Statistica Sinica **32** (2022), 1295-1316
doi:https://doi.org/10.5705/ss.202019.0324

ROBUSTNESS AND TRACTABILITY FOR NONCONVEX M-ESTIMATORS

Ruizhi Zhang[1], Yajun Mei[2], Jianjun Shi[2] and Huan Xu[3]

[1] *University of Nebraska-Lincoln,* [2] *Georgia Institute of Technology*
and [3] *Alibaba Inc.*

Abstract: We investigate two important properties of M-estimators, namely, robustness and tractability, in a linear regression setting, when the observations are contaminated by some arbitrary outliers. Specifically, robustness is the statistical property that the estimator should always be close to the true underlying parameters, *regardless of the distribution of the outliers*, and tractability refers to the computational property that the estimator can be computed efficiently, even if the objective function of the M-estimator is *nonconvex.* In this article, by examining the empirical risk, we show that under some sufficient conditions, many M-estimators enjoy nice robustness and tractability properties simultaneously when the percentage of outliers is small. We extend our analysis to the high-dimensional setting, where the number of parameters is greater than the number of samples, $p \gg n$, and prove that when the proportion of outliers is small, the penalized M-estimators with the L_1 penalty enjoy robustness and tractability simultaneously. Our research provides an analytic approach to determine the effects of outliers and tuning parameters on the robustness and tractability of some families of M-estimators. Simulations and case studies are presented to illustrate the usefulness of our theoretical results for M-estimators under Welsch's exponential squared loss and Tukey's bisquare loss.

Key words and phrases: Computational tractability, gross error, high-dimensionality, nonconvexity, robust regression, sparsity.

1. Introduction

M-estimation plays an essential role in linear regression, owing to its robustness and flexibility. From a statistical viewpoint, it has been shown that many M-estimators enjoy desirable robustness properties in the presence of outliers, and asymptotic normality when the data are normally distributed without outliers. Some general theoretical properties and reviews of robust M-estimators can be found in Bai, Rao and Wu (1992), Huber and Ronchetti (2009), Cheng and Huang (2010), Hampel et al. (2011), and El Karoui et al. (2013). In the high-dimensional setting, where the dimensionality is greater than the number of

Corresponding author: Ruizhi Zhang, The Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA. E-mail: rzhang35@unl.edu.

samples, penalized M-estimators have been widely used to tackle the challenges of outliers, and have been used for sparse recovery and variable selection; see Lambert-Lacroix and Zwald (2011), Li, Peng and Zhu (2011), Wang et al. (2013), and Loh (2017). However, it is often not easy to compute the M-estimators from a computational tractability perspective, because optimization problems over nonconvex loss functions are usually involved. Moreover, the tractability issue may become more challenging when the data are contaminated by some arbitrary outliers, which is essentially the situation that robust M-estimators are designed to address.

This study simultaneously investigates two important properties of M-estimators, *robustness* and *tractability*, simultaneously under *the gross error model.* Specifically, we assume the data-generation model is $y_i = \langle \theta_0, x_i \rangle + \epsilon_i$, where $y_i \in \mathbb{R}, x_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$, and the noise term ϵ_i is from Huber's gross error model (Huber (1964)): $\epsilon_i \sim (1-\delta) f_0 + \delta g$, for $i = 1, \ldots, n$. Here, f_0 denotes the probability density function (pdf) of the noise of the normal samples, which has desirable properties such as a zero mean and a finite variance; g denotes the pdf of the outliers (contaminations), which can be arbitrary, and may also depend on the explanatory variable x_i, for $i = 1, \ldots, n$. Note that we do not require the mean of g to be zero. The parameter $\delta \in [0, 1]$ denotes the percentage of contaminations, also known as the contamination ratio in the robust statistics literature. The gross error model indicates that for the ith sample, the residual term ϵ_i is generated from the pdf f_0 with probability $1 - \delta$, and from the pdf g with probability δ. Note that the residual ϵ_i is independent of x_i and other x_js when it is from the pdf f_0, but can be dependent on the variable x_i when it is from the pdf g.

In the first part of this paper, we start with the low-dimensional case when the dimension $p \ll n$. We consider the robust M-estimation with a constraint on the ℓ_2 norm of θ. Mathematically, we study the following optimization problem:

$$\begin{aligned} &\underset{\theta}{\text{Minimize:}} \quad \hat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \langle \theta, x_i \rangle), \\ &\text{subject to:} \quad \|\theta\|_2 \le r. \end{aligned} \tag{1.1}$$

Here, $\rho : \mathbb{R} \to \mathbb{R}$ is the loss function, and is often *nonconvex.* We consider the problem with the ℓ_2 constraint for three reasons. First, it is well known that the constrained optimization problem in (1.1) is equivalent to the unconstrained optimization problem with an ℓ_2 regularizer. Therefore, it is related to the ridge regression, which alleviates multicollinearity among the regression predictors. Second, considering the problem of (1.1) in a compact ball with radius r guarantees the existence of the global optimal, which is necessary for establishing the tractability properties of the M-estimator. Finally, by working on a constrained optimization problem, we avoid technical complications and establish the uniform convergence theorems of the empirical risk and population risk. Note that constrained M-estimators are widely used and studied in the literature; see Geyer (1994), Mei, Bai and Montanari (2018), and Loh (2017) for more details. To be consistent with the assumptions used in the literature, in the current work, we assume r is a constant and the true parameter θ_0 is inside the ball.

In the second part, we extend our research to the high-dimensional case, where $p \gg n$ and the true parameter θ_0 is sparse. To achieve sparsity in the resulting estimator, we consider the penalized M-estimator with the ℓ_1 regularizer:

$$\begin{aligned} &\underset{\theta}{\text{Minimize:}} \quad \hat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - \langle \theta, x_i \rangle) + \lambda_n ||\theta||_1, \qquad (1.2)\\ &\text{subject to:} \quad \|\theta\|_2 \le r. \end{aligned}$$

Note that the corresponding penalized M-estimator with the ℓ_2 constraint is related to the elastic net, which overcomes the limitations of the lasso-type regularization (Zou and Hastie (2005)).

In both parts, we show that (in the finite-sample setting) the M-estimator obtained from (1.1) or (1.2) is robust in the sense that all stationary points of the empirical risk function $\hat{R}_n(\theta)$ or $\hat{L}_n(\theta)$ are bounded in the neighborhood of the true parameter θ_0 when the proportion of outliers is small. In addition, we show that with a high probability, there is a unique stationary point of the empirical risk function that is the global minimizer of (1.1) or (1.2) for some general (possibly nonconvex) loss functions ρ. This implies that the M-estimator can be computed efficiently. To illustrate our general theoretical results, we study some specific M-estimators, namely, Welsch's exponential squared loss (Dennis Jr and Welsch (1978)) and Tukey's bisquare loss (Beaton and Tukey (1974)), and explicitly discuss how the tuning parameter and the percentage of outliers affect the robustness and tractability of the corresponding M-estimators.

Our research makes several fundamental contributions to the field of robust statistics and nonconvex optimization. First, we demonstrate the uniform convergence results for the gradient and Hessian of the empirical risk to the population risk under the gross error model. Second, we provide a nonasymptotic upper bound of the estimation error for general M-estimators that nearly achieves the minimax error bound in Chen, Gao and Ren (2016). Third, we investigate the computational tractability of general nonconvex M-estimators under the gross error model. The results show that when the contamination ratio δ is small, there is only one unique stationary point of the empirical risk function. Therefore, efficient algorithms such as gradient descent or proximal gradient descent can be guaranteed to converge to a unique global minimizer, irrespective of the initialization. Our general results also imply the following interesting statement: the percentage of outliers impacts the *tractability* of nonconvex M-estimators. In essence, the estimation and the corresponding optimization problem become more complicated in terms of the solution quality and computational efficiency when more outliers appear. While the former is expected, that more outliers make M-estimators more difficult to compute numerically is an interesting and somewhat surprising discovery. Our simulation results and case study also verify this phenomenon.

Related works

Since Huber's pioneering work on robust M-estimators (Huber (1964)), many M-estimators with different choices of loss functions have been proposed, including Huber's loss (Huber (1964)), Andrew's sine loss (Andrews et al. (1972)), Tukey's bisquare loss (Beaton and Tukey (1974)), and Welsch's exponential squared loss (Dennis Jr and Welsch (1978)), among others. From a statistical perspective, several works have investigate the robustness of M-estimators, for example, the large breakdown point (Donoho and Huber (1982); Mizera and Müller (1999); Alfons, Croux and Gelper (2013)), finite influent function (Hampel et al. (2011)) and asymptotic normality (Maronna and Yohai (1981); Lehmann and Casella (2006); El Karoui et al. (2013)). Recently, regularized M-estimators have received much attention in high-dimensional contexts. Lambert-Lacroix and Zwald (2011) proposed a robust variable selection method by combing Huber's loss and the adaptive lasso penalty. Li, Peng and Zhu (2011) show that the nonconcave penalized M-estimation method can perform parameter estimation and variable selection simultaneously. Welsch's exponential squared loss combined with the adaptive lasso penalty is used by Wang et al. (2013) to construct a robust estimator for sparse estimation and variable selection. Chang, Roberts and Welsh (2018) proposed a robust estimator by combining Tukey's bisquare loss with the adaptive lasso penalty. Loh and Wainwright (2015) proved that under mild conditions, any stationary point of the nonconvex objective function is close to the true underlying parameters. However, these statistical works do not discuss the computational tractability of the M-estimators, even though many of the loss functions are nonconvex.

During the last several years, nonconvex optimization has attracted fast-growing interests owing to its ubiquitous applications in machine learning and deep learning, such as dictionary learning (Mairal et al. (2009)), phase retrieval (Candes, Li and Soltanolkotabi (2015)), orthogonal tensor decomposition (Anandkumar et al. (2014)), and training deep neural networks (Bengio (2009)). It is well known that there is no efficient algorithm that can guarantee finding a global optimal solution for general nonconvex optimization.

Fortunately, in the context of estimating nonconvex M-estimators for high-dimensional linear regression (*without outliers*), under some mild statistical assumptions, Loh (2017) establishes the uniqueness of the stationary point of the nonconvex M-estimator when using some nonconvex bounded regularizers instead of the ℓ_1 regularizer. By investigating the uniform convergence of gradient and Hessian of the empirical risk, Mei, Bai and Montanari (2018) prove that with a high probability, there exists one unique stationary point of the regularized empirical risk function with the ℓ_1 regularizer. Thus, regardless of the initial points, many computationally efficient algorithms, such as the gradient descent or proximal gradient descent algorithms, can be applied, and are guaranteed to converge to the global optimizer, which implies the high tractability of the M-estimator. However, their analysis is restricted to the standard linear regression setting without outliers. In particular, they assume that the distribution of the noise terms in the linear regression model should have some desirable properties, such as have a zero mean, be sub-Gaussian, and be independent of the feature vector x, which might not hold when the data are contaminated by outliers. To the best of our knowledge, no studies have analyzed the computational tractability properties of nonconvex M-estimators when the data are contaminated by arbitrary outliers, despite M-estimators having being developed to handle outliers in the linear regression. Our research is the first to fill this significant gap in the tractability of nonconvex M-estimators. We prove that under mild assumptions, many M-estimators can tolerate a small number of arbitrary outliers in the sense of keeping the tractability, even if the loss functions are nonconvex.

Notation. Given $\mu, \nu \in \mathbb{R}^p$, their standard inner product is defined by $\langle \mu, \nu \rangle = \sum_{i=1}^{p} \mu_i \nu_i$. The ℓ_p norm of a vector x is denoted by $||x||_p$. The p-by-p identity matrix is denoted by $I_{p\times p}$. Given a matrix $M \in \mathbb{R}^{m\times m}$, let $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote the largest and smallest eigenvalues of M, respectively. The operator norm of M is denoted by $||M||_{op}$, which is equal to $\max(\lambda_{\max}(M), -\lambda_{\min}(M))$ when $M \in \mathbb{R}^{m\times m}$. Let $B_q^p(a,r) = \{x \in \mathbb{R}^p : ||x-a||_q \leq r\}$ be the ℓ_q ball in the $\mathbb{R}^p$ space with center a and radius r. Moreover, let $B_q^p(r)$ be the ℓ_q ball in the $\mathbb{R}^p$ space with center $\mathbf{0}$ and radius r. Given a random variable X with pdf f, we denote the corresponding expectation by $\mathbf{E}_f$. We often omit the density function subscript f when it is clear from the context, and the expectation is taken for all variables.

The rest of this paper is organized as follows. In Section 2, we present the theorems about the robustness and tractability of general M-estimators under the low-dimensional setup, when the dimension p is much smaller than n. Then, in Section 3, we consider the penalized M-estimator with the ℓ_1 regularizer in a high-dimensional regression when $p \gg n$. The ℓ_2 error bounds of the estimation and the scenario in which the M-estimator has nice tractability are provided. In Section 4, we discuss two special families of robust estimators, constructed using Welsch's exponential loss and Tukey's bisquare loss as examples, to illustrate our general theorems of robustness and tractability of M-estimators. Simulation results and a case study are presented in Section 5 and Section 6, respectively, to illustrate the robustness and tractability properties when the data are contaminated by outliers. Concluding remarks are given in Section 7. We relegate all proofs and supporting lemmas to the Supplementary Material.

2. M-estimators in the Low-Dimensional Regime

In this section, we investigate two critical properties of M-estimators, namely *robustness*, and *tractability*, in the setting of a linear regression with arbitrary outliers in a low-dimensional regime, where the dimension p is much smaller than the number of samples n. In terms of robustness, we show that under some mild conditions, any stationary point of the objective function in (1.1) is well bounded in a neighborhood of the true parameter θ_0. Moreover, the neighborhood shrinks when the proportion of outliers decreases. In terms of tractability, we show that when the proportion of outliers is small and the sample size is large, with a high probability, there is a *unique stationary point* of the empirical risk function, which is the global optimum (and hence the corresponding M-estimator). Consequently, many first-order methods are guaranteed to converge to the global optimum, irrespective of the initialization. In particular, we show that the gradient descent algorithm converges to the global optimum exponentially, for any initializations.

Before presenting our main theorems, we make the following mild assumptions on the loss function ρ, explanatory or feature vectors x_i, and idealized noise distribution f_0. We define the score function $\psi(z) := \rho'(z)$.

Assumption 1.

(a) *The score function* $\psi(z)$ *is twice differentiable and odd in* z *with* $\psi(z) \geq 0$, *for all* $z \geq 0$. *Moreover, we assume* $\max\{||\psi(z)||_\infty, ||\psi'(z)||_\infty, ||\psi''(z)||_\infty\} \leq L_\psi$.

(b) *The feature vector x_i is independent and identically distributed with zero mean and is τ^2-sub-Gaussain; that is, $\mathbf{E}[e^{\langle\lambda,x_i\rangle}] \leq \exp((1/2)\tau^2||\lambda||_2^2)$, for all $\lambda \in \mathbb{R}^p$.*

(c) *The feature vector x_i spans all possible directions in $\mathbb{R}^p$; that is, $\mathbf{E}[x_i x_i^T] \succeq \gamma\tau^2 I_{p\times p}$, for some $0 < \gamma \leq 1$.*

(d) *The idealized noise distribution $f_0(\epsilon)$ is symmetric. Define $h(z) := \int_{-\infty}^{\infty} f_0(\epsilon)$ $\psi(z+\epsilon)d\epsilon$, and $h(z)$ satisfies $h(z) > 0$, for all $z > 0$ and $h'(0) > 0$.*

Assumption (a) requires the smoothness of the loss function in the objective function, which is crucial to study the tractability of the estimation problem. Assumption (b) assumes a sub-Gaussian design of the observed feature matrix. Assumption (c) assumes that the covariance matrix of the feature vector is positive semidefinite. Note that the condition on $h(z)$ is mild. It is not difficult to show that it is satisfied if the idealized noise distribution $f_0(\epsilon)$ is strictly positive for all ϵ and decreasing for $\epsilon > 0$, for example, if $f_0 =$ pdf of $N(0,\sigma^2)$.

Before presenting our main results, we first define the population risk as follows:

$$R(\theta) = \mathbf{E}\hat{R}_n(\theta) = \mathbf{E}[\rho(Y - \langle\theta, X\rangle)]. \tag{2.1}$$

Conceptually, we analyze the population risk first, and then build a link between the population risk and the empirical risk, which solves the original estimation problem. Theorem 1 summarizes the results for the population risk function $R(\theta)$ in (2.1).

Theorem 1. *Assume that Assumption 1 holds and that the true parameter θ_0 satisfies $||\theta_0||_2 \leq r/3$.*

(a) *There exists a constant $\eta_0 = (\delta/(1-\delta))C_1$ such that any stationary point θ^* of $R(\theta)$ satisfies $||\theta^* - \theta_0||_2 \leq \eta_0$, where δ is the contamination ratio, and C_1 is a positive constant that depends only on $\gamma, r, \tau, \psi(z)$, and the pdf f_0, but does not depend on the outlier pdf g.*

(b) *When δ is small, there exists a constant $\eta_1 = C_2 - C_3\delta > 0$, where C_2, C_3 are two positive constants that depend only on $\gamma, r, \tau, \psi(z)$, and the pdf f_0, but do not depend on the outlier pdf g, such that*

$$\lambda_{\min}(\nabla^2 R(\theta)) > 0, \tag{2.2}$$

for every θ with $||\theta_0 - \theta||_2 < \eta_1$.

(c) *There is a unique stationary point of $R(\theta)$ in the ball $B_2^p(0, r)$, as long as $\eta_0 < \eta_1$, for a given contamination ratio δ.*

It is useful to add some remarks to better understand Theorem 1. First, recall that the noise term ϵ_i follows the gross error model: $\epsilon_i \sim (1-\delta)f_0 + \delta g$, where the outlier pdf g may also depend on x_i. While the true parameter θ_0 may no longer be the stationary point of the population risk function $R(\theta)$, Theorem 1 implies that the stationary points of $R(\theta)$ will always be bounded in a neighborhood of the true parameter θ_0 when the percentage of contamination δ is small. This indicates the robustness of M-estimators in the population case.

Second, Theorem 1 asserts that when there are no outliers, that is, $\delta = 0$, the stationary point is indeed the true parameter θ_0. In addition, because the constant η_0 in (a) is an increasing function of δ, whereas the constant η_1 in (b) is a decreasing function of δ, stationary points of $R(\theta)$ may disperse from the true parameter θ_0, and the strongly convex region around θ_0 will be decreasing, as the contamination ratio δ increases. This indicates the difficulty of optimization for cases with large contamination ratios.

Third, part (c) follows directly from part (a) and (b). Note that $\eta_0(\delta = 0) = 0 < \eta_1(\delta = 0) = C_2$. Thus there exists a positive δ^* such that $\eta_0 < \eta_1$, for any $\delta < \delta^*$. A simple lower bound on δ^* is $C_3/(C_1 + C_2 + C_3)$, because $C_1\delta < (1-\delta)(C_2 - C_3\delta)$ whenever $0 \leq \delta \leq C_3/(C_1 + C_2 + C_3)$.

Our next step is to link the empirical risk function (and the corresponding M-estimator) to the population version. To this end, we introduce Lemma 1, which shows the global uniform convergence theorem of the sample gradient and Hessian. For brevity, it is presented in the Supplementary Material.

We are now ready to present our main result about M-estimators by investigating the empirical risk function $\hat{R}_n(\theta)$.

Theorem 2. *Assume Assumption* 1 *holds and $||\theta_0||_2 \leq r/3$. We use the same notation η_0 and η_1 as in Theorem* 1. *Then, for any $\pi > 0$, there exist constants C, $C_\pi = C_0(C_h \vee \log(r\tau/\pi) \vee 1)$, where C is a constant greater than C_π, C_0 is a universal constant, C_h is a constant depending on $\gamma, r, \tau, \psi(z)$, and $h(z)$, but is independent of π, p, n, δ, and g, such that as $n \geq Cp\log n$, the following statements hold with probability at least $1 - \pi$:*

(a) *for all $||\theta - \theta_0||_2 > \eta_0 + (1/(1-\delta))\zeta$,*

$$\langle \theta - \theta_0, \nabla \widehat{R}_n(\theta) \rangle > 0, \tag{2.3}$$

where ζ is a constant that does not depend on δ.

(b) *for all* $||\theta - \theta_0||_2 < \eta_1$,

$$\lambda_{\min}(\nabla^2 \widehat{R}_n(\theta)) > 0. \tag{2.4}$$

Thus, as long as $\eta_0 + (1/(1-\delta))\zeta < \eta_1$, $\widehat{R}_n(\theta)$ *has a unique stationary point, which lies in the ball* $B_2^p(\theta_0, \eta_0 + (1/(1-\delta))\zeta)$. *This is the unique global optimal solution of* (1.1)*; denote this unique stationary point by* $\widehat{\theta}_n$.

(c) *There exists a positive constant* κ *that depends on* $\pi, \gamma, r, \psi, \delta$, *and* f_0, *but is independent of* n, p, *and* g, *such that*

$$||\widehat{\theta}_n - \theta_0||_2 \leq \eta_0 + \frac{4\tau}{\kappa}\sqrt{\frac{C_\pi p \log n}{n}}. \tag{2.5}$$

(d) *There exist constants* $C_1, C_2, h_{\max}$ *that depend on* $\pi, \gamma, r, \psi, \delta$, *and* f_0, *but that are independent of* n, p, *and* g, *such that the gradient descent with fixed step size* $h \leq h_{\max}$ *converges exponentially fast to the global minimizer; that is, for any initialization* $\theta_n(0) \in B_2^p(0, r)$,

$$\|\theta_n(k) - \widehat{\theta}_n\|_2^2 \leq C_1(1 - C_2 h)^k \|\theta_n(0) - \widehat{\theta}_n\|_2^2. \tag{2.6}$$

A few remarks are in order. First, the constant C_π is the same constant in Lemma 1, which gauranntees the uniform convergence of the sample gradient and Hessian when $n \geq C_\pi p \log n$. C is a constant that depends on C_π and is larger than C_π, which means additional samples are required to ensure the results in Theorem 2 compared to the sample size in Lemma 1. Second, because η_0, ζ are independent of n, p, and g, Theorem 2(a) asserts that the M-estimator that minimizes $\widehat{R}_n(\theta)$ is always bounded in the ball $B_2^p(\theta_0, \eta_0 + (1/(1-\delta))\zeta)$, regardless of g (and hence the outliers observed). This indicates the robustness of the M-estimator; that is, the estimates are not severely skewed by a small number of "bad" outliers. Next, when the contamination ratio δ is small such that $\eta_0 + (1/(1-\delta))\zeta < \eta_1$, there is a unique stationary point of $\widehat{R}_n(\theta)$. In fact, as shown in the Supplementary Material, when $\delta = 0$, we always have $\eta_0 + \zeta < \eta_1$, which implies that the condition $\eta_0 + (1/(1-\delta))\zeta < \eta_1$ always holds for some small value of δ. Therefore, although the original optimization problem (1.1) is nonconvex and the sample contains some arbitrary outliers, the optimal solution of $\widehat{R}_n(\theta)$ can be computed efficiently using most off-the-shelf first-order algorithms, such as the gradient descent or stochastic gradient descent. Specifically, in Theorem 2, we show with high probability that the gradient descent algorithm converges to the global optimal solution exponentially, regardless of the initializations. This indicates the tractability of the M-estimator. Interestingly, as in the population risk case, the tractability is closely related to the number of outliers; the problem is easier to optimize when the data contain fewer outliers. Finally, when the number of samples $n \gg p \log n$, the estimation error bound is $O(\delta + \sqrt{p \log n/n})$, which nearly achieves the minimax lower bound of $O(\delta + \sqrt{p/n})$ in Chen, Gao and Ren (2016).

3. Penalized M-estimator in the High-Dimensional Regime

In this section, we investigate the tractability and robustness of the penalized M-estimator in the high-dimension region, where the dimension of the parameter p is much greater than the number of samples n. Specifically, we consider the same data-generation model $y_i = \langle \theta_0, x_i \rangle + \epsilon_i$, where $y_i \in \mathbb{R}, x_i \in \mathbb{R}^p$, and the noise term ϵ_i is from Huber's gross error model (Huber (1964)): $\epsilon_i \sim (1-\delta)f_0 + \delta g$. Moreover, we assume $p \gg n$ and that the true parameter θ_0 is sparse.

We consider the ℓ_1-regularized M-estimator under an ℓ_2-constraint on θ:

$$\begin{aligned} &\underset{\theta}{\text{Minimize:}} \quad \hat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - \langle \theta, x_i \rangle) + \lambda_n ||\theta||_1, \\ &\text{subject to:} \quad ||\theta||_2 \le r. \end{aligned} \tag{3.1}$$

Before presenting our main theorem, we need additional assumptions on the feature vector x.

Assumption 2. *The feature vector x has a pdf in $\mathbb{R}^p$. In addition, there exists a constant $M > 1$ that is independent of the dimension p, such that $||x||_\infty \le M\tau$, almost surely.*

Remark 1. For unbounded subGaussian feature vectors, Theorem 3 can be supplemented by taking a truncation at $M = C\sqrt{\log(np)}$. Then, the conclusions still hold, with an additional $\log(np)$ term. Thus, for simplicity of the statement of Theorem 3, we consider the case when Assumption 2 holds.

In the Supplementary Material, we present Lemma 2, which shows the uniform convergence of the gradient and the Hessian under Huber's contamination model in the high-dimensional setting, where $p \gg n$. Then, we are ready for our main theorem.

Theorem 3. *Assume that Assumption 1 and Assumption 2 hold, and the true parameter θ_0 satisfies $||\theta_0||_2 \le r/3$ and $||\theta_0||_0 \le s_0$. Then, there exist constants C, C_0, C_1 that are dependent on $(\rho, L_\psi, \tau^2, r, \gamma, \pi)$, but independent of (δ, s_0, n, p, M), such that as $n \ge C s_0 \log p$ and $\lambda_n \ge 2C_0 M\sqrt{\log p/n} + 2\delta L_\psi \tau$, the following hold with probability at least $1 - \pi$:*

(a) *All stationary points of problem* (3.1) *are in* $B_2^p(\theta_0, \eta_0 + (\sqrt{s_0}/(1-\delta))\lambda_n C_1)$.

(b) *As long as* n *is large enough such that* $n \geq C s_0 \log^2 p$ *and the contamination ratio* δ *is small such that* $(\eta_0 + (1/(1-\delta))\sqrt{s_0}\lambda_n C_1) \leq \eta_1$, *the problem* (3.1) *has a unique local stationary point, which is also the global minimizer.*

The proof of Theorem 3 is based on several lemmas, which are postponed to the Supplementary Material. We believe that some of our lemmas are of interest in their own right. Theorem 3 implies that the estimation error of the penalized M-estimator is bounded as $O(\delta + \sqrt{s_0 \log p/n})$, which achieves the minimax estimation rate (Chen, Gao and Ren (2016)). Moreover, it implies that the penalized M-estimator has good tractability when the percentage of outliers δ is small.

Remark 2. In Theorem 3, we show there is a unique local stationary point for the problem (3.1) if $(\eta_0 + (1/(1-\delta))\sqrt{s_0}\lambda_n C_2) \leq \eta_1$ and n is large. Thus, many first-order algorithms can be guaranteed to converge to the global optimal when the initialization is in the ball $B_2^p(\theta, \eta_1)$. However, owing to the complexity of analyzing the restricted empirical risk $\hat{L}_n(\theta)$, we leave as an open problem the convergence analysis of such fast algorithms for any initializations in the ball $B_2^p(r)$.

4. Example

In this section, we use some examples to illustrate our general theoretical results about the robustness and tractability of M-estimators. In the first subsection, we consider the low-dimensional regime, and study a family of M-estimators with a specific loss function, known as Welsch's exponential squared loss (Dennis Jr and Welsch (1978); Rey (2012); Wang et al. (2013)). In the second subsection, we consider the high-dimensional regime, and study the penalized M-estimator with Tukey's bisquare loss (Beaton and Tukey (1974)). In both subsections, we derive explicit expressions of the two critical radii η_0, η_1, and discuss the robustness and tractability of the corresponding M-estimators.

4.1. M-estimators with Welsch's exponential squared loss

In this subsection, we illustrate the general results presented in Section 2 by considering a family of M-estimators with a specific nonconvex loss function known as Welsch's exponential squared loss (Dennis Jr and Welsch (1978); Rey (2012); Wang et al. (2013)),

$$\rho_\alpha(t) = \frac{1 - \exp(-\alpha t^2/2)}{\alpha}, \tag{4.1}$$

where $\alpha \geq 0$ is a tuning parameter. The corresponding M-estimator is obtained by solving the optimization problem

$$\begin{aligned} \min_\theta \hat{R}_n(\theta) &:= \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(y_i - \langle \theta, x_i \rangle), \\ \text{subject to } \ & ||\theta||_2 \leq r. \end{aligned} \tag{4.2}$$

The nonconvex loss function $\rho_\alpha(t)$ in (4.1) has been used in other contexts, such as robust estimation and robust hypothesis testing, owing to their many nice properties; see Ferrari and Yang (2010) and Qin and Priebe (2017). First, it is a smooth function of both α and t, and the gradient and Hessian are well defined. Second, when α goes to zero, $\rho_\alpha(t)$ converges to $t^2/2$. Thus, the least squares estimator is a special case of the M-estimator obtained from (4.4). Third, for fixed $\alpha > 0$, $\rho_\alpha(t), \rho'_\alpha(t)$, and $\rho''_\alpha(t)$ are all bounded. Intuitively, this implies that the impact of outlier observations of y_i is controlled, and thus the corresponding statistical procedure is robust.

We now study the robustness and tractability of the M-estimator of (4.2) based on our framework in Theorem 2. In order to emphasize the effects of the tuning parameter α and the contamination ratio δ on the robustness and tractability properties, we consider a simplified assumption on the feature vector x_i and the pdf of the idealized residual f_0.

Assumption 3.

(a) The feature vector x_i has an i.i.d. multivariate Gaussian distribution $N(0, \tau^2 I_{p\times p})$.

(b) The idealized noise pdf $f_0(\epsilon)$ has a Gaussian distribution $N(0, \sigma^2)$.

(c) Assume the true parameter $||\theta_0||_2 \leq r/3$.

Now, we are ready to present Corollary 1, which is a direct application of Theorem 2.

Corollary 1. *Assume Assumption 3 holds and $||\theta_0||_2 \leq r/3$. For any $\pi > 0$, there exists a constant C such that as $n \geq Cp \log n$, the following statements hold with probability at least $1 - \pi$:*

(a) All stationary points of problem (4.2) are in $B_2^p(\theta_0, \eta_0 + (1/(1-\delta))\zeta)$.

(b) The empirical risk function $\widehat{R}_n(\theta)$ is strongly convex in the ball $B_2^p(\theta_0, \eta_1)$.

(c) As long as $\eta_0 + (1/(1-\delta))\zeta < \eta_1$, $\widehat{R}_n(\theta)$ *has a unique stationary point, which is the unique global optimal solution of* (1.1).

Here,

$$\begin{aligned}\zeta &= \frac{1}{13.5\sqrt{3\alpha}(1+\alpha\sigma^2)^{3/2}\tau},\\ \eta_0(\delta,\alpha) &= \frac{\delta}{1-\delta}\sqrt{\frac{e}{\alpha}}\frac{4(1+\alpha\sigma^2)^{3/2}}{\tau}e^{32\alpha r^2\tau^2/(3(1+\alpha\sigma^2))},\\ \eta_1(\delta,\alpha) &= \frac{1}{9\sqrt{3\alpha}(1+\alpha\sigma^2)^{3/2}\tau}\left[1-\delta(1+3(1+\alpha\sigma^2)^{3/2})\right].\end{aligned}$$

A special case of Corollary 1 with $\alpha = 0$ reduces to the least squares estimator. On the one hand, with $\alpha = 0$, we have $\eta_1(\delta, \alpha = 0) = +\infty$, for any $\delta > 0$. Thus, the corresponding risk function is strongly convex in the entire region of $B_2^p(0, r = 10)$, and hence is always tractable. On the other hand, because $\eta_0(\delta, \alpha = 0) = +\infty$, the solution of the optimization problem in (4.4) can be arbitrarily in the ball $B_2^p(0, r = 10)$, even when the proportion of outliers is small. Thus, it is not robust to outliers. This supports the well-known fact that the least squares estimator is easy to compute, but is very sensitive to outliers.

Additionally, for another special case with $\delta = 0$ and $\alpha > 0$, we have $\eta_0(\delta = 0, \alpha) = 0$ and $\zeta < \eta_1(\delta = 0, \alpha)$. This implies that Welsch's estimator has nice tractability when there are no outliers. However, when the percentage of outliers δ is increasing, $\eta_1(\delta, \alpha)$ decreases, implying that the presence of outliers reduces the tractability of the M-estimator.

4.2. Penalized M-estimators with Tukey's bisquare loss

In this subsection, we illustrate the general results presented in Section 3 by studying Tukey's bisquare loss function (Beaton and Tukey (1974)):

$$\rho_\alpha(t) = \begin{cases} \frac{1}{6}\alpha^2\left[1-\left(1-\left(\frac{t}{\alpha}\right)^2\right)^3\right], & \text{if } |t| > \alpha,\\ 0, & \text{if } |t| > \alpha,\end{cases} \tag{4.3}$$

where $\alpha > 0$ is a tuning parameter. The corresponding penalized M-estimator is obtained by solving the optimization problem

$$\min_\theta \hat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^n \rho_\alpha(y_i - \langle\theta, x_i\rangle) + \lambda_n||\theta||_1, \tag{4.4}$$

$$\text{subject to } \ ||\theta||_2 \leq r.$$

Note that the loss function $\rho_\alpha(t)$ in (4.3) is nonconvex. For fixed $\alpha > 0$, $\rho'_\alpha(t)$ and $\rho''_\alpha(t)$ are both bounded. We now study the robustness and tractability of the penalized M-estimator of (4.4) based on our framework in Theorem 3. When α goes to ∞, $\rho_\alpha(t)$ converges to $t^2/2$. Thus, the penalized M-estimator obtained by (4.4) reduces to the lasso estimator, which can be computed easily. However, the lasso is also known to be very sensitive to outliers (Alfons, Croux and Gelper (2013)). On the other hand, when α increases, the estimator becomes more robust, but may lose tractability, owing to the nonconvexity of the function $\rho_\alpha(t)$ and the presence of outliers.

In order to emphasize the relation between the tuning parameter α and the contamination ratio δ, we consider a simplified assumption on the feature vector x_i and the pdf of the idealized residual f_0.

Assumption 4.

(a) The feature vector x_i has an i.i.d. multivariate uniform distribution $[-\tau, \tau]^p$.

(b) The idealized noise pdf $f_0(\epsilon)$ has a Gaussian distribution $N(0, \sigma^2)$.

(c) The true parameter $||\theta_0||_2 \leq r/3$.

Assumption 4 and Theorem 3 yield Corollary 2, which characterizes the robustness and tractability of the penalized M-estimator with Tukey's exponential squared loss in (4.3).

Corollary 2. *Assume that Assumption 4 holds, and that the true parameter θ_0 satisfies $||\theta_0||_2 \leq r/3$. Then, for any $\pi \in (0,1)$, there exists a constant C_π such that if choosing $\lambda_n = 2C_\pi\tau\sqrt{\log p/n} + 2\alpha\tau\delta$, as $n \gg s_0 \log p$, the following hold with probability at least $1 - \pi$:*

(a) All stationary points of problem (4.4) are in $B_2^p(\theta_0, (1+2\tau)\eta_0)$.

(b) The empirical risk function $\hat{L}_n(\theta)$ are strong convex in the ball $B_2^p(\theta_0, \eta_1)$.

(c) As long as n is large enough and the contamination ratio δ is small such that $(1+2\tau)\,\eta_0 \leq \eta_1$, the problem (4.4) has a unique local stationary point, which is also the global minimizer.

Here,

$$\eta_0(\delta, \alpha) = \frac{\delta}{1-\delta}\frac{28\sqrt{2\pi}}{\tau\sigma^3\alpha^2}e^{(\alpha^2+64\tau^2r^2)/\sigma^2}, \tag{4.5}$$

$$\eta_1(\delta, \alpha) = \frac{(1-\delta)M(\alpha,\sigma)\tau^2 - 4\delta}{2\sqrt{3}\tau}\alpha, \tag{4.6}$$

where $M(\alpha, \sigma) = 2\alpha \int_0^1 (1-t)(1+t)(1-5t^2) f_0(\alpha t) dt$ *is a positive number when* $\alpha > 0, \sigma > 0$.

The special case of Corollary 2 with $\alpha \to \infty$ reduces to the lasso estimator. On the one hand, with $\alpha = \infty$, we have $\eta_1(\delta, \alpha = 0) = +\infty$, for any $\delta > 0$. This means that the corresponding risk function is strongly convex in the entire region of $B_2^p(0, r = 10)$, and hence is always tractable. On the other hand, because $\eta_0(\delta, \alpha \to \infty) \to +\infty$, the solution of the optimization problem in (4.4) can be arbitrarily in the ball $B_2^p(0, r = 10)$, even when the proportion of outliers is small. Thus, it is not robust to outliers. This supports the well-known fact that the lasso estimator is easy to compute, but is very sensitive to outliers.

Additionally, for another special case with $\delta = 0$ and $\alpha > 0$, we have $\eta_0(\delta = 0, \alpha) = 0$, which means the true parameter θ_0 is the unique stationary point of the risk function. This implies that Tukey's estimator has nice tractability when there are no outliers. However, when the percentage of outliers δ is increasing, $\eta_1(\delta, \alpha)$ decreases, implying that the presence of more outliers reduces the tractability of the M-estimator.

5. Simulation Results

In this section, we report simulation results using Welsch's exponential loss and Tukey's bisquare loss when the data are contaminated, using synthetic data. We first generate covariates $x_i \sim N(0, I_{p\times p})$ and responses $y_i = \langle \theta_0, x_i \rangle + \epsilon_i$, where $||\theta_0||_2 = 1$. We consider the case when the residual term ϵ_i is from the gross error model with contamination ratio δ; that is, $\epsilon_i \sim (1-\delta)N(0,1) + \delta N(\mu_i, 3^2)$, where $\mu_i = ||x_i||_2^2 + 1$. The outlier distribution is chosen to highlight the effects of outliers when they are dependent on x_i and have a nonzero mean.

In the first part, we consider the low-dimensional case, when the dimension $p = 10$. Specifically, we generate $n = 100$ pairs of data $(y_i, x_i)_{i=1,\ldots,n}$ with dimension $p = 10$ and with different choices of contamination ratios δ. We use the projected gradient descent to solve the optimization problem in (4.2) with Welsch's loss and $r = 10$. To make the iteration points be inside the ball, we project the points back into $B_2^p(0, r = 10)$ if they fall outside of the ball. The step size is fixed as one. In order to test the tractability of the M-estimator, we run the gradient descent algorithm with 20 random initial values in the ball $B_2^p(0, r = 10)$ to determine whether the algorithm can converge to the same

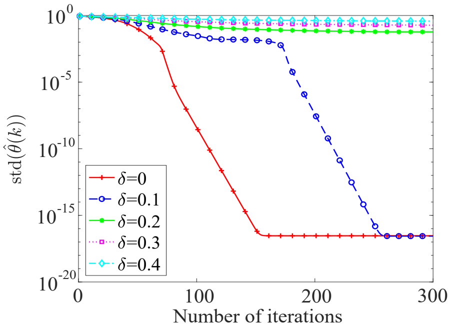

Figure 1. The value of $\mathrm{std}(\hat{\theta}(k))$ for different δ. Y-axis is with log scale.

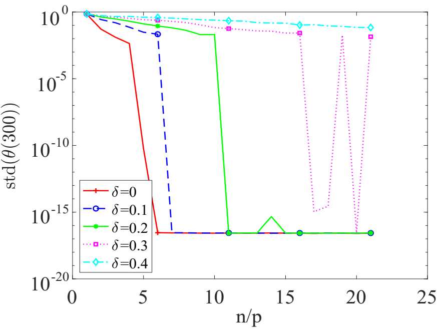

Figure 2. The value of $\mathrm{std}(\hat{\theta}(300))$ for different δ. Y-axis is with log scale

stationary point. Denoting $\hat{\theta}(k)$ as the kth iteration point, we then plot the empirical standard deviation of each iteration $\mathrm{std}(\hat{\theta}(k)) = \mathrm{Tr}(\widehat{\mathrm{Var}}(\hat{\theta}(k)))$ for 20 different initializations. Figure 1 shows the convergence of the gradient descent algorithm for Welsch's exponential loss with $\alpha = 0.1$ under the gross error model, with different δ. From Figure 1, we observe that when the proportion of outliers is small (i.e., $\delta \le 0.1$), the algorithm converges to the same stationary point fast. However, when the contamination ratio δ becomes larger, the algorithm may not converge to the same point for different initial points, indicating a loss of tractability for the same objective function with an increasing proportion of outliers. These observations are consistent with Theorem 2, which asserts that the M-estimator is tractable when the contamination ratio δ is small. Then, in Figure 2, we show the empirical standard deviation at the $k = 300$ iteration $\mathrm{std}(\hat{\theta}(300))$ when $p = 20$ and the ratio of n/p varies from 1 to 21. The figure shows that when the sample size n is small, the gradient descent may not converge to the same stationary point. However, when n is large enough, for a small proportion of outliers δ, the algorithm does converge to the same stationary point, which implies the uniqueness of the stationary point.

To illustrate the robustness of the M-estimator, we generate 100 realizations of (Y, X) and run the gradient descent algorithm with different initial values. The average estimation errors between the M-estimator and the true parameter θ_0 are presented in Figure 3. As we can see, when $\delta = 0$, all estimators have small estimation errors, which is expected because those M-estimators are consistent without outliers (Huber (1964); Huber and Ronchetti (2009)). However, for the M-estimator with $\alpha = 0$, that is, the least squares estimator, the estimation error increases dramatically as the proportion of outliers increases. This confirms that the least squares estimator is not robust to outliers.

When $\alpha = 0.1$, the overall estimation error does not increase much, even with 40% outliers, which clearly demonstrates the robustness of the M-estimator. Note that when α is increased from 0.1 to 0.3, although the estimator error is still very small for $\delta \leq 0.2$, it increases dramatically when δ is greater than 0.2. We believe that two reasons contribute to this phenomenon. First, robustness starts to decrease when α becomes too large. Second, and more importantly, the algorithm fails to find the global optimum owing to the multiple stationary points when α is large. Thus for each α, there exists a critical bound of δ such that the estimator will be robust and tractable when the proportion of outliers is smaller than that bound.

In the second part, we present our results in the high-dimensional region when $p = 200$ and $n = 200$. Data (y_i, x_i) are generated from the same gross error model in the previous simulation study, with the true parameter θ_0 a sparse vector with $s = 10$ nonzero entries. All nonzero entries are set to $1/\sqrt{10}$. We use the proximal gradient descent algorithm to solve problem (3.1) with Tukey's bisquare loss. As before, we project the points back into $B_2^p(0, r = 10)$ if they fall outside of the ball. We set the fixed step size as 0.1 and the L_1 regularization parameter $\lambda = \sqrt{\log(p)/n}$. We first illustrate the robustness of the penalized M-estimator using Tukey's loss with the tuning parameter $\alpha = 4, 5, 10, 20, 500$. We generate 100 realizations of (Y, X) and run the proximal gradient descent algorithm. The average estimation errors between the penalized M-estimator and the true parameter are reported in Figure 4. First, note that as α is large, Tukey's loss is similar to the squared loss. Thus, the penalized M-estimator with $\alpha = 500$ performs similarly to the lasso. From Figure 4, we can see it has the smallest estimation error when $\delta = 0$, but has the largest estimation error when $\delta \geq 0.1$. Moreover, when α is small, the corresponding estimation error does not increase much, even if $\delta = 0.4$. These results imply the robustness of the penalized robust M-estimator.

Next, we illustrate the tractability of the penalized M-estimator by showing $\text{std}(\hat{\theta}(k))$ for 20 initializations of the proximal gradient descent algorithm with Tukey's loss and $\alpha = 20$ under the gross error model, with different δ. Figure 5 shows the result for $p = 200$ and $n = 200$, and Figure 6 shows the result for $p = 400$ and $n = 400$. From the two plots, we observe an interesting phenomenon: the proximal gradient descent converges to the same stationary points, even when the percentage of outliers $\delta = 0.4$. This result seems to contradict the result for the low-dimensional case, where $\alpha = 0.4$ can make the algorithm converge to different stationary points. Thus, a more accurate analysis on the tractability property of the penalized M-estimators is needed.

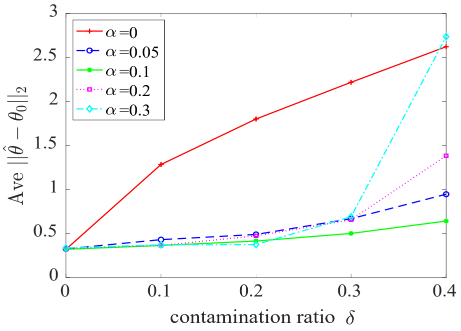

Figure 3. The estimation error for different α and δ.

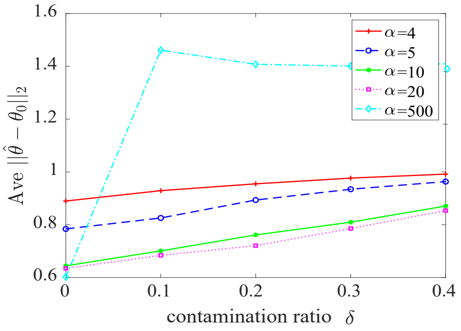

Figure 4. The estimation error for different α and δ.

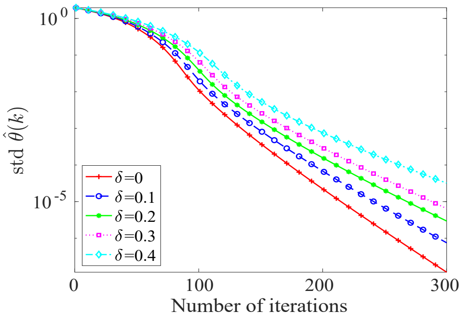

Figure 5. The value of std$(\hat{\theta}(k))$ for different δ; n=p=200.

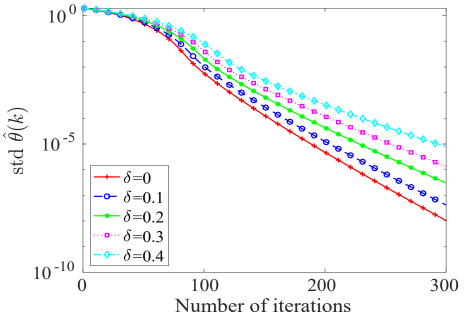

Figure 6. The value of std$(\hat{\theta}(k))$ for different δ; n=p=400.

6. Case Study

In this section, we present a case study of the robust regression problem for the Airfoil Self-Noise data set (Brooks, Pope and Marcolini (2014)), which is available from the UCI Machine Learning Repository. The data set was processed by NASA and is commonly used in regression studies to learn the relation between the airfoil self-noise and five explanatory variables. Specifically, the data set contains five explanatory variables: Frequency (in Hertz), Angle of attack (in degrees), Chord length (in meters), Free-stream velocity (in meters per second), and Suction side displacement thickness (in meters). There are 1,503 observations in the data set. The response variable is Scaled sound pressure level (in decibels). In this section, the five explanatory variables are scaled to have a zero mean and unit variance. Then, we corrupt the response by adding noise ϵ from the same gross error model as the previous section: $\epsilon_i \sim (1-\delta)N(0,1)+\delta N(\mu_i, 3^2)$, with $\mu_i = ||x_i||_2^2 + 1$.

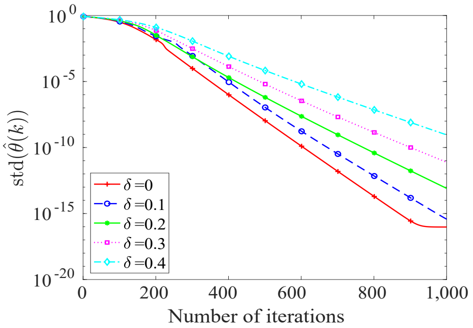

Figure 7. The convergence of the gradient descent algorithm for different δ. Y-axis is with log scale.

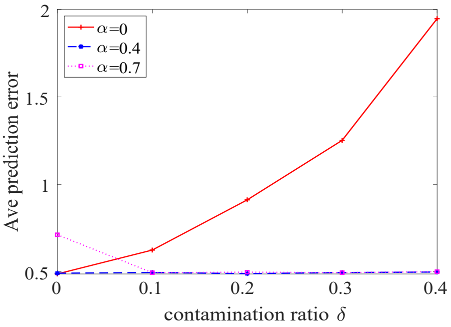

Figure 8. The prediction error for different α and δ.

We apply the M-estimator using Welsch's exponential loss (Dennis Jr and Welsch (1978)) to the data set to validate the tractability and robustness of the corresponding M-estimator. First, we run 100 Monte Carlo simulations. At each time, we split the data set of 1,503 pairs of observations into a training data set of size 1,000 and a testing data set of size 503. Then for the training data set, we use the gradient descent method with 20 different initial values to update the iteration points.

Figure 7 shows the empirical standard deviation of each iteration std($\hat{\theta}(k)$) with $\alpha = 0.3$ and step size 0.5. Clearly, when δ is smaller than 0.3, the gradient descent converges to the same local minimizer, which implies the uniqueness of the stationary point. This result demonstrates the nice tractability of the M-estimator under the gross error model when the proportion of outliers is small. Then, using the optimal point as the M-estimator, we calculate the prediction error, which is the mean squared error on the testing data. Figure 8 shows the average prediction error on the testing data. As we can see, the prediction error with $\alpha = 0$ increases dramatically when the percentage of outliers increases. In contrast, the prediction error of the M-estimator with $\alpha = 0.4$ is stable, even with a large percentage of outliers. This illustrates the robustness of M-estimators for some positive α.

7. Conclusion

We have investigated the robustness and computational tractability of general (nonconvex) M-estimators in both classical low-dimensional and modern high-dimensional regimes. In terms of *robustness*, in the low-dimensional regime, we show that the estimation error of the M-estimator is $O(\delta + \sqrt{p \log n / n})$, which nearly achieves the minimax lower bound of $O(\delta + \sqrt{p/n})$ in Chen, Gao and Ren (2016). In the high-dimensional regime, we show that the estimation error of the penalized M-estimator has an estimation error $O(\delta + \sqrt{s_0 \log p/n})$, which achieves the minimax estimation rate (Chen, Gao and Ren (2016)).

In terms of *tractability*, our theoretical results imply that under sufficient conditions, when the percentage of arbitrary outliers is small, the general M-estimator could have good computational tractability because it has only one unique stationary point, even if the loss function is nonconvex. Therefore, M-estimators can tolerate a certain level of outliers while maintaining both their estimation accuracy and computational efficiency. Both simulations and a real-data case study validate our theoretical results about the robustness and tractability of M-estimators in the presence of outliers.

Supplementary Material

The online Supplementary Material contains proofs for the lemmas and main theorems.

Acknowledgments

The authors thank three anonymous reviewers and the Associate Editor for their thoughtful and constructive comments that greatly improved the presentation of this article. Y. Mei's research was supported in part by NSF grants 1830344 and 2015405.

References

Alfons, A., Croux, C. and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics* **7**, 226–248.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M. and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research* **15**, 2773–2832.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and W.Tukey, J. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton.

Bai, Z., Rao, C. R. and Wu, Y. (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica* **2**, 237–254.

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16**, 147–185.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning* **2**, 1–127.

Brooks, T., Pope, S. and Marcolini, M. (2014). Airfoil Self-Noise. *UCI Machine Learning Repository.*

Candes, E. J., Li, X. and Soltanolkotabi, M. (2015). Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis* **39**, 277–299.

Chang, L., Roberts, S. and Welsh, A. (2018). Robust Lasso regression using Tukey's biweight criterion. *Technometrics* **60**, 36–47.

Chen, M., Gao, C. and Ren, Z. (2016). A general decision theory for Huber's ϵ-contamination model. *Electronic Journal of Statistics* **10**, 3752–3774.

Cheng, G. and Huang, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *The Annals of Statistics* **38**, 2884–2915.

Dennis Jr, J. E. and Welsch, R. E. (1978). Techniques for nonlinear least squares and robust regression. *Communications in Statistics-Simulation and Computation* **7**, 345–359.

Donoho, D. L. and Huber, P. J. (1982). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (Edited by P. J. Bickel, K. Doksum and J. L. Hodges), 157–184. Chapman and Hall/CRC, Boca Raton.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C. and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences* **110**, 14557–14562.

Ferrari, D. and Yang, Y. (2010). Maximum Lq-likelihood estimation. *The Annals of Statistics* **38**, 753–783.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *The Annals of Statistics* **22**, 1993–2010.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (2011). *Robust Statistics: The Approach Based on Influence Functions*. Wiley-Interscience, New York.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73–101.

Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics*. Wiley, New York.

Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics* **5**, 1015–1053.

Lehmann, E. L. and Casella, G. (2006). *Theory of Point Estimation*. Springer, New York.

Li, G., Peng, H. and Zhu, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica* **21**, 391–419.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics* **45**, 866–896.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* **16**, 559–616.

Mairal, J., Bach, F., Ponce, J. and Sapiro, G. (2009). Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 689–696. ACM, New York.

Maronna, R. A. and Yohai, V. J. (1981). Asymptotic behavior of general M-estimates for regression and scale with random carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **58**, 7–20.

Mei, S., Bai, Y. and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *The Annals of Statistics* **46**, 2747–2774.

Mizera, I. and Müller, C. H. (1999). Breakdown points and variation exponents of robust M-estimators in linear models. *The Annals of Statistics* **27**, 1164–1177.

Qin, Y. and Priebe, C. E. (2017). Robust hypothesis testing via L_q-likelihood. *Statistica Sinica* **27**, 1793–1813.

Rey, W. J. (2012). *Introduction to Robust and Quasi-Robust Statistical Methods.* Springer-Verlag, Heidelberg.

Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association* **108**, 632–643.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

Ruizhi Zhang

The Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA.

E-mail: rzhang35@unl.edu

Yajun Mei

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

E-mail: yajun.mei@isye.gatech.edu

Jianjun Shi

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

E-mail: jianjun.shi@isye.gatech.edu

Huan Xu

Alibaba Inc., Bellevue, WA 98004, USA.

E-mail: huan.xu@alibaba-inc.com

(Received September 2019; accepted November 2020)