)\}. \tag{19}$$

This is particularly attractive from a theoretical standpoint, because the oracle result implies that local stationary points inherit all the properties of the lower-dimensional oracle estimator $\widehat{\beta}_S^{\mathcal{O}}$, which is covered by previous theory.

Note that $\widehat{\beta}_S^{\mathcal{O}}$ is truly an oracle estimator, since it requires knowledge of both the actual support set S of β^* *and* of β^* itself; the optimization of the loss function is taken only over a small neighborhood around β^*. In cases where $\mathcal{L}_n$ is

convex or global optima of $\mathcal{L}_n$ that are supported on S lie in the ball of radius r centered around β^*, the constraint $\|\beta - \beta^*\|_2 \leq r$ may be omitted. If $\mathcal{L}_n$ satisfies Assumption 2, the oracle program (19) is guaranteed to be convex, as stated in the following simple lemma, proved in Appendix F.1.

LEMMA 1. *Suppose $\mathcal{L}_n$ satisfies Assumption 2 and $n \geq \frac{2\tau}{\alpha} k \log p$. Then $\mathcal{L}_n$ is strongly convex over the region $S_r := \{\beta \in \mathbb{R}^p : \operatorname{supp}(\beta) \subseteq S, \|\beta - \beta^*\|_2 \leq r\}$.*

In particular, the oracle estimator $\widehat{\beta}_S^{\mathcal{O}}$ is guaranteed to be unique.

Our central result of this section shows that when the regularizer is (μ, γ)-amenable and the loss function satisfies the local RSC condition in Assumption 2, stationary points of the M-estimator (2) within the local neighborhood of β^* are in fact unique and equal to the oracle estimator (19). We also require a condition on the minimum signal strength, which we denote by $\beta^*_{\min} := \min_{j \in S} |\beta^*_j|$. For simplicity, we state the theorem as a probabilistic result for sub-Gaussian covariates and the unweighted M-estimator (4); similar results could be derived for generalized M-estimators under weaker distributional assumptions.

THEOREM 2. *Suppose the loss function $\mathcal{L}_n$ is given by the M-estimator (4) and is twice differentiable in the ℓ_2-ball of radius r around β^*. Suppose the regularizer ρ_λ is (μ, γ)-amenable. Under the same conditions of Theorem 1, suppose in addition that $\|\beta^*\|_1 \leq \frac{R}{2}$ and $\frac{160\lambda k}{2\alpha - \mu} < R$, and $\beta^*_{\min} \geq C\sqrt{\frac{\log k}{n}} + \gamma\lambda$. Suppose $n \geq c_0 \max\{k^2, k \log p\}$. With probability at least $1 - c\exp(-c' \min\{k, \log p\})$, any stationary point $\widetilde{\beta}$ of the program (2) such that $\|\widetilde{\beta} - \beta^*\|_2 \leq r$ satisfies $\operatorname{supp}(\widetilde{\beta}) \subseteq S$ and $\widetilde{\beta}_S = \widehat{\beta}_S^{\mathcal{O}}$.*

The proof of Theorem 2 builds upon the machinery developed in the recent paper [Loh and Wainwright (2014)]. However, the argument here is slightly simpler, because we only need to prove the oracle result for stationary points within a radius r of β^*. For completeness, we include a proof of Theorem 2 in Section C.2, highlighting the modifications that are necessary to obtain the statement in the present paper.

REMARK. Several other papers [Fan and Peng (2004), Bradic, Fan and Wang (2011), Li, Peng and Zhu (2011)] have established oracle results of a similar flavor, but only in cases where the M-estimator takes the form described in Section 2.1 and the loss is convex. Furthermore, the results of previous authors only concern global optima and/or guarantee the existence of local optima with the desired oracle properties. Hence, our conclusions are at once more general and more complex, since we need a more careful treatment of possible local optima.

In fact, since the oracle program (19) is essentially a k-dimensional optimization problem, Theorem 2 allows us to apply previous results in the literature concerning the asymptotic behavior of low-dimensional M-estimators to simultaneously analyze the asymptotic distribution of $\widehat{\beta}_S^{\mathcal{O}}$ and $\widetilde{\beta}$. Huber (1973) studied asymptotic properties of M-estimators when the loss function is convex, and established asymptotic normality assuming $\frac{p^3}{n} \to 0$, a result which was improved upon by Yohai and Maronna (1979). Portnoy (1985) and Mammen (1989) extended these results to nonconvex M-estimators. Fewer results exist concerning generalized M-estimators: Bai and Wu (1997) and He and Shao (1996) established asymptotic normality for a fairly general class of estimators, but the assumption is that p is fixed and $n \to \infty$. He and Shao (2000) extended their results to the case where p is also allowed to grow and proved asymptotic normality when $\frac{p^2 \log p}{n} \to 0$, assuming a convex loss.

Although the overall M-estimator may be highly nonconvex, the *restricted* program (19) defining the oracle estimator is nonetheless convex (cf. Lemma 1 above). Hence, the standard convex theory for M-estimators with a diverging number of parameters applies without modification. Since the regularity conditions existing in the literature that guarantee asymptotic normality vary substantially depending on the form of the loss function, we only provide a sample corollary for a specific (unweighted) case, as an illustration of the types of results on asymptotic normality that may be derived from Theorem 2.

COROLLARY 2. *Suppose the loss function $\mathcal{L}_n$ is given by the M-estimator (4) and the regularizer ρ_λ is (μ, γ)-amenable. Under the same conditions of Theorem 2, suppose in addition that $\ell \in C^3$, $\mathbb{E}[\ell''(\varepsilon_i)] \in (0, \infty)$, and $k \geq C \log n$. Let $\widetilde{\beta}$ be any stationary point of the program (2) such that $\|\widetilde{\beta} - \beta^*\|_2 \leq r$. If $\frac{k \log^3 k}{n} \to 0$, then $\|\widetilde{\beta} - \beta^*\|_2 = \mathcal{O}_P(\sqrt{\frac{k}{n}})$. If $\frac{k^2 \log k}{n} \to 0$, then for any bounded sequence $\{v_n\} \subseteq \mathbb{R}^p$, we have*

$$\frac{\sqrt{n}}{\sigma(v_n)} \cdot v_n^T \big(\widetilde{\beta} - \beta^*\big) \xrightarrow{d} N(0, 1),$$

where

$$\sigma^2(v) := \frac{1}{\mathbb{E}[\ell''(\varepsilon_i)] \cdot \mathbb{E}[(\ell'(\varepsilon_i))^2]} \cdot v^T \left(\frac{X^T X}{n} \right) v.$$

The proof of Corollary 2 is provided in Appendix E.2. Analogous results may be derived for other loss functions considered in this paper under slightly different regularity assumptions, by modifying appropriate low-dimensional results with diverging dimensionality [e.g., Mammen (1989), Portnoy (1985)].

4. Optimization. We now discuss how our statistical theory gives rise to a useful two-step algorithm for optimizing the resulting high-dimensional M-estimators. We first present some theory for the composite gradient descent algorithm, including rates of convergence for the regularized problem. We then describe our new two-step algorithm, which is guaranteed to converge to a stationary point within the local region where the RSC condition holds.

4.1. *Composite gradient descent.* In order to obtain stationary points of the program (2), we use composite gradient descent [Nesterov (2007)]. Denoting $\overline{\mathcal{L}}_n(\beta) := \mathcal{L}_n(\beta) - q_\lambda(\beta)$, we may rewrite the program as

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \le R} \{\overline{\mathcal{L}}_n(\beta) + \lambda\|\beta\|_1\}.$$

Then the composite gradient iterates are given by

$$\beta^{t+1} \in \arg\min_{\|\beta\|_1 \le R} \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 + \frac{\lambda}{\eta}\|\beta\|_1 \right\}, \tag{20}$$

where η is the stepsize parameter. Defining the soft-thresholding operator $S_{\lambda/\eta}(\beta)$ componentwise according to

$$S^j_{\lambda/\eta} := \operatorname{sign}(\beta_j)\left(|\beta_j| - \frac{\lambda}{\eta} \right)_+,$$

a simple calculation shows that the iterates (20) take the form

$$\beta^{t+1} = S_{\lambda/\eta}\left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right). \tag{21}$$

The following theorem guarantees that the composite gradient descent algorithm will converge at a linear rate to a point near β^* as long as the initial point β^0 is chosen close enough to β^*. We will require the following assumptions on $\mathcal{L}_n$, where

$$\mathcal{T}'(\beta_1, \beta_2) := \mathcal{L}_n(\beta_1) - \mathcal{L}_n(\beta_2) - \langle \nabla \mathcal{L}_n(\beta_2), \beta_1 - \beta_2 \rangle$$

denotes the Taylor remainder. In the statements below, we assume $\alpha', \alpha'' > 0$ and $\tau', \tau'' \ge 0$.

ASSUMPTION 4 (Local RSC′ and RSM conditions). Suppose $\mathcal{L}_n$ satisfies the restricted strong convexity condition

$$\mathcal{T}'(\beta_1, \beta_2) \ge \alpha'\|\beta_1 - \beta_2\|_2^2 - \tau' \frac{\log p}{n}\|\beta_1 - \beta_2\|_1^2, \qquad \forall \beta_1, \beta_2 \in \mathbb{B}_r(\beta^*). \tag{22}$$

In addition, suppose $\mathcal{L}_n$ satisfies the restricted smoothness (RSM) condition

$$\mathcal{T}'(\beta_1, \beta_2) \le \alpha''\|\beta_1 - \beta_2\|_2^2 + \tau'' \frac{\log p}{n}\|\beta_1 - \beta_2\|_1^2, \qquad \forall \beta_1, \beta_2 \in \mathbb{R}^p. \tag{23}$$

Note that the definition of $\mathcal{T}'$ in Assumption 4 differs slightly from the definition of the related Taylor difference used in Assumption 2. However, one may verify the RSC$'$ condition (22) in exactly the same way as we verify Assumption 2 via the mean value theorem argument of Section 3.2, so we do not repeat the proofs here. The restricted smoothness condition (23) is fairly mild and is easily seen to hold with $\tau'' = 0$ when the loss function ℓ appearing in the definition of the M-estimator has a bounded second derivative. We will also assume for simplicity that q_λ is convex, as is the case for the SCAD and MCP regularizers; the theorem may be extended to situations where q_λ is nonconvex, given an appropriate quadratic bound on the Taylor remainder of q_λ.

We have the following theorem, proved in Appendix C.3. It guarantees that as long as the initial point β^0 of the composite gradient descent algorithm is chosen close enough to β^*, the log of the ℓ_2-error between iterates β^t and a global minimizer $\widehat{\beta}$ of the regularized M-estimator (2) will decrease linearly with t up to the order of the statistical error $\|\widehat{\beta} - \beta^*\|_2$.

THEOREM 3. *Suppose $\mathcal{L}_n$ satisfies the local RSC$'$ condition (22) and the RSM condition (23), and suppose ρ_λ is μ-amenable with $\mu < 2\alpha$ and q_λ is convex. Suppose the regularization parameters satisfy the scaling*

$$C \max\left\{ \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \tau \sqrt{\frac{\log p}{n}} \right\} \le \lambda \le \frac{C'\alpha}{R}.$$

Also suppose $\widehat{\beta}$ is a global optimum of the objective (2) over $\mathbb{B}_{r/2}(\beta^*)$, and $\eta \ge 2\alpha''$ and

$$n \ge \frac{4(2\tau' + \tau'')}{\alpha' - \mu/2 + \eta/2} \cdot \frac{\alpha' - \mu/2}{\alpha' - \mu/2 + \eta/2} \cdot \frac{r^2}{4} \cdot R^2 \log p. \tag{24}$$

If $\beta^0 \in \mathbb{B}_{r/2}(\beta^*)$, successive iterates of the composite gradient descent algorithm satisfy

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

where $\delta^2 \ge \frac{c'\|\widehat{\beta} - \beta^*\|_2^2}{1-\kappa}$ is a tolerance parameter, $\kappa \in (0, 1)$, and $T^*(\delta) = \frac{c'' \log(1/\delta^2)}{\log(1/\kappa)}$.

REMARK. It is not obvious a priori that even if β^0 is chosen within a small constant radius of β^*, successive iterates will also remain close by. Indeed, the hard work to establish this fact is contained in the proof of Lemma 3 in Appendix C.3. Furthermore, note that we cannot expect a global convergence guarantee to hold in general, since the only assumption on $\mathcal{L}_n$ is the local version of RSC. Hence, a local convergence result such as the one stated in Theorem 3 is the best we can hope for in this scenario. In the simulations of Section 5, we see cases where

initializing the composite gradient descent algorithm outside the local basin of attraction where the RSC condition holds causes iterates to converge to a stationary point outside the local region, and the resulting stationary point is *not* consistent for β^*. Hence, the assumption concerning the proximity of β^0 to β^* in Theorem 3 is necessary in order to ensure good behavior of the optimization trajectory for nonconvex robust estimators.

4.2. *Two-step estimators*. As discussed in Section 3 above, whereas different choices of ℓ with bounded derivative yield estimators that are asymptotically unbiased and satisfy the same ℓ_2-bounds up to constant factors, certain M-estimators may be more desirable from the point of view of asymptotic efficiency. When ℓ is nonconvex, we can no longer guarantee fast global convergence of the composite gradient descent algorithm—indeed, the algorithm may even converge to statistically inconsistent local optima. Nonetheless, Theorem 3 guarantees that composite gradient descent will converge quickly to a desirable stationary point if the initial point is chosen within a constant radius of the true regression vector. We now propose a new two-step algorithm that may be applied to optimize high-dimensional robust M-estimators. Even when the regression function is nonconvex, our algorithm will always converge to a stationary point that is statistically consistent for β^*.

Two-step procedure.

(1) Run composite gradient descent using a convex regression function ℓ with convex ℓ_1-penalty, such that ℓ' is bounded.

(2) Use the output of step (1) to initialize composite gradient descent on the desired high-dimensional M-estimator.

According to our results on statistical consistency (cf. Theorem 1 and Corollary 1), step (1) will produce a global optimum $\widehat{\beta}^1$ such that $\|\widehat{\beta}^1 - \beta^*\|_2 \leq c\sqrt{\frac{k \log p}{n}}$, as long as the regression function ℓ is chosen appropriately.[2] Under the scaling $n \geq Cr^2 \cdot k \log p$, we then have $\|\widehat{\beta}^1 - \beta^*\|_2 \leq r$. Hence, by Theorem 3, composite gradient descent initialized at $\widehat{\beta}^1$ in step (2) will converge to a stationary point of the M-estimator at a linear rate. By our results of Section 3, the final output $\widehat{\beta}^2$ in step (2) is then statistically consistent and agrees with the local oracle estimator if we use a (μ, γ)-amenable penalty.

REMARK. Our proposed two-step algorithm bears some similarity to classical algorithms used for locating optima of robust regression estimators in low-dimensional settings. Recall the notion of a one-step M-estimator [Bickel (1975)],

[2]The rate of convergence may be sublinear in the initial iterations [Nesterov (2007)], but we are still guaranteed to have convergence, provided ℓ is convex.

which is obtained by taking a single step of the Newton–Raphson algorithm starting from a properly chosen initial point. Yohai (1987) and Simpson, Ruppert and Carroll (1992) study asymptotic properties of one-step GM- and MM-estimators in the setting where p is fixed, and show that the resulting regression estimators may simultaneously enjoy high-breakdown and high-efficiency properties. Welsh and Ronchetti (2002) present a finer-grained analysis of the asymptotic distribution and influence function of one-step M-estimators as a function of the initialization. Most directly related is the suggestion of Hampel et al. (1986) for optimizing redescending M-estimators using a one-step procedure initialized using a least median of squares estimator, in order to overcome the problem of nonconvexity and possibly multiple local optima; however, the method is mostly justified heuristically. Although each step of our two-step method involves running a composite gradient descent algorithm fully until convergence, the overall goal is still to produce an estimator at the end of the second step that is more efficient and has better theoretical properties than the solution of the first step alone.

The simulations in the next section demonstrate the efficacy of our two-step algorithm and the importance of step (1) in obtaining a proper initialization to step (2).

5. Simulations. In this section, we expound upon concrete instances of our theoretical results and provide simulation results. Data are generated i.i.d. from the linear model

$$y_i = x_i^T \beta^* + \varepsilon_i, \qquad \forall 1 \le i \le n.$$

5.1. *Statistical consistency.* In the first set of simulations, we verify the ℓ_2-consistency of high-dimensional robust M-estimators when data are drawn from various distributions.

We begin our discussion with a lemma that demonstrates the failure of the Lasso to achieve the minimax $\mathcal{O}(\sqrt{\frac{k \log p}{n}})$ rate when the ε_i's are drawn from an α-stable distribution with $\alpha < 2$. Recall that a variable X_0 has an α-stable distribution with scale parameter γ if the characteristic function of X_0 is given by

$$\mathbb{E}\big[\exp(itX_0)\big] = \exp\big(-\gamma^\alpha |t|^\alpha\big), \qquad \forall t \in \mathbb{R}, \tag{25}$$

and $\alpha \in (0, 2]$ [Nolan (2015)]. In particular, the standard normal distribution is an α-stable distribution with $(\alpha, \gamma) = (2, \frac{1}{\sqrt{2}})$, and the standard Cauchy distribution is an α-stable distribution with $(\alpha, \gamma) = (1, 1)$. The lemma is proved in Appendix F.2.

LEMMA 2. *Suppose X is a sub-Gaussian matrix and ε is an i.i.d. vector of α-stable random variables with scale parameter* 1. *Suppose $\lambda \asymp \sqrt{\frac{\log p}{n}}$. If $\alpha < 2$*

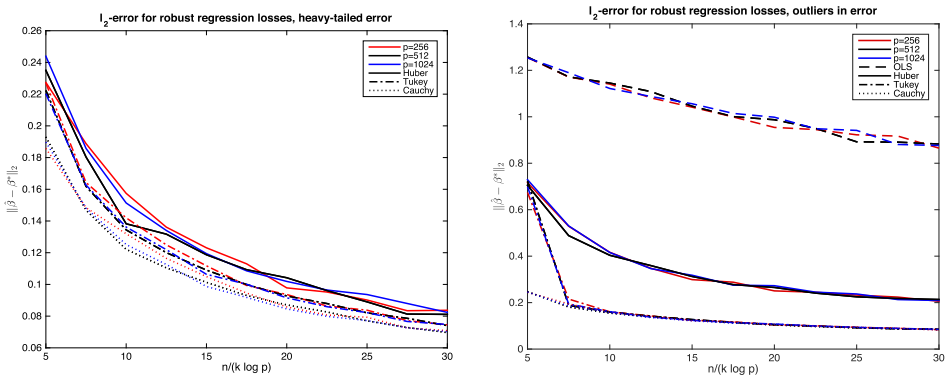
and $\log p = o(n^{\frac{2-\alpha}{\alpha}})$, *then*

$$\mathbb{P}\left(\left\|\frac{X^T\varepsilon}{n}\right\|_\infty \geq \lambda\right) \geq c_\alpha > 0,$$

where $c_\alpha \leq 1$ *is a constant that depends only on the sub-Gaussian parameter of the rows of* X *and does* not *scale with the problem dimensions. In particular, if* ε *is an i.i.d. vector of Cauchy random variables, the Lasso estimator is* inconsistent.

In contrast, as established in Theorem 1 and the propositions of Section 3.2, replacing the ordinary least squares loss by an appropriate robust loss function yields estimators that are consistent at the usual $\mathcal{O}(\sqrt{\frac{k\log p}{n}})$ rate.

In our first set of simulations, we generated ε_i's from a Cauchy distribution with scale parameter 0.1, and the x_i's from a standard normal distribution. We ran simulations for three problem sizes: $p = 256, 512$ and 1024, with sparsity level $k \approx p^{1/3}$. In each case, we set $\beta^* = (\frac{1}{\sqrt{k}}, \ldots, \frac{1}{\sqrt{k}}, 0, \ldots, 0)$. Figure 1(a) shows the results when the loss function $\mathcal{L}_n$ is equal to the Huber, Tukey and Cauchy robust losses, and the regularizer is the ℓ_1-penalty. The estimator $\widehat{\beta}$ was obtained using the composite gradient descent algorithm described in Section 4.1 in the case of the Huber loss, and the two-step algorithm described in Section 4.2 in the cases of the Tukey and Cauchy losses, with the output of the Huber estimator used to



(a) (b)

FIG. 1. *Plots showing statistical consistency of* ℓ_1*-penalized robust regression functions. Curves correspond to the Huber (solid), Tukey (dash-dotted), Cauchy (dotted) and ordinary least squares (dashed) losses, and are color-coded according to the problem sizes* $p = 256$ (*red*), 512 (*black*) *and* 1024 (*blue*). (a) *Plots for a heavy-tailed Cauchy error distribution. The Huber, Tukey and Cauchy robust losses all yield statistically consistent results.* (b) *Plots for a mixture of normals error distribution with* 30% *large-variance outliers. Since the error distribution is sub-Gaussian, the ordinary least squares loss also yields a statistically consistent estimator at minimax rates; however, the robust regression losses provide a significant improvement in the prefactor.*

initialize the second step of the algorithm. In each case, we set the regularization parameters $\lambda = 0.3\sqrt{\frac{\log p}{n}}$ and $R = 1.1\|\beta^*\|_1$, and averaged the results over 50 randomly generated data sets. As shown in the figure, the ℓ_1-penalized robust regression functions all yield statistically consistent estimators. Furthermore, the curves for different problem sizes align when the ℓ_2-error is plotted against the rescaled sample size $\frac{n}{k \log p}$, agreeing with the theoretical bound in Theorem 1.

We also ran simulations when the ε_i's were generated from a mixture of normals, representing a contaminated distribution with a constant fraction of outliers. With probability 0.7, the value of ε_i was distributed according to $N(0, (0.1)^2)$, and was otherwise drawn from a $N(0, 10^2)$ distribution. Figure 1(b) shows the results of the simulations. Again, we see that the robust regression functions all give rise to statistically consistent estimators with ℓ_2-error scaling as $\mathcal{O}(\sqrt{\frac{k \log p}{n}})$. We also include the plots for the standard Lasso estimator with the ordinary least squares objective. Since the distribution of ε_i is sub-Gaussian for the mixture distribution, the Lasso estimator is also ℓ_2-consistent; however, we see that the robust loss functions improve upon the ℓ_2-error of the Lasso by a constant factor.

Finally, we ran simulations to test the statistical consistency of generalized M-estimators under relaxed distributional assumptions on the covariates. We generated x_i's from a sub-exponential distribution, given by independent chi-square variables with 10 degrees of freedom, and recentered to have mean zero. The ε_i's were drawn from a Cauchy distribution with scale parameter 0.1. We ran trials for problem sizes $p = 128, 256, 512$ and 1024, with $k \approx p^{1/3}$ and $\beta^* = (\frac{1}{\sqrt{k}}, \ldots, \frac{1}{\sqrt{k}}, 0, \ldots, 0)$. We used the ℓ_1-penalized Mallows estimator described in Proposition 3, with $b = 3$, $B = I_p$ and ℓ equal to the Huber loss function, and optimized the function using the composite gradient descent algorithm with random initializations, with the regularization parameters $\lambda = 0.3\sqrt{\frac{\log p}{n}}$ and $R = 1.1\|\beta^*\|_1$. Figure 2 shows the result of the simulations, from which we observe that the Mallows estimator is indeed statistically consistent, as predicted by Theorem 1 and Proposition 3. We also plot the results for ℓ_1-penalized Huber regression. It is not difficult to see from the proof of Theorem 1 that $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty$ is also of the order $\mathcal{O}(\sqrt{\frac{k \log p}{n}})$ when the x_i's are sub-exponential, but with a larger prefactor than the Mallows loss. We observe in Figure 2 that the Huber loss indeed appears to yield a statistically consistent estimator as well, but at a relatively slower rate (for $p = 128, 256$, and 512). In our simulations, we needed a slightly larger value $\lambda = \sqrt{\frac{\log p}{n}}$ for the Huber loss in order to achieve statistical consistency.

5.2. *Convergence of optimization algorithm.* Next, we ran simulations to verify the convergence behavior of the composite gradient descent algorithm described in Section 4. We set $p = 512$, $k \approx p^{1/3}$, and $n \approx 15k \log p$, and generated ε_i's from a Cauchy distribution with scale parameter 0.1, and the x_i's from a standard normal distribution. We set $\beta^* = (\frac{1}{\sqrt{k}}, \ldots, \frac{1}{\sqrt{k}}, 0, \ldots, 0)$. We then simulated

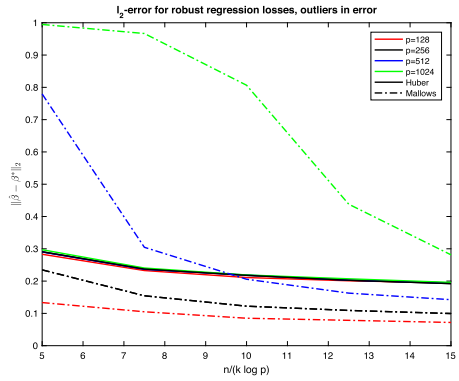
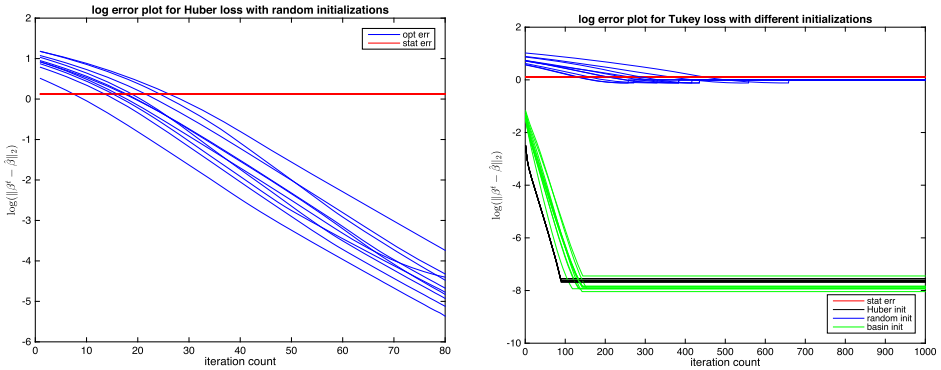


FIG. 2. *Plot showing simulation results for the ℓ_1-penalized Mallows generalized M-estimator with a Huber loss function, when covariates are drawn from a sub-exponential distribution and errors are drawn from a heavy-tailed Cauchy distribution. Results for the ℓ_1-penalized Huber loss are shown for comparison. Although both estimators appear to be statistically consistent, the Mallows estimator exhibits better performance. The plot agrees with the behavior predicted by Theorem 1 and Proposition 3.*

the solution paths for the Huber and Cauchy loss functions with an ℓ_1-penalty, with regularization parameters $\lambda = 0.3\sqrt{\frac{\log p}{n}}$ and $R = 1.1\|\beta^*\|_1$. Panel (a) of Figure 3 shows solution paths for the composite gradient descent algorithm with the Huber



(a) (b)

FIG. 3. *Plots showing optimization trajectories for composite gradient descent applied to various high-dimensional robust regression functions. Solution paths are shown in blue and measured with respect to β^*, and the statistical error is plotted in red.* (a) *Solution paths for the ℓ_1-penalized convex Huber loss with* 10 *random initializations.* (b) *Solution paths for the ℓ_1-penalized nonconvex Tukey loss with* 10 *random initializations from the ℓ_1-penalized Huber output* (*black*); *slight perturbations of β^* within the local region where the loss satisfies RSC* (*green*); *and random initializations* (*blue*). *The black and green trajectories converge at a linear rate to a unique stationary point in the local region. The blue iterates converge to an entirely different stationary point.*

loss from 10 different starting points, chosen randomly from a $N(0, 6^2 I_p)$ distribution. An estimate of the global optimum $\widehat{\beta}$ was obtained from preliminary runs of the optimization error, and the log optimization error $\log(\|\beta^t - \widehat{\beta}\|_2)$ for each of the initializations was computed accordingly. In addition, we plot the statistical error $\log(\|\widehat{\beta} - \beta^*\|_2)$ in red for comparison. As seen in the plot, the log errors decay roughly linearly in t. Since the ℓ_1-penalized Huber objective is convex, our theory guarantees sublinear convergence of the iterates initially and then linear convergence locally around β^* within the radius $\frac{r}{2}$, as specified by Theorem 3. Indeed, our plots suggest nearly linear convergence even outside the local RSC region. All iterates converge to the unique global optimum $\widetilde{\beta}$ (the apparent bifurcation is due to the small nonzero error tolerance provided in our implementation of the algorithm as a criterion for convergence).

Figure 3(b) shows solution paths using the ℓ_1-penalized Tukey loss. We plot the composite gradient descent iterates for 10 different starting points chosen by the output of composite gradient descent applied to the ℓ_1-penalized Huber loss (black) with random initializations; 10 randomly chosen starting points given by β^* plus a $N(0, (0.05)^2 I_p)$ perturbation (green); and 10 randomly chosen starting points drawn from a $N(0, 3^2 I_p)$ distribution (blue). The simulation results reveal a linear rate of convergence for composite gradient descent iterates in the first two cases, as predicted by Theorem 3, since the initial iterates lie within the local region around β^* where the Tukey loss satisfies the RSC condition. All of the black and green trajectories converge to the same unique stationary point in the local region. In the third case, however, the rate of convergence of composite gradient descent iterates is slower, and the iterates actually converge to a different stationary point further away from β^*. This emphasizes the cautionary message that stationary points may indeed exist for nonconvex robust regression functions that are *not* consistent for the true regression vector, and first-order optimization algorithms may converge to these undesirable stationary points if initialized improperly.

5.3. *Nonconvex regularization*. Finally, we ran simulations to verify the oracle results described in Section 3.3. Figure 4 shows side-by-side comparisons for robust regression using the Huber and Cauchy loss functions with the SCAD penalty, with parameter $a = 2.5$. We ran simulations for $p = 256, 512$, and 1024, with $k \approx p^{1/3}$ and $\beta^* = (\frac{1}{\sqrt{k}}, \ldots, \frac{1}{\sqrt{k}}, 0, \ldots, 0)$. The ε_i's were drawn from a Cauchy distribution with scale parameter 0.1, and the x_i's were drawn from a standard normal distribution. The ℓ_1-penalized Huber loss was used to select an initial point for the composite gradient descent algorithm, as prescribed by the two-step algorithm; in all cases, we set the regularization parameters to be $\lambda = \sqrt{\frac{\log p}{n}}$ and $R = 1.1\|\beta^*\|_1$. Panel (a) plots the ℓ_2-error versus the rescaled sample size $\frac{n}{k \log p}$, from which we see that both SCAD-penalized objective functions yield statistically consistent estimators. Panel (b) plots the fraction of trials (out of 50) for which the recovered support of the estimator agrees with the true support of β^*.

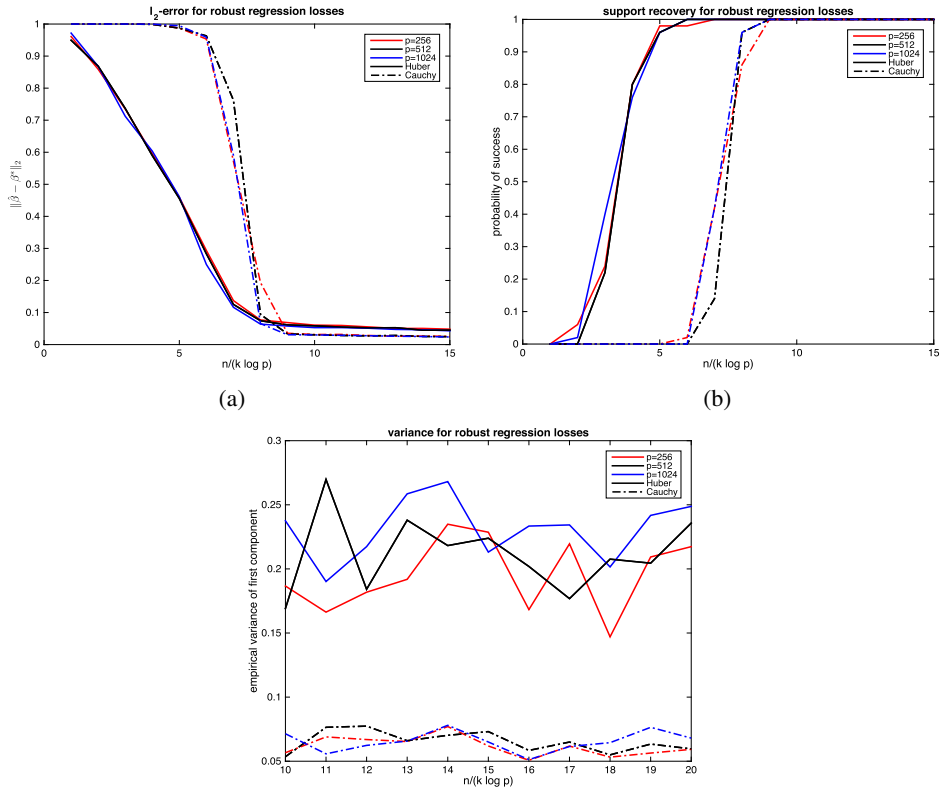


FIG. 4. *Simulation results for robust regression with a nonconvex SCAD regularizer, using a Huber loss (solid lines) and Cauchy loss (dashed lines), for $p = 256$ (red), $p = 512$ (black) and $p = 1024$ (blue). (a) Plot showing ℓ_2-error as a function of the rescaled sample size $\frac{n}{k \log p}$. Both regularizers yield statistically consistent estimators, as predicted by Theorem 1. (b) Plot showing variable selection consistency. The probability of success in recovering the support transitions sharply from 0 to 1 as a function of the sample size, agreeing with the theoretical predictions of Theorem 2. The transition threshold corresponds with the sharp drop in ℓ_2-error seen in panel (a), since $\widetilde{\beta}$ agrees with the oracle result. (c) Plot showing the empirical variance of $\sqrt{n} \cdot e_1^T(\widetilde{\beta} - \beta^*)$, the rescaled first component in the error vector. As predicted by the asymptotic normality result of Corollary 2, the empirical variance remains roughly constant for sufficiently large sample sizes.*

As we see, the families of curves for different loss functions stack up when the horizontal axis is rescaled according to $\frac{n}{k \log p}$. Furthermore, the probability of correct support recovery transitions sharply from 0 to 1 in panel (b), as predicted by Theorem 2. Note that the transition point for the Cauchy loss in panel (b), which happens for $\frac{n}{k \log p} \approx 8$, also corresponds to a sharp drop in the ℓ_2-error in panel (a), since $\widetilde{\beta}$ is then equal to the low-dimensional oracle estimator. Panel (c) plots the empirical variance of $\sqrt{n} \cdot e_1^T(\widetilde{\beta} - \beta^*)$, the first component of the error vec-

tor rescaled by $\sqrt{n}$. We see that the variance for the Cauchy loss is uniformly smaller than the variance for the Huber loss—indeed, the Cauchy loss corresponds to the MLE of the error distribution. Furthermore, the curves for each loss function roughly align for different problem sizes, and the variance is roughly constant for increasing n, as predicted by Corollary 2. Note that Corollary 2 requires third-order differentiability, so it does not directly address the Huber loss. However, the empirical variance of the Huber estimators is also roughly constant, suggesting that a version of Corollary 2 might also exist for the Huber loss function.

6. Discussion. We have studied penalized high-dimensional robust estimators for linear regression. Our results show that under a local RSC condition satisfied by many robust regression M-estimators, stationary points within the region of restricted curvature are actually statistically consistent estimators of the true regression vector, and even under heavy-tailed errors or outlier contamination, these estimators enjoy the same convergence rate as ℓ_1-penalized least squares regression with sub-Gaussian errors. Furthermore, we show that when the penalty is chosen from an appropriate family of nonconvex, amenable regularizers, the stationary point within the local RSC region is unique and agrees with the local oracle solution. This allows us to establish asymptotic normality of local stationary points under appropriate regularity conditions, and in some cases conclude that the regularized M-estimator is asymptotically efficient. Finally, we propose a two-step M-estimation procedure for obtaining local stationary points when the M-estimator is nonconvex, where the first step consists of optimizing a convex problem in order to obtain a sufficiently close initialization for a final run of composite gradient descent in the second step.

Several open questions remain that provide interesting avenues for future work. First, although the side constraint $\|\beta\|_1 \leq R$ in the regularized M-estimation program (2) is required in our proofs to ensure that stationary points obey a cone condition, it is unclear whether this side condition is necessary. Indeed, since we are only concerned with stationary points within a small radius r of β^*, the additional ℓ_1-constraint may be redundant. It would be useful to remove the appearance of R for practical problems, since we would then only need to tune the parameter λ. Second, as a consequence of the oracle result in Theorem 2, local stationary points inherit other properties of the oracle solution $\widehat{\beta}_S^{\mathcal{O}}$ in addition to asymptotic normality, such as breakdown behavior and properties of the influence function. It would be interesting to explore these properties for robust M-estimators with a diverging number of parameters. A potentially harder problem would be to derive bounds on the measures of robustness for stationary points of regularized robust estimators when the oracle result does not hold (i.e., for ℓ_1-penalized robust M-estimators). As suggested by the reviewers, two other interesting areas of research would be to explore the case where β^* only satisfies weak sparsity, or when endogeneity is present in the data due to ultra-high-dimensionality. Lastly, whereas our results on asymptotic normality allow us to draw conclusions regarding the asymptotic

variance of the local oracle solution, it would be valuable to derive nonasymptotic bounds on the variance of high-dimensional robust M-estimators. By trading off the nonasymptotic bias and variance, one could then determine the form of a robust regression function that is optimal in some sense.

Acknowledgments. The author thanks the Associate Editor and three anonymous reviewers for encouraging and helpful feedback in revising the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical consistency and asymptotic normality for high-dimensional robust M-estimators." (DOI: 10.1214/16-AOS1471SUPP; .pdf). We provide detailed technical proofs for the results stated in the main body of the paper.

REFERENCES

AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. MR3097609

BAI, Z. D. and WU, Y. (1997). General M-estimation. *J. Multivariate Anal.* **63** 119–135. MR1491570

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

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434. MR0386168

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

BOX, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40** 318–335. MR0058937

BRADIC, J., FAN, J. and WANG, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 325–349. MR2815779

CLARKE, F. H. (1983). *Optimization and Nonsmooth Analysis*. Wiley, New York. MR0709590

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J., LI, Q. and WANG, Y. (2014). Robust estimation of high-dimensional mean regression. Preprint. Available at arXiv:1410.2150.

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

FREEDMAN, D. A. and DIACONIS, P. (1982). On inconsistent M-estimators. *Ann. Statist.* **10** 454–461. MR0653520

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation Ph.D. thesis, Univ. of California, Berkeley.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. Wiley, New York. MR0829458

HE, X. and SHAO, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630. MR1425971

HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73** 120–135. MR1766124

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*. Springer, New York. MR1639875

LI, G., PENG, H. and ZHU, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statist. Sinica* **21** 391–419. MR2796868

LOH, P. (2016). Supplement to "Statistical consistency and asymptotic normality for high-dimensional robust M-estimators." DOI:10.1214/16-AOS1471SUPP.

LOH, P. and WAINWRIGHT, M. J. (2014). Support recovery without incoherence: A case for nonconvex regularization. Preprint. Available at arXiv:1412.5632.

LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

LOZANO, A. C. and MEINSHAUSEN, N. (2013). Minimum distance estimation for robust high-dimensional regression. Preprint. Available at arXiv:1307.3227.

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400. MR0981457

MARONNA, R. A., MARTIN, R. D. and YOHAI, V. J. (2006). *Robust Statistics*: *Theory and Methods*. Wiley, Chichester. MR2238141

MENDELSON, S. (2014). Learning without concentration for general loss functions. Preprint. Available at arXiv:1410.3192.

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

NESTEROV, Y. (2007). Gradient methods for minimizing composite objective function. CORE Discussion Papers No. 2007076, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE).

NOLAN, J. P. (2015). *Stable Distributions—Models for Heavy Tailed Data*. Birkhauser, Boston.

PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

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

ROUSSEEUW, P. J. and LEROY, A. M. (2005). *Robust Regression and Outlier Detection*. Wiley, New York.

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

SHEVLYAKOV, G., MORGENTHALER, S. and SHURYGIN, A. (2008). Redescending M-estimators. *J. Statist. Plann. Inference* **138** 2906–2917. MR2526216

SIMPSON, D. G., RUPPERT, D. and CARROLL, R. J. (1992). On one-step GM estimates and stability of inferences in linear regression. *J. Amer. Statist. Assoc.* **87** 439–450. MR1173809

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

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics* 448–485. Stanford Univ. Press, Stanford, CA. MR0120720

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

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

WELSH, A. H. and RONCHETTI, E. (2002). A journey in single steps: Robust one-step M-estimation in linear regression. *J. Statist. Plann. Inference* **103** 287–310. MR1896997

YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. MR0888431

YOHAI, V. J. and MARONNA, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7** 258–268. MR0520237

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

DEPARTMENTS OF ELECTRICAL AND
COMPUTER ENGINEERING AND STATISTICS
UNIVERSITY OF WISCONSIN-MADISON
1415 ENGINEERING DRIVE
MADISON, WISCONSIN 53706
USA
E-MAIL: loh@ece.wisc.edu