Statistica Sinica Preprint No: SS-2023-0329	
Title	Empirical Risk Minimization for Losses without Variance
Manuscript ID	SS-2023-0329
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0329
Complete List of Authors	Guanhua Fang, Ping Li and Gennady Samorodnitsky
Corresponding Authors	Guanhua Fang
E-mails	fanggh@fudan.edu.cn
Notice: Accepted author version.	

Statistica Sinica

Empirical Risk Minimization for Losses without Variance

Guanhua Fang[1], Ping Li[2], Gennady Samorodnitsky[3]

[1] *Fudan University* [2] *VecML Inc.* [3] *Cornell University*

Abstract: This paper considers an empirical risk minimization problem under heavy-tailed settings, where data does not have finite variance, but only has $(1+\varepsilon)$-th moment with $\varepsilon \in (0,1)$. Instead of using an estimation procedure based on truncated observed data, we choose the optimizer by minimizing the risk value. Those risk values can be robustly estimated via using the remarkable Catoni's method. Thanks to the structure of Catoni-type influence functions, we are able to establish excess risk upper bounds via using generalized generic chaining methods. Moreover, we take computational issues into consideration. We especially theoretically investigate two types of optimization methods, robust gradient descent algorithm and empirical risk-based methods. With an extensive numerical study, we find that the optimizer based on empirical risks via Catoni-style estimation indeed shows better performance than other baselines. It indicates that estimation directly based on truncated data may lead to unsatisfactory results.

Key words and phrases: Catoni estimator, Chaining method, Excess risk, Heavy tail, Robust gradient descent

1. Introduction

Modern data usually exhibit heavy-tail phenomena. For example, in the financial market (Bradley and Taqqu, 2003; Ahn et al., 2012), the returns are usually not normally distributed. In telecommunications (Crovella and Taqqu, 1999; Glaz et al., 2001), the data sources may sometimes experience a burst of extreme events. In network analysis, the distributions of indegrees, outdegrees, and sizes of connected components might be heavy-tailed (Meusel et al., 2014). In past decades, theoretical analysis of heavy-tailed data has increasingly become a hot topic in the machine learning field, including multi-armed bandits (Bubeck et al., 2013), reinforcement learning (Zhuang and Sui, 2021), mean-estimation problems (Minsker, 2018; Lugosi and Mendelson, 2019), etc. Among those, the empirical risk minimization theory for heavy-tailed data has not been fully explored yet, especially in the situation when the variance does not exist. In this paper, we provide a corresponding theoretical framework standing on the remarkable estimator introduced by Catoni (2012).

Empirical risk minimization (a small sample of the important works include Vapnik (1991); Van de Geer and van de Geer (2000); Bartlett et al. (2005); Zhang et al. (2017)) is one of the basic and fundamental principles in statistical learning problems. The general setting can be described as follows. Let X be a random variable taking values in a measurable space $\mathcal{X}$ and let $\mathcal{F}$ be a set of functions defined on $\mathcal{X}$. For each fixed function $f \in \mathcal{F}$, we define the risk $m_f = \mathbb{E}[f(X)]$ and let $m^* = \inf_{f\in\mathcal{F}} m_f$ be the optimal risk. Given a set of samples of n random variables $X_1, \ldots, X_n$ which are identically and independently distributed (i.i.d.) as X, one aims at finding a function that leads to the smallest risk. That is, our goal is to find $\hat{f} := \arg\min_{f\in\mathcal{F}} \mathbb{E}[f(X)|X_1, \ldots, X_n]$ which is the best we can do. To this end, the standard method is to use an empirical risk minimizer,

$$f_{ERM} = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i), \tag{1.1}$$

and the risk $m_{ERM} = \mathbb{E}[f_{ERM}(X)|X_1, \ldots, X_n]$, where the expectation is taken with respect to a new sample X and is conditioning on $X_1, \ldots, X_n$, is used to quantify the performance of f_{ERM} (Brownlees et al., 2015).

Generic Example. To be more concrete, we consider a general prediction task. The training data is $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ which are i.i.d.. One wishes to predict the response Y given a new observation Z. A predictor is a function g whose quality is evaluated under a pre-determined loss function ℓ. The risk of predictor g is defined as $\mathbb{E}[\ell(g(Z), Y)]$. Given a class $\mathcal{G}$ of functions g, the empirical risk minimization procedure returns a function that minimizes $\frac{1}{n}\sum_{i=1}^{n} \ell(g(Z_i), Y_i)$ over $\mathcal{G}$. Adopting the notion in this paper, we can treat (Z_i, Y_i) as X_i, $\ell(g(\cdot), \cdot)$ as $f(\cdot)$, and $\mathbb{E}[\ell(g(Z), Y)]$ as

m_f.

In this work, we specifically consider the heavy-tailed setting in the infinite variance scenario, that is, $\mathbb{E}[(f(X))^2]$ may not exist for some $f \in \mathcal{F}$. Instead, we assume a weak moment condition,

$$\mathbb{E}[|f(X)|^{1+\varepsilon}] \leq v$$

for any $f \in \mathcal{F}$, where $\varepsilon \in (0,1)$ and v is a fixed positive constant. (Note that $\varepsilon = 1$ refers to the finite variance case.) Under the current setting, f_{ERM} defined in (1.1) is not reliable and is sensitive to outliers (Lerasle and Oliveira, 2011; Diakonikolas et al., 2020). Moreover, it has been shown that the f_{ERM} is no longer rate-optimal (Catoni, 2012) in the presence of weak moment condition. Therefore, we need to look for a robust estimation method and develop the corresponding learning theory.

Literature Review. In the classical ERM literature, it is commonly assumed that loss function f is bounded (Bartlett and Mendelson, 2006; Yi et al., 2022) or noise terms (input values) are bounded (Liu and Tao, 2014; Zhu and Zhou, 2021). Later, an ERM theory was extended to unbounded cases within sub-Gaussian classes, (Lecué and Mendelson, 2013, 2016). Moreover, progress is made to address ERM theory under heavy-tailed settings where the tails of noise decay much slower than sub-Gaussian rate (Brownlees et al., 2015; Hsu and Sabato, 2016; Roy et al., 2021).

In the line of recent literature in robust empirical risk minimization theory, the main approaches can be divided into two categories. The first category of approach is based on using truncated loss (Zhang and Zhou, 2018; Xu et al., 2020; Chen et al., 2021; Xu et al., 2023). That is, it introduces a (non-linear) truncation function ϕ and aims to find the following optimizer, $\hat{f} = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i))$. For example, $\phi(x)$ may take as

$$\phi(x) = \begin{cases} \frac{1}{M}x^2 & |x| \leq M, \\ |x| & |x| > M \end{cases}$$

in Huber (2011);

$$\phi(x) = \begin{cases} \log(1 + x + x^2/2) & x \geq 0, \\ -\log(1 - x + x^2/2) & x < 0 \end{cases}$$

in Catoni (2012);

$$\phi(x) = \begin{cases} \log(1 + \sum_{k=1}^{m} |x|^k/k!) & x \geq 0, \\ -\log(1 + \sum_{k=1}^{m} |x|^k/k!) & x < 0 \end{cases}$$

in Xu et al. (2020). This type of approach is intuitive and is also computationally friendly since it only requires an extra truncation procedure compared to the classical empirical risk minimization. However, the drawback is that the final estimator $\hat{f}$ is not necessarily the minimizer in terms of the risk values.

The second category of technique is based on robust estimation of loss (Brownlees et al., 2015). That is,

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{\mu}_f, \quad \text{where } \hat{\mu}_f \text{ is a robust estimation of loss } m_f.$$

The advantage of such method is that it indeed finds the best function which minimizes the excess risk based on training samples. However, computation of $\hat{f}$ could be extremely prohibitive when functional space $\mathcal{F}$ is large. In this work, we develop new theories under weak moment condition $0 < \varepsilon < 1$ via adopting the second type of approach. Additionally, we propose feasible computational schemes to overcome the optimization obstacles in statistical learning or deep learning (DL) problems. A short summary of our **two types of contributions** is given in Table 1 and Table 2.

Robust ERM Theory		
	$\varepsilon \geq 1$	$0 < \varepsilon < 1$
Truncation	Zhang and Zhou (2018); Xu et al. (2020)	Chen et al. (2021); Xu et al. (2023)
Risk-based	Brownlees et al. (2015)	Ours

Table 1: A brief summary of robust empirical risk minimization theory.

Technical Overview. The theory of bounding excess risk, $m_{\hat{f}} - m^*$, relies on the following points, (1) constructing a good estimator ($\hat{\mu}_f$) of risk m_f and (2) obtaining sharper uniform deviation bounds of $\sup_{f \in \mathcal{F}} |m_f - \hat{\mu}_f|$.

Computational Approach		
Robust Gradient Descent	Algorithm 1	slow as parameter dimension goes large
Empirical Risk-based Methods	Algorithm 2	parameter dimension-free
	Algorithm 3	easy-implementable with DL framework

Table 2: Advantages and disadvantages of three algorithms.

For the first point, we borrow the idea from the remarkable Catoni's estimator (Catoni, 2012). For any $f \in \mathcal{F}$, we consider the following estimator $\hat{\mu}_f$ to approximate $m_f = \mathbb{E}[f(X)]$ such that $\hat{\mu}_f$ is the root of the non-linear equation

$$0 = \hat{r}_f(\mu) = \frac{1}{n\alpha}\sum_{i=1}^{n} \phi(\alpha(f(X_i) - \mu)). \tag{1.2}$$

The influence function ϕ is non-linear and is assumed to satisfy

$$-\log(1 - x + C_\varepsilon |x|^{1+\varepsilon}) \le \phi(x) \le \log(1 + x + C_\varepsilon |x|^{1+\varepsilon}), \tag{1.3}$$

C_ε is a constant depending on ε and α is a tuning parameter which can be dependent on the sample size n. Via using (1.2), we define the new estimator as

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{\mu}_f. \tag{1.4}$$

To understand the excess risk $m_{\hat{f}} - m^*$, it is required to deal with the second point. This is due to the following fact

$$m_{\hat{f}} - m^* = (m_{\hat{f}} - \hat{\mu}_{\hat{f}}) + (\hat{\mu}_{\hat{f}} - m^*) \leq 2 \sup_{f \in \mathcal{F}} |m_f - \hat{\mu}_f|,$$

where we use the fact that $\hat{f}$ and $f^*(:= \arg\min_{f\in\mathcal{F}} m_f)$ belong to $\mathcal{F}$.

For this point, we follow the ideas of Brownlees et al. (2015). The right-hand side of the above inequality depends on developing a new theory of suprema of some empirical process. We summarize the high-level idea here. We first establish a sharp concentration bound of $|m_f - \hat{\mu}_f|$ when f equals f^*. The calculation of such bound is more involved and sophisticated than that in case $\varepsilon = 1$. Next, thanks to the optimality of $\hat{f}$ and observations in Brownlees et al. (2015), we find that it suffices to study term $\frac{1}{\alpha}\mathbb{E}[\phi(\alpha(\hat{f}(X) - \mu))]$ with the choice of $\mu = \mu_0 := m_{f^*} + \epsilon_0$ where ϵ_0 is a small constant. Finally, we find the bounds of $\sup_f |X_f(\mu_0) - X_{f^*}(\mu_0)|$, where

$$X_f(\mu) := \frac{1}{n}\sum_{i=1}^{n}[\frac{1}{\alpha}\phi(\alpha(f(X_i) - \mu)) - \frac{1}{\alpha}\mathbb{E}[\phi(\alpha(f(X_i) - \mu))]],$$

via using generalized Talagrand's chaining method (Talagrand, 1996) to conclude the proof. We here would like to highlight the main technical difference from Brownlees et al. (2015) that we **cannot directly apply** the Hoeffeding inequality / Bernstein inequality or use the classical Talagrand's chaining result since the data does not have the finite variance in our settings. Thanks to the special construction of the influence function, we find that ϕ satisfies the Hölder condition (see later explanation in (4.22)). It helps to simplify the upper bounds of $|X_f(\mu) - X_{f^*}(\mu)|$ which further allows us to establish Bernstein-type concentration inequality. By introducing the generalized $\gamma_{\beta,\varepsilon}$ functional for $\varepsilon \in (0,1)$, we are able to establish the new chaining results.

On the other hand, for the computational feasibility, we propose the empirical risk-based methods, via introducing a novel way for calculating the robust gradient based on nice properties of Catoni-style influence function. Specifically, we treat

$$\nu_i := \frac{\phi'(\alpha(f(X_i) - \hat{\mu}_f))}{\sum_i \phi'(\alpha(f(X_i) - \hat{\mu}_f))}$$

as the weight of i-th sample so that sample i will get lower weight as $f(X_i)$ deviates from current risk estimate $\hat{\mu}_f$. Then weighted average (i.e., $\sum \nu_i \partial f(X_i)$ where ∂f represents the derivative of f with respect to model parameters) is adopted as the new gradient direction. It is then shown that the iterates will converge to the local maximum of $\hat{\mu}_f$. In addition, we further accelerate the algorithm by finding the approximate value of $\hat{\mu}_f$ instead of solving Catoni-style non-linear equation. All the above steps can be easily embedded into any popular deep learning optimizers for the practical use. Readers can find more details in Section 5. In comparison with the robust gradient descent methods (Holland and Ikeda, 2019; Holland and Haress, 2021), we do not require to solve non-linear equations coordinatewisely and it largely increases the computational efficiency. Therefore, our method is parameter dimension-free in the sense that the relative computational time (with respect to vanilla gradient descent) is independent of the number of model parameters. By comparing with other truncated loss based methods (Zhang and Zhou, 2018; Chen et al., 2021), our algorithm returns an estimator which is close to a (local) maximum point, while existing algorithms provide no theoretical results of gaps or relationships between their proposed estimator and the excess risk minimizer.

Notations. We use $\phi(x)$ to represent the Catoni-style influence function. We use n to denote the sample size, and use α as a tuning parameter to let $\delta \in (0,1)$ be a fixed confidence level. Symbol f is some generic loss function, ∇f represents the first-order derivative of function f and $(\nabla f)[j]$ denotes its j-th coordinate. $\mathbb{E}$ and $\mathbb{P}$ are used to denote the generic expectation and probability, respectively. $Y_n = O(X_n)$, $Y_n = o(X_n)$ and $Y_n = \Theta(X_n)$ represents that $Y_n \leq CX_n$, $Y_n/X_n \to 0$, and $Y_n \leq CX_n, X_n \leq CY_n$ for some constant C. For a random variable X and $r \geq 1$ we use the notation $\|X\|_{\phi_r}$ for the Orlicz norm $\|X\|_{\phi_r} = \inf\big\{c > 0 : \, E\exp\big\{|X|^r/c^r\big\} \leq 2$ and $\|\cdot\|$ as the usual l_2-norm.

Organization. The rest of the paper is organized as follows. In Section 2, we provide a background of Catoni's estimator and corresponding technical prerequisites for its generalizations. In Section 3, we describe the main steps of how to derive a tight upper bound of excess risk, $m_{\hat{f}} - m^*$. In Section 4, we present our main theories via using empirical process theory to finalize the bound. In Section 5, we propose a new empirical risk-based gradient descent algorithm and discuss its advantages over existing methods. Multiple experimental results, shown in Section 6, corroborate our new theory and validate the effectiveness of proposed computational approaches. The concluding remarks are given in Section 7. Additional details, simulations, and technical proofs are given in the supplementary.

2. Catoni's Estimator with $\varepsilon \in (0, 1)$

In this section, we present a generalized Catoni's estimator under weak moment condition. To start with, we first recap the classical Catoni's estimator. Considering a sequence of i.i.d random variables $\{X_i\}_{i=1}^n$ be such that $\mathbb{E}(X_1) = \mu$ and $\mathbb{E}|X_1 - \mu|^2 \leq v$, Catoni (Catoni, 2012) introduced a robust mean estimator, $\hat{\mu}$, which is the solution to the non-linear equation, $\sum_{i=1}^n \phi(\alpha(X_i - \mu)) = 0$. Here ϕ is an influence function that is non-decreasing and satisfies $-\log(1 - x + x^2/2) \leq \phi(x) \leq \log(1 + x + x^2/2)$. For general $\varepsilon \in (0, 1)$, using the same idea, we similarly consider $\phi : \mathbb{R} \to \mathbb{R}$ to be a non-decreasing influence function such that for all $x \in \mathbb{R}$

$$-\log(1 - x + C_\varepsilon|x|^{1+\varepsilon}) \leq \phi(x) \leq \log(1 + x + C_\varepsilon|x|^{1+\varepsilon}), \qquad (2.5)$$

where C_ε is a constant depending on the moment order ε. We suggest to choose $C_\varepsilon = \left(\frac{\varepsilon}{1+\varepsilon}\right)^{\frac{1+\varepsilon}{2}} \left(\frac{1-\varepsilon}{\varepsilon}\right)^{\frac{1-\varepsilon}{2}}$. See supplementary for more explanations.

Similarly, we define the generalized Catoni's M-estimator $\widetilde{\mu}_c$ as a solution to the equation

$$\sum_{i=1}^{n} \phi\Big(\alpha(X_i - \mu)\Big) = 0 \qquad (2.6)$$

using an influence function ϕ satisfying (2.5). If the solution is not unique, choose $\widetilde{\mu}_c$ to be the median solution. We here make some requirements on n, α. That is, in the sequel, we always assume

$$C_\varepsilon 2^\varepsilon \alpha^\varepsilon < 1/2; \qquad (2.7)$$

$$2^\varepsilon \alpha^{1+\varepsilon} C_\varepsilon v + \frac{\log(2/\delta)}{n} \leq \frac{\varepsilon}{2(1+\varepsilon)} (\frac{1}{(1+\varepsilon)C_\varepsilon})^{1/\varepsilon}; \qquad (2.8)$$

$$2^\varepsilon C_\varepsilon \alpha^\varepsilon (v+1) + \frac{\log(2/\delta)}{\alpha n} < 1 \qquad (2.9)$$

hold. Note that inequalities (2.7)-(2.9) are (α, δ)-dependent. We therefore call (2.7)-(2.9) as (α, δ)-condition. In fact, the condition is mild since that (7) - (9) are satisfied when n is large and α is small for any fixed δ.

Remark 1. In (2.7)-(2.9), δ is a confidence parameter. In practical applications, δ can be simply taken as 0.05.

Theorem 1. *Let $\{X_i\}_{i=1}^n$ be i.i.d. random variables with mean μ and $\mathbb{E}|X_1 - \mu|^{1+\varepsilon} \le v$. Let $\delta \in (0,1)$ and $\varepsilon \in (0,1)$. Assume that (α, δ)-condition holds, then we have the Catoni's M-estimator $\widetilde{\mu}_c$ satisfies*

$$|\widetilde{\mu}_c - \mu| \le 2(2^\varepsilon C_\varepsilon \alpha^\varepsilon v + \frac{\log(2/\delta)}{\alpha n}). \tag{2.10}$$

with probability $1-\delta$. Especially, we take $\alpha = \left(\frac{\log(2/\delta)}{nC_\varepsilon v}\right)^{\frac{1}{1+\varepsilon}} 2^{\frac{-\varepsilon}{1+\varepsilon}}$, it holds

$$|\widetilde{\mu}_c - \mu| \le 4(C_\varepsilon v)^{\frac{1}{1+\varepsilon}} 2^{\frac{\varepsilon}{1+\varepsilon}} \left(\frac{\log(2/\delta)}{n}\right)^{\frac{\varepsilon}{1+\varepsilon}}. \tag{2.11}$$

Remark 2. By treating $f(X_i)$ as X_i, m_f as μ and $\hat{\mu}_f$ as $\tilde{\mu}_c$ in Theorem 1, we then get the upper bound of $|\hat{\mu}_f - m_f|$ for any fixed $f \in \mathcal{F}$.

Remark 3. Here we would like to point out that generalizing Catoni's estimator to the case of $0 < \varepsilon < 1$ is not a new idea. Estimation bound of $|\tilde{\mu}_c - \mu|$ is also considered in the existing literature (Chen et al., 2021; Bhatt et al., 2022b). Theorem 1 is a special case of Theorem 3.2 in Bhatt et al. (2022b) with a simpler presentation.

3. Bounding $m_{\hat{f}} - m^*$

Given the results in Section 2, we provide the main procedures of how to get the upper bound of $m_{\hat{f}} - m^*$ in this section.

3.1 Finite $\mathcal{F}$

To start with, we first consider $\mathcal{F}$ to be a discrete family as a warm-up. We let $|\mathcal{F}|$ be the cardinality of $\mathcal{F}$, i.e., the number of functions $f \in \mathcal{F}$. First of all, note that

$$\begin{aligned} m_{\hat{f}} - m^* &= (m_{\hat{f}} - \hat{\mu}_{\hat{f}}) + (\hat{\mu}_{\hat{f}} - m^*) \\ &\leq (m_{\hat{f}} - \hat{\mu}_{\hat{f}}) + (\hat{\mu}_{f^*} - m^*) \qquad (3.12) \\ &\leq 2 \sup_{f \in \mathcal{F}} |m_f - \hat{\mu}_f|. \qquad (3.13) \end{aligned}$$

In the above, (3.12) uses the definition that $\hat{f}$ is the minimizer of $\hat{\mu}_f$ and (3.13) uses the fact that $\hat{f}, f^*$ belong to $\mathcal{F}$. Furthermore, recalling that $\hat{\mu}_f$ is the solution to $0 = \frac{1}{n\alpha} \sum_{i=1}^n \phi(\alpha(f(X_i) - \mu))$, we can apply Theorem 1 to $|\hat{\mu}_f - m_f|$ for each $f \in \mathcal{F}$. By union bound, we have the following result.

Theorem 2. *Let $\{X_i\}_{i=1}^n$ be a set of i.i.d. random variables. Assume $(\alpha, \delta/|\mathcal{F}|)$-condition holds, $\sup_{f\in\mathcal{F}} \mathbb{E}|f(X_1)|^{1+\varepsilon} \leq v$ and $\delta \in (0,1)$. We have*

$$m_{\hat{f}} - m^* \leq 4(2^\varepsilon C_\varepsilon \alpha^\varepsilon v + \frac{\log(2|\mathcal{F}|/\delta)}{\alpha n}) \qquad (3.14)$$

with probability at least $1 - \delta$.

We can optimize the bound in the right-hand side of (3.14) by choosing $\alpha = \left(\frac{\log(2|\mathcal{F}|/\delta)}{2^\varepsilon n v C_\varepsilon}\right)^{\frac{1}{1+\varepsilon}}$. Then we have

$$m_{\hat{f}} - m^* \leq 8\left(\frac{\log(\frac{2|\mathcal{F}|}{\delta})}{n}\right)^{\frac{\varepsilon}{1+\varepsilon}} (\frac{1}{2^\varepsilon v C_\varepsilon})^{-\frac{1}{1+\varepsilon}}.$$

3.2 General $\mathcal{F}$

However, in the general case, $\mathcal{F}$ may not be finite. In other words, $|\mathcal{F}|$ is not well-defined. In this section, we seek an alternative approach to bounding $m_{\hat{f}} - m^*$. A few additional useful quantities are given as follows. For convenience, we define

$$\hat{r}_f(\mu) := \frac{1}{n\alpha}\sum_{i=1}^{n}\phi(\alpha(f(X_i)-\mu)) \text{ and } \bar{r}_f(\mu) := \frac{1}{\alpha}\mathbb{E}[\phi(\alpha(f(X)-\mu))].$$

That is, $\bar{r}_f(\mu)$ is the population version of $\hat{r}_f(\mu)$. We further define $X_f(\mu) := \hat{r}_f(\mu) - \bar{r}_f(\mu)$, $\bar{\mu}_f$ to be the solution to $\bar{r}_f(\mu) = 0$ and set

$$A_\alpha(\delta) = 2(2^\varepsilon C_\varepsilon \alpha^\varepsilon v + \frac{\log(2/\delta)}{\alpha n}). \tag{3.15}$$

Throughout the rest of the paper, we assume ϕ to be L_ε-Lipschitz continuous. We additionally require that α satisfies

$$C_\varepsilon \alpha^\varepsilon 2^{2+\varepsilon}(2^\varepsilon C_\varepsilon \alpha^\varepsilon v + 2\eta)^\varepsilon < 1, \tag{3.16}$$

which is called η-*condition*. Here η is a positive number that may be specified from place to place. We typically use η specified in (3.17) below.

Remark 4. Here η-condition is mild. It is not hard to see that the left-hand side of (3.16) is an increasing function of α. Moreover, α is usually taken as $(1/n)^{1/(1+\varepsilon)}$ in applications. Hence, η-condition holds once sample size n is sufficiently large.

By a closer examination of the structure of $\bar{r}_f(\mu)$, we claim the following two lemma.

Lemma 1. *Let $\mu_0 := m_{f^*} + A_\alpha(\delta)$. With probability at least $1-2\delta$, it holds*

$$\bar{r}_{\hat{f}}(\mu_0) \leq 2L_\varepsilon A_\alpha(\delta) + Q(\mu_0, \delta) =: \eta, \tag{3.17}$$

where $Q(\mu, \delta)$ is the $1-\delta$ quantile of $\sup_{f\in\mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)|$, i.e., the minimum possible q satisfying that

$$\mathbb{P}(\sup_{f\in\mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)| \leq q) \geq 1-\delta.$$

Lemma 2. *If η-condition holds with η being given in (3.17), then with probability at least $1-2\delta$, it holds*

$$m_{\hat{f}} \leq \mu_0 + 2^{1+\varepsilon} C_\varepsilon \alpha^\varepsilon v + 2\eta. \tag{3.18}$$

By Lemma 1 - 2 and suitable re-arrangements, we then have

$$\begin{aligned} m_{\hat{f}} - m_{f^*} &\leq 2^{1+\varepsilon} C_\varepsilon \alpha^\varepsilon v + A_\alpha(\delta) + 4L_\varepsilon A_\alpha(\delta) + 2Q(\mu_0,\delta) \\ &\leq 6L_\varepsilon(2^\varepsilon C_\varepsilon \alpha^\varepsilon v + \frac{\log(2/\delta)}{\alpha n}) + 2Q(\mu_0,\delta). \end{aligned} \tag{3.19}$$

In the next section, we focus on working with $Q(\mu_0,\delta)$. The final bounds are given as in (4.23) and (4.25). By the careful choice of α's order as $n^{-1/(1+\varepsilon)}$, the excess risk is $O((\frac{1}{n})^{\varepsilon/(1+\varepsilon)})$.

4. Working with $Q(\mu_0,\delta)$

In this section, our goal is to find the upper bound of $Q(\mu_0,\delta)$, which requires to compute tail probability $\mathbb{P}(\sup_{f\in\mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)| > t)$. It relies on the generic chaining technique (Talagrand, 1996). To introduce the main results, we need to describe the geometric complexity of class $\mathcal{F}$ under different distances. In particular, the L_p distance is defined as $d_p(f, f') = (\mathbb{E}[|f(X) - f'(X)|^p])^{1/p}$, the L_∞ distance is given as $D(f, f') = \text{ess-sup}_{x\in\mathcal{X}}|f(x)-f'(x)|$, and the L_p-distance of the sample version is $d_{X,p}(f, f') = (\frac{1}{n}\sum_{i=1}^{n}|f(X_i) - f'(X_i)|^p)^{1/p}$. For any metric space T equipped with a distance d, we define $\text{diam}_d(T)$ to be the diameter of space T under metric d. We additionally define the quantity $\gamma_{\beta,\varepsilon}(T, d)$ $(\beta = 1, 2)$ as

$$\gamma_{\beta,\varepsilon}(T, d) = \inf_{\mathcal{A}_n}\sup_{t\in T}\sum_{n\geq 0} 2^{n/\beta}(\Delta_d(A_n(t)))^{(1+\varepsilon)/2}, \tag{4.20}$$

where $(\mathcal{A}_n)$ is an increasing sequence of partitions of T and is *admissible* if for all $n \geq 0$, $|\mathcal{A}_n| \leq 2^{2^n}$. For any $t \in T$, $A_n(t)$ is the unique element of $\mathcal{A}_n$ that contains t. $\Delta_d(A)$ denotes the diameter of the set $A \subset T$ given metric d. Sometimes, we may also write $\Delta(A)$ instead of $\Delta_d(A)$ when there is no ambiguity. Here functional $\gamma_{\beta,\varepsilon}(T, d)$ is the generalized version of the classical $\gamma_\beta(T, d)$ (when $\varepsilon = 1$) which appears in the literature of generic chaining methods (Talagrand, 1996). Although $\gamma_{\beta,\varepsilon}(T, d)$ is hard to compute based on its definition, an upper bound on $\gamma_{\beta,\varepsilon}(T, d)$ can be obtained in the next lemma.

Lemma 3. *For any $0 < \varepsilon \leq 1$ and β, there exists a constant $C_{\beta,\varepsilon}$ such that*

$$\gamma_{\beta,\varepsilon}(T, d) \leq C_{\beta,\varepsilon}\int_0^\infty \epsilon^{(\varepsilon-1)/2}(\log N(T, d, \epsilon/2))^{1/\beta}d\epsilon, \tag{4.21}$$

where $N(T, d, \epsilon)$ is the ϵ-covering of space (T, d).

Lemma 3 gives us a way to upper bound $\gamma_{\beta,\varepsilon}(T, d)$, whose value is finite once the covering number is integrable in the sense of (4.21).

Remark 5. When the diameter of space T is not greater than 1, then $\gamma_{\beta,\varepsilon}(T, d)$ is increasing as ε decreases. By (4.21), the upper bound of $\gamma_{\beta,\varepsilon}(T, d)$ may differ from that of the classical $\gamma_\beta(T, d)$ up to a constant when the space is VC-class.

4.1 Bound 1 on $Q(\mu, \delta)$

In the rest of paper, we focus on a class of Catoni influence functions that satisfy a *Hölder condition*, that is,

$$|\phi(x_1) - \phi(x_2)| \leq C_{3\varepsilon} \cdot |x_1 - x_2|^{(1+\varepsilon)/2} \text{ for any } x_1, x_2 \in \mathbb{R}. \tag{4.22}$$

The above Hölder requirement is mild that we can easily construct influence functions to satisfy (4.22). See details in the supplementary. We take $A'_\alpha(\delta) = 2(C_\varepsilon \alpha^\varepsilon 2^{2+\varepsilon} v + \frac{\log(2/\delta)}{\alpha n})$ and and have the following result.

Theorem 3. *Let $\{X_i\}_{i=1}^n$ be a set of i.i.d. random variables with $\sup_{f\in\mathcal{F}} \mathbb{E}|f(X_1)|^{(1+\varepsilon)} \leq v$. Let $\delta \in (0, 1)$ and suppose that the (α, δ)-condition and the η-condition with $\eta = 2L_\varepsilon A'_\alpha(\delta) + Q_1(\delta)$ and $Q_1(\delta)$ defined in (4.24) below hold, and assume that the influence function satisfies the Hölder condition (4.22) .*

Then we have

$$m_{\hat{f}} - m^* \le 6L_\varepsilon \left(C_\varepsilon \alpha^\varepsilon 2^{2+\varepsilon} v + \frac{\log(2/\delta)}{\alpha n} \right) + 2Q_1(\delta) \tag{4.23}$$

with probability at least $1 - 2\delta$*. Here*

$$Q_1(\delta) = 384 C_{3\varepsilon} \log 2 \, \log(2/\delta) \left(\frac{2\alpha^{(\varepsilon-1)/2}}{3n} \gamma_{1,\varepsilon}(\mathcal{F}, D) + \sqrt{\frac{\alpha^{\varepsilon-1}}{n}} \gamma_{2,\varepsilon}(\mathcal{F}, d_p) \right) \tag{4.24}$$

with $p = 1 + \varepsilon$*.*

Remark 6. In particular, if we take $\alpha = (\frac{\log(2/\delta)}{nv})^{\frac{1}{1+\varepsilon}}$, then

$$m_{\hat{f}} - m^* = O_p\left((\frac{1}{n})^{\frac{\varepsilon}{1+\varepsilon}} \cdot \left((\log(2/\delta))^{\frac{\varepsilon}{1+\varepsilon}} v^{1/(1+\varepsilon)} + (\frac{\log(2/\delta)}{v})^{\frac{\varepsilon-1}{2(1+\varepsilon)}} \gamma_{2,\varepsilon}(\mathcal{F}, d_p) \right) \right).$$

Here the term $\gamma_{1,\varepsilon}(\mathcal{F}, D)$ disappears since it contains an n-dependent term of a smaller order.

4.2 Bound 2 on $Q(\mu, \delta)$

Theorem 3 can be used when the sup-norm D is finite. In this section, we establish a second upper bound on $Q(\mu, \delta)$, which does not rely on D. Instead, we use the sample L_p-distance ($p = 1 + \varepsilon$) and define

$$\Gamma_\delta := \min\{c : \mathbb{P}(\gamma_{2,\varepsilon}(\mathcal{F}, d_{X,p}) > c) \le \frac{\delta}{8}\}$$

to measure the span of space $(\mathcal{F}, d_{X,p})$.

Theorem 4. *Let* $\{X_i\}_{i=1}^n$ *be a set of i.i.d. random variables with* $\sup_{f \in \mathcal{F}} \mathbb{E}|f(X_1)|^{1+\varepsilon} \le v$*. Let* $\delta \in (0,1)$ *and suppose that the* (α, δ)*-condition and the* η*-condition* *with* $\eta = 2L_\varepsilon A'_\alpha(\delta) + Q_2(\delta)$ *and* $Q_2(\delta)$ *defined in* (4.26) *below, hold and assume that the influence function satisfies the Hölder condition* (4.22) *. Then, with probability at least* $1 - 2\delta$*, we have*

$$m_{\hat{f}} - m^* \leq 6L_\varepsilon \left(C_\varepsilon \alpha^\varepsilon 2^{2+\varepsilon} v + \frac{\log(2/\delta)}{\alpha n} \right) + 2Q_2(\delta) \quad (4.25)$$

$$\text{with} \qquad Q_2(\delta) = K \max\{\Gamma_\delta, (diam_{d_p}(\mathcal{F}))^{(1+\varepsilon)/2}\} \sqrt{\frac{\log(8/\delta)}{n\alpha^{1-\varepsilon}}} \quad (4.26)$$

for a constant K *which may depend on* ε*. Here* $p = 1 + \varepsilon$ *in* $diam_{d_p}$*.*

5. Computations

In this section, we discuss the computational issues of finding the optimizer $\hat{f} \in \mathcal{F}$. In the sequel, to facilitate our life, we only consider the case that $\mathcal{F}$ is a parametric family. That is, $\mathcal{F} = \{f_w : w \in \mathbb{R}^d\}$, d is the dimension of parameter vector w. Since there are infinitely many candidates for w, it is not desired to directly solve $\hat{\mu}_{f_w}$ from (1.4) for all $f_w \in \mathcal{F}$ to get $f_{\hat{w}}$. Therefore, we want to design an algorithm that returns $f_{\tilde{w}}$ such that $m_{f_{\tilde{w}}}$ shares similar properties as $m_{f_{\hat{w}}}$.

We know that $w^* = \arg\min_w m_{f_w} = \arg\min_w \mathbb{E}[f_w(X)]$. To solve w^*, it is popular to use gradient descent methods (Chong and Zak, 2004; Lemaréchal, 2012; Ruder, 2016) once we know the explicit formula of $\mathbb{E}[f_w(X)]$. The gradient of $\mathbb{E}[f_w(X)]$ is $\nabla\mathbb{E}[f_w(X)]$ which is equal to $\mathbb{E}[\nabla f_w(X)]$ provided that $|\nabla f_w(X)|$ is integrable. In the rest, we discuss two types of tractable methods for computing $\hat{f}$.

5.1 On Robust Gradient Descent Method

We first present a computational method via using robust gradients. The core idea of such a method is to estimate $\mathbb{E}[\nabla f_w(X)]$ via using robust techniques. This type of approach is interesting since robust gradient can be embedded into any machine learning optimizer. A similar type of algorithm is also considered in the literature (Holland and Ikeda, 2019; Holland and Haress, 2021) when the moment order $\varepsilon \geq 1$. The procedure is described in Algorithm 1.

To be self-contained, we also provide the complete theoretical analysis of Algorithm 1. The assumptions on function $f_w \in \mathcal{F}$ are given as below.

A0. It holds that $\mathbb{E}[|\nabla f_w(X)[j]|^{1+\varepsilon}] \leq v$ for any $f_w \in \mathcal{F}$ and $j \in [d]$. Parameter w lies in a bounded space with D_w being its diameter.

A1. For any $B > 0$, there exists a constant R_B such that it holds $|\nabla f_{w_1}(X)[j] - \nabla f_{w_2}(X)[j]| \leq R_B \|w_1 - w_2\|$ for any $\|X\| \leq B$, $j \in [d]$.

A2. Let $F(w)$ be $\mathbb{E}[f_w(X)]$ and $\nabla F(w)$ be $\mathbb{E}[\nabla f_w(X)]$. It is assumed that $F(w)$ and $\nabla F(w)$ are L_f-lipschitz continuous.

Assumption A0 ensures that each coordinate of the gradient satisfies the weaker moment condition and that the parameter space is bounded.

Assumption A1 here is mild and is weaker than the bounded Lipschitz condition, that is, $|\nabla f_{w_1}(X) - \nabla f_{w_2}(X)| \leq R\|w_1 - w_2\|$ for any X. Assumption A2 is a smoothness condition on loss function in the population version.

Remark 7. Consider the linear regression problem that $f_w(X) = (Y - w^T Z)^2$ with $X := (Z, Y)$. Thus it holds $\nabla f_w(X)[j] = (w^T Z - Y)Z[j]$ for $j \in [d]$. It is easy to see that $|\nabla f_{w_1}(X)[j] - \nabla f_{w_2}(X)[j]| \leq B^2\|w_1 - w_2\|$ whenever $\|X\| \leq B$.

Remark 8. For any $\eta \in [0, 1]$, we define $B_\eta := \inf_B\{\mathbb{P}(\|X\| \geq B)\} \leq \eta$. By Assumption A1, we then know $|\nabla f_{w_1}(X) - \nabla f_{w_2}(X)| \leq R_{B_\eta}\|w_1 - w_2\|$ holds with probability at least $1 - \eta$.

To provide convergent properties of Algorithm 1, we consider to choose a twice-differentiable influence function ϕ satisfying

$$\phi'(x) \geq \frac{1}{2}, \text{for any } |x| \leq x_c \text{ for some positive constant } x_c \tag{5.28}$$

$$\phi'(x) \equiv 0, \text{when } |x| \geq x_{cut} \text{ for some positive constant } x_{cut}. \tag{5.29}$$

Here (5.28) automatically holds for twice continuously-differentiable ϕ's and (5.29) guarantees ϕ to be bounded. For tuning parameter α, we can choose it to be n^{-c_0} with $0 < c_0 < 1$ so that it holds

$$\alpha(R_{B_{1/2}} + 1)D_w \leq x_c, \tag{5.30}$$

$$384\eta/\alpha < \bar{A}_\alpha(\delta), \tag{5.31}$$

where $\bar{A}_\alpha(\delta) := 2A_\alpha(\frac{\delta}{d(D_w(R_{B_\eta}+L_f)\alpha/(64A\eta))^d})$. Here $\bar{A}_\alpha(\delta)$ can be viewed as the uniform error for controlling the difference between the empirical gradient and its expectation. In the linear regression as described in Remark 7 with bounded Z and heavy-tailed Y, we can take $\eta = O(1/n)$ and $R_{B_\eta} = O(1)$. We can set $\alpha = n^{-1/(1+\varepsilon)}$ and $\bar{A}_\alpha(\delta)$ becomes $O(n^{-\frac{\varepsilon}{(1+\varepsilon)}}(d\log n + \log(1/\delta))^{\frac{\varepsilon}{(1+\varepsilon)}})$ to make (5.30) - (5.31) held for large n.

Theorem 5. *Under Assumptions A0-A2 and let $\delta \geq \exp\{-n/4\}$, $\eta \geq \frac{\log(1/\delta)}{n}$, and $\gamma_t \equiv \gamma \leq \frac{4}{9L_f}$, we also choose ϕ and α such that* (5.28) - (5.31) *hold. Then Algorithm 1 will return an estimator $\hat{w}$, with $1-3\delta$ probability, it holds that*

$$T_{stop} \leq \frac{18(m_{f_{w^{(0)}}} - m_{f_{w^*}})}{5\gamma d(\bar{A}_\alpha(\delta))^2},$$

where $T_{stop} := \min\{t : \|\nabla F(w^{(t)})\| \leq \sqrt{d}\bar{A}_\alpha(\delta)\}$.

If $F(w)$ is additionally assumed to be κ-strictly convex, then it holds with probability $1-3\delta$ that

$$\|w^{(t+1)} - w^*\| \leq (1 - \sqrt{\frac{2\gamma\kappa L_f}{\kappa + L_f}})^{t+1}\|w^{(0)} - w^*\| + \sqrt{d}\bar{A}_\alpha(\delta)\frac{\gamma}{1 - \sqrt{\frac{2\gamma\kappa L_f}{\kappa+L_f}}}$$

for any $0 \leq t < T_{stop}$.

Remark 9. The proof technique of Theorem 5 is similar to that of Theorem 5 in Holland and Ikeda (2019). However, our case $0 < \varepsilon < 1$ requires more involved computation than case $\varepsilon \geq 1$ does. In addition, results in Holland and Ikeda (2019) are established under stronger conditions, i.e., $\sup_X |\nabla f_{w_1}(X) - \nabla f_{w_2}(X)| \leq R\|w_1 - w_2\|$, which is not generally true when the support of X is unbounded.

5.2 Acceleration: Finding $\hat{f}_w$ via risk minimization

Unfortunately, computing an robust estimate of $\mathbb{E}[\nabla f_w(X)]$ is generally inefficient when dimension d goes extremely large. We need to go around with this issue by seeking other types of approaches.

Our ultimate goal is to find out the minimizer, $\arg\min_w m_{f_w}$. It is necessary to find the solution of $\nabla_w m_{f_w} = \mathbf{0}$. However, the explicit formula of m_{f_w} is unknown to us. We then look for the solution of $\nabla_w \hat{\mu}_{f_w} = \mathbf{0}$ instead. Recall the following identity,

$$0 = \frac{1}{n\alpha} \sum_i^n \phi(\alpha(f_w(X_i) - \hat{\mu}_{f_w})). \tag{5.32}$$

Taking derivative with respect to w on both sides, we have

$$\begin{aligned} \mathbf{0} &= \frac{1}{n\alpha} \frac{\partial \{\sum_i^n \phi(\alpha(f_w(X_i) - \mu_{f_w}))\}}{\partial w} \\ &= \frac{1}{n} \sum_i \phi'(\alpha(f_w(X_i) - \hat{\mu}_{f_w}))\big(\frac{\partial f_w(X_i)}{\partial w} - \nabla_w \hat{\mu}_{f_w}\big). \end{aligned} \tag{5.33}$$

With re-arrangement, we arrive at

$$\nabla_w \hat{\mu}_{f_w} = \frac{\sum_i \phi'(\alpha(f_w(X_i) - \hat{\mu}_{f_w}))\frac{\partial f_w(X_i)}{\partial w}}{\sum_i \phi'(\alpha(f_w(X_i) - \hat{\mu}_{f_w}))}. \tag{5.34}$$

By gradient descent, $w^{(t+1)} = w^{(t)} - \gamma_t \nabla_w \hat{\mu}_{f_w^{(t)}}$, we can find the stationary point $\tilde{w}$ such that $\nabla_w \hat{\mu}_{f_w}|_{w=\tilde{w}} = 0$. Therefore, we have the following proposition.

Proposition 1. *If $f_{\hat{w}}$ is the optimizer as defined in* (1.4)*, then it holds that* $\nabla_w \hat{\mu}_{f_w}|_{w=\hat{w}} = \mathbf{0}$.

For the above reasons, we naturally have the following empirical risk-based gradient descent algorithm, i.e., Algorithm 2. Some remarks are explained here. Since ϕ is non-decreasing, then $\phi'(\cdot)$ is always non-negative. We can view $\frac{\phi'(\alpha(f_{w(t)}(X_i) - \hat{\mu}_{f_w}))}{\sum_i \phi'(\alpha(f_{w(t)}(X_i) - \hat{\mu}_{f_{w(t)}}))}$ as the weight of i-th sample at iteration t. Therefore, larger $f_{w^{(t)}}(X_i)$ has smaller weight, thanks to the construction of ϕ. Compared with the robust gradient descent method (Algorithm 1), we only need to solve the non-linear equation for **one time** in each iteration instead of computing robust gradient coordinate by coordinate. Hence Algorithm 2 is more computationally friendly. In the field of robust statistics, a similar type of algorithm is considered in Mathieu and Minsker (2021).

Note that Algorithm 2 requires to solve the non-linear equation (5.35). To further accelerate the whole estimation procedure, we compute an approximate of $\hat{\mu}_{f_{w^{(t)}}}$ instead of computing it exactly. At $(t+1)$-th step, such approximation is obtained via the following recursive formula

$$\hat{\mu}^{(t+1)} = \hat{\mu}^{(t)} + \sum_i \nu_i^{(t)} (f_{w^{(t+1)}}(X_i) - f_{w^{(t)}}(X_i)), \tag{5.37}$$

where $\hat{\mu}^{(t)}$ is viewed as the proxy of $\hat{\mu}_{f_{w^{(t)}}}$ and $\nu_i^{(t)} := \frac{\phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}_{f_w}))}{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}_{f_{w^{(t)}}}))}$ is the weight of sample i. The full procedure is summarized in Algorithm 3. Note that weights $\{\nu_i^{(t)}\}$'s have been used *twice* for computing gradient $g^{(t)}$ and approximated risk $\hat{\mu}^{(t+1)}$. Therefore, we call this method as **double-weighted** gradient descent algorithm.

We further prove the convergent property of Algorithm 3. That is, for any $\varrho \approx (\sqrt{d}\alpha^{1+\varepsilon})^{1/3}$, we can show that it takes no longer than $O(1/\varrho^2)$ steps to return an estimator whose gradient norm is not larger than ϱ.

Theorem 6. *Under Assumptions A0 - A2, we choose ϕ and α such that* (5.28) - (5.31) *hold and let step size $\gamma_t \equiv \gamma \leq 1/(5L_f)$ in Algorithm 3. For any ϱ satisfying $\varrho \geq \tilde{\varrho} := \left(\tilde{C}\alpha^{(1+\varepsilon)}\sqrt{d}(\log n)\frac{\hat{\mu}_{f_{w^{(0)}}}}{\gamma}\right)^{1/3}$, we define $T_{end}(\varrho) :=$* $\min\{t : \|\nabla_w \hat{\mu}_{f_{w^{(t)}}}\| \leq \varrho\}$. *Then it holds $T_{end}(\rho) \leq \frac{\hat{\mu}_{f_{w^{(0)}}}}{\gamma\varrho^2}$ with probability going to 1 as $n \to \infty$, where $\tilde{C}$ is a constant depending on the second derivative of influence function $\phi(x)$ and ε.*

Remark 10. Especially, we take $\alpha = \Theta(n^{-1/(1+\varepsilon)})$. Then $\tilde{\varrho} = O\left(\alpha^{1+\varepsilon}\sqrt{d}(\log n)\right)^{1/3} = O((\sqrt{d}\log n/n)^{1/3}) = o(1)$ once $d = o(n/\log n)^2$. Hence the algorithm returns a solution close to a local stationary point with high probability.

Due to the space limitations, more discussions on the proposed algorithms can be found in the supplementary I.

6. Numerical Experiments

In this section, we provide multiple simulation results to show the usefulness and superiority of our proposed empirical risk-based method. Specifically, we compare the following six algorithms. **ERM-wide:** Algorithm 2 with choice of $\phi(x) = \phi_{wide}(x)$; **ERM-narrow:** Algorithm 2 with choice of $\phi(x) = \phi_{narrow}(x)$; **Grad-wide:** Algorithm 1 with choice of $\phi(x) = \phi_{wide}(x)$; **Grad-narrow:** Algorithm 1 with choice of $\phi(x) = \phi_{narrow}(x)$; **Mean:** Algorithm 1 via replacing step 5 (Robust gradient) by computing $g^{(t)} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_{w^{(t)}}(X_i)$; **Trim:** Algorithm 1 via replacing step 5 (Robust gradient) by computing

$$g^{(t)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Trunc}(\nabla f_{w^{(t)}}(X_i), B),$$

where $\mathrm{Trunc}(X, B) = X\mathbf{1}\{|X| \leq B\}$. Here

$$\phi_{wide}(x) := \begin{cases} \log(1 + x + C_\varepsilon |x|^{1+\varepsilon}) & x \geq 0 \\ -\log(1 - x + C_\varepsilon |x|^{1+\varepsilon}) & x < 0. \end{cases}$$

and

$$\phi_{narrow}(x) := \begin{cases} \log\left(1 - \frac{\varepsilon}{1+\varepsilon}((1+\varepsilon)C_\varepsilon)^{-\frac{1}{\varepsilon}}\right) & \text{if } x \leq -((1+\varepsilon)C_\varepsilon)^{-\frac{1}{\varepsilon}}, \\ \log(1 + x + C_\varepsilon|x|^{1+\varepsilon}) & \text{if } -((1+\varepsilon)C_\varepsilon)^{-\frac{1}{\varepsilon}} \leq x \leq 0, \\ -\log(1 - x + C_\varepsilon|x|^{1+\varepsilon}) & \text{if } 0 < x \leq ((1+\varepsilon)C_\varepsilon)^{-\frac{1}{\varepsilon}}, \\ -\log\left(1 - \frac{\varepsilon}{1+\varepsilon}((1+\varepsilon)C_\varepsilon)^{-\frac{1}{\varepsilon}}\right) & \text{if } x \geq ((1+\varepsilon)C_\varepsilon)^{-\frac{1}{\varepsilon}}. \end{cases}$$

$\phi_{wide}(x)$ and $\phi_{narrow}(x)$ are the widest and narrowest influence functions satisfying (2.5). Please refer to Bhatt et al. (2022a) for more detailed explanations of "widest" and "narrowest". The choices of tuning parameters in Sections 6.1 - 6.2 are given as follows. $\alpha = v^{-1/(1+\varepsilon)}(\varepsilon)^{-1/(1+\varepsilon)}2^{(1+\varepsilon)/\varepsilon}C_\varepsilon^{-1/(1+\varepsilon)}$ $(\frac{\log(2T_{max}/\delta)}{n})^{1/(1+\varepsilon)}$ for ERM-wide, ERM-narrow, Grad-wide, and Grad-narrow methods and $B = v^{1/(1+\varepsilon)}(\frac{n}{\log(2dT_{max}/\delta)})^{1/(1+\varepsilon)}$, where T_{max} is the maximum iteration number. We further fix $v = 1$ and $T_{max} = 1000$.

6.1 Regression

We first consider a regression problem, where we in particular assume that $Y_i = X_i^T w_* + \xi_i$, where ξ_i's are symmetrized Pareto random variables. That is, $\xi_i = (2u_i - 1)\tilde{\xi}_i$ with $\tilde{\xi}_i \sim_{i.i.d.} F_{pareto}(x)$ and $u_i = \text{Bernoulli}(0.5)$, $F_{pareto}(x) = 1 - \frac{1}{x^{1+a}}$ and a is the shape (tail) parameter. We further choose dimension $d \in \{2, 4, 8, 16, 32\}$ and set $a \in \{0.5, 1, 2\}$ when $p = 2$ or $a \in \{0.5, 1, 1.5\}$ when $p = 1.5$. For each setting, we set the number of samples to be 2000, randomly generate w_* (each entry is sampled from

$\{-1, 1\}$), and replicate it for 50 times. The averages of estimation errors ($\|\hat{w} - w_*\|_2$'s) are reported in Figure 1.

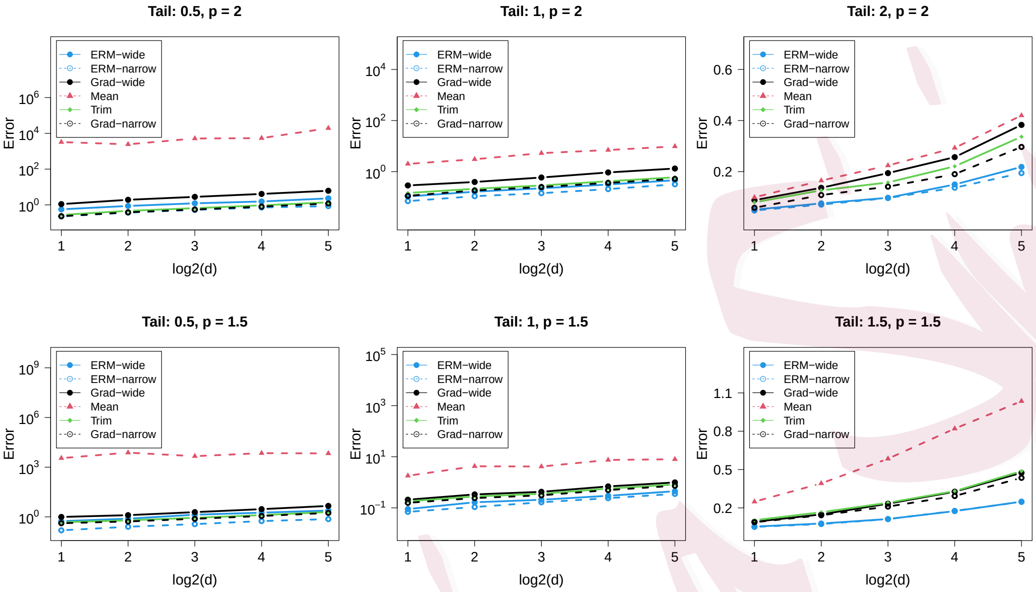

Figure 1: The comparison between six methods in regression problems under different dimension values and shape parameters.

6.2 Comparison with geometric Median

Geometric median (Minsker, 2015; Hsu and Sabato, 2016) is another type of popular method for estimation problems in heavy-tailed settings. This approach is the generalization of "Median of mean" estimator (Bubeck et al., 2013). For a regression problem, the main framework of the geometric median can be described as follows. (i) We divide data into M subsets. (ii) Compute estimators $w_1, \ldots, w_M$ from M subsets. (iii) For each $i \in [M]$, we compute the distance between $d_{ij} = \|w_i - w_j\|$ $(j \neq i, j \in [M])$. Compute r_i be the median of d_{ij}'s. (iv) Compute $i_* := \arg\min_i r_i$ and return $\hat{w} = w_{i_*}$.

We then compare our proposed ERM gradient method (ERM-narrow) with such geometric median estimator under the choice of dimension $d \in \{5, 10, 20\}$ and $a \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.2, 1.6, 2\}$. The choices of sample size and w^* are the same as before. The results are given in Figure 2.

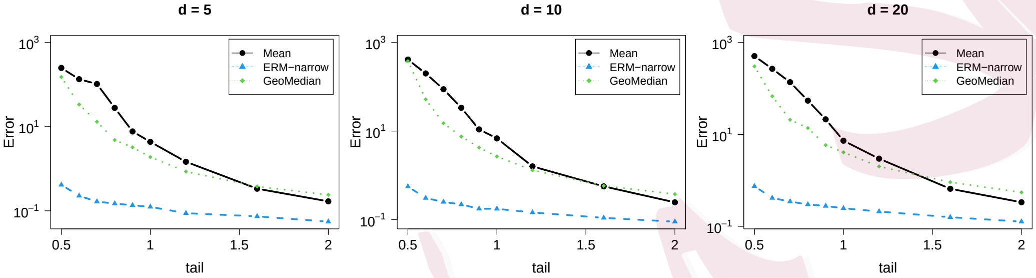

Figure 2: Comparison between geometric median methods and proposed ERM gradient method.

6.3 Combining with Deep Learning Framework

We implement our proposed double-weighted gradient descent (Algorithm 3) in the Pytorch platform and compare that with vanilla gradient descent, clipped-norm gradient descent and trimmed gradient descent methods.

Remark 11. In the clipped-norm gradient descent, the update formula is $w^{(t+1)} = w^{(t)} - \gamma_t g^{(t)}/\|g^{(t)}\|$, where $g^{(t)} = \frac{1}{n}\sum_i \frac{\partial f_{w^{(t)}}(X_i)}{\partial w}$. In the trimmed gradient descent, we compute the trimmed loss of sample i, i.e., $f_{trim}(X_i) =$

$\min\{f(X_i), B\}$ with B being a tuning parameter. Then the update formula is $w^{(t+1)} = w^{(t)} - \gamma_t g^{(t)}_{trim}$, where $g^{(t)}_{trim} = \frac{1}{n}\sum_i \frac{\partial f_{trim,w^{(t)}}(X_i)}{\partial w}$.

The data $y_i = f(x_i) + \epsilon_i$, where ϵ_i's are symmetrized Pareto random variables with the shape parameter taken in $\{1.2, 1.4, 1.6, 1.8, 2.0\}$. The underlying generating function f is chosen as $f(x) = \sin(x)$ [called as "Sin"], $f(x) = \sin(x)\exp\{x\}$ [called as "Sin_exp"], and $f(x) = x^2\cos(x)$ [called as "Cos_x2"], respectively. The considered neural network $\hat{f}$ is chosen to be a two-layer ReLU network with 128 hidden units.

In the training stage, we choose 2000 points x_i's from $[-\pi, \pi]$. In the testing stage, we randomly sample 100 points x_j's from Normal$(0, 1)$. The average prediction error is reported, i.e., $\frac{1}{100}\sum_{j=1}^{100} |\hat{f}(x_j) - f(x_j)|$. For the choice of α, we set it to be $c \cdot \alpha_0$, where α_0 is defined as the inverse of 95 % quantile of $f_{w^{(0)}}(X_i)$. For B in the trimmed gradient descent, we set it to $B = c/\alpha_0$. The hyper-parameter c is tuned via using the grid search with $c \in \{0.05, 0.1, 0.25, 1, 2, 5, 10, 20\}$. Each case is replicated for 50 times and the corresponding result is given in Figure 3.

Lastly, we report the relative computational time (i.e., $\frac{t_{dw}}{t_{van}}$, where t_{dw} is the time for our double-weighted method training on single data set and t_{van} is similarly defined for vanilla gradient descent) in Table 3.

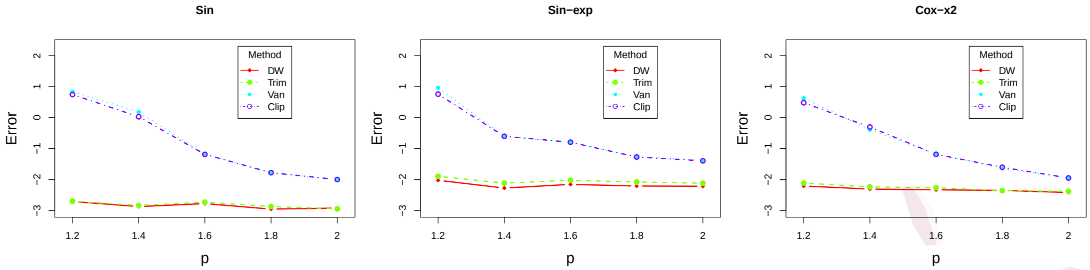

Figure 3: Prediction Error under Different Underlying Functions. The prediction error is reported under log-scale. (Lower is better.)

	”Sin”	”Sin_exp”	”Cos_x2”
mean	1.343	1.307	1.281
std	0.1528	0.1769	0.1048

Table 3: Table for relative computational time between the proposed double-weighted algorithm and the vanilla gradient descent method.

6.4 Findings

From Figure 1, we can see that the proposed ERM gradient outperforms other baselines when the tails of data distributions are heavy. This suggests that finding the optimizers by minimizing the risk values indeed improves the performances. The trim method has a similar performance as the robust gradient descent method. It indicates that estimating robust gradient coordinate-wisely is equivalent to trimming gradient in practice. The choice of influence function ϕ matters the final outcomes. Choosing narrow function ϕ_{narrow} will lead to a more robust estimator. Based on Figure 2, we can find that the geometric median method is sub-optimal compared with the proposed method. Although the median-type method also has reasonable theoretical guarantees, it is not a satisfactory algorithm in practice. Lastly, from Figure 3, we can see that our proposed double-weighted gradient descent method can be well embedded into Pytorch deep learning framework. It can achieve good prediction errors via neural network approximation. Additionally, Table 3 reveals that the computational time of the proposed algorithm is also comparable to the classical gradient descent method. Additional simulation studies are considered and the results can be found in the supplementary.

7. Conclusions

In this paper, we consider empirical risk minimization problem with heavy-tailed data, which is an important area in learning theory. We assume a weaker moment condition that data does not have finite variance, but only has $(1+\varepsilon)$-th moment with $0 < \varepsilon < 1$. In contrast to using truncation-based method, we directly work on estimating the excess risk values, where we adopt Catoni's method for robust estimation. The final optimizer is returned via minimizing the estimated excess risk. For such optimizer, we establish the excess risk bounds by studying the properties of Catoni's influence functions and using the generic chaining techniques. On the other hand, we propose an empirical risk-type gradient algorithms to address the computational challenges. The proposed algorithm gives a computationally friendly way to compute the robust gradient and also leads to better performance in terms of estimation errors. Compared with other competing baselines, our method is shown to be more robust under different types of outliers and data contamination. Our findings unveil that estimator based on minimizing risk values can be practically better than truncation methods. Our method is also shown to be easy-to-implement in Pytorch platform. Therefore it might be interesting and promising to study the proposed methodology in different large deep learning models in future work.

Acknowledgment. The authors thank the Associate Editor and two anonymous referees for their constructive suggestions and comments to improve the quality of our paper. The earlier version of this work was done when the authors were employed by Baidu USA. Guanhua Fang is partly supported by the National Natural Science Foundation of China (nos. 12301376) and Shanghai Educational Development Foundation (23CGA02).

References

Ahn, S., J. H. Kim, and V. Ramaswami (2012). A new class of models for heavy tailed distributions in finance and insurance risk. *Insurance: Mathematics and Economics 51*(1), 43–52.

Bartlett, P. L., O. Bousquet, and S. Mendelson (2005). Local rademacher complexities.

Bartlett, P. L. and S. Mendelson (2006). Empirical minimization. *Probability theory and related fields 135*(3), 311–334.

Bhatt, S., G. Fang, P. Li, and G. Samorodnitsky (2022a). Minimax m-estimation under adversarial contamination. In *International Conference on Machine Learning*, pp. 1906–1924.

Bhatt, S., G. Fang, P. Li, and G. Samorodnitsky (2022b). Nearly optimal catoni's m-estimator for infinite variance. In *International Conference on Machine Learning*, pp. 1925–1944.

Bradley, B. O. and M. S. Taqqu (2003). Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pp. 35–103. Elsevier.

Brownlees, C., E. Joly, and G. Lugosi (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics 43*(6), 2507–2536.

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

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, Volume 48, pp. 1148–1185.

Chen, P., X. Jin, X. Li, and L. Xu (2021). A generalized catoni's m-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. *Electronic Journal of Statistics 15*(2), 5523–5544.

Chong, E. K. and S. H. Zak (2004). *An introduction to optimization.* John Wiley & Sons.

Crovella, M. E. and M. S. Taqqu (1999). Estimating the heavy tail index from scaling properties. *Methodology and computing in applied probability 1*(1), 55–79.

Diakonikolas, I., D. M. Kane, and A. Pensia (2020). Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems 33*, 1830–1840.

Glaz, J., J. I. Naus, and S. Wallenstein (2001). *Scan statistics.* Springer.

Holland, M. and M. Haress (2021). Learning with risk-averse feedback under potentially heavy tails. In *International Conference on Artificial Intelligence and Statistics*, pp. 892–900.

Holland, M. and K. Ikeda (2019). Better generalization with less data using robust gradient descent. In *International Conference on Machine Learning*, pp. 2761–2770. PMLR.

Hsu, D. and S. Sabato (2016). Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research 17*(1), 543–582.

Huber, P. J. (2011). Robust statistics. In *International encyclopedia of statistical science*, pp. 1248–1251. Springer.

Lecué, G. and S. Mendelson (2013). Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*.

Lecué, G. and S. Mendelson (2016). Performance of empirical risk minimization in linear aggregation. *Bernoulli 22*(3), 1520–1534.

Lemaréchal, C. (2012). Cauchy and the gradient method. *Doc Math Extra 251*(254), 10.

Lerasle, M. and R. I. Oliveira (2011). Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.

Liu, T. and D. Tao (2014). On the robustness and generalization of cauchy regression. In *2014 4th IEEE International Conference on Information Science and Technology*, pp. 100–105.

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

Mathieu, T. and S. Minsker (2021). Excess risk bounds in robust empirical risk minimization. *Information and Inference: A Journal of the IMA 10*(4), 1423–1490.

Meusel, R., S. Vigna, O. Lehmberg, and C. Bizer (2014). Graph structure in the web—revisited: a trick of the heavy tail. In *Proceedings of the 23rd international conference on World Wide Web*, pp. 427–432.

Minsker, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli 21*(4), 2308–2335.

Minsker, S. (2018). Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics 46*(6A), 2871–2903.

Roy, A., K. Balasubramanian, and M. A. Erdogdu (2021). On empirical risk minimization with dependent and heavy-tailed data. *Advances in Neural Information Processing Systems 34*.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Talagrand, M. (1996). Majorizing measures: the generic chaining. *The Annals of Probability 24*(3), 1049–1103.

Van de Geer, S. A. and S. van de Geer (2000). *Empirical Processes in M-estimation*, Volume 6. Cambridge university press.

Vapnik, V. (1991). Principles of risk minimization for learning theory. *Advances in neural information processing systems 4*.

Xu, L., F. Yao, Q. Yao, and H. Zhang (2023). Non-asymptotic guarantees for robust statistical learning under infinite variance assumption. *Journal of Machine Learning Research 24*(92), 1–46.

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

Yi, M., R. Wang, and Z.-M. Ma (2022). Characterization of excess risk for locally strongly convex population risk. *Advances in Neural Information Processing Systems 35*, 21270–21285.

Zhang, H., M. Cisse, Y. N. Dauphin, and D. Lopez-Paz (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, L. and Z.-H. Zhou (2018). l1-regression with heavy-tailed distributions. *Advances in Neural Information Processing Systems 31*.

Zhu, Z. and W. Zhou (2021). Taming heavy-tailed features by shrinkage. In *International Conference on Artificial Intelligence and Statistics*, pp. 3268–3276. PMLR.

Zhuang, V. and Y. Sui (2021). No-regret reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pp. 3385–3393. PMLR.

Algorithm 1 Robust Gradient Descent Method

```
Input: Observations: ${X_i, i \in \{1, \dots, n\}}$. A bounded Catoni influence function $\phi$.
Output: Estimated parameter: $\tilde{w}$
Initialization: Randomly choose $w^{(0)}$ from $\mathbb{R}^d$ and set time index $t = 0$.
while not converged do
    [Robust gradient] Compute robust gradient $g^{(t)}$ by solving
        $\sum_{i=1}^{n} \phi(\alpha(\nabla f_{w^{(t)}}(X_i)[j] - g^{(t)}[j])) = 0 \qquad (5.27)$
    coordinate-wisely for $j \in [d]$.
    Update parameter by $w^{(t+1)} = w^{(t)} - \gamma_t g^{(t)}$.
    Increase time index $t = t + 1$.
end while
Return parameter estimate $\tilde{w} = w^{(t)}$.
```

Algorithm 2 Empirical Risk Gradient Descent

```
**Input:** Observations: $\{X_i, i \in \{1, \dots, n\}\}$.
**Output:** Estimated parameter: $\tilde{w}$
**Initialization:** Randomly choose $w^{(0)}$ from $\mathbb{R}^d$ and set time index $t = 0$.
**while** not converged **do**
    Find $\hat{\mu}_{f_{w^{(t)}}}$ by solving
        $0 = \frac{1}{n\alpha} \sum_{i=1}^{n} \phi(\alpha(f_{w^{(t)}}(X_i) - \mu)).$   (5.35)
    Compute gradient $g^{(t)}$ by
        $g^{(t)} = \nabla_w \hat{\mu}_{f_w^{(t)}} = \frac{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}_{f_w^{(t)}})) \frac{\partial f_{w^{(t)}}(X_i)}{\partial w}}{\sum_i \phi'(\alpha(f_{w^{(t)}}(X_i) - \hat{\mu}_{f_w^{(t)}}))}.$   (5.36)
    Update parameter by $w^{(t+1)} = w^{(t)} - \gamma_t g^{(t)}$.
    Increase time index $t = t + 1$.
**end while**
Return parameter estimate $\tilde{w} = w^{(t)}$.
```

Algorithm 3 Double-weighted Gradient Descent

```
Input: Observations: $\{X_i, i \in \{1, \ldots, n\}\}$; A catoni-influence function $\phi$; Stopping threshold $\tilde{\varrho}$.
Output: Estimated parameter: $\tilde{w}$
Initialization: Set time index $t = 0$. Randomly choose $w^{(0)}$ from $\mathbb{R}^d$, choose an $\mu^{(0)} \in \mathbb{R}^+$ such that $\hat{\mu}^{(0)}$ is an $\alpha$-approximate solution to
    $0 = \frac{1}{n} \sum_{i=1}^{n} \phi(\alpha(f_{w^{(0)}}(X_i) - \mu)).$    (5.38)
    Choose weights to be $\nu_i^{(0)} = \frac{\phi'(\alpha(f_{w^{(0)}}(X_i) - \hat{\mu}^{(0)}))}{\sum_i \phi'(\alpha(f_{w^{(0)}}(X_i) - \hat{\mu}^{(0)}))}$.
while not converged do
    Compute gradient $g^{(t)}$ by $g^{(t)} = \sum \nu_i^{(t)} \frac{\partial f_{w^{(t)}}(X_i)}{\partial w}$.
    Update parameter by $w^{(t+1)} = w^{(t)} - \gamma_t g^{(t)}$.
    Find $\hat{\mu}^{(t+1)}$ by computing
        $\hat{\mu}^{(t+1)} = \hat{\mu}^{(t)} + \sum_i \nu_i^{(t)} (f_{w^{(t+1)}}(X_i) - f_{w^{(t)}}(X_i)).$    (5.39)
    Compute weights $\nu_i^{(t+1)} = \frac{\phi'(\alpha(f_{w^{(t+1)}}(X_i) - \hat{\mu}^{(t+1)}))}{\sum_i \phi'(\alpha(f_{w^{(t+1)}}(X_i) - \hat{\mu}^{(t+1)}))}$.
    Increase time index $t = t + 1$.
end while
Return parameter estimate $\tilde{w} = w^{(t)}$.
```