Statistica Sinica Preprint No: SS-2019-0324	
Title	Robustness and Tractability for Non-convex M-estimators
Manuscript ID	SS-2019-0324
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0324
Complete List of Authors	Ruizhi Zhang, Yajun Mei, Jianjun Shi and Huan Xu
Corresponding Author	Ruizhi Zhang
E-mail	rzhang35@unl.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

Robustness and Tractability for Non-convex M-estimators

Ruizhi Zhang*, Yajun Mei**, Jianjun Shi**, Huan Xu***

*University of Nebraska-Lincoln, **Georgia Institute of Technology, ***Alibaba Inc.

Abstract: We investigate two important properties of M-estimators, namely, robustness and tractability, in the linear regression setting, when the observations are contaminated by some arbitrary outliers. Specifically, robustness means the statistical property that the estimator should always be close to the true underlying parameters *regardless of the distribution of the outliers*, and tractability indicates the computational property that the estimator can be computed efficiently, even if the objective function of the M-estimator is *non-convex*. In this article, by learning the landscape of 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 further 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 L_1 penalty will enjoy robustness and tractability simultaneously. Our research provides an analytic approach to see the effects of outliers and tuning parameters on the robustness and tractability of some families of M-estimators. Simulation 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, non-convexity, robust regression, sparsity

1. Introduction

M-estimation plays an essential role in linear regression due to its robustness and flexibility. From the statistical viewpoint, it has been shown that many M-estimators enjoy desirable robustness properties in the presence of outliers, as well as asymptotic normality when the data are normally distributed without outliers. Some general theoretical properties and review of robust M-estimators can be found in Bai et al. (1992); Huber and Ronchetti (2009); Cheng et al. (2010); Hampel et al. (2011); El Karoui et al. (2013). In the high-dimensional setting, where the dimensionality is greater than the number of 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 et al. (2011); Wang et al. (2013); Loh (2017). However, it is often not easy to compute the M-estimators from the computational tractability perspective since optimization problems over non-convex 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 where robust M-estimators are designed to tackle.

This paper aims to investigate 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 $i = 1, \cdots, n$, and the noise term ϵ_i's are from Huber's gross error model (Huber, 1964): $\epsilon_i \sim (1-\delta)f_0 + \delta g$, for $i = 1, \cdots, n$. Here, f_0 denotes the probability density function (pdf) of the noise of the normal samples, which has the desirable properties, such as zero mean and 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 $i = 1, \cdots, n$. One thing to notice is that we do not require the mean of g to be 0. The parameter $\delta \in [0, 1]$, denotes the percentage of the contaminations, which is also known as the contamination ratio in robust statistics literature. The gross error model indicates that for the i^{th} 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. It is important to point out that the residual ϵ_i is independent of x_i and other x_j's 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:

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

Here, $\rho : \mathbb{R} \to \mathbb{R}$ is the loss function, and is often *non-convex*. We consider the problem with the ℓ_2 constraint due to three reasons: first, it is well known 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 can alleviate multicollinearity amongst regression predictors. Second, by considering the problem of (1.1) in a compact ball with radius r, it guarantees the existence of the global optimal, which is necessary for establishing the tractability properties of the M-estimator. Finally, by working on the constrained optimization problem, we can avoid technical complications and establish the uniform convergence theorems of the empirical risk and population risk. Besides, constrained M-estimators are widely used and studied in the literature. See Geyer et al. (1994); Mei et al. (2018); 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 of 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 the sparsity in the resulting estimator, we consider the penalized M-estimator with the ℓ_1 regularizer:

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

Note 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 will 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 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. Besides, we will show that with a high probability, there is a unique stationary point of the empirical risk function, which is the global minimizer of (1.1) or (1.2) for some general (possibly non-convex) 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 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 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 non-convex 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, which nearly achieve the minimax error bound in Chen et al. (2016). Third, we investigate the computational tractability of the general non-convex M-estimators under the gross error model. The result shows 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 non-convex M-estimators. In a nutshell, the estimation and the corresponding optimization problem become more complicated in terms of solution quality and computational efficiency when more outliers appear. While the former is well expected, we find the latter – that more outliers make M-estimators more difficult to

compute numerically – an interesting and somewhat surprising discovery. Our simulation results and case study also verify this phenomenon.

Related works

Since Huber's pioneer work on robust M-estimators (Huber, 1964), many M-estimators with different choices of loss functions have been proposed, e.g., Huber's loss (Huber, 1964), Andrew's sine loss (Andrews et al., 1972), Tukey's Bisquare loss (Beaton and Tukey, 1974), Welsch's exponential squared loss (Dennis Jr and Welsch, 1978), to name a few. From the statistical perspective, much research has been done to investigate the robustness of M-estimators such as large breakdown point (Donoho and Huber, 1983; Mizera and Müller, 1999; Alfons et al., 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, in the high-dimensional context, regularized M-estimators have received much attention. Lambert-Lacroix and Zwald (2011) proposed a robust variable selection method by combing Huber's loss and adaptive lasso penalty. Li et al. (2011) show that the nonconcave penalized M-estimation method can perform parameter estimation and variable selection simultaneously. Welsch's exponential squared loss combined with adaptive lasso penalty is used by Wang et al. (2013) to

construct a robust estimator for sparse estimation and variable selection. Chang et al. (2018) proposed a robust estimator by combining Tukey's bisquare loss with adaptive lasso penalty. Loh and Wainwright (2015) proved that under mild conditions, any stationary point of the non-convex objective function would close to the true underlying parameters. However, those statistical works did not discuss the computational tractability of the M-estimators, even though many of these loss functions are non-convex.

During the last several years, non-convex optimization has attracted fast-growing interests due to its ubiquitous applications in machine learning and deep learning, such as dictionary learning (Mairal et al., 2009), phase retrieval (Candes et al., 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 to find the global optimal solution for general non-convex optimization.

Fortunately, in the context of estimating non-convex 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 non-convex M-estimator when using some non-convex bounded regularizers instead of ℓ_1 regularizer. By investigating the uniform convergence of gradient and Hessian of the empirical risk, Mei, Bai, Montanari,

et al. (2018) prove that with a high probability, there exists one unique stationary point of the regularized empirical risk function with ℓ_1 regularizer. Thus regardless of the initial points, many computational efficient algorithms such as gradient descent or proximal gradient descent algorithm could 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 the distribution of the noise terms in the linear regression model should have some desirable properties such as zero mean, sub-gaussian, and independent of feature vector x, which might not hold when the data are contaminated with outliers. To the best of our knowledge, no research has been done on analyzing the computational tractability properties of the non-convex M-estimators when data are contaminated by arbitrary outliers, although the very reason why M-estimators are proposed is to handle outliers in linear regression in the robust statistics literature. Our research is the first to fill the significant gap in the tractability of non-convex 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 non-convex.

Notations. 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 $\lambda_{\max}(M), \lambda_{\min}(M)$ denote the largest and the smallest eigenvalue 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 probability density function f, we denote the corresponding expectation by $\mathbf{E}_f$. We will often omit the density function subscript f when it is clear from the context, the expectation is taken for all variables.

Organization. The rest of this article 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 dimension p is much smaller than n. Then in Section 3, we consider the penalized M-estimator with ℓ_1 regularizer in the high-dimensional regression when $p \gg n$. The ℓ_2 error bounds of the estimation and the scenario when the M-estimator has nice tractability are provided. In Section 4, we discuss two special families of robust estimator constructed by 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 linear regression with arbitrary outliers in the 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) will be 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 initialization. In particular, we will show the gra-

dient descent algorithm can converge to the global optimum exponentially for any initializations.

Before presenting our main theorems, we make the following mild assumptions on the loss function ρ, the explanatory or feature vectors x_i, and the 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 are i.i.d with zero mean and τ^2-sub-Gaussain, that is $\mathbf{E}[e^{\langle \lambda, x_i \rangle}] \leq \exp(\frac{1}{2}\tau^2||\lambda||_2^2)$, 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 the sub-Gaussian design of the observed feature matrix; Assumption (c) assumes that the covariance matrix of the feature vector is positive semidefinite. We remark 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 $\epsilon > 0$, e.g., if $f_0 =$ pdf of $N(0, \sigma^2)$.

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

$$R(\theta) = \mathbf{E}\hat{R}_n(\theta) = \mathbf{E}[\rho(Y - \langle \theta, X \rangle)]. \tag{2.3}$$

The high level idea is to analyze the population risk first, and then we build a link between the population risk and the empirical risk, which solves the original estimation problem. Theorem 1 below summarizes the results for the population risk function $R(\theta)$ in (2.3).

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

(a) *There exists a constant $\eta_0 = \frac{\delta}{1-\delta} C_1$ 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 only depends 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 exist a constant $\eta_1 = C_2 - C_3\delta > 0$, where C_2, C_3 are two positive constants that only depend on $\gamma, r, \tau, \psi(z)$ and the pdf f_0 but not depend on the outlier pdf g, such that*

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

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 for better understanding 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 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, i.e., $\delta = 0$, the stationary point is indeed the true parameter θ_0. In addition, since 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 is increasing. This indicates the difficulty of optimization for large contamination ratio cases.

Third, part (c) is a direct result 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 $C_3/(C_1 + C_2 + C_3)$, since $C_1\delta < (1 - \delta)(C_2 - C_3\delta)$ 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) with the population version. To this end, we need to introduce Lemma 1, which shows the global uniform convergence theorem of the sample gradient and Hessian. Due to the page limit, it will be 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* $||\theta_0||_2 \leq r/3$. *Let us 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* $\gamma, r, \tau, \psi(z), h(z)$ *but 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* $||\theta - \theta_0||_2 > \eta_0 + \frac{1}{1-\delta}\zeta$,

$$\langle \theta - \theta_0, \nabla \widehat{R}_n(\theta) \rangle > 0, \tag{2.5}$$

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

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

$$\lambda_{\min}(\nabla^2 \widehat{R}_n(\theta)) > 0. \tag{2.6}$$

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

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

$$||\widehat{\theta}_n - \theta_0||_2 \leq \eta_0 + \frac{4\tau}{\kappa}\sqrt{\frac{C_\pi p \log n}{n}}. \tag{2.7}$$

(d) *There exist constants* $C_1, C_2, h_{\max}$ *that depend on* $\pi, \gamma, r, \psi, \delta, f_0$ *but 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, i.e., for any initialization* $\theta_n(0) \in B_2^p(0, r)$,

$$\|\theta_n(k) - \widehat{\theta}_n\|_2^2 \leq C_1(1 - C_2 h)^k \|\theta_n(0) - \widehat{\theta}_n\|_2^2, \tag{2.8}$$

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 depends on C_π and 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, since η_0, ζ are independent of n, p and g, Theorem 2(a) asserts that the M-estimator which minimizes $\widehat{R}_n(\theta)$ is always bounded in the ball $B_2^p(\theta_0, \eta_0 + \frac{1}{1-\delta}\zeta)$, regardless of g (and hence the outliers observed). This indicates the robustness of the M-estimator, i.e., the estimates are not severely skewed by a small amount of "bad" outliers. Next, when the contamination ratio δ is small such that $\eta_0 + \frac{1}{1-\delta}\zeta < \eta_1$, there is a unique stationary point of $\widehat{R}_n(\theta)$. In fact, as will be shown in the Supplementary Material, when $\delta = 0$, we always have $\eta_0 + \zeta < \eta_1$, which implies the condition $\eta_0 + \frac{1}{1-\delta}\zeta < \eta_1$ will always hold for some small value of δ. Therefore, although the original optimization problem (1.1) is non-convex and the sample contains some arbitrary outliers, the optimal solution of $\widehat{R}_n(\theta)$ can be computed efficiently via most off-the-shelf first-order algorithms such as gradient descent or stochastic gradient descent. Specifically, in Theorem 2, we show with high probability, 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 contains fewer outliers. Finally, when the number of samples $n \gg p \log n$, the estimation error bound is as the order of $O(\delta + \sqrt{\frac{p \log n}{n}})$, which nearly achieves the minimax lower bound of $O(\delta + \sqrt{\frac{p}{n}})$ in Chen et al. (2016).

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

In this section, we investigate the tractability and the robustness of the penalized M-estimator in the high-dimension region where the dimension of 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 are from Huber's gross error model (Huber, 1964): $\epsilon_i \sim (1-\delta) f_0 + \delta g$. Moreover, we assume $p \gg n$ and the true parameter θ_0 is sparse.

We consider the ℓ_1-regularized M-estimation under a ℓ_2-constraint on θ:

$$\begin{aligned} \underset{\theta}{\text{Minimize:}} \quad & \hat{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \langle \theta, x_i \rangle) + \lambda_n ||\theta||_1, \\ \text{subject to:} \quad & \|\theta\|_2 \leq r. \end{aligned} \tag{3.9}$$

Before presenting our main theorem, we need additional assumptions

on the feature vector x.

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

Remark 1. For unbounded sub-Gaussian feature vectors, Theorem 3 below 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 will present Lemma 2, which shows the uniform convergence of gradient and Hessian under the 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 $||\theta_0||_2 \leq r/3$ and $||\theta_0||_0 \leq s_0$. Then there exist constants C, C_0, C_1 that are dependent on $(\rho, L_\psi, \tau^2, r, \gamma, \pi)$ but independent on (δ, s_0, n, p, M) such that as $n \geq Cs_0 \log p$ and $\lambda_n \geq 2C_0 M\sqrt{\frac{\log p}{n}} + 2\delta L_\psi \tau$, the following hold with probability as least $1 - \pi$:*

(a) *All stationary points of problem (3.9) are in $B_2^p(\theta_0, \eta_0 + \frac{\sqrt{s_0}}{1-\delta}\lambda_n C_1)$*

(b) *As long as n is large enough such that $n \geq Cs_0 \log^2 p$ and the contamination ratio δ is small such that $(\eta_0 + \frac{1}{1-\delta}\sqrt{s_0}\lambda_n C_1) \leq \eta_1$, the problem (3.9) 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 the estimation error of the penalized M-estimator is bounded as the order of $O(\delta + \sqrt{\frac{s_0 \log p}{n}})$, which achieves the minimax estimation rate (Chen et al., 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 (3.9) if $(\eta_0 + \frac{1}{1-\delta}\sqrt{s_0}\lambda_n C_2) \leq \eta_1$ 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, due to the complicity of analyzing the restricted empirical risk $\hat{L}_n(\theta)$, we still leave an open problem about 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 will derive the explicit expression of the two critical radius η_0, η_1, and discuss the robustness and tractability of the corresponding M-estimators.

4.1 M-estimators via 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 non-convex loss function known as Welsch's exponential squared loss (Dennis Jr and Welsch, 1978; Rey, 2012; Wang et al., 2013),

$$\rho_\alpha(t) = \frac{1 - \exp(-\alpha t^2/2)}{\alpha}, \tag{4.10}$$

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

$$\begin{aligned} \min_{\theta} \quad & \hat{R}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(y_i - \langle \theta, x_i \rangle), \\ \text{subject to} \quad & ||\theta||_2 \leq r. \end{aligned} \tag{4.11}$$

The non-convex loss function $\rho_\alpha(t)$ in (4.10) has been used in other contexts such as robust estimation and robust hypothesis testing, see Ferrari and Yang (2010); Qin and Priebe (2017), as it has many nice properties. First, it is a smooth function of both α and t, and the gradient and Hessian are well-defined. Second, when α goes to 0, $\rho_\alpha(t)$ will converge to $t^2/2$. Thus, the least square estimator is a special case of the M-estimator obtained from (4.16). Third, for fixed $\alpha > 0$, $\rho_\alpha(t), \rho'_\alpha(t), \rho''_\alpha(t)$ are all bounded. Intuitively, this implies that the impact of outlier observations of y_i will be controlled and thus the corresponding statistical procedure will be robust.

We now study the robustness and tractability of the M-estimator of (4.11) based on our framework in Theorem 2. In order to emphasize on the effects of the tuning parameter α and the contamination ratio δ on the robustness property and tractability property, we consider a simplified assumption on the feature vector x_i and the pdf of idealized residual f_0.

Assumption 3. (a) *The feature vector x_i are i.i.d multivariate Gaussian*

distribution $N(0, \tau^2 I_{p\times p})$.

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

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

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

Corollary 1. *Assume Assumption 3 holds and* $||\theta_0||_2 \leq r/3$. *For any* $\pi > 0$, *there exist 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 (4.11) are in* $B_2^p(\theta_0, \eta_0 + \frac{1}{1-\delta}\zeta)$.

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

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

Here

$$\zeta = \frac{1}{13.5\sqrt{3\alpha}(1+\alpha\sigma^2)^{3/2}\tau}, \tag{4.12}$$

$$\eta_0(\delta,\alpha) = \frac{\delta}{1-\delta}\sqrt{\frac{e}{\alpha}}\frac{4(1+\alpha\sigma^2)^{3/2}}{\tau}e^{\frac{32\alpha r^2\tau^2}{3(1+\alpha\sigma^2)}}, \tag{4.13}$$

$$\eta_1(\delta,\alpha) = \frac{1}{9\sqrt{3\alpha}(1+\alpha\sigma^2)^{3/2}\tau}\left[1-\delta(1+3(1+\alpha\sigma^2)^{3/2})\right]. \tag{4.14}$$

It is interesting to see the special case of Corollary 1 with $\alpha = 0$, which reduces to the least square estimator. On the one hand, with $\alpha = 0$, 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 it is always tractable. On the other hand, since $\eta_0(\delta, \alpha = 0) = +\infty$, the solution of the optimization problem in (4.16) 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 the outliers. This recovers the well-known fact: the least square 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 the Welsch's estimator has nice tractability when there are no outliers. However, when the percentage of outlier δ is increasing, $\eta_1(\delta, \alpha)$ will decrease, which implies more outliers will reduce the tractability of the M-estimator.

4.2 Penalized M-estimators via Tukey's Bisquare Loss

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

$$\rho_\alpha(t) = \begin{cases} \frac{1}{6}\alpha^2 \left[1 - (1 - (t/\alpha)^2)^3\right], & \text{if } |t| \le \alpha \\ 0, & \text{if } |t| > \alpha. \end{cases} \tag{4.15}$$

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

$$\begin{aligned} \min_{\theta} \quad & \hat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(y_i - \langle \theta, x_i \rangle) + \lambda_n ||\theta||_1, \\ \text{subject to} \quad & ||\theta||_2 \leq r. \end{aligned} \tag{4.16}$$

Note the loss function $\rho_\alpha(t)$ in (4.15) is non-convex. For fixed $\alpha > 0$, $\rho'_\alpha(t), \rho''_\alpha(t)$ are all bounded. We now study the robustness and tractability of the penalized M-estimator of (4.16) based on our framework in Theorem 3. When α goes to ∞, $\rho_\alpha(t)$ will converge to $t^2/2$. Thus, the penalized M-estimator obtained by (4.16) reduces to the LASSO estimator, which can be computed easily. However, LASSO is also known to be very sensitive to the outliers (Alfons et al., 2013). On the other hand, when α increases, the estimator becomes more robust but may lose tractability due to the non-convexity of the function $\rho_\alpha(t)$ as well as the presence of the outliers.

In order to emphasize on 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 idealized residual f_0.

Assumption 4. **(a)** *The feature vector x_i are i.i.d multivariate uniform distribution $[-\tau, \tau]^p$.*

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

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

With Assumption 4 and Theorem 3, we can get the following Corollary 2, which characterizes the robustness and tractability of the penalized M-estimator with Tukey's exponential squared loss in (4.15):

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

(a) *All stationary points of problem (4.16) 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 (4.16) has a unique local stationary point which is also the global minimizer.*

Here

$$\eta_0(\delta,\alpha) = \frac{\delta}{1-\delta}\frac{28\sqrt{2\pi}}{\tau\sigma^3\alpha^2}e^{\frac{\alpha^2+64\tau^2r^2}{\sigma^2}}, \tag{4.17}$$

$$\eta_1(\delta,\alpha) = \frac{(1-\delta)M(\alpha,\sigma)\tau^2 - 4\delta}{2\sqrt{3}\tau}\alpha, \tag{4.18}$$

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* $\alpha > 0, \sigma > 0$.

It is interesting to see the special case of Corollary 2 with $\alpha \to \infty$, which 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 it is always tractable. On the other hand, since $\eta_0(\delta, \alpha \to \infty) \to +\infty$, the solution of the optimization problem in (4.16) 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 the outliers. This recovers the well-known fact: 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 the Tukey's estimator has nice tractability when there are no outliers. However, when the percentage of outlier δ is increasing, $\eta_1(\delta, \alpha)$ will decrease, which implies more outliers may reduce the tractability of the M-estimator.

5. Simulation Results

In this section, we report the simulation results by using Welsch's exponential loss and Tukey's bisquare loss when the data are contaminated, using the synthetic data. We first generate covariates $x_i \sim N(0, I_{p\times p})$ and re-

sponses $y_i = \langle \theta_0, x_i \rangle + \epsilon_i$, where $||\theta_0||_2 = 1$. We consider the case when the residual term ϵ_i have gross error model with contamination ratio δ, i.e., $\epsilon_i \sim (1-\delta)N(0,1) + \delta N(\mu_i, 3^2)$ 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 has non-zero mean.

In the first part, we consider the low-dimensional case when the dimension $p = 10$. Specifically, we generate $n = 100$ pairs of data $(y_i, x_i)_{i=1,..,n}$ with dimension $p = 10$ and with different choices of contamination ratios δ. We use projected gradient descent to solve the optimization problem in (4.11) with Welsch's loss and $r = 10$. To make the iteration points be inside the ball, we will project the points back into $B_2^p(0, r = 10)$ if they fall out of the ball. The step size is fixed as 1. In order to test the tractability of the M-estimator, we run gradient descent algorithm with 20 random initial values in the ball $B_2^p(0, r = 10)$ to see whether the gradient descent algorithm can converge to the same stationary point or not. Denote $\hat{\theta}(k)$ as the k^{th} iteration points, we then plot the empirical standard deviation of each iteration $\text{std}(\hat{\theta}(k)) = \text{Tr}(\widehat{\text{Var}}(\hat{\theta}(k)))$ among those 20 different initializations. Figure 1 shows the convergence of the gradient descent algorithm for the Welsch's exponential loss with the choice of $\alpha = 0.1$ under the gross error model with different δ. From Figure 1 we observe when the proportion of

outliers is small (i.e., $\delta \leq 0.1$,) gradient descent could converge to the same stationary point fast. However, when the contamination ratio δ becomes larger, gradient descent may not converge to the same point for different initial points, indicating the loss of tractability for the same objective function with increasing proportion of outliers. Those observations are consistent to our Theorem 2, which asserts the M-estimator is tractable when the contamination ratio δ is small. Then, we further show the empirical standard deviation at the $k = 300$ iteration $\text{std}(\hat{\theta}(300))$ when $p = 20$ and the ratio of n/p varies from 1 to 21 in Figure 2. From Figure 2, we can see 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 small proportion of outlier δ, the algorithm will 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 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 are well expected as those M-estimators are consistent without outliers (Huber, 1964; Huber and Ronchetti, 2009). However, for the M-estimator with $\alpha = 0$, i.e., the

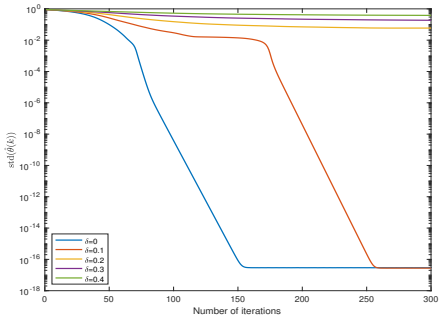

Figure 1: The value of $\mathrm{std}(\hat{\theta}(k))$ for different δ. Y-axis is with log scale.

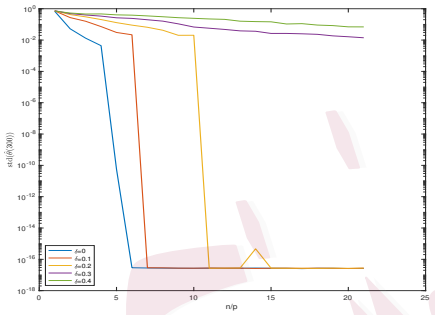

Figure 2: The value of $\mathrm{std}(\hat{\theta}(300))$ for different δ. Y-axis is with log scale

least square estimator, the estimation error will increase dramatically as the proportion of outliers increases. This confirms that the least square estimator is not robust to the outliers.

Meanwhile, when $\alpha = 0.1$, the overall estimation error does not increase much even with 40% outliers, which clearly demonstrate the robustness of the M-estimator. Note that when α is further increased from 0.1 to 0.3, although the estimator error is still very small for $\delta \leq 0.2$, it will increase dramatically when δ is greater than 0.2. We believe that two reasons contribute to this phenomenon: robustness starts to decrease when α becomes too large. More importantly, the algorithm fails to find the global optimum due to 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 efficiently when the proportion of outliers is smaller than that bound.

In the second part, we present our results in the high-dimensional region when $p = 200, n = 200$. 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 be $1/\sqrt{10}$. We use proximal gradient descent algorithm to solve problem (3.9) with Tukey's bisquare loss. Similarly, we will project the points back into $B_2^p(0, r = 10)$ if they fall out 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 by Tukey's loss with different choices of tuning parameter $\alpha = 4, 5, 10, 20, 500$. We generate 100 realizations of (Y, X) and run proximal gradient descent algorithm. The average estimation errors between the penalized M-estimator and the true parameter are reported in Figure 4. First, note as α is large, Tukey's loss will be similar to the square loss. Thus, the penalized M estimator with $\alpha = 500$ will have a similar performance as 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 will not increase a lot even if $\delta = 0.4$. These results imply

the robustness of the penalized robust M-estimator.

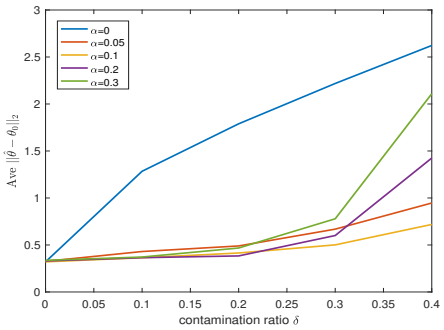

Figure 3: The estimation error for different α and δ.

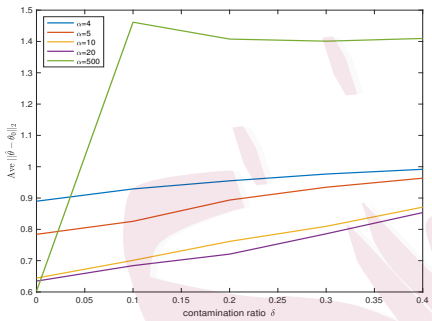

Figure 4: The estimation error for different α and δ.

Next, we will illustrate the tractability of the penalized M-estimator by showing $\text{std}(\hat{\theta}(k))$ among 20 initializations of the proximal gradient descent algorithm for the Tukey's loss with the choice of $\alpha = 20$ under the gross error model with different δ. Figure 5 shows the result with $p = 200, n = 200$ and Figure 6 shows the result with $p = 400, n = 400$. From the two plots, we observe an interesting phenomenon: the proximal gradient descent will converge 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, more accurate analysis on the tractability property of the penalized M-estimators are needed.

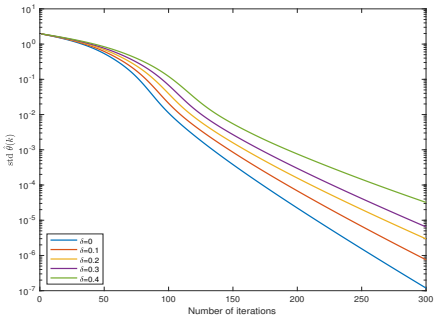

Figure 5: The value of std($\hat{\theta}(k)$) for different δ. n=p=200.

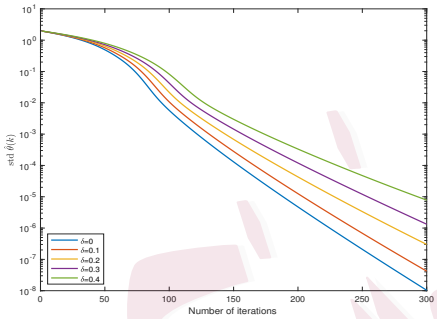

Figure 6: The value of std($\hat{\theta}(k)$) for different δ. n=p=400.

6. Case study

In this section, we present a case study of the robust regression problem for the Airfoil Self-Noise dataset (Brooks et al., 2014), which is available on UCI Machine Learning Repository. The dataset was processed by NASA and is commonly used for regression study to learn the relation between the airfoil self-noise and five explanatory variables. Specifically, the dataset contain the following 5 explanatory variables: Frequency (in Hertzs), 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 1503 observations in the dataset. The response variable is Scaled sound pressure level (in decibels). In this section, the five explanatory variables are scaled to have 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 $\mu_i = ||x_i||_2^2 + 1$.

We consider the M-estimator using Welsch's exponential loss (Dennis Jr and Welsch, 1978) on the dataset to validate the tractability and the robustness of the corresponding M-estimator. First, we run 100 Monte Carlo simulations. At each time, we split the dataset which consists of 1503 pairs of data into a training dataset of size 1000 and a testing dataset of size 503. Then for the training dataset, we use gradient descent method with 20 different initial values to update the iteration points.

Figure 7 shows the average distance between each iteration point and the optimal point with the choice of $\alpha = 0.7$ and step size 0.5. Clearly, when δ is smaller than 0.3, gradient descent will converge 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 square error on the testing data. Figure 8 shows the average prediction error on the testing data. As we can see, the prediction error with the choice of $\alpha = 0$ will increase dramatically when the percentage of outliers increases. In contrast, the prediction errors of M-estimators with

$\alpha = 0.4$ is stable even with a large percentage of outliers. This illustrates the robustness of M-estimators for some positive α.

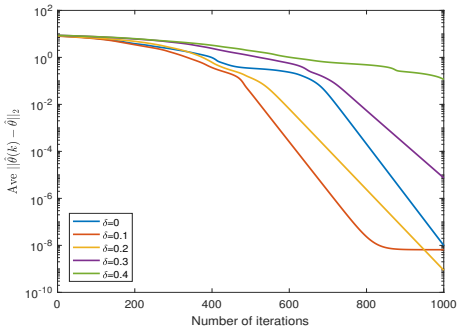

Figure 7: The convergence of gradient descent algorithm for different δ. Y-axis is with log scale.

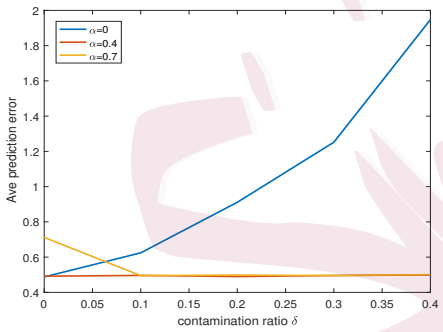

Figure 8: The prediction error for different α and δ

7. Conclusions

In this paper, we investigate the robustness and computational tractability of general (non-convex) M-estimators in both classical low-dimensional regime and modern high-dimensional regime. In terms of *robustness*, in the low-dimensional regime, we show the estimation error of the M-estimator is as the order of $O(\delta + \sqrt{\frac{p \log n}{n}})$, which nearly achieves the minimax lower bound of $O(\delta + \sqrt{\frac{p}{n}})$ in Chen et al. (2016). In the high-dimensional regime, we show the estimation error of the penalized M-estimator has the esti-

mation error as the order of $O(\delta + \sqrt{\frac{s_0 \log p}{n}})$, which achieves the minimax estimation rate (Chen et al., 2016).

In terms of *tractability*, our theoretical results imply under sufficient conditions, when the percentage of arbitrary outliers is small, the general M-estimator could have good computational tractability since it has only one unique stationary point, even if the loss function is non-convex. Therefore, M-estimators can tolerate certain level of outliers by keeping both estimation accuracy and computation efficiency. Both simulation and real data case study are conducted to validate our theoretical results about the robustness and tractability of M-estimators in the presence of outliers.

References

Alfons, A., C. Croux, and S. Gelper (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics 7*(1), 226–248.

Anandkumar, A., R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research 15*(1), 2773–2832.

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

Bai, Z., C. R. Rao, and Y. Wu (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica 2*(1), 237–254.

Beaton, A. E. and J. W. Tukey (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics 16*(2), 147–185.

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

Brooks, T., S. Pope, and M. Marcolini (2014). Uci machine learning repository.

Candes, E. J., X. Li, and M. Soltanolkotabi (2015). Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis 39*(2), 277–299.

Chang, L., S. Roberts, and A. Welsh (2018). Robust lasso regression using tukey's biweight criterion. *Technometrics 60*(1), 36–47.

REFERENCES

Chen, M., C. Gao, and Z. Ren (2016). A general decision theory for huber's ϵ-contamination model. *Electronic Journal of Statistics 10*(2), 3752–3774.

Cheng, G., J. Z. Huang, et al. (2010). Bootstrap consistency for general semiparametric m-estimation. *The Annals of Statistics 38*(5), 2884–2915.

Dennis Jr, J. E. and R. E. Welsch (1978). Techniques for nonlinear least squares and robust regression. *Communications in Statistics-Simulation and Computation 7*(4), 345–359.

Donoho, D. L. and P. J. Huber (1983). The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157–184.

El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences 110*(36), 14557–14562.

Ferrari, D. and Y. Yang (2010). Maximum lq-likelihood estimation. *The Annals of Statistics 38*(2), 753–783.

Geyer, C. J. et al. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics 22*(4), 1993–2010.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (2011). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

Huber, P. J. (1964). Robust estimation of a location parameter. *The annals of mathematical statistics 35*(1), 73–101.

REFERENCES

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

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

Lehmann, E. L. and G. Casella (2006). *Theory of point estimation*. Springer Science & Business Media.

Li, G., H. Peng, and L. Zhu (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*(2), 866–896.

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

Mairal, J., F. Bach, J. Ponce, and G. Sapiro (2009). Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pp. 689–696. ACM.

Maronna, R. A. and V. J. Yohai (1981). Asymptotic behavior of general m-estimates for regression and scale with random carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 58*(1), 7–20.

Mei, S., Y. Bai, A. Montanari, et al. (2018). The landscape of empirical risk for nonconvex

losses. *The Annals of Statistics 46*(6A), 2747–2774.

Mizera, I. and C. H. Müller (1999). Breakdown points and variation exponents of robust *m*-estimators in linear models. *The Annals of Statistics 27*(4), 1164–1177.

Qin, Y. and C. E. Priebe (2017). Robust hypothesis testing via lq-likelihood. *Statistica Sinica 27*(4), 1793–1813.

Rey, W. J. (2012). *Introduction to robust and quasi-robust statistical methods*. Springer Science & Business Media.

Wang, X., Y. Jiang, M. Huang, and H. Zhang (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association 108*(502), 632–643.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

The Department of Statistics, University of Nebraska-Lincoln

E-mail: rzhang35@unl.edu

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

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

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

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

Alibaba Inc.

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