Electronic Journal of Statistics
Vol. 15 (2021) 5523–5544
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1911

A generalized Catoni's $\mathbb{M}$-estimator under finite α-th moment assumption with $\alpha \in (1,2)$

Peng Chen [1,2,3] **Xinghu Jin** [2,3] **Xiang Li** [2,3] **and Lihu Xu** [*2,3]

[1] *Department of Mathematics, College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China.*
e-mail: chenpengmath@nuaa.edu.cn

[2] *Department of Mathematics, Faculty of Science and Technology, University of Macau, Av. Padre Tomás Pereira, Taipa Macau, China.*

[3] *UM Zhuhai Research Institute, Zhuhai, China.*
e-mail: jinxinghu764@163.com; yc07904@um.edu.mo; lihuxu@um.edu.mo

Abstract: We generalize Catoni's M-estimator, put forward in [3] by Catoni under finite variance assumption, to the case in which distributions can have finite α-th moment with $\alpha \in (1,2)$. Our approach, inspired by the Taylor-like expansion developed in [4], is via slightly modifying the influence function φ in [3]. A deviation bound is established for this generalized estimator, and coincides with that in [3] as $\alpha \uparrow 2$. Experiment shows that our M-estimator performs better than the empirical mean, the smaller the α is, the better the performance will be. As an application, we study an ℓ_1 regression considered by Zhang et al. [19], who assumed that samples have finite variance, under finite α-th moment assumption with $\alpha \in (1,2)$.

MSC2020 subject classifications: Primary 62G05; secondary 62G35.
Keywords and phrases: Catoni's M-estimator, empirical mean, α-th moment, deviation bound, ℓ_1 regression.

Received October 2020.

Contents

arXiv: 2010.05008
*Corresponding author

1. Introduction

Let $X_1, \cdots, X_n$ be a sequence of samples drawn from a distribution, its empirical mean estimator is defined by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

The empirical mean $\bar{X}$ has an optimal minimax mean square error among all mean estimators, but its deviation is suboptimal for heavy tail distribution [3].

Catoni put forward in his seminal paper [3] a new M-estimator for heavy-tailed samples with finite variances, by solving the following equation about θ:

$$\sum_{i=1}^{n} \varphi\left(\beta(X_i - \theta)\right) = 0$$

with

$$-\log\left(1 - x + \frac{|x|^2}{2}\right) \leq \varphi(x) \leq \log\left(1 + x + \frac{|x|^2}{2}\right),$$

where $\beta > 0$ is a parameter to be tuned and φ is non-decreasing and called influence function. The deviation performance of this estimator is much better than $\bar{X}$. Catoni's idea has been broadly applied to many research problems, see for instance [1, 15, 5, 6, 7, 11, 12, 17]. The finite variance assumption plays an important role in Catoni's analysis, but it rules out many interesting distributions such as Pareto law [10, 16, 4, 8], which describes the distributions of wealth and social networks.

We generalize Catoni's M-estimator to the case in which samples can have finite α-th moment with $\alpha \in (1, 2)$. Our approach is by replacing Catoni's influence function with the one satisfying

$$-\log\left(1 - x + \frac{|x|^\alpha}{\alpha}\right) \leq \varphi(x) \leq \log\left(1 + x + \frac{|x|^\alpha}{\alpha}\right).$$

The choice of the new φ is inspired by the Taylor-like expansion developed in [4]. By an argument very similar to Catnoi's, we obtain a deviation upper bound which coincides with that in [3] as $\alpha \uparrow 2$ (see Theorem 2.1 and Remark 2.1 below). Experiment shows that our generalized M-estimator performs better than the empirical mean estimator, the smaller the α is, the better the performance will be.

Catoni's argument for establishing the M-estimator in [3] can be divided into two steps. The one is to find two deterministic values θ_- and θ_+, both depending on a parameter β to be tuned later, such that the M-estimator $\hat{\theta}$ falls between θ_- and θ_+ with high probability. The θ_- and θ_+ were obtained explicitly by solving two quadratic algebraic equations $B_-(\theta) = 0$ and $B_+(\theta) = 0$ respectively, whereas in our setting the corresponding equations are not quadratic and the solutions do not have explicit forms. Alternatively, we first prove that $B_-(\theta) = 0$ has a largest solution, while $B_+(\theta) = 0$ has a smallest one, and then use them as a replacement of θ_- and θ_+ in our analysis. The other is to show that as one chooses $\beta > 0$ sufficiently small, the difference between θ_- and θ_+ can be as small as we wish, whence the estimator can be localized in a small interval with high probability. As in [3], we also need to choose a sufficiently small β (depending on α) to make our estimator fall in a small interval whose two end points are the above special solutions. As $\alpha \uparrow 2$, our result coincides with that in [3].

As an application of our generalized estimator, we consider the ℓ_1- regression with heavy-tailed samples studied by Zhang et al. [19] who assumed the samples have finite variance. The linear regression considered in [19] aims to find the minimizer θ^* of the optimization problem as follows:

$$\min_{\theta\in\mathbf{\Theta}} R_{\ell_1}(\theta) \qquad \text{with} \qquad R_{\ell_1}(\theta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbf{\Pi}}\left[|\mathbf{x}^T\theta - y|\right],$$

where $\mathbf{\Pi}$ is a probability distribution, and $\mathbf{\Theta} \subseteq \mathbb{R}^d$ is the set in which θ^* is located. In practice, $\mathbf{\Pi}$ is not known, one usually draws a data set $\mathcal{T} = (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ from $\mathbf{\Pi}$ and considers the following empirical optimization problem:

$$\min_{\theta\in\mathbf{\Theta}} \widehat{R}_{\ell_1}(\theta) \quad \text{with} \quad \widehat{R}_{\ell_1}(\theta) = \frac{1}{n}\sum_{i=1}^{n} |\mathbf{x}_i^T\theta - y_i|.$$

The theoretical guarantees for bounded or sub-Gaussian distributed $\mathbf{\Pi}$ have been discussed in many papers, see for instance [2, 9, 18].

Inspired by Catoni's work, Zhang et al. considered the case that $\mathbf{\Pi}$ is heavy-tailed with finite variance and proposed a new minimization problem

$$\min_{\theta\in\mathbf{\Theta}} \widehat{R}_{\varphi,\ell_1}(\theta) \quad \text{with} \quad \widehat{R}_{\varphi,\ell_1}(\theta) = \frac{1}{n\beta}\sum_{i=1}^{n} \varphi\left(\beta|y_i - \mathbf{x}_i^T\theta|\right),$$

where φ is the same as that in [3] and $\beta > 0$ is a parameter to be tuned. A new estimator was established from this minimization problem and an error bound was obtained. When the sample size n tends to infinity, this error bound tends to zero.

Thanks to the analysis of Section 2 below, we extend the results in [19] to the case in which samples can have finite α-th moment with $\alpha \in (1, 2)$, our approach is by replacing the original φ with the one in Section 2 and solving the corresponding minimization problem. We establish a similar error bound for our estimator and prove that it tends to zero as $n \to \infty$.

The paper is organized as follows. In Section 2, we give the deviation analysis for the generalized M-estimator and show that the M-estimator has a performance better than the empirical mean. In Section 3, we state the upper bounds and the corresponding lower bounds on the empirical mean. In the last section, under finite α-th moment assumption with $\alpha \in (1, 2)$, we discuss the ℓ_1-regression of heavy-tailed distributions.

2. A generalized Catoni's M-estimator and its deviation analysis

Let $(X_i)_{i=1}^n$ be a sequence of i.i.d. samples drawn from some unknown probability distribution $\mathbf{\Pi}$ on $\mathbb{R}$. We assume that there exists some $\alpha \in (1,2)$ such that

$$\mathbb{E}|X_1|^\alpha < \infty.$$

Further denote

$$m = \mathbb{E}[X_1], \qquad v = \mathbb{E}|X_1 - m|^\alpha.$$

Inspired by Catoni's idea in [3] and the Taylor-like expansion develop in [4], we consider a non-decreasing function $\varphi : \mathbb{R} \to \mathbb{R}$ such that

$$-\log\left(1 - x + \frac{|x|^\alpha}{\alpha}\right) \leq \varphi(x) \leq \log\left(1 + x + \frac{|x|^\alpha}{\alpha}\right), \quad \forall x \in \mathbb{R}. \tag{2.1}$$

We claim that such φ exists. Indeed, to prove the existence, it suffices to show

$$-\log\left(1 - x + \frac{|x|^\alpha}{\alpha}\right) \leq \log\left(1 + x + \frac{|x|^\alpha}{\alpha}\right), \quad \forall x \in \mathbb{R}. \tag{2.2}$$

To prove (2.2), we only need to show

$$\log\left[\left(1 + \frac{|x|^\alpha}{\alpha} + x\right)\left(1 + \frac{|x|^\alpha}{\alpha} - x\right)\right] \geq 0.$$

By symmetry, we can restrict to $x \geq 0$. When $x \in [0,1]$, since $\alpha \in (1,2)$, we have

$$\left(1 + \frac{|x|^\alpha}{\alpha}\right)^2 - x^2 = 1 + \frac{2|x|^\alpha}{\alpha} + \frac{|x|^{2\alpha}}{\alpha^2} - x^2 \geq 1 + \frac{2}{\alpha}\left(|x|^\alpha - |x|^2\right) \geq 1,$$

which implies

$$\log\left[\left(1 + \frac{|x|^\alpha}{\alpha} + x\right)\left(1 + \frac{|x|^\alpha}{\alpha} - x\right)\right] = \log\left[\left(1 + \frac{|x|^\alpha}{\alpha}\right)^2 - x^2\right] \geq 0.$$

When $x \geq 1$, since the functions

$$x \mapsto 1 + x + \frac{x^\alpha}{\alpha} \quad \text{and} \quad x \mapsto 1 - x + \frac{x^\alpha}{\alpha}$$

are both increasing and strictly positive, so that their product is also increasing and $(1 + x + \frac{x^\alpha}{\alpha})(1 - x + \frac{x^\alpha}{\alpha}) \geq \frac{2}{\alpha} + \frac{1}{\alpha^2} > 1$ for all $x \geq 1$. Thus, we know the inequality (2.2) holds.

The *widest* possible choice of φ (see Figure 1) compatible with these inequalities is

$$\varphi(x) = \begin{cases} \log\left(1 + x + \frac{x^\alpha}{\alpha}\right), & x \geq 0, \\ -\log\left(1 - x + \frac{|x|^\alpha}{\alpha}\right), & x < 0. \end{cases}$$

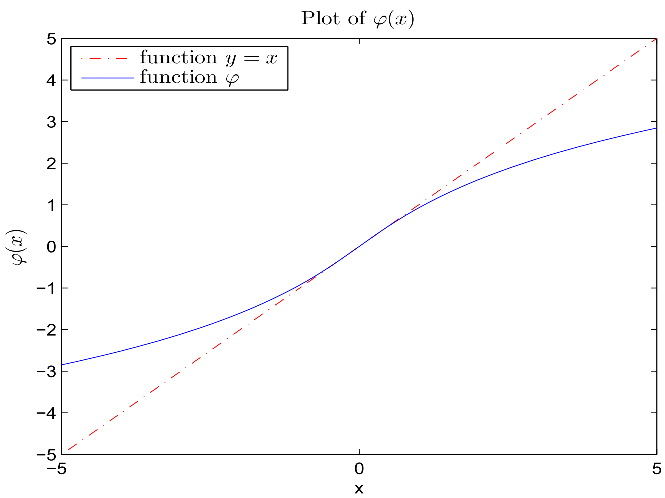

FIG 1. *widest possible choice of φ*

Let β be some strictly positive real parameter that will be chosen later and denote the estimator of the mean m by $\hat{\theta}$, which is the solution to the equation

$$\sum_{i=1}^{n} \varphi\left(\beta\left(X_i - \hat{\theta}\right)\right) = 0.$$

For further use, we denote

$$r(\theta) = \frac{1}{\beta n} \sum_{i=1}^{n} \varphi\left(\beta\left(X_i - \theta\right)\right), \quad \theta \in \mathbb{R}. \tag{2.3}$$

It is easy to see $r(\theta)$ is a non-increasing random variable since φ is non-decreasing.

Let us briefly explain the way in which we look for the estimator $\hat{\theta}$ as the following. We firstly find two deterministic values θ_- and θ_+, both depending on β, such that $r(\theta_-) > 0 > r(\theta_+)$ with high probability, from the non-decreasing property of r, we know that $\theta_- < \hat{\theta} < \theta_+$ holds with high probability. Secondly, we show that as we choose $\beta > 0$ sufficiently small, the difference between θ_- and θ_+ can be as small as we wish, whence the estimator can be localized in a small interval with high probability.

Lemma 2.1. *Keep the same notation and assumptions as above. Then, for any $\theta \in \mathbb{R}$ and $1 < p, q < \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have*

$$\mathbb{E}\left[\exp\left(\beta n r(\theta)\right)\right] \leq \exp\left(n\beta(m-\theta) + \frac{n\beta^{\alpha}}{\alpha}\left(p^{\alpha-1}v + q^{\alpha-1}|m-\theta|^{\alpha}\right)\right) \tag{2.4}$$

and

$$\mathbb{E}\left[\exp\left(-\beta n r(\theta)\right)\right] \le \exp\left(-n\beta(m-\theta) + \frac{n\beta^\alpha}{\alpha}\left(p^{\alpha-1}v + q^{\alpha-1}|m-\theta|^\alpha\right)\right). \tag{2.5}$$

Proof. Notice that $\alpha > 1$, for any $x > 0$, the function $x \mapsto x^\alpha$ is convex. Then, for any $a, b > 0$, we have

$$(a+b)^\alpha \le p^{\alpha-1}a^\alpha + q^{\alpha-1}b^\alpha. \tag{2.6}$$

Then, noting that X_i, $i = 1, \cdots, n$ are i.i.d., by (2.1), we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\beta n r(\theta)\right)\right] =& \mathbb{E}\left[\exp\left[\sum_{i=1}^n \varphi\left(\beta(X_i-\theta)\right)\right]\right]\\ &= \left(\mathbb{E}\left[\exp\left[\varphi\left(\beta(X_1-\theta)\right)\right]\right]\right)^n\\ &\le \left(\mathbb{E}\left[1+\beta(X_1-\theta)+\frac{\beta^\alpha}{\alpha}|X_1-\theta|^\alpha\right]\right)^n,\end{aligned}$$

and noting that $\alpha \in (1,2)$, by (2.6), we further have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\beta n r(\theta)\right)\right] &\le \left[1+\beta(m-\theta)+\frac{\beta^\alpha}{\alpha}\mathbb{E}|X_1-m+m-\theta|^\alpha\right]^n\\ &\le \left[1+\beta(m-\theta)+\frac{\beta^\alpha}{\alpha}\left(p^{\alpha-1}v+q^{\alpha-1}|m-\theta|^\alpha\right)\right]^n\\ &\le \exp\left(n\beta(m-\theta)+\frac{n\beta^\alpha}{\alpha}\left(p^{\alpha-1}v+q^{\alpha-1}|m-\theta|^\alpha\right)\right),\end{aligned}$$

where the last inequality is by the inequality $1 + x \le e^x$ for any $x \in \mathbb{R}$, (2.4) is proved and the inequality (2.5) can be proved in the same way. The proof is complete. □

According to (2.4) and (2.5), for any $\epsilon \in (0, \frac{1}{2})$, we denote

$$B_+(\theta) = m-\theta+\frac{\beta^{\alpha-1}}{\alpha}\left(p^{\alpha-1}v+q^{\alpha-1}|m-\theta|^\alpha\right)+\frac{\log\left(\epsilon^{-1}\right)}{n\beta}, \tag{2.7}$$

$$B_-(\theta) = m-\theta-\frac{\beta^{\alpha-1}}{\alpha}\left(p^{\alpha-1}v+q^{\alpha-1}|m-\theta|^\alpha\right)-\frac{\log\left(\epsilon^{-1}\right)}{n\beta}. \tag{2.8}$$

Lemma 2.2. *Keep the same notation and assumptions as above. Then, for any $\theta \in \mathbb{R}$ and $1 < p, q < \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have*

$$\mathbb{P}\left(r(\theta) < B_+(\theta)\right) \ge 1-\epsilon \tag{2.9}$$

and

$$\mathbb{P}\left(r(\theta) > B_-(\theta)\right) \ge 1-\epsilon. \tag{2.10}$$

In particular, for any $\theta \in \mathbb{R}$, we have

$$\mathbb{P}\left(B_-(\theta) < r(\theta) < B_+(\theta)\right) \ge 1-2\epsilon. \tag{2.11}$$

Proof. By Markov inequality and (2.4), we have

$$
\begin{aligned}
&\mathbb{P}\left(r(\theta) \geq B_+(\theta)\right)\\
=&\mathbb{P}\left(\exp\left(n\beta r(\theta)\right) \geq \exp\left(n\beta B_+(\theta)\right)\right)\\
\leq&\frac{\mathbb{E}\left[\exp\left(n\beta r(\theta)\right)\right]}{\exp\left(n\beta\left(m-\theta+\frac{\beta^{\alpha-1}}{\alpha}\left(p^{\alpha-1}v+q^{\alpha-1}|m-\theta|^{\alpha}\right)+\frac{\log(\epsilon^{-1})}{n\beta}\right)\right)}\\
\leq&\frac{\exp\left(n\beta(m-\theta)+\frac{n\beta^{\alpha}}{\alpha}\left(p^{\alpha-1}v+q^{\alpha-1}|m-\theta|^{\alpha}\right)\right)}{\exp\left(n\beta(m-\theta)+\frac{n\beta^{\alpha}}{\alpha}\left(p^{\alpha-1}v+q^{\alpha-1}|m-\theta|^{\alpha}\right)+\log\left(\epsilon^{-1}\right)\right)}=\epsilon,
\end{aligned}
$$

the inequality (2.9) is proved. With the help of (2.5), the inequality (2.10) can be proved in the same way. The estimate (2.11) immediately follows from (2.9) and (2.10). ☐

Now, we can give the main result in this section, which can give a deviation upper bound for the M-estimator $\hat{\theta}$.

Theorem 2.1. *Keep the same notation and assumptions as above. For any* $\epsilon \in \left(0, \frac{1}{2}\right)$ *and* $c > 1$ *be a constant, let us choose the positive integer* n *satisfying*

$$
n \geq \left(\frac{c^{\alpha}}{\alpha(c-1)}\right)^{\frac{1}{\alpha-1}} \frac{\alpha q \log\left(\epsilon^{-1}\right)}{\alpha-1}, \tag{2.12}
$$

and let $\beta = \left(\frac{\alpha\log\left(\epsilon^{-1}\right)}{(\alpha-1)p^{\alpha-1}vn}\right)^{\frac{1}{\alpha}}$. *Then, the inequality*

$$
\left|m-\hat{\theta}\right| \leq v^{\frac{1}{\alpha}}\left(\frac{\alpha p\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\frac{\alpha-1}{\alpha}}\left(1-\frac{1}{\alpha}\left(\frac{cq\alpha\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\alpha-1}\right)^{-1} := \eta \tag{2.13}
$$

holds with probability at least $1-2\epsilon$.

Remark 2.1. *In Theorem 2.1, if we choose* $c = 2$ *and* $q = \sqrt{n}$*, when* n *tends to infinity, we get that*

$$
p = (1-\frac{1}{q})^{-1} = 1 + \frac{1}{\sqrt{n}-1} \sim 1
$$

and

$$
\eta \sim v^{\frac{1}{\alpha}}\left(\frac{\alpha\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\frac{\alpha-1}{\alpha}},
$$

while the condition on n *is*

$$
\sqrt{n} \geq \left(\frac{2^{\alpha}}{\alpha}\right)^{\frac{1}{\alpha-1}} \frac{\alpha\log\left(\epsilon^{-1}\right)}{\alpha-1}.
$$

When $\alpha = 2$, our result coincides with that in [3, Proposition 2.4] up to a constant.

Proof. The key point of the proof is to find two values θ_+ and θ_- such that $B_+(\theta_+) \leq 0$ and $B_-(\theta_-) \geq 0$. Once we find such θ_+ and θ_-, Lemma 2.2 and the monotonicity of $r(\theta)$ will then give us a high probability bound.

Recall $B_+(\theta)$ in Eq. (2.7), we know $B_+(\theta) > 0$ when $\theta \leq m$. Denote $\theta_+ = m + \eta_+$, we are looking for a positive value of η_+ such that

$$B_+(\theta_+) = -\eta_+ + \frac{\beta^{\alpha-1}}{\alpha}\left[p^{\alpha-1}v + q^{\alpha-1}\eta_+^{\alpha}\right] + \frac{\log\left(\epsilon^{-1}\right)}{n\beta} \leq 0,$$

that is, $a + b\eta_+^{\alpha} \leq \eta_+$ with $a = \frac{(\beta p)^{\alpha-1}}{\alpha}v + \frac{\log\left(\epsilon^{-1}\right)}{n\beta}$ and $b = \frac{(\beta q)^{\alpha-1}}{\alpha}$. This can also be written as

$$\frac{a}{1 - b\eta_+^{\alpha-1}} \leq \eta_+ \quad \text{and} \quad b\eta_+^{\alpha-1} < 1. \tag{2.14}$$

Notice the function $\frac{1}{1-bx^{\alpha-1}}$ is increasing in x when $bx^{\alpha-1} < 1$. So we can find a $c > 0$ such that $b\eta_+^{\alpha-1} < b(ca)^{\alpha-1} < 1$ and $\eta_+ = \frac{a}{1-b(ca)^{\alpha-1}}$. Now, to find a positive value η_+ satisfies (2.14), it suffice to find a positive value η_+ satisfies

$$b(ca)^{\alpha-1} < 1 \quad \text{and} \quad \eta_+ = \frac{a}{1 - b(ca)^{\alpha-1}} < ca. \tag{2.15}$$

A simple calculation shows that the second inequality is equivalent to $b(ca)^{\alpha-1} < \frac{c-1}{c}$ which implies the first inequality and $c > 1$. Hence, the restrictive conditions finally transform into

$$\eta_+ = \frac{a}{1 - b(ca)^{\alpha-1}} \quad \text{and} \quad b(ca)^{\alpha-1} < \frac{c-1}{c}. \tag{2.16}$$

Now, according to (2.15), choosing $\beta = \left(\frac{\alpha\log(\epsilon^{-1})}{(\alpha-1)p^{\alpha-1}nv}\right)^{\frac{1}{\alpha}}$ which minimizes a, we have

$$a = v^{\frac{1}{\alpha}}\left(\frac{\alpha p\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\frac{\alpha-1}{\alpha}} \quad \text{and} \quad b = \frac{q^{\alpha-1}}{\alpha}\left(\frac{\alpha\log\left(\epsilon^{-1}\right)}{(\alpha-1)p^{\alpha-1}vn}\right)^{\frac{\alpha-1}{\alpha}}.$$

For $c > 1$, if n satisfies (2.12), we have

$$b(ca)^{\alpha-1} = \frac{1}{\alpha}\left(\frac{\alpha q\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\alpha-1} c^{\alpha-1} \leq \frac{1}{\alpha}\frac{\alpha(c-1)}{c^{\alpha}}c^{\alpha-1} = \frac{c-1}{c} < 1,$$

so (2.16) is satisfied. Therefore, we have

$$\theta_+ - m = \eta_+ = \frac{a}{1 - b(ca)^{\alpha-1}}$$

$$=v^{\frac{1}{\alpha}}\left(\frac{\alpha p\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\frac{\alpha-1}{\alpha}}\left(1-\frac{1}{\alpha}\left(\frac{cq\alpha\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\alpha-1}\right)^{-1}.$$

Moreover, denote $\theta_-=m-\eta_-$, we are looking for a positive value of η_- such that

$$B_-(\theta_-)=\eta_--\frac{\beta^{\alpha-1}}{\alpha}\left(p^{\alpha-1}v+q^{\alpha-1}\eta_-^{\alpha}\right)-\frac{\log\left(\epsilon^{-1}\right)}{n\beta}\geq 0.$$

Then, the same argument as above implies

$$m-\theta_-=v^{\frac{1}{\alpha}}\left(\frac{\alpha p\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\frac{\alpha-1}{\alpha}}\left(1-\frac{1}{\alpha}\left(\frac{cq\alpha\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\alpha-1}\right)^{-1}.$$

By (2.11), we know that the following event holds with probability at least $1-2\epsilon$:

$$r(\theta_-)>0 \quad \text{and} \quad r(\theta_+)<0.$$

Since $r(\theta)$ is a continuous function and non-increasing, $r(\theta)=0$ has a solution $\hat{\theta}$ between θ_- and θ_+ such that

$$\theta_-\leq\hat{\theta}\leq\theta_+$$

holds with probability at least $1-2\epsilon$, that is, $\mathbb{P}\left(\theta_-\leq\hat{\theta}\leq\theta_+\right)\geq 1-2\epsilon$, which implies that the inequality

$$\left|\hat{\theta}-m\right|\leq v^{\frac{1}{\alpha}}\left(\frac{\alpha p\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\frac{\alpha-1}{\alpha}}\left(1-\frac{1}{\alpha}\left(\frac{cq\alpha\log\left(\epsilon^{-1}\right)}{(\alpha-1)n}\right)^{\alpha-1}\right)^{-1}$$

holds with probability at least $1-2\epsilon$. □

The empirical mean estimator is defined by

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

we postpone to study deviation bounds for the empirical mean $\bar{X}$ in Section 3 below.

In Figures 2-5, we compare the bound on the deviations of the M-estimator $\hat{\theta}$ with the deviations of the empirical mean $\bar{X}$, when the sample distribution is a Pareto distribution with shape parameter $\frac{2+\alpha}{2}$ and scale parameter $\left(\frac{2+\alpha}{2-\alpha}\right)^{-\frac{1}{\alpha}}$ (see, e.g., [10, Chapter 23]), that is,

$$\mathbb{P}\left(X_1\geq x\right) \quad = \quad 2^{-1}\left(\frac{2+\alpha}{2-\alpha}\right)^{-\frac{2+\alpha}{2\alpha}}x^{-\frac{2+\alpha}{2}}, \quad x\geq\left(\frac{2+\alpha}{2-\alpha}\right)^{-\frac{1}{\alpha}},$$

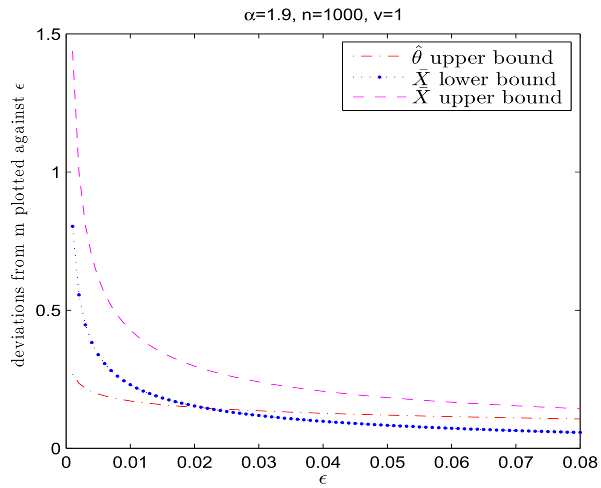

Fig 2. *Deviations of $\hat{\theta}$ from the sample mean, compared with those of empirical mean*

$$\mathbb{P}\left(X_1 \le x\right) \quad = \quad 2^{-1}\left(\frac{2+\alpha}{2-\alpha}\right)^{-\frac{2+\alpha}{2\alpha}}(-x)^{-\frac{2+\alpha}{2}}, \quad x \le -\left(\frac{2+\alpha}{2-\alpha}\right)^{-\frac{1}{\alpha}}.$$

By the definition, it is easy to verify that $m = \mathbb{E}X_1 = 0$ and $v = \mathbb{E}\left|X_1 - m\right|^{\alpha} = 1$. We can get figures for the upper bound of $\hat{\theta}$, the upper bound and lower bound of $\bar{X}$. It is obvious from Figures 2-5 that the $\hat{\theta}$ has a better performance when ϵ is small enough. We can also see that the smaller the α is, the better the performance of $\hat{\theta}$ will be comparing with that of $\bar{X}$. The parameters for Figures 2-5 are in Table 1 and $c = 2$, $q = \sqrt{n}$, where $0.001 : 0.001 : 0.08$ means the range of ϵ is from 0.001 to 0.08 with step-size 0.001. The ranges of ϵ in Table 1 satisfy (3.2) and the values of n in Table 1 satisfy (2.12) and (3.2).

Table 1
Parameters in Figures 2-5

	α	ϵ	n
Figure 2	1.9	0.001:0.001:0.08	1000
Figure 3	1.7	0.001:0.001:0.08	2000
Figure 4	1.5	0.001:0.001:0.08	6000
Figure 5	1.3	0.001:0.001:0.08	7000

3. The deviation upper and lower bounds of the empirical mean estimator

3.1. Upper bounds

Lemma 3.1. *Let $(X_i)_{i=1}^n$ be a sequence of random variables independently drawn from some distribution $\mathbf{\Pi}$ with mean m and α-th central moment v.*

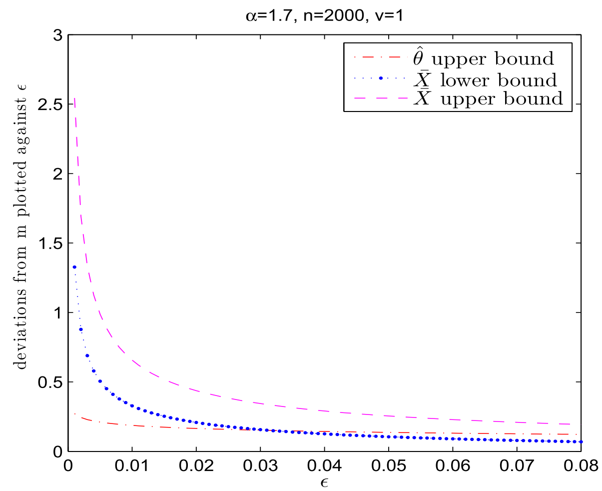

FIG 3. *Deviations of $\hat{\theta}$ from the sample mean, compared with those of empirical mean*

Then, denote the empirical mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, we have

$$\mathbb{P}\left(\left|\bar{X} - m\right| \geq \left(\frac{v}{\epsilon n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\right) \leq 2\epsilon.$$

Proof. Noticing that $(X_i - m)_{i=1}^n$ are i.i.d. random variables with mean zero, by [13, Theorem 2], we have

$$\mathbb{E}\left|\sum_{i=1}^n [X_i - m]\right|^\alpha \leq 2\sum_{i=1}^n \mathbb{E}\left|X_i - m\right|^\alpha = 2nv,$$

which implies

$$\mathbb{P}\left(\left|\bar{X} - m\right| \geq \left(\frac{v}{\epsilon n^{\alpha-1}}\right)^\alpha\right) \leq \frac{\mathbb{E}\left|\bar{X} - m\right|^\alpha}{\frac{v}{\epsilon n^{\alpha-1}}} \leq \frac{\frac{1}{n^\alpha}\mathbb{E}\left|\sum_{i=1}^n [X_i - m]\right|^\alpha}{\frac{v}{\epsilon n^{\alpha-1}}} \leq 2\epsilon,$$

the desired result follows. □

3.2. Lower bounds

In contrast to Lemma 3.1, the following lemma gives a lower bound for the deviations of the empirical mean for some specific distributions.

Lemma 3.2. *For any value of the α-th central moment v, any deviation $\eta > 0$, there is some distribution $\mathbf{\Pi}$ with mean zero and α-th central moment v such that*

$$\mathbb{P}\left(\bar{X} \geq \eta\right) = \mathbb{P}\left(\bar{X} \leq -\eta\right) \geq \frac{v}{3n^{\alpha-1}\eta^\alpha}\left(1 - \frac{v}{n^\alpha \eta^\alpha}\right)^{n-1}, \qquad (3.1)$$

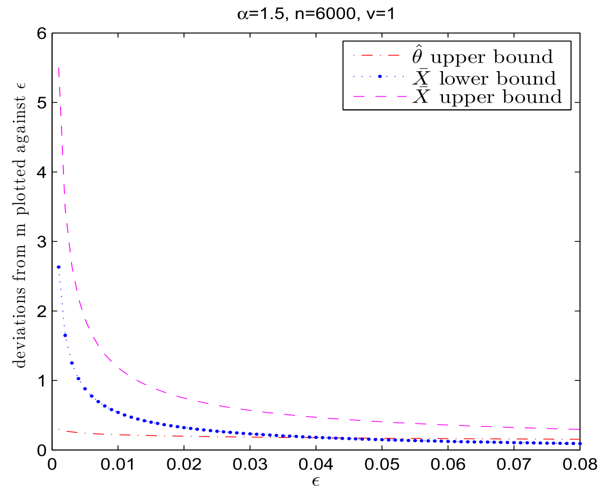

FIG 4. *Deviations of $\hat{\theta}$ from the sample mean, compared with those of empirical mean*

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *with* $(X_i)_{i=1}^n$ *independently drawn from the distribution* $\mathbf{\Pi}$*. Furthermore, if*

$$\epsilon < (3e)^{-1} \quad \textit{and} \quad n \geq 2, \tag{3.2}$$

the inequality

$$\left|\bar{X} - m\right| \geq \left(\frac{v}{3n^{\alpha-1}\epsilon}\right)^{\frac{1}{\alpha}} \left(1 - \frac{3e\epsilon}{n}\right)^{\frac{n-1}{\alpha}}$$

holds with probability at least 2ϵ*.*

Proof. Let us consider the random variable X, which has the following distribution:

$$\mathbb{P}(X = 0) = 1 - \frac{v}{n^\alpha \eta^\alpha}, \quad \mathbb{P}(X = n\eta) = \mathbb{P}(X = -n\eta) = \frac{v}{3n^\alpha \eta^\alpha}$$

and

$$\begin{aligned} \mathbb{P}\left(X \in (x, \infty)\backslash\{n\eta\}\right) &= \frac{q}{2\gamma} x^{-\gamma}, \quad x \in (p, \infty)\backslash\{n\eta\} \\ \mathbb{P}\left(X \in (-\infty, x)\backslash\{-n\eta\}\right) &= \frac{q}{2\gamma} |x|^{-\gamma}, \quad x \in (-\infty, -p)\backslash\{-n\eta\}, \end{aligned}$$

where $\gamma \in (\alpha, 2)$, $p = \left(\frac{\gamma-\alpha}{\gamma}\right)^{\frac{1}{\alpha}} n\eta$ and $q = \frac{\gamma v}{3}\left(\frac{\gamma-\alpha}{\gamma}\right)^{\frac{\gamma}{\alpha}} (n\eta)^{\gamma-\alpha}$. It is easy to check that $\mathbb{E}X = 0$ and

$$\mathbb{E}|X|^\alpha = (n\eta)^\alpha \frac{v}{3n^\alpha \eta^\alpha} + (n\eta)^\alpha \frac{v}{3n^\alpha \eta^\alpha} + \frac{q}{\gamma - \alpha} p^{\alpha-\gamma} = \frac{v}{3} + \frac{v}{3} + \frac{v}{3} = v.$$

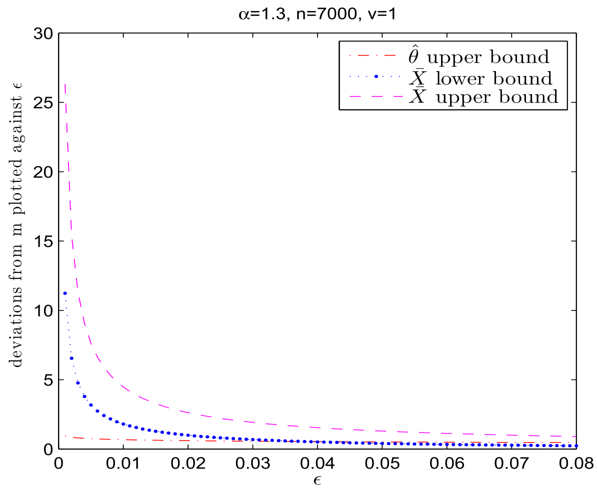

FIG 5. *Deviations of $\hat{\theta}$ from the sample mean, compared with those of empirical mean*

Let $(X_i)_{i=1}^n$ be i.i.d., which have the same distribution as X. Then,

$$\mathbb{P}\left(\bar{X} \geq \eta\right) = \mathbb{P}\left(\bar{X} \leq -\eta\right) \geq \mathbb{P}\left(\bar{X} = \eta\right) \geq \frac{v}{3n^{\alpha-1}\eta^\alpha}\left(1 - \frac{v}{n^\alpha\eta^\alpha}\right)^{n-1},$$

so (3.1) is proved.

Taking $\eta = \left(\frac{v}{3n^{\alpha-1}\epsilon}\right)^{\frac{1}{\alpha}}\left(1 - \frac{3e\epsilon}{n}\right)^{\frac{n-1}{\alpha}}$, we have

$$\frac{v}{3n^{\alpha-1}\eta^\alpha}\left(1 - \frac{v}{n^\alpha\eta^\alpha}\right)^{n-1} = \epsilon\left(1 - \frac{3e\epsilon}{n}\right)^{-(n-1)}\left(1 - \frac{3\epsilon}{n\left(1 - \frac{3e\epsilon}{n}\right)^{n-1}}\right)^{n-1}.$$

If $\epsilon < (3e)^{-1}$, then $\left(1 - \frac{3e\epsilon}{x}\right)^{x-1} \geq \left(1 - \frac{1}{x}\right)^{x-1}$. For any $x \geq 1$, we denote $f(x) = \left(1 - \frac{1}{x}\right)^{x-1}$, then

$$\begin{aligned} f'(x) &= \left(1 - \frac{1}{x}\right)^{x-1}\left(\log\left(1 - \frac{1}{x}\right) + \frac{(x-1)}{x^2\left(1 - \frac{1}{x}\right)}\right) \\ &= \left(1 - \frac{1}{x}\right)^{x-1}\left(\log\left(1 - \frac{1}{x}\right) + \frac{1}{x}\right). \end{aligned}$$

Noting that $\left(1 - \frac{1}{x}\right)^{x-1} > 0$, let $g(x) = \log\left(1 - \frac{1}{x}\right) + \frac{1}{x}$ for $x \geq 1$, then we have

$$g'(x) = \frac{1}{x^2\left(1 - \frac{1}{x}\right)} - \frac{1}{x^2} = \frac{1}{x^2(x-1)} > 0$$

and

$$\lim_{x\to\infty} g(x) = \lim_{x\to\infty}\left[\log\left(1 - \frac{1}{x}\right) + \frac{1}{x}\right] = 0,$$

which imply $g(x) \leq 0$, so we have $f'(x) \leq 0$ for $x \geq 1$. Moreover, we have

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \left(1 - \frac{1}{x}\right)^{x-1} = e^{-1},$$

which implies $\left(1 - \frac{3e\epsilon}{n}\right)^{n-1} \geq e^{-1}$. Therefore, we have

$$\frac{v}{3n^{\alpha-1}\eta^{\alpha}} \left(1 - \frac{v}{n^{\alpha}\eta^{\alpha}}\right)^{n-1} \geq \epsilon \left(1 - \frac{3e\epsilon}{n}\right)^{-(n-1)} \left(1 - \frac{3e\epsilon}{n}\right)^{n-1} = \epsilon.$$

The proof is complete. □

4. ℓ_1-regression for heavy-tailed samples having finite α-th moment with $\alpha \in (1, 2)$

The linear regression considered in [19] aims to find the unknown minimizer θ^* of the following minimization problem:

$$\min_{\theta\in\mathbf{\Theta}} R_{\ell_1}(\theta) \quad \text{with} \quad R_{\ell_1}(\theta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbf{\Pi}} \left[\left|\mathbf{x}^T\theta - y\right|\right], \tag{4.1}$$

where $\mathbf{\Pi}$ is the population's distribution, and $\mathbf{\Theta} \subseteq \mathbb{R}^d$ is the set in which θ^* is located. In practice, $\mathbf{\Pi}$ is not known, one usually draws a data set $\mathcal{T} = (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ from $\mathbf{\Pi}$ and consider the following empirical optimization problem:

$$\min_{\theta\in\mathbf{\Theta}} \widehat{R}_{\ell_1}(\theta) \quad \text{with} \quad \widehat{R}_{\ell_1}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \left|\mathbf{x}_i^T\theta - y_i\right|.$$

Inspired by Catoni's work, Zhang et al. [19] considered the case that $\mathbf{\Pi}$ is heavy tailed with finite variance and proposed a new minimization problem

$$\min_{\theta\in\mathbf{\Theta}} \widehat{R}_{\varphi,\ell_1}(\theta) \quad \text{with} \quad \widehat{R}_{\varphi,\ell_1}(\theta) = \frac{1}{n\beta}\sum_{i=1}^{n} \varphi\left(\beta\left|y_i - \mathbf{x}_i^T\theta\right|\right), \tag{4.2}$$

where φ is the same as that in [3] and $\beta > 0$ is to be determined later.

Thanks to the analysis of Section 2, we extend the results in [19] to the case in which samples can have finite α-th moment with $\alpha \in (1, 2)$, the approach is by replacing the original φ with (2.1).

4.1. Main results of this section

Before stating the main results, we first give some definitions and assumptions.

Definition 4.1. *Let $(\mathbf{\Theta}, d)$ be a metric space, and $\mathbf{K}$ be a subset of $\mathbf{\Theta}$. Then a subset $\mathcal{N} \subseteq \mathbf{K}$ is called an δ-net of $\mathbf{K}$ if for every $\theta \in \mathbf{K}$, we can find a $\tilde{\theta} \in \mathcal{N}$ such that $d\left(\theta, \tilde{\theta}\right) \leq \delta$. The covering number is the minimal cardinality of the δ-net of $\mathbf{\Theta}$ and denoted by $N(\mathbf{\Theta}, \delta)$.*

We shall assume:

Assumption A1 (i) The domain $\mathbf{\Theta}$ is totally bounded, that is, for any $\delta > 0$, there exists a finite δ-net of $\mathbf{\Theta}$.

(ii) The expectation of the α-th moment of $\mathbf{x}$ is bounded, that is,

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathbf{\Pi}}\left[|\mathbf{x}|^{\alpha}\right] < \infty.$$

(iii) The ℓ_α-risk of all $\theta \in \mathbf{\Theta}$ is bounded, that is,

$$\sup_{\theta\in\mathbf{\Theta}} R_{\ell_\alpha}(\theta) = \sup_{\theta\in\mathbf{\Theta}} \mathbb{E}_{(\mathbf{x},y)\sim\mathbf{\Pi}}\left[|y - \mathbf{x}^T\theta|^{\alpha}\right] < \infty.$$

Then, we can state the second theorem, which will be proved in subsection 4.2.

Theorem 4.1. *Let θ^* and $\hat{\theta}$ be the minimizers of (4.1) and (4.2), respectively. Under **Assumption A1**, for any $\delta > 0$, for any $\epsilon \in \left(0, \frac{1}{2}\right)$, with probability at least $1 - 2\epsilon$, we have*

$$\begin{aligned} & R_{\ell_1}\left(\hat{\theta}\right) - R_{\ell_1}\left(\theta^*\right) \\ \le & 2\delta\mathbb{E}|\mathbf{x}_1| + \left(\frac{2^{\alpha-1}\delta^{\alpha}}{\alpha}\mathbb{E}|\mathbf{x}_1|^{\alpha} + \frac{2^{\alpha-1}+1}{\alpha}\sup_{\theta\in\mathbf{\Theta}} R_{\ell_\alpha}(\theta)\right)\beta^{\alpha-1} + \frac{1}{n\beta}\log\frac{N\left(\mathbf{\Theta},\delta\right)}{\epsilon^2}. \end{aligned}$$

Furthermore, let

$$\beta = \left(\frac{1}{n}\log\frac{N\left(\mathbf{\Theta},\delta\right)}{\epsilon^2}\right)^{\frac{1}{\alpha}},$$

we have

$$\begin{aligned} & R_{\ell_1}\left(\hat{\theta}\right) - R_{\ell_1}\left(\theta^*\right) \\ \le & \left(\frac{2^{\alpha-1}\delta^{\alpha}}{\alpha}\mathbb{E}|\mathbf{x}_1|^{\alpha} + \frac{2^{\alpha-1}+1}{\alpha}\sup_{\theta\in\mathbf{\Theta}} R_{\ell_\alpha}(\theta) + 1\right)\left(\frac{1}{n}\log\frac{N\left(\mathbf{\Theta},\delta\right)}{\epsilon^2}\right)^{\frac{\alpha-1}{\alpha}} \\ & + 2\delta\mathbb{E}|\mathbf{x}_1|. \end{aligned} \tag{4.3}$$

In order to compute the covering number, we further assume:

Assumption A2 The domain $\mathbf{\Theta} \subseteq \mathbb{R}^d$, and its radius is bounded by a constant r, that is,

$$|\theta| \le r, \quad \forall\theta \in \mathbf{\Theta}.$$

Then, we have the following corollary, which will be proved in subsection 4.2.

Corollary 4.1. *Keep the same notation and assumptions in Theorem 4.1. In addition, we suppose the **Assumption A2** holds. Then, for any $\epsilon \in \left(0, \frac{1}{2}\right)$, with probability at least $1 - 2\epsilon$, we have*

$$R_{\ell_1}\left(\hat{\theta}\right) - R_{\ell_1}\left(\theta^*\right)$$

$$
\begin{aligned}
&\leq \left(\frac{2^{\alpha-1}}{\alpha n^{\alpha}}\mathbb{E}|\mathbf{x}_1|^{\alpha} + \frac{2^{\alpha-1}+1}{\alpha}\sup_{\theta\in\boldsymbol{\Theta}} R_{\ell_\alpha}(\theta) + 1\right)\left(\frac{1}{n}\left(d\log(6nr) + \log\frac{1}{\epsilon^2}\right)\right)^{\frac{\alpha-1}{\alpha}} \\
&\quad + \frac{2}{n}\mathbb{E}|\mathbf{x}_1| \\
&= O\left(\left(\frac{d\log n}{n}\right)^{\frac{\alpha-1}{\alpha}}\right).
\end{aligned}
$$

4.2. Proof of Theorem 4.1 and Corollary 4.1

Before proving the Theorem 4.1, we first give the following auxiliary lemmas.

Lemma 4.1. *Keep the same notation and assumptions as in Theorem 4.1. Then, for any $\epsilon \in (0,1)$, the following inequality*

$$
\widehat{R}_{\varphi,\ell_1}(\theta^*) - R_{\ell_1}(\theta^*) \leq \frac{\beta^{\alpha-1}}{\alpha} R_{\ell_\alpha}(\theta^*) + \frac{1}{n\beta}\log\frac{1}{\epsilon}
$$

holds with probability at least $1-\epsilon$.

Proof. Noticing that $(\mathbf{x}_i, y_i)$, $i = 1, \cdots, n$, are i.i.d., by (2.1), we have

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(n\beta \widehat{R}_{\varphi,\ell_1}(\theta^*)\right)\right] &= \mathbb{E}\left[\exp\left(\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\theta^*\right|\right)\right)\right] \\
&= \left[\mathbb{E}\left[\exp\left(\varphi\left(\beta\left|y_1 - \mathbf{x}_1^T\theta^*\right|\right)\right)\right]\right]^n \\
&\leq \left[\mathbb{E}\left[1 + \beta\left|y_1 - \mathbf{x}_1^T\theta^*\right| + \frac{\beta^{\alpha}\left|y_1 - \mathbf{x}_1^T\theta^*\right|^{\alpha}}{\alpha}\right]\right]^n,
\end{aligned}
$$

then, by the inequality $1 + x \leq e^x$ for all $x \in \mathbb{R}$, we have

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(n\beta \widehat{R}_{\varphi,\ell_1}(\theta^*)\right)\right] &\leq \left[1 + \beta R_{\ell_1}(\theta^*) + \frac{\beta^{\alpha}}{\alpha} R_{\ell_\alpha}(\theta^*)\right]^n \\
&\leq \exp\left(n\beta R_{\ell_1}(\theta^*) + \frac{n\beta^{\alpha}}{\alpha} R_{\ell_\alpha}(\theta^*)\right).
\end{aligned}
$$

Therefore, by Markov inequality, we have

$$
\begin{aligned}
&\mathbb{P}\left(n\beta \widehat{R}_{\varphi,\ell_1}(\theta^*) \geq n\beta R_{\ell_1}\left(\hat{\theta}\right) + \frac{n\beta^{\alpha}}{\alpha} R_{\ell_\alpha}(\theta^*) + \log\frac{1}{\epsilon}\right) \\
=&\mathbb{P}\left(\exp\left(n\beta \widehat{R}_{\varphi,\ell_1}(\theta^*)\right) \geq \exp\left(n\beta R_{\ell_1}(\theta^*) + \frac{n\beta^{\alpha}}{\alpha} R_{\ell_\alpha}(\theta^*) + \log\frac{1}{\epsilon}\right)\right) \\
\leq&\frac{\mathbb{E}\left[\exp\left(n\beta \widehat{R}_{\varphi,\ell_1}(\theta^*)\right)\right]}{\exp\left(n\beta R_{\ell_1}(\theta^*) + \frac{n\beta^{\alpha}}{\alpha} R_{\ell_\alpha}(\theta^*) + \log\frac{1}{\epsilon}\right)} \leq \epsilon.
\end{aligned}
$$

The proof is complete. ∎

Lemma 4.2. *For any* $\delta > 0$*, let* $\mathcal{N}(\mathbf{\Theta}, \delta)$ *be an* δ*-net of* $\mathbf{\Theta}$ *with cardinality* $N(\mathbf{\Theta}, \delta)$*. Then, for any* $\epsilon \in (0,1)$*, with probability at least* $1-\epsilon$*, the following inequality*

$$\begin{aligned}
& -\frac{1}{n\beta}\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\tilde{\theta}\right| - \beta\delta|\mathbf{x}_i|\right) \\
\leq\ & -R_{\ell_1}\left(\tilde{\theta}\right) + \delta\mathbb{E}|\mathbf{x}_1| + \frac{(2\beta)^{\alpha-1}}{\alpha}\sup_{\theta\in\mathbf{\Theta}} R_{\ell_\alpha}(\theta) + \frac{(2\beta)^{\alpha-1}\delta^\alpha}{\alpha}\mathbb{E}\left|\mathbf{x}_1\right|^\alpha \\
& + \frac{1}{n\beta}\log\frac{N(\mathbf{\Theta},\delta)}{\epsilon}
\end{aligned}$$

holds for all $\tilde{\theta} \in \mathcal{N}(\mathbf{\Theta}, \delta)$*.*

Proof. For a fixed $\tilde{\theta} \in \mathcal{N}(\mathbf{\Theta}, \delta)$, noticing that $(\mathbf{x}_i, y_i)$, $i = 1, \cdots, n$, are i.i.d., by (2.1), we have

$$\begin{aligned}
& \mathbb{E}\left[\exp\left(-\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\tilde{\theta}\right| - \beta\delta|\mathbf{x}_i|\right)\right)\right] \\
=\ & \left[\mathbb{E}\left[\exp\left(-\varphi\left(\beta\left|y_1 - \mathbf{x}_1^T\tilde{\theta}\right| - \beta\delta|\mathbf{x}_1|\right)\right)\right]\right]^n \\
\leq\ & \left[\mathbb{E}\left[1 - \beta\left|y_1 - \mathbf{x}_1^T\tilde{\theta}\right| + \beta\delta|\mathbf{x}_1| + \frac{\beta^\alpha\left|\left|y_1 - \mathbf{x}_1^T\tilde{\theta}\right| - \delta|\mathbf{x}_1|\right|^\alpha}{\alpha}\right]\right]^n \\
=\ & \left[1 - \beta R_{\ell_1}\left(\tilde{\theta}\right) + \beta\delta\mathbb{E}|\mathbf{x}_1| + \frac{\beta^\alpha}{\alpha}\mathbb{E}\left[\left|\left|y_1 - \mathbf{x}_1^T\tilde{\theta}\right| - \delta|\mathbf{x}_1|\right|^\alpha\right]\right]^n,
\end{aligned}$$

then, by (2.6) with $p = q = 2$, and the inequality $1 + x \leq e^x$ for all $x \in \mathbb{R}$, we have

$$\begin{aligned}
& \mathbb{E}\left[\exp\left(-\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\tilde{\theta}\right| - \beta\delta|\mathbf{x}_i|\right)\right)\right] \\
\leq\ & \left[1 - \beta R_{\ell_1}\left(\tilde{\theta}\right) + \beta\delta\mathbb{E}|\mathbf{x}_1| + \frac{\beta^\alpha 2^{\alpha-1}}{\alpha}R_{\ell_\alpha}\left(\tilde{\theta}\right) + \frac{\beta^\alpha\delta^\alpha 2^{\alpha-1}}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha\right]^n \\
\leq\ & \exp\left[n\left(-\beta R_{\ell_1}\left(\tilde{\theta}\right) + \beta\delta\mathbb{E}|\mathbf{x}_1| + \frac{\beta^\alpha 2^{\alpha-1}}{\alpha}R_{\ell_\alpha}\left(\tilde{\theta}\right) + \frac{\beta^\alpha\delta^\alpha 2^{\alpha-1}}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha\right)\right].
\end{aligned}$$

Therefore, by Markov inequality, we have

$$\begin{aligned}
& \mathbb{P}\left(-\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\tilde{\theta}\right| - \beta\delta|\mathbf{x}_i|\right) \geq \log\frac{1}{\epsilon'} + \right. \\
& \qquad \left. n\left(-\beta R_{\ell_1}\left(\tilde{\theta}\right) + \beta\delta\mathbb{E}|\mathbf{x}_1| + \frac{\beta^\alpha 2^{\alpha-1}}{\alpha}R_{\ell_\alpha}\left(\tilde{\theta}\right) + \frac{\beta^\alpha\delta^\alpha 2^{\alpha-1}}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha\right)\right) \\
\leq\ & \frac{\mathbb{E}\left[\exp\left(-\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\tilde{\theta}\right| - \beta\delta|\mathbf{x}_i|\right)\right)\right]}{\exp\left[n\left(-\beta R_{\ell_1}\left(\tilde{\theta}\right) + \beta\delta\mathbb{E}|\mathbf{x}_1| + \frac{\beta^\alpha 2^{\alpha-1}}{\alpha}R_{\ell_\alpha}\left(\tilde{\theta}\right) + \frac{\beta^\alpha\delta^\alpha 2^{\alpha-1}}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha\right) + \log\frac{1}{\epsilon'}\right]}
\end{aligned}$$

$$\leq \epsilon',$$

where $\epsilon' \in (0,1)$, which will be chosen later. Hence, for a fixed $\tilde{\theta} \in \mathcal{N}(\mathbf{\Theta}, \delta)$, with probability at most ϵ', we have

$$\begin{aligned}
&-\frac{1}{n\beta}\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\tilde{\theta}\right| - \beta\delta|\mathbf{x}_i|\right)\\
\geq & -R_{\ell_1}\left(\tilde{\theta}\right) + \delta\mathbb{E}|\mathbf{x}_1| + \frac{(2\beta)^{\alpha-1}}{\alpha}R_{\ell_\alpha}\left(\tilde{\theta}\right) + \frac{(2\beta)^{\alpha-1}\delta^\alpha}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha + \frac{1}{n\beta}\log\frac{1}{\epsilon'}.
\end{aligned}$$

Therefore, since the set $\mathcal{N}(\mathbf{\Theta}, \delta)$ has $N(\mathbf{\Theta}, \delta)$ elements, we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{\tilde{\theta}\in\mathcal{N}(\mathbf{\Theta},\delta)}\left\{-\frac{1}{n\beta}\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\tilde{\theta}\right| - \beta\delta|\mathbf{x}_i|\right) \leq -R_{\ell_1}\left(\tilde{\theta}\right) + \delta\mathbb{E}|\mathbf{x}_1|\right.\right.\\
&\qquad\left.\left. + \frac{(2\beta)^{\alpha-1}}{\alpha}R_{\ell_\alpha}\left(\tilde{\theta}\right) + \frac{(2\beta)^{\alpha-1}\delta^\alpha}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha + \frac{1}{n\beta}\log\frac{1}{\epsilon'}\right\}\right)\\
\geq & 1 - N(\mathbf{\Theta}, \delta)\,\epsilon'.
\end{aligned}$$

Finally, taking $\epsilon' = \frac{\epsilon}{N(\mathbf{\Theta},\delta)}$, with probability at least $1-\epsilon$, the following inequality

$$\begin{aligned}
&-\frac{1}{n\beta}\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\tilde{\theta}\right| - \beta\delta|\mathbf{x}_i|\right)\\
\leq & -R_{\ell_1}\left(\tilde{\theta}\right) + \delta\mathbb{E}|\mathbf{x}_1| + \frac{(2\beta)^{\alpha-1}}{\alpha}R_{\ell_\alpha}\left(\tilde{\theta}\right)\\
& + \frac{(2\beta)^{\alpha-1}\delta^\alpha}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha + \frac{1}{n\beta}\log\frac{N(\mathbf{\Theta},\delta)}{\epsilon}\\
\leq & -R_{\ell_1}\left(\tilde{\theta}\right) + \delta\mathbb{E}|\mathbf{x}_1| + \frac{(2\beta)^{\alpha-1}}{\alpha}\sup_{\theta\in\mathbf{\Theta}}R_{\ell_\alpha}(\theta)\\
& + \frac{(2\beta)^{\alpha-1}\delta^\alpha}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha + \frac{1}{n\beta}\log\frac{N(\mathbf{\Theta},\delta)}{\epsilon}
\end{aligned}$$

holds for all $\tilde{\theta} \in \mathcal{N}(\mathbf{\Theta}, \delta)$. The proof is complete. □

Based on Lemma 4.2, we have the following lemma.

Lemma 4.3. *Keep the same notation and assumptions as in Theorem 4.1. Then, for any $\delta > 0$, for any $\epsilon \in (0,1)$, the following inequality*

$$\begin{aligned}
&R_{\ell_1}\left(\hat{\theta}\right) - \widehat{R}_{\varphi,\ell_1}\left(\hat{\theta}\right)\\
\leq & 2\delta\mathbb{E}|\mathbf{x}_1| + \frac{(2\beta)^{\alpha-1}}{\alpha}\sup_{\theta\in\mathbf{\Theta}}R_{\ell_\alpha}(\theta) + \frac{(2\beta)^{\alpha-1}\delta^\alpha}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha + \frac{1}{n\beta}\log\frac{N(\mathbf{\Theta},\delta)}{\epsilon}
\end{aligned}$$

holds with probability at least $1-\epsilon$.

Proof. Since $\hat{\theta} \in \boldsymbol{\Theta}$, there exists a $\tilde{\theta} \in \mathcal{N}(\boldsymbol{\Theta}, \delta)$ such that

$$\left|\hat{\theta} - \tilde{\theta}\right| \le \delta,$$

which implies

$$\left|y_i - \mathbf{x}_i^T \hat{\theta}\right| \ge \left|y_i - \mathbf{x}_i^T \tilde{\theta}\right| - \left|\mathbf{x}_i^T (\tilde{\theta} - \hat{\theta})\right| \ge \left|y_i - \mathbf{x}_i^T \tilde{\theta}\right| - \delta|\mathbf{x}_i|. \tag{4.4}$$

Then, since $\varphi(\cdot)$ is non-decreasing, we have

$$\widehat{R}_{\varphi,\ell_1}\left(\hat{\theta}\right) = \frac{1}{n\beta}\sum_{i=1}^{n} \varphi\left(\beta\left|y_i - \mathbf{x}_i^T \hat{\theta}\right|\right) \ge \frac{1}{n\beta}\sum_{i=1}^{n} \varphi\left(\beta\left|y_i - \mathbf{x}_i^T \tilde{\theta}\right| - \beta\delta|\mathbf{x}_i|\right),$$

by Lemma 4.2, with probability at least $1-\epsilon$, we have

$$\begin{aligned}\widehat{R}_{\varphi,\ell_1}\left(\hat{\theta}\right) \ge R_{\ell_1}\left(\tilde{\theta}\right) - \Bigg[&\delta\mathbb{E}|\mathbf{x}_1| + \frac{(2\beta)^{\alpha-1}}{\alpha}\sup_{\theta\in\boldsymbol{\Theta}} R_{\ell_\alpha}(\theta) \\ &+ \frac{(2\beta)^{\alpha-1}\delta^\alpha}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha + \frac{1}{n\beta}\log\frac{N(\boldsymbol{\Theta},\delta)}{\epsilon}\Bigg].\end{aligned}$$

Moreover, by (4.4) and triangle inequality, we have

$$R_{\ell_1}\left(\hat{\theta}\right) - R_{\ell_1}\left(\tilde{\theta}\right) = \mathbb{E}\left[\left|\mathbf{x}_1^T\hat{\theta} - y_1\right| - \left|\mathbf{x}_1^T\tilde{\theta} - y_1\right|\right] \le \mathbb{E}\left[\left|\mathbf{x}_1^T\hat{\theta} - \mathbf{x}_1^T\tilde{\theta}\right|\right] \le \delta\mathbb{E}|\mathbf{x}_1|,$$

which further implies that with probability at least $1-\epsilon$, the inequality

$$\begin{aligned}\widehat{R}_{\varphi,\ell_1}\left(\hat{\theta}\right) \ge R_{\ell_1}\left(\hat{\theta}\right) - \Bigg[&2\delta\mathbb{E}|\mathbf{x}_1| + \frac{(2\beta)^{\alpha-1}}{\alpha}\sup_{\theta\in\boldsymbol{\Theta}} R_{\ell_\alpha}(\theta) \\ &+ \frac{(2\beta)^{\alpha-1}\delta^\alpha}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha + \frac{1}{n\beta}\log\frac{N(\boldsymbol{\Theta},\delta)}{\epsilon}\Bigg]\end{aligned}$$

holds. The proof is complete. □

Now, we can give the proof of Theorem 4.1.

Proof of Theorem 4.1. Recall

$$\widehat{R}_{\varphi,\ell_1}(\theta) = \frac{1}{n\beta}\sum_{i=1}^{n}\varphi\left(\beta\left|y_i - \mathbf{x}_i^T\theta\right|\right),$$

since $\hat{\theta}$ is the minimizer of (4.2), we have

$$\widehat{R}_{\varphi,\ell_1}\left(\hat{\theta}\right) - \widehat{R}_{\varphi,\ell_1}(\theta^*) \le 0,$$

which implies

$$R_{\ell_1}\left(\hat{\theta}\right) - R_{\ell_1}(\theta^*)$$

$$
\begin{aligned}
&= \left(R_{\ell_1}\left(\hat{\theta}\right) - \widehat{R}_{\varphi,\ell_1}\left(\hat{\theta}\right)\right) + \left(\widehat{R}_{\varphi,\ell_1}\left(\hat{\theta}\right) - \widehat{R}_{\varphi,\ell_1}\left(\theta^*\right)\right) + \left(\widehat{R}_{\varphi,\ell_1}\left(\theta^*\right) - R_{\ell_1}\left(\theta^*\right)\right) \\
&\leq \left(R_{\ell_1}\left(\hat{\theta}\right) - \widehat{R}_{\varphi,\ell_1}\left(\hat{\theta}\right)\right) + \left(\widehat{R}_{\varphi,\ell_1}\left(\theta^*\right) - R_{\ell_1}\left(\theta^*\right)\right).
\end{aligned}
$$

By Lemma 4.3 and Lemma 4.1, we immediately obtain the desired result. □

Now we are at the position to give the proof of Corollary 4.1.

Proof of Corollary 4.1. For any $\delta \in (0,1]$, by [14, Corollary 4.2.13] we have

$$
N\left(B_1, \delta\right) \leq \left(1 + \frac{2}{\delta}\right)^d \leq \left(\frac{3}{\delta}\right)^d,
$$

where $B_1 = \left\{x \in \mathbb{R}^d : |x| \leq 1\right\}$. Since $\boldsymbol{\Theta} \subseteq B_r$, we have

$$
\log N\left(\boldsymbol{\Theta}, \delta\right) \leq \log N\left(B_r, \frac{\delta}{2}\right) \leq d \log \frac{6r}{\delta}. \tag{4.5}
$$

Therefore, by (4.3) with $\delta = \frac{1}{n}$, we have

$$
\begin{aligned}
&R_{\ell_1}\left(\hat{\theta}\right) - R_{\ell_1}\left(\theta^*\right) \\
&\leq \left(\frac{2^{\alpha-1}\delta^\alpha}{\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha + \frac{2^{\alpha-1}+1}{\alpha}\sup_{\theta\in\boldsymbol{\Theta}} R_{\ell_\alpha}(\theta) + 1\right)\left(\frac{1}{n}\left(d\log\frac{6r}{\delta} + \log\frac{1}{\epsilon^2}\right)\right)^{\frac{\alpha-1}{\alpha}} \\
&\quad + 2\delta\mathbb{E}|\mathbf{x}_1| \\
&= \left(\frac{2^{\alpha-1}}{\alpha n^\alpha}\mathbb{E}|\mathbf{x}_1|^\alpha + \frac{2^{\alpha-1}+1}{\alpha}\sup_{\theta\in\boldsymbol{\Theta}} R_{\ell_\alpha}(\theta) + 1\right)\left(\frac{1}{n}\left(d\log(6nr) + \log\frac{1}{\epsilon^2}\right)\right)^{\frac{\alpha-1}{\alpha}} \\
&\quad + \frac{2}{n}\mathbb{E}|\mathbf{x}_1|.
\end{aligned}
$$

The proof is complete. □

Acknowledgments

LX is supported in part by NSFC grant (No. 12071499), Macao S.A.R grant FDCT 0090/2019/A2 and University of Macau grant MYRG2018-00133-FST. We are grateful to the referee whose numerous comments and suggestions have helped to greatly improve the presentation of this paper.

References

[1] Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory.* [J]. **59**(11), pp. 7711-7717. MR3124669

[2] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli.* [J]. **4**(3), pp. 329-375. MR1653272

[3] Catoni O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *In Annales de l'IHP Probabilitis et statistiques.* [J]. **48**(4), pp. 1148-1185. MR3052407

[4] Chen, P., Nourdin, I. and Xu, L. (2020). Stein's method for asymmetric α-stable distributions, with application to the stable CLT. *Journal of Theoretical Probability.* [J]. **34**, pp. 1382-1407. MR4289888

[5] Fan, J., Liu, H. and Wang, W. (2018). Large covariance estimation through elliptical factor models. *Annals of statistics.* [J]. **46**(4), pp. 1383-1414. MR3819104

[6] Huber, M. (2018). Robust estimation of the mean with bounded relative standard deviation. *In International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* (pp. 271-284). Springer, Cham. MR4112627

[7] Huber, M. (2019). An optimal (ϵ, δ)-randomized approximation scheme for the mean of random variables with bounded relative variance. *Random Structures and Algorithms.* [J]. **55**(2), pp. 356-370. MR3983786

[8] Jin, X., Li, X. and Lu J. (2020). A kernel bound for non-symmetric stable distribution and its applications. *Journal of Mathematical Analysis and Applications.* [J]. **488**(2), 124063. MR4081550

[9] Koltchinskii, V. (2011). Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008 (Vol. 2033). Springer Science and Business Media. MR2829871

[10] King, M. (2017). *Statistics for Process Control Engineers: A Practical Approach.* John Wiley and Sons.

[11] Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics.* [J]. **19**(5), pp. 1145-1190. MR4017683

[12] Sun, Q., Zhou, W. X. and Fan, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association.* [J]. **115**(529), pp. 254-265. MR4078461

[13] von Bahr, B. and Esseen, C. G. (1965). Inequalities for the r-th Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics.* [J]. **36**(1), pp. 299-303. MR0170407

[14] Vershynin, R. (2018). *High dimensional probability. An introduction with applications in Data Science.* Cambridge University Press. MR3837109

[15] Wei, X. and Minsker, S. (2017). Estimation of the covariance structure of heavy-tailed distributions. *In Advances in Neural Information Processing Systems* (pp. 2859-2868).

[16] Xu, L. (2019). Approximation of stable law in Wasserstein-1 distance by Stein's method. *The Annals of Applied Probability.* [J]. **29**(1), pp. 458-504. MR3910009

[17] Xu, Y., Zhu, S., Yang, S., Zhang, C., Jin, R. and Yang, T. (2020). Learning with non-convex truncated losses by SGD. *In Uncertainty in Artificial Intelligence* (pp. 701-711). PMLR.

[18] Zhang, L., Yang, T. and Jin, R. (2017). Empirical Risk Minimization

for Stochastic Convex Optimization: $O(1/n)$-and $O(1/n^2)$-type of Risk Bounds. *In Conference on Learning Theory* (pp. 1954-1979). PMLR.

[19] Zhang, L. and Zhou, Z. H. (2018). ℓ_1-regression with Heavy-tailed Distributions. *In Advances in Neural Information Processing Systems*, pp. 1076-1086.