Statistica Sinica Preprint No: SS-2024-0268	
Title	Asymptotic Normality of Robust Risk Minimizers
Manuscript ID	SS-2024-0268
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0268
Complete List of Authors	Stanislav Minsker
Corresponding Authors	Stanislav Minsker
E-mails	stas.minsker@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0268

Statistica Sinica

Asymptotic Normality of Robust Risk Minimizers

Stanislav Minsker

University of Southern California

Abstract: This paper investigates the asymptotic properties of algorithms that can be viewed as robust analogues of the classical empirical risk minimization. These strategies are based on replacing the usual empirical average with a robust proxy of the mean, such as a variant of the median-of-means estimator. It is well known by now that the excess risk of resulting estimators often converges to zero at optimal rates under much weaker assumptions than those required by their classical counterparts. However, less is known about the asymptotic properties of the estimators themselves, for instance, whether robust analogues of the maximum likelihood estimators are asymptotically efficient. We take a step towards answering these questions and show that for a wide class of parametric problems, minimizers of the appropriately defined robust proxy of the risk converge to the minimizers of the true risk at the same rate, and often have the same asymptotic variance, as the estimators obtained by minimizing the usual empirical risk. Finally, we discuss the computational aspects of the problem and demonstrate the numerical performance of the methods under consideration in numerical experiments.

Key words and phrases: Robust estimation, median-of-means estimator, asymptotic normality, adversarial contamination.

1. Introduction.

The concept of robustness addresses stability of statistical estimators under various forms of perturbations, such as the presence of corrupted/atypical observations ("outliers") in the data. The questions related to robustness in the framework of statistical learning theory have seen a surge in interest, both from the theoretical and practical perspectives, and resulted in the development of novel algorithms. These new robust algorithms are characterized by the fact that they provably work under minimal assumptions on the underlying data-generating mechanism, often requiring the existence of moments of low order only. Majority of the existing works focused on the upper bounds for the risk of the estimators (such as the classification or prediction error) produced by the algorithms, while in this paper we are interested in the asymptotic properties of the estimators themselves. The asymptotic viewpoint allows one to gauge efficiency of the estimators and understand the magnitude of constants appearing in the bounds, as opposed to just studying the form of dependence of the bounds on the parameters of interest (sample size, dimension, etc.) The mean estimators at the core of the approach under consideration are non-linear and are defined as solutions of optimization problems, which makes the analysis more technical. Navigation through the technical details and development of the tools such as Bahadur-type representations needed to tackle the non-linearities occupies a large part of the analysis. Therefore, the main contributions of the paper are technical in nature.

Next, we introduce the mathematical framework used in the exposition. Let $(S, \mathcal{S})$ be a measurable space, and let $X \in S$ be a random variable with distribution P. Suppose that $X_1, \ldots, X_N$ are i.i.d. copies of X. Moreover, assume that $\mathcal{L} = \left\{\ell(\theta, \cdot),\ \theta \in \Theta \subseteq \mathbb{R}^d\right\}$ is a class of measurable functions from S to $\mathbb{R}$ indexed by an open subset of $\mathbb{R}^d$. Population versions of many estimation problems in statistics and statistical learning, such as maximum likelihood estimation and regression, can be formulated as risk minimization of the form

$$\mathbb{E}\,\ell(\theta, X) \to \min_{\theta \in \Theta}. \tag{1.1}$$

In particular, when $\{p_\theta,\ \theta \in \Theta\}$ is a family of probability density functions with respect to some σ-finite measure μ and $\ell(\theta, \cdot) = -\log p_\theta(\cdot)$, the resulting problem corresponds to maximum likelihood estimation. In what follows, we will set $L(\theta)$ to be the risk associated with the parameter θ, namely $L(\theta) = \mathbb{E}\ell(\theta, X)$. Throughout the paper, we will assume that the minimum in problem (1.1) is attained at a unique point $\theta_0 \in \Theta$. The true distribution P is typically unknown, and an estimator of θ_0 is obtained via minimizing the *empirical risk*, namely,

$$\widetilde{\theta}_N := \operatorname*{argmin}_{\theta\in\Theta} L_N(\theta), \tag{1.2}$$

where $L_N(\theta) := \frac{1}{N}\sum_{j=1}^{N} \ell(\theta, X_j)$. If the marginal distributions of the process $\{\ell(\theta,\cdot),\ \theta\in\Theta\}$ are heavy-tailed, meaning that they possess finite moments of low order only, then the error $|L_N(\theta) - L(\theta)|$ can be large with non-negligible probability, motivating the need for alternative proxies for the risk $L(\theta)$. Another scenario of interest corresponds to the *adversarial contamination* framework, where the initial dataset of cardinality N' is merged with a set of $\mathcal{O} < N'$ outliers generated by an adversary who has complete knowledge of the underlying distribution and an opportunity to inspect the data, and the combined dataset of cardinality $N = N' + \mathcal{O}$ is presented to the algorithm responsible for constructing the estimator of θ_0. In what follows, the proportion of outliers will be denoted by $\kappa := \frac{\mathcal{O}}{N}$. Similarly to the heavy-tailed scenario, the empirical loss $L_N(\theta)$ is not a robust proxy for $\mathbb{E}\ell(\theta, X)$ in this case, therefore estimation and inference results based on minimizing $L_N(\theta)$ may be unreliable. One may approach the problem of estimating θ_0 robustly from different angles. One class of popular methods consists of robust versions of the gradient descent algorithm for the optimization problem (1.1), where the gradient $\nabla L(\theta_k)$ is estimated on each iteration k; for example, this approach has been explored by Prasad et al. (2020); Chen et al. (2017); Alistarh et al. (2018), among others. Another technique (the one that we investigate in this paper) is based on replacing the average $L_N(\cdot)$ by a robust proxy of $L(\theta)$. Its advantage is the fact that we only need to estimate a real-valued quantity $L(\theta)$, as opposed to the high-dimensional gradient vector $\nabla L(\theta)$. On the other hand, favorable properties, such as convexity, that are "inherited" by the formulation (1.2) from (1.1), are usually lost in this case. Several representative papers that explore this direction include the works by Audibert et al. (2011); Lerasle and Oliveira (2011); Brownlees et al. (2015); Lugosi and Mendelson (2019b); Lecué and Lerasle (2020); Cherapanamjeri et al. (2019); Mathieu and Minsker (2021); also, see an excellent survey paper by Lugosi and Mendelson (2019a). Instead of the empirical risk $L_N(\theta)$, these works employ robust estimators of the risk such as the median-of-means estimator (Nemirovski and Yudin, 1983; Alon et al., 1996; Devroye et al., 2016) or Catoni's estimator and its variants (Catoni, 2012; Li et al., 2022). In this paper, we study estimators based on the modification of the median-of-means principle introduced by Minsker (2019) combined with the idea behind the so-called "median-of-means tournaments" (Lugosi and Mendelson, 2019b) and the closely related "min-max" robust estimators (Audibert et al., 2011; Lecué and Lerasle, 2020). The latter are based on an observation that θ_0 can be alternatively obtained via

$$\theta_0 = \underset{\theta\in\Theta}{\operatorname{argmin}} \max_{\theta'\in\Theta} \left(L(\theta) - L(\theta')\right). \tag{1.3}$$

Therefore, an estimator of θ_0 can be constructed by replacing the difference $L(\theta,\theta') := L(\theta) - L(\theta')$ by its robust proxy constructed as follows. Let $k \leqslant N/2$ be an integer, and assume that $G_1,\ldots,G_k$ are disjoint subsets of the index set $\{1,\ldots,N\}$ of cardinality $|G_j| = n \geqslant \lfloor N/k \rfloor$ each. For $\theta \in \Theta$, let

$$\bar{L}_j(\theta) := \frac{1}{n}\sum_{i\in G_j} \ell(\theta, X_i)$$

be the empirical risk evaluated over the subsample indexed by G_j. Assume that $\rho : \mathbb{R} \mapsto \mathbb{R}_+$ is a convex, even function that is increasing on $(0,\infty)$ and such that its (right) derivative is bounded. Let $\{\Delta_n\}_{n\geqslant 1}$ be a non-decreasing positive sequence of "scaling factors" such that $\Delta_n = o(\sqrt{n})$ and $\Delta_\infty := \lim_{n\to\infty}\Delta_n \in (0,\infty]$, and define

$$\widehat{L}(\theta,\theta') \in \underset{z\in\mathbb{R}}{\operatorname{argmin}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\,\frac{\bar{L}_j(\theta) - \bar{L}_j(\theta') - z}{\Delta_n}\right). \tag{1.4}$$

For example, the choice $\Delta_n \asymp \log(n)$ suffices for all results of the paper to hold (in fact, it suffices for Δ_∞ to be a sufficiently large constant); we will make a remark regarding the practical aspects of setting Δ_n below. The estimator $\widehat{L}(\theta,\theta')$ is what we referred to as the robust proxy of $L(\theta,\theta')$, where robustness is justified by the fact that the error $\left|\widehat{L}(\theta,\theta') - L(\theta,\theta')\right|$ satisfies non-asymptotic exponential deviation bounds under minimal assumptions on the tails of the random variables $\ell(\theta, X) - \ell(\theta', X)$ and the ability of $\widehat{L}(\theta,\theta')$ to resist adversarial outliers. For example, Theorem 3 in (Minsker, 2019) essentially states that whenever $\Delta_n \gtrsim \mathrm{Var}^{1/2}\left(\ell(\theta, X) - \ell(\theta', X)\right)$ and for all $s \lesssim k$,

$$\left|\widehat{L}(\theta,\theta') - L(\theta,\theta')\right| \lesssim \sigma(\theta,\theta')\sqrt{\frac{s}{N}} + \Delta_n\left(\frac{k}{N} + \frac{\mathcal{O}\sqrt{n}}{N}\right)$$

with probability at least $1 - e^{-s}$, assuming that $\mathbb{E}|\ell(\theta, X) - \ell(\theta', X)|^3 < \infty$ and where $\lesssim$ denotes the inequality up to absolute numerical constants; similar guarantees also hold uniformly over $\theta, \theta' \in \Theta$; note that setting $\Delta_n \asymp \sigma(\theta,\theta')$ yields the most robust estimator. Given the robust proxy $\widehat{L}(\theta,\theta')$ of $L(\theta,\theta')$, an analogue of the classical empirical risk minimizer $\widetilde{\theta}_N$ can be obtained via

$$\widehat{\theta}_{n,k} = \operatorname*{argmin}_{\theta\in\Theta} \sup_{\theta'\in\Theta} \widehat{L}(\theta,\theta'). \tag{1.5}$$

Simple sufficient conditions for the existence of $\widehat{\theta}_{n,k}$ are discussed in the supplementary material; in principle, one could consider near-minimizers instead, however, we avoid this route due to the extra layer of technicalities it brings. The idea behind considering differences of the risks and defining θ_0 via (1.3) is related to the fact that the estimators (1.4) of

$L(\theta)$, unlike their traditional counterparts $L_N(\theta)$, are non-linear: if we set $\widehat{L}(\theta) = \operatorname{argmin}_{z\in\mathbb{R}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\,\frac{\bar{L}_j(\theta)-z}{\Delta_n}\right)$, then $\widehat{L}(\theta,\theta') \neq \widehat{L}(\theta) - \widehat{L}(\theta')$.

Related approaches based on direct minimization of $\widehat{L}(\theta)$ have been previously investigated by Brownlees et al. (2015); Holland and Ikeda (2017); Lecué et al. (2020); Mathieu and Minsker (2021), where the main object of interest was the excess risk $\mathcal{E}(\widehat{\theta}_{n,k}) := L(\widehat{\theta}_{n,k}) - L(\theta_0)$. It has been recognized however that non-linearity of $\widehat{L}(\theta)$ often results in sub-optimal rates, while the tournament-type procedures avoid these shortcomings. In the present work, we will be interested in the asymptotic behavior of the error $\widehat{\theta}_{n,k} - \theta_0$, rather than the excess risk: in particular, we will establish asymptotic normality of the sequence $\sqrt{N}\left(\widehat{\theta}_{n,k} - \theta_0\right)$ and demonstrate that robust estimators can still be efficient under essentially the same set of sufficient conditions as required by the standard M-estimators (van der Vaart, 2000). The nonlinear nature of the estimator $\widehat{L}(\theta,\theta')$ makes the proofs significantly more technical compared to the classical theory of M-estimators based on usual empirical risk minimization. To tackle these challenges, our arguments rely on Bahadur-type representations for $\widehat{L}(\theta,\theta')$ whose remainder terms admit tight uniform bounds.

1.1 Notation.

Absolute constants will be denoted c, c_1, C, C_1, C', etc., and may take different values in different parts of the paper. Given $a, b \in \mathbb{R}$, we will write $a \wedge b$ for $\min(a, b)$ and $a \vee b$ for $\max(a, b)$. For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, define

$$\operatorname*{argmin}_{y \in \mathbb{R}^d} f(y) := \{y \in \mathbb{R}^d : f(y) \leqslant f(x) \text{ for all } x \in \mathbb{R}^d\},$$

and $\|f\|_\infty := \operatorname{ess\,sup}\{|f(y)| : y \in \mathbb{R}^d\}$. Moreover, $\mathrm{Lip}(f)$ will stand for the Lipschitz constant of f; if $d = 1$ and f is m times differentiable, $f^{(m)}$ will denote the m-th derivative of f. For a function $g(\theta, x)$ mapping $\mathbb{R}^d \times \mathbb{R}$ to $\mathbb{R}$, $\partial_\theta g$ will denote the vector of partial derivatives with respect to the coordinates of θ; similarly, $\partial^2_\theta g$ will denote the matrix of second partial derivatives. For $x \in \mathbb{R}^d$, $\|x\|$ will stand for the Euclidean norm of x, $\|x\|_\infty := \max_j |x_j|$, and for a matrix $A \in \mathbb{R}^{d \times d}$, $\|A\|$ will denote the spectral norm of A. We will frequently use the standard big-O and small-o notation, as well as their in-probability siblings o_P and O_P. For vector-valued sequences $\{x_j\}_{j \geqslant 1}$, $\{y_j\}_{j \geqslant 1} \subset \mathbb{R}^d$, asymptotic relations $x_j = o(y_j)$ and $x_j = O(y_j)$ are assumed to hold coordinate-wise. We will write $x_j \ll y_j$ if $x_j = o(y_j)$ and $x_j \gg y_j$ if $y_j = o(x_j)$. For a square matrix $A \in \mathbb{R}^{d \times d}$, $\operatorname{tr} A := \sum_{j=1}^d A_{j,j}$ denotes the trace of A. Given a function $g : \mathbb{R} \mapsto \mathbb{R}$, measure Q and $1 \leqslant p < \infty$, we set $\|g\|^p_{L_p(Q)} := \int_{\mathbb{R}} |g(x)|^p dQ$. For i.i.d. random variables

$X_1, \ldots, X_N$ distributed according to P, $P_N := \frac{1}{N}\sum_{j=1}^N \delta_{X_j}$ will stand for the empirical measure; here, $\delta_X(g) := g(X)$. The expectation with respect to a probability measure Q will be denoted $\mathbb{E}_Q$; if the measure is not specified, it will be assumed that the expectation is taken with respect to P, the distribution of X. Given $f : S \mapsto \mathbb{R}^d$, we will write Qf for $\int f dQ \in \mathbb{R}^d$, assuming that the last integral is calculated coordinate-wise. For $\theta, \theta' \in \Theta$, let $\sigma^2(\theta, \theta') = \mathrm{Var}\left(\ell(\theta, X) - \ell(\theta', X)\right)$ and for $\Theta' \subseteq \Theta$, define $\sigma^2(\Theta') := \sup_{\theta,\theta' \in \Theta'} \sigma^2(\theta, \theta')$.

Finally, we will adopt the convention that the infimum over the empty set is equal to $+\infty$. Additional notation and auxiliary results are introduced on demand.

2. Statements of the main results.

We begin by listing the assumptions on the model; these conditions are similar to the standard assumptions made in the parametric estimation framework (van der Vaart, 2000; van der Vaart and Wellner, 1996). The first assumption lists the requirements for the loss function ρ (note that the choice of this function is completely determined by the statistician).

Assumption 1. *The function $\rho : \mathbb{R} \mapsto \mathbb{R}$ is convex, even, and such that*

(i) $\rho'(z) = z$ *for* $|z| \leqslant 1$ *and* $\rho'(z) = \mathrm{const}$ *for* $z \geqslant 2$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0268

(ii) $z - \rho'(z)$ *is nondecreasing;*

(iii) $\rho^{(5)}$ *is bounded and Lipschitz continuous.*

An example of a function ρ satisfying required assumptions is given by "smoothed" Huber's loss defined as follows. Let

$$H(y) = \frac{y^2}{2} I\{|y| \leqslant 3/2\} + \frac{3}{2}\left(|y| - \frac{3}{4}\right) I\{|y| > 3/2\}$$

be the usual Huber's loss. Moreover, let ψ be the mollifier

$$\psi(x) = C \exp\left(-\frac{4}{1-4x^2}\right)\left\{|x| \leqslant \frac{1}{2}\right\}$$

where C is chosen so that $\int_{\mathbb{R}} \psi(x)dx = 1$. Then ρ given by the convolution $\rho(x) = (h * \psi)(x)$ satisfies Assumption 1.

Remark 1. The classical median-of-means estimator (Nemirovski and Yudin, 1983; Alon et al., 1996) corresponds to the choice $\rho(x) = |x|$ that does not satisfy smoothness assumptions imposed above. Asymptotic behavior of the estimators corresponding to this loss is left as an open problem; numerical evidence suggesting that asymptotic normality does not hold in this case is presented in (Minsker and Yao, 2025).

Assumption 2. *The Hessian* $\partial^2_\theta L(\theta_0)$ *exists and is strictly positive definite.*

This assumption ensures that in a sufficiently small neighborhood of θ_0, $c(\theta_0)\|\theta-\theta_0\|^2 \leqslant L(\theta) - L(\theta_0) \leqslant C(\theta_0)\|\theta-\theta_0\|^2$ for some $0 < c(\theta_0) \leqslant C(\theta_0) <$

∞. The following two conditions allow one to control the "complexity" of the class $\{\ell(\theta,\cdot),\ \theta\in\Theta\}$.

Assumption 3. *For every $\theta\in\Theta$, the map $\theta'\mapsto\ell(\theta',x)$ is differentiable at θ for P-almost all x (where the exceptional set of measure 0 can depend on θ), with derivative $\partial_\theta\ell(\theta,x)$. Moreover, $\forall\theta\in\Theta$, the envelope function $\mathcal{V}(x;\delta):=\sup_{\|\tilde\theta-\theta\|\leqslant\delta}\left\|\partial_\theta\ell(\tilde\theta,x)\right\|$ of the class $\left\{\partial_\theta\ell(\tilde\theta,\cdot):\|\tilde\theta-\theta\|\leqslant\delta\right\}$ satisfies $\mathbb{E}\mathcal{V}^2(X;\delta)<\infty$ for sufficiently small $\delta=\delta(\theta)$.*

An immediate implication of this assumption is the fact that the function $\theta\mapsto\ell(\theta,x)$ is locally Lipschitz. It other words, for any $\theta\in\Theta$, there exists a ball $B(\theta,r(\theta))$ of radius $r(\theta)$ such that for all $\theta_1,\,\theta_2\in B(\theta,r(\theta))$, $|\ell(\theta_1,x)-\ell(\theta_2,x)|\leqslant\mathcal{V}(x;r(\theta))\|\theta_1-\theta_2\|$. In particular, this condition suffices to prove consistency of the estimators considered in this work and is similar to the classical assumptions used in the analysis of M-estimators, e.g. see the book by van der Vaart (2000). The final assumption that we impose allows us to treat non-compact parameter spaces. Essentially, we require that the estimator $\widehat\theta_{n,k}$ defined via (1.5) belongs to a compact set of sufficiently large diameter with high probability, namely,

$$\lim_{R\to\infty}\limsup_{n,k\to\infty}\mathbb{P}\left(\left\|\widehat\theta_{n,k}-\theta_0\right\|\geqslant R\right)=0\text{ and}$$

The following condition is sufficient for the display above to hold:

Assumption 4. *Let $X_1, \dots, X_n$ be i.i.d. Given $t, R > 0$ and a positive integer n, define*

$$B(n, R, t) := \mathbb{P}\left(\inf_{\theta\in\Theta,\, \|\theta-\theta_0\|\geqslant R} \frac{1}{n}\sum_{j=1}^{n} \ell(\theta, X_j) < \mathbb{E}\ell(\theta_0, X) + t \right).$$

Then $\lim_{R\to\infty}\limsup_{n\to\infty} B(n,R,t) = 0$ *for some* $t > 0$.

Let us emphasize that the data $X_1, \dots, X_n$ in Assumption 4 do not contain outliers. Requirements similar to this assumption are commonly imposed in the classical framework of M-estimation, (e.g see van der Vaart, 2000). Of course, when Θ is compact, Assumption 4 holds automatically; another general scenario when Assumption 4 is true occurs if the class $\{\ell(\theta,\cdot):\ \theta\in\Theta\}$ is Glivenko-Cantelli (van der Vaart and Wellner, 1996). Otherwise, it can usually be verified on a case-by-case basis. For instance, consider the framework of linear regression, where the data consist of i.i.d. copies of the random couple $(Z, Y) \in \mathbb{R}^d \times \mathbb{R}$ such that $Y = \langle Z, \theta_*\rangle + \varepsilon$ for some $\theta_* \in \mathbb{R}^d$ and a noise variable ε that is independent of Z and has variance σ^2. Moreover, assume that Z is centered and has positive definite covariance matrix Σ. In this case, $\ell(\theta, Z, Y) = (Y - \langle Z, \theta\rangle)^2$, and it is easy to see that $\frac{1}{n}\sum_{j=1}^{n} \ell(\theta, Z_j, Y_j) = \frac{1}{n}\left(\|\vec{\varepsilon}\|^2 + \|\mathbb{Z}(\theta-\theta_*)\|^2 - 2\langle\vec{\varepsilon}, \mathbb{Z}(\theta_*-\theta)\rangle\right)$, where $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ and $\mathbb{Z} \in \mathbb{R}^{n\times d}$ has $Z_1, \dots, Z_n$ as rows. Cauchy-Schwarz inequality combined with a simple relation $2|ab| \leqslant a^2/2 + 2b^2$ that holds for all $a, b \in \mathbb{R}$ yield that

$$\frac{1}{n}\sum_{j=1}^{n} \ell(\theta, Z_j, Y_j) \geqslant \frac{1}{2n}\|\mathbb{Z}(\theta-\theta_*)\|^2 - \frac{1}{n}\|\vec{\varepsilon}\|^2,$$

hence $\inf_{\|\theta-\theta_*\|\geqslant R} \frac{1}{n}\sum_{j=1}^{n} \ell(\theta, Z_j, Y_j) \geqslant \frac{R^2}{2}\inf_{\|u\|=1}\langle \Sigma_n u, u\rangle - \frac{1}{n}\|\vec{\varepsilon}\|^2$ where $\Sigma_n = \frac{1}{n}\sum_{j=1}^{n} Z_j Z_j^T$ is the sample covariance matrix. Since $\inf_{\|u\|=1}\langle \Sigma_n u, u\rangle \geqslant \lambda_{\min}(\Sigma) - \|\Sigma_n - \Sigma\| = \lambda_{\min}(\Sigma) - o_P(1)$ and $\frac{1}{n}\|\vec{\varepsilon}\|^2 = O_p(1)$, it is easy to conclude that Assumption 4 holds; here, we used the fact that $\|\Sigma_n - \Sigma\| = o_P(1)$ in view of the law of large numbers.

We are ready to state the main results regarding consistency and asymptotic normality of the estimator (1.5). Recall the adversarial contamination framework defined in section 1. In all statements below, we assume that the sequences $\{k_j\}_{j\geqslant 1}$ and $\{n_j\}_{j\geqslant 1}$, corresponding the the number of subgroups and their cardinality respectively, are non-decreasing and converge to ∞ as $j \to \infty$, and that the total sample size is $N_j := k_j n_j$.

Theorem 1. *Let assumptions 1, 2, 3 and 4 be satisfied. Suppose that the number of outliers $\mathcal{O}_j$ is such that $\limsup\limits_{j\to\infty} \frac{\mathcal{O}_j}{k_j} \leqslant c$ for a sufficiently small absolute constant $c > 0$. Then the estimator $\widehat{\theta}_{n_j,k_j}$ defined in (1.5) is consistent: $\widehat{\theta}_{n_j,k_j} \to \theta_0$ in probability as $j \to \infty$.*

We remark that the contamination framework considered in Theorem 1 is quite general: for instance, in the framework if linear regression, $X =$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0268

$(Z, Y) \in \mathbb{R}^d \times \mathbb{R}$, hence outliers can occur among both the predictor Z and response variable Y. On the other hand, many classical robust regression methods, such as Huber's regression, only allow the outliers among the responses. The following theorem constitutes the main contribution of the paper.

Theorem 2. *Let assumptions 1, 2, 3 and 4 be satisfied, and suppose that the number of outliers* $\mathcal{O}_j$ *is such that* $\limsup_{j\to\infty} \frac{\mathcal{O}_j}{k_j} \leqslant c$ *for a sufficiently small absolute constant* $c > 0$. *Moreover, assume that* $\{\alpha_{n_j,k_j}\}_{j\geqslant 1}$ *is a non-increasing sequence such that*

$$\alpha^2_{n_j,k_j} \geqslant \frac{1}{n_j k_j} \text{ and } \alpha^2_{n_j,k_j} \gg \frac{\mathcal{O}_j}{k_j}\frac{1}{\sqrt{n_j}}.$$

Then

$$\lim_{M\to\infty} \limsup_{n_j,k_j\to\infty} \mathbb{P}\left(\|\widehat{\theta}_{n_j,k_j} - \theta_0\| \geqslant M \cdot \alpha_{n_j,k_j} \right) = 0.$$

In addition, if the sample is free of adversarial contamination (that is, $\mathcal{O}_j = 0$*), then*

$$\sqrt{N_j}\left(\widehat{\theta}_{n_j,k_j} - \theta_0\right) \xrightarrow{d} N\left(0, D^2(\theta_0)\right) \text{ as } j \to \infty,$$

where $D^2(\theta_0) = \left[\partial^2_\theta L(\theta_0)\right]^{-1} \Sigma \left[\partial^2_\theta L(\theta_0)\right]^{-1}$ *and* $\Sigma = \mathbb{E}\left[\partial_\theta \ell(\theta_0, X) \partial_\theta \ell(\theta_0, X)^T\right]$.

This result goes one step further compared to Theorem 1 and establishes the rate of convergence of $\widehat{\theta}_{n,k}$ to θ_0. Moreover, it implies that in the "ideal," outlier-free scenario, $\alpha_{n,k} = \frac{1}{\sqrt{nk}} = \frac{1}{\sqrt{N}}$ is the standard parametric rate (and the rate is strictly slower if $\mathcal{O}_j \geqslant 1$), and that no loss of asymptotic efficiency occurs compared to the standard M-estimator based on empirical risk minimization. For example, maximum likelihood estimator corresponds to the case when $\{p_\theta, \ \theta \in \Theta\}$ is a family of probability density functions with respect to some σ-finite measure μ and $\ell(\theta, \cdot) = -\log p_\theta(\cdot)$. If it holds that

$$-\partial_\theta^2 \, \mathbb{E} \log p_{\theta_0}(X) = I(\theta_0) := \mathbb{E}\left[\partial_\theta \log p_{\theta_0}(X) \partial_\theta \log p_{\theta_0}(X)^T\right],$$

then it follows that $\widehat{\theta}_{n_j,k_j}$ is asymptotically equivalent to the maximum likelihood estimator. The proof of Theorem 2 is presented in section 3.2 below, while the proof of Theorem 1 is outlined in section S2 of the supplementary material.

Remark 2. One may wonder whether the second claim of Theorem 2 remains valid in the presence of outliers (that is, $\mathcal{O}_j > 0$). To the best of our knowledge, this is not the case. One possible path to constructing estimators that remain asymptotically normal in the presence of adversarial contamination is to consider an approach based on the gradient descent algorithm applied to the optimization problem (1.1), where the gradient $\nabla L(\theta_k)$ is robustly estimated on each iteration k; we refer the reader to the list of references investigating such methods and listed in section 1. Investigation of the asymptotic properties of such methods is an interesting direction for future research.

2.1 Computational aspects.

Here, we briefly discuss some of the more practical aspects of the proposed estimators, including the choice of the scaling factors Δ_n. Note that, while $\widehat{L}(\theta,\theta')$ itself is defined as a minimizer of a convex function, it is not a convex-concave function itself, and the problem (1.5) is not guaranteed to be convex-concave or have a unique solution. However, the gradient of $\widehat{L}(\theta,\theta')$, both with respect to θ and θ', is easily computable: as $\sum_{j=1}^{k}\rho'\left(\sqrt{n}\,\frac{\bar{L}_j(\theta)-\bar{L}_j(\theta')-\widehat{L}(\theta,\theta')}{\Delta_n}\right) = 0$, differentiating this expression yields that

$$\partial_\theta \widehat{L}(\theta,\theta') = \frac{\sum_{j=1}^{k}\partial_\theta \bar{L}_j(\theta)\rho''\left(\sqrt{n}\,\frac{\bar{L}_j(\theta)-\bar{L}_j(\theta')-\widehat{L}(\theta,\theta')}{\Delta_n}\right)}{\sum_{j=1}^{k}\rho''\left(\sqrt{n}\,\frac{\bar{L}_j(\theta)-\bar{L}_j(\theta')-\widehat{L}(\theta,\theta')}{\Delta_n}\right)}.$$

Due to this fact, gradient descent-ascent type methods for solving the problems closely related to (1.5) have been proposed and have shown good performance in extended simulation studies; we refer the reader to (Lecué and Lerasle, 2020; Mathieu and Minsker, 2021) for the details.

The problem of choosing the scaling factor for robust estimators of location has been studied since the seminal work of Huber (1964). Here, we suggest setting Δ_n in a data-dependent way using the "median absolute deviation" (MAD) estimator; this idea has been suggested and numerically tested in (Mathieu and Minsker, 2021). We start with $\Delta_n := \Delta_{n,0}$ being a fixed number (e.g., $\Delta_{n,0} = 1$). Given an approximate solution (θ_t, θ'_t), e.g., obtained via the gradient descent-ascent iteration, set $\widehat{M}(\theta_t,\theta'_t) :=$ median $\left(\bar{L}_1(\theta_t,\theta'_t),\ldots,\bar{L}_k(\theta_t,\theta'_t)\right)$, and

$$\mathrm{MAD}(\theta_t,\theta'_t) = \text{median}\left(\left|\bar{L}_1(\theta_t,\theta'_t)-\widehat{M}(\theta_t,\theta'_t)\right|,\ldots,\left|\bar{L}_k(\theta_t,\theta'_t)-\widehat{M}(\theta_t,\theta'_t)\right|\right).$$

Finally, define $\widehat{\Delta}_{n,t+1} := \frac{\mathrm{MAD}(\theta_t,\theta'_t)}{\Phi^{-1}(3/4)}$, where Φ is the distribution function of the standard normal law and the normalizing factor comes from the fact that for a sample from the normal distribution $N(\mu,\sigma^2)$, the expected value of MAD equals $\Phi^{-1}(3/4)\sigma$. The scaling factor can be updated again after a fixed number of iterations. Our theoretical results do not allow for a data-dependent choice of Δ_n however, and it would be an interesting avenue for further investigation. We include a simple proof-of-concept numerical simulation in section S8 of the supplementary material.

3. Proofs.

The proof of Theorem 2 uses characterization of $\widehat{\theta}_{n,k}$ as the solution of the min-max problem, and follows a standard pattern of consequently establishing consistency, rate of convergence and finally the asymptotic normality. The arguments are quite general and can be extended beyond the classes that satisfy Lipschitz property imposed by Assumption 3. Since $\widehat{L}(\theta_1,\theta_2)$ is defined implicitly as a solution of the convex minimization problem, we rely on the Bahadur-type linear representation of $\widehat{L}(\theta_1,\theta_2)-L(\theta_1,\theta_2)$ with uniform control of the remainder terms.

3.1 Preliminaries.

Below, we state several results that our proofs frequently rely upon.

Lemma 1. *Let $F : \mathbb{R} \mapsto \mathbb{R}$ be a function such that F'' is bounded and Lipschitz continuous. Moreover, suppose that $\xi_1, \ldots, \xi_n$ are independent centered random variables such that $\mathbb{E}|\xi_j|^2 < \infty$ for all j, and that Z_j, $j = 1, \ldots, n$ are independent with normal distribution $N(0, \mathrm{Var}(\xi_j))$. Then*

$$\left|\mathbb{E}F\left(\sum_{j=1}^n \xi_j\right) - \mathbb{E}F\left(\sum_{j=1}^n Z_j\right)\right| \leqslant C(F)\sum_{j=1}^n \mathbb{E}\left[\xi_j^2 \cdot \min(|\xi_j|, 1)\right].$$

In particular, if $\mathbb{E}|\xi_j|^{2+\tau} < \infty$ for some $\tau \in (0, 1]$ and all j, then

$$\left|\mathbb{E}F\left(\sum_{j=1}^n \xi_j\right) - \mathbb{E}F\left(\sum_{j=1}^n Z_j\right)\right| \leqslant C(F)\sum_{j=1}^n \mathbb{E}|\xi_j|^{2+\tau}.$$

The proof is given in section S3 of the supplementary material.

Lemma 2. *Let $\mathcal{F} = \left\{f_\theta,\ \theta \in \Theta' \subseteq \mathbb{R}^d\right\}$ be a class of functions that is Lipschitz in parameter, meaning that $|f_{\theta_1}(x) - f_{\theta_2}(x)| \leqslant M(x)\|\theta_1 - \theta_2\|$. Moreover, assume that $\mathbb{E}M^p(X) < \infty$ for some $p \geqslant 1$. Finally, suppose that $X_1, \ldots, X_n$ are i.i.d. Then*

$$\mathbb{E}\sup_{\theta_1,\theta_2\in\Theta'}\left(\frac{1}{\sqrt{n}}\left|\sum_{j=1}^n \left(f_{\theta_1}(X_j) - f_{\theta_2}(X_j) - P(f_{\theta_1} - f_{\theta_2})\right)\right|\right)^p$$

$$\leqslant C(p) d^{p/2} \mathrm{diam}^p(\Theta', \|\cdot\|) \mathbb{E}\|M\|^p_{L_2(P_n)}$$

and

$$\mathbb{E} \sup_{\theta\in\Theta'} \left(\frac{1}{\sqrt{n}} \left| \sum_{j=1}^n (f_\theta(X_j) - Pf_{\theta_1}) \right| \right)^p$$
$$\leqslant C(p) \left(d^{p/2} \mathrm{diam}^p(\Theta', \|\cdot\|) \mathbb{E}\|M\|^p_{L_2(P_n)} + \mathbb{E}^{1\wedge\frac{p}{2}} \left| f_{\theta_0}(X) - Pf_{\theta_0} \right|^{2\vee p} \right)$$

for any $\theta_0 \in \Theta'$.

The proof is outlined in section S4 of the supplementary material. The following result that can be viewed as a weak Bahadur representation of $\widehat{L}(\theta, \theta_0)$ is one of the key technical components that the proof of Theorem 2 relies on. Recall that $r(\theta_0) > 0$ is such that for all θ_1, $\theta_2 \in B(\theta, r(\theta_0))$, $|\ell(\theta_1, x) - \ell(\theta_2, x)| \leqslant \mathcal{V}(x; r(\theta_0)) \|\theta_1 - \theta_2\|$ (see the paragraph following Assumption 3 for more details).

Lemma 3. *Assume that adversarial contamination framework, and let* $\mathcal{O}$ *denote the number of outliers. Let* $\mathcal{L} = \{\ell(\theta, \cdot),\ \theta \in \Theta\}$ *be a class of functions, and, given* $\theta_0 \in \Theta$*, set* $\sigma^2(\delta) := \sup_{\|\theta-\theta_0\|\leqslant\delta} \mathrm{Var}\left(\ell(\theta, X) - \ell(\theta_0, X)\right)$. *Moreover, let Assumption 3 hold. Then for every* $\delta \leqslant r(\theta_0)$*, the following representation holds uniformly over* $\|\theta - \theta_0\| \leqslant \delta$:

$$\sqrt{N} \left(\widehat{L}(\theta, \theta_0) - L(\theta, \theta_0) \right)$$

$$= \frac{\Delta_n}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta,\theta_0) - L(\theta,\theta_0)\right)\right)} \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0) - L(\theta,\theta_0)\right)\right)$$

$$+ \mathcal{R}_{n,k}(\theta), \quad (3.1)$$

where

$$\sup_{\|\theta-\theta_0\|\leqslant\delta} |\mathcal{R}_{n,k}(\theta)| \leqslant C(d,\theta_0)\left(\delta^2 \frac{s^2}{\sqrt{k}} + \sqrt{k}\delta^3 + \frac{\mathcal{O}^2}{k^{3/2}}\right)$$

with probability at least $1 - \frac{3}{s}$.

The proof is contained in section S5 of the supplementary material.

3.2 Proof of Theorem 2.

As in the proof of Theorem 1, we will omit subscript j and write "k, n" instead of "k_j, n_j" to denote the increasing sequences of the number of subgroups and their cardinalities. The argument is divided into two steps. The first step consists in establishing the fact that the estimator $\widehat{\theta}_{n,k}$ converges to θ_0 at $\sqrt{N}$-rate, while on the second step we prove asymptotic normality by "zooming" to the resolution level $N^{-1/2}$; this proof pattern is quite standard in the empirical process theory (van der Vaart and Wellner, 1996).

Step one. Similar to the proof of Theorem 1, we set

$$\widehat{\theta}(\theta') := \operatorname*{argmax}_{\theta\in\Theta} \widehat{L}(\theta',\theta) = \operatorname*{argmin}_{\theta\in\Theta} \widehat{L}(\theta,\theta')$$

and define $\widehat{\theta}^{(1)}_{n,k} := \widehat{\theta}_{n,k}$ and $\widehat{\theta}^{(2)}_{n,k} := \widehat{\theta}(\widehat{\theta}^{(1)}_{n,k})$. We present a detailed argument establishing the convergence rate for $\widehat{\theta}^{(1)}_{n,k}$, and outline the modifications necessary to establish the result for $\widehat{\theta}^{(2)}_{n,k}$. Our goal can be equivalently stated as showing that

$$\lim_{M\to\infty}\limsup_{n,k\to\infty}\mathbb{P}\Big(\|\widehat{\theta}^{(1)}_{n,k}-\theta_0\|\geqslant 2^M\alpha_{n,k}\Big)=0. \tag{3.2}$$

Define

$$S_{N,j}:=\left\{\theta:\ 2^{j-1}\alpha_{n,k}<\|\theta-\theta_0\|\leqslant 2^j\alpha_{n,k}\right\},$$
$$\bar{S}_{N,j}:=\left\{\theta:\ 0\leqslant\|\theta-\theta_0\|\leqslant 2^j\alpha_{n,k}\right\},$$

and observe that

$$\|\widehat{\theta}^{(1)}_{n,k}-\theta_0\|\geqslant 2^M\alpha_{n,k}\implies\inf_{\theta\in S_{N,j}}\Big(\widehat{L}(\theta,\widehat{\theta}(\theta))-\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))\Big)\leqslant 0\text{ for some }j>M,$$

where $\widehat{\theta}(\theta'):=\operatorname{argmax}_{\theta\in\Theta}\widehat{L}(\theta',\theta)$. As $\widehat{L}(\theta,\widehat{\theta}(\theta))\geqslant\widehat{L}(\theta,\theta_0)$ for any θ, the inequality $\|\widehat{\theta}^{(1)}-\theta_0\|\geqslant 2^M\alpha_{n,k}$ implies that $\inf_{\theta\in S_{N,j}}\Big(\widehat{L}(\theta,\theta_0)-\widehat{L}(\theta_0,\widehat{\theta}(\theta_0)\Big)\leqslant 0$ for some $j>M$, which in turn entails that

$$\inf_{\theta\in S_{N,j}}\Big(\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)-\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))+L(\theta_0,\widehat{\theta}(\theta_0))\Big)$$
$$\leqslant L(\theta_0,\widehat{\theta}(\theta_0))-\inf_{\theta\in S_{N,j}}L(\theta,\theta_0)$$

for some $j>M$. Since $L(\theta_0,\widehat{\theta}(\theta_0))-\inf_{\theta\in S_{N,j}}L(\theta,\theta_0)\leqslant 0$ by the definition of θ_0, the previous display yields that

$$\sup_{\theta\in S_{N,j}}\left|\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)-\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))+L(\theta_0,\widehat{\theta}(\theta_0))\right| \\ \geqslant \inf_{\theta\in S_{N,j}} L(\theta,\theta_0)-L(\theta_0,\widehat{\theta}(\theta_0)) \geqslant \inf_{\theta\in S_{N,j}} L(\theta,\theta_0),$$

which further implies that either

$$\sup_{\theta\in S_{N,j}}\left|\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)\right| \geqslant \inf_{\theta\in S_{N,j}}\frac{L(\theta,\theta_0)}{2},$$

or $\left|\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))-L(\theta_0,\widehat{\theta}(\theta_0))\right| \geqslant \inf_{\theta\in S_{N,j}}\frac{L(\theta,\theta_0)}{2}$. Let $0<\eta_1\leqslant r(\theta_0)$ be small enough so that $L(\theta)-L(\theta_0)\geqslant c\|\theta-\theta_0\|^2$ for θ such that $\|\theta-\theta_0\|\leqslant\eta_1$ (existence of η_1 follows from Assumption 2), and observe that $\mathbb{P}\left(\|\widehat{\theta}^{(1)}_{n,k}-\theta_0\|\geqslant\eta_1\right)\to 0$ as $n,k\to\infty$ due to consistency of the estimator under assumptions of the theorem. We then have

$$\mathbb{P}\left(\|\widehat{\theta}^{(1)}_{n,k}-\theta_0\|\geqslant 2^M\alpha_{n,k}\right)\leqslant\mathbb{P}\left(\|\widehat{\theta}^{(1)}_{n,k}-\theta_0\|\geqslant\eta_1\right) \\ +\mathbb{P}\left(\left|\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))-L(\theta_0,\widehat{\theta}(\theta_0))\right|\geqslant c\,2^{2M}\alpha^2_{n,k}\right) \\ +\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ 2^j\alpha_{n,k}\leqslant\eta_1}\sup_{\theta\in S_{N,j}}\left|\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)\right|\geqslant c\,2^{2j-2}\alpha^2_{n,k}\right). \quad (3.3)$$

We will now estimate the second and third terms on the right-hand side of the display above, starting with the third term.

- **Estimating** $\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ 2^j\alpha_{n,k}\leqslant\eta_1}\sup_{\theta\in S_{N,j}}\left|\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)\right|\geqslant c\,2^{2j-2}\alpha^2_{n,k}\right)$.

Let us invoke Lemma 3 applied to the class $\left\{\ell(\theta,\cdot)-\ell(\theta_0,\cdot),\ \theta\in\bar{S}_{N,j}\right\}$. Together with the union bound applied over $M<j\leqslant J_{\max}:=\lfloor\log(\sqrt{N}\eta_1)\rfloor+1$

with $s_j := j^2$, it implies that for all $\theta \in S_{N,j}$, $M+1 \leqslant j \leqslant J_{\max}$,

$$\sqrt{N}\left(\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right) = \frac{\Delta_n}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - \bar{L}_1(\theta_0) - L(\theta,\theta_0)\right)\right)} \\ \times \frac{1}{\sqrt{k}}\sum_{i=1}^{k}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_i(\theta) - \bar{L}_i(\theta_0) - L(\theta,\theta_0)\right)\right) + \mathcal{R}_{n,k,j}(\theta), \quad (3.4)$$

where

$$\sup_{\theta \in \bar{S}_{N,j}} |\mathcal{R}_{n,k,j}(\theta)| \leqslant C(d,\theta_0)\left(\frac{2^{2j}}{N}\frac{j^4}{\sqrt{k}} + \sqrt{k}\frac{2^{3j}}{N^{3/2}} + \frac{\mathcal{O}^2}{k^{3/2}}\right)$$

uniformly over all $M \leqslant j \leqslant J_{\max}$ with probability at least $1 - 3\sum_{j:j\geqslant M+1} j^{-2} \geqslant 1 - \frac{C}{M}$. Let $\mathcal{E}$ denote the event of probability at least $1 - \frac{C}{M}$ on which the previous representation holds. Moreover, observe that, in view of Lemma 1, for η_1 small enough and N large enough,

$$\sup_{\|\theta-\theta_0\| \leqslant \eta_1}\left|\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - \bar{L}_1(\theta_0) - L(\theta,\theta_0)\right)\right) - \rho''(0)\right| \leqslant \frac{\rho''(0)}{2} = \frac{1}{2}.$$

Taking this fact into account and noting that (i) $\frac{2^j}{\sqrt{N}}\frac{j^4}{\sqrt{k}} + \sqrt{k}\frac{2^{2j}}{N} \leqslant \tilde{c}2^j$ for any $j \leqslant J_{\max}$ and any $\tilde{c} > 0$ given that n is large enough and that the relation (ii) $\frac{\mathcal{O}^2}{k^{3/2}} = o(\alpha_{n,k}^2\sqrt{N})$ follows from assumptions of the theorem, we see that the remainder term $\mathcal{R}_{n,k,j}(\theta)$ is smaller than $\tilde{c}2^{2j}\left(\frac{1}{\sqrt{N}} + \alpha_{n,k}^2\sqrt{N}\right)$ on event $\mathcal{E}$, hence

$$\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ \frac{2^j}{\sqrt{N}}\leqslant\eta_1}\ \sup_{\theta\in S_{N,j}}\left|\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right| \geqslant c\,2^{2j-2}\alpha_{n,k}^2\right) \leqslant \frac{C}{M}$$

$$+ \sum_{j:j\geqslant M+1,\ 2^j\alpha_{n,k}\leqslant \eta_1} \mathbb{P}\left(\sup_{\theta\in S_{N,j}} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_i(\theta) - \bar{L}_i(\theta_0) - L(\theta,\theta_0)\right)\right)\right| \geqslant c_1 2^{2j}\alpha_{n,k}^2\sqrt{N} \right)$$

where we used the fact that whenever $\tilde{c}$ is small enough,

$$c2^{2j-2}\alpha_{n,k}^2 - \tilde{c}2^{2j}\left(\frac{1}{N} + \alpha_{n,k}^2\right) \geqslant c_1 2^{2j}\alpha_{n,k}^2 \text{ for } c_1 > 0.$$

Invoking Lemma 1 again, we see that (assuming that $\bar{L}_1(\cdot)$ is based on a contamination-free sample)

$$\sup_{\theta\in\bar{S}_{N,j}} \left| \mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - \bar{L}_1(\theta_0) - L(\theta,\theta_0)\right)\right)\right| \leqslant C\frac{2^{2j}}{N}.$$

Let us denote $\rho'_{n,i}(\theta,\theta_0) = \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_i(\theta) - \bar{L}_i(\theta_0) - L(\theta,\theta_0)\right)\right)$, $i = 1,\ldots,k$ for brevity. Moreover, let $\tilde{\rho}'_{n,i}(\theta,\theta_0)$ be a version of $\rho'_{n,i}(\theta,\theta_0)$ based on a contamination-free i.i.d. sample $\tilde{X}_1,\ldots,\tilde{X}_N$ such that $\tilde{X}_j = X_j$ for $j\notin J$ where $J\subset\{1,\ldots,N\}$ contains the indices of the outliers among $X_1,\ldots,X_N$. As (i) $\left|\frac{1}{\sqrt{k}}\sum_{i=1}^k(\rho'_{n,i}(\theta,\theta_0) - \tilde{\rho}'_{n,i}(\theta,\theta_0))\right| \leqslant 2\|\rho'\|_\infty\frac{\mathcal{O}}{\sqrt{k}}$, (ii) $\frac{\mathcal{O}}{\sqrt{k}} \ll \alpha_{n,k}^2\sqrt{N}$ by assumption, and (iii) $\sqrt{k}\frac{2^{2j}}{N} \leqslant c_2\frac{2^{2j}}{\sqrt{N}} \leqslant c_2 2^{2j}\alpha_{n,k}^2\sqrt{N}$ for any $c_2 > 0$ and sufficiently large n, it is easy to check that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\theta\in S_{N,j}} \left| \frac{1}{\sqrt{k}}\sum_{i=1}^{k}\rho'_{n,i}(\theta,\theta_0)\right| \geqslant c_1 2^{2j}\alpha_{n,k}^2\sqrt{N}\right) \\
&\quad\leqslant \mathbb{P}\left(\sup_{\theta\in S_{N,j}} \left| \frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left(\tilde{\rho}'_{n,i}(\theta,\theta_0) - \mathbb{E}\tilde{\rho}'_{n,i}(\theta,\theta_0)\right)\right| \geqslant c_2 2^{2j}\alpha_{n,k}^2\sqrt{N}\right) \\
&\qquad\leqslant \frac{1}{c_2 2^{2j}\alpha_{n,k}^2\sqrt{N}}\mathbb{E}\sup_{\theta\in S_{N,j}}\left| \frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left(\tilde{\rho}'_{n,i}(\theta,\theta_0) - \mathbb{E}\tilde{\rho}'_{n,i}(\theta,\theta_0)\right)\right|
\end{aligned}$$

where we used Markov's inequality on the last step. To bound the expected supremum, we proceed in exactly the same fashion using symmetrization, contraction and desymmetrization inequalities as in the proof of Lemma 3 (see the supplementary material), and deduce that

$$\mathbb{E}\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left(\tilde\rho'_{n,i}(\theta,\theta_0)-\mathbb{E}\tilde\rho'_{n,i}(\theta,\theta_0)\right)\right| \\ \leqslant \frac{C}{\Delta_n}\mathbb{E}\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(\ell(\theta,\tilde X_j)-\ell(\theta_0,\tilde X_j)-L(\theta,\theta_0)\right)\right|.$$

The right side of the display above can be bounded by $\frac{C(d,\theta_0)}{\Delta_n}\frac{2^j}{\sqrt{N}}$ (using Lemma 2), implying that

$$\mathbb{P}\left(\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left(\tilde\rho'_{n,i}(\theta,\theta_0)-\mathbb{E}\tilde\rho'_{n,i}(\theta,\theta_0)\right)\right|\geqslant c_2 2^{2j}\alpha^2_{n,k}\sqrt{N}\right)\leqslant\frac{C_1(d,\theta_0)}{\Delta_n}\frac{1}{2^j},$$

where we used the fact that $\alpha^2_{n,k}\geqslant\frac{1}{N}$. Therefore,

$$\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ \frac{2^j}{\sqrt{N}}\leqslant\eta_1}\sup_{\theta\in S_{N,j}}\left|\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)\right|\geqslant c\,2^{2j-2}\alpha^2_{n,k}\right) \\ \leqslant\frac{C}{M}+\frac{C_1(d,\theta_0)}{\Delta_n}\sum_{j\geqslant M}2^{-j}\leqslant\frac{C}{M}+\frac{C_1(d,\theta_0)}{\Delta_n}2^{-M+1}\to 0\text{ as }M\to\infty$$

whenever n,k are large enough.

- **Estimating** $\mathbb{P}\left(\left|\widehat{L}(\theta_0,\widehat\theta(\theta_0))-L(\theta_0,\widehat\theta(\theta_0))\right|\geqslant c\,2^{2M}\alpha^2_{n,k}\right)$.

In view of (3.3), it only remains to show that

$$\mathbb{P}\left(\sqrt{N}\left|\widehat{L}(\theta_0,\widehat\theta(\theta_0))-L(\theta_0,\widehat\theta(\theta_0))\right|\geqslant c\,2^{2M}\alpha^2_{n,k}\right)\to 0\text{ as }n,k\to\infty. \quad (3.5)$$

To this end, it suffices to repeat the argument presented above, with several simplifications. First, we will start by proving that

$$\lim_{M\to\infty}\limsup_{n,k\to\infty}\mathbb{P}\Big(\|\widehat{\theta}(\theta_0)-\theta_0\|\geqslant 2^M\alpha_{n,k}\Big)=0.$$

We have already shown in the course of the proof of Theorem 1 that $\widehat{\theta}(\theta_0)$ is a consistent estimator of θ_0, so that $\mathbb{P}\Big(\|\widehat{\theta}(\theta_0)-\theta_0\|\geqslant\eta_2\Big)\to 0$ for any $\eta_2>0$. If $\|\widehat{\theta}(\theta_0)-\theta_0\|\geqslant 2^M\alpha_{n,k}^2$, then $\widehat{\theta}(\theta_0)\in S_{N,j}$ for some $j>M$, implying that $\sup_{\theta\in S_{N,j}}\widehat{L}(\theta_0,\theta)\geqslant\widehat{L}(\theta_0,\theta_0)=0$, which entails the inequality $\sup_{\theta\in S_{N,j}}\Big(\widehat{L}(\theta_0,\theta)-L(\theta_0,\theta)\Big)\geqslant-\sup_{\theta\in S_{N,j}}L(\theta_0,\theta)=\inf_{\theta\in S_{N,j}}L(\theta,\theta_0)\geqslant c\,2^{2j-2}\alpha_{n,k}^2$ whenever $2^j\alpha_{n,k}\leqslant\eta_2$ and η_2 is small enough. Therefore,

$$\begin{aligned}\mathbb{P}\Big(\|\widehat{\theta}(\theta_0)-\theta_0\|\geqslant 2^M\alpha_{n,k}\Big)&\leqslant\mathbb{P}\Big(\|\widehat{\theta}(\theta_0)-\theta_0\|\geqslant\eta_2\Big)\\&\quad+\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ 2^j\alpha_{n,k}\leqslant\eta_2}\sup_{\theta\in S_{N,j}}\Big|\widehat{L}(\theta_0,\theta)-L(\theta_0,\theta)\Big|\geqslant c\,2^{2j-2}\alpha_{n,k}^2\right).\end{aligned}$$

The probability of the union is estimated as before using Lemma 3, implying that it converges to 0 as $M\to\infty$. To complete the proof of (3.5), observe that

$$\begin{aligned}\mathbb{P}\Big(\Big|\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))-L(\theta_0,\widehat{\theta}(\theta_0))\Big|>c\,2^{2M}\alpha_{n,k}^2\Big)&\leqslant\mathbb{P}\Big(\|\widehat{\theta}(\theta_0)-\theta_0\|\geqslant 2^M\alpha_{n,k}\Big)\\&\quad+\mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant 2^M\alpha_{n,k}}\Big|\widehat{L}(\theta_0,\theta)-L(\theta_0,\theta)\Big|\geqslant c\,2^{2M}\alpha_{n,k}^2\right)\end{aligned}$$

and that

$$\mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant 2^M\alpha_{n,k}}\Big|\widehat{L}(\theta_0,\theta)-L(\theta_0,\theta)\Big|\geqslant c\,2^{2M}\alpha_{n,k}^2\right)\leqslant\frac{C}{M}+\frac{C(d,\theta_0)}{\Delta_n}2^{-M}\to 0$$

as $M \to \infty$, which follows from the representation (3.4) in the same fashion as before. This completes the proof of relation (3.2). To establish that

$$\lim_{M\to\infty} \limsup_{n,k\to\infty} \mathbb{P}\Big(\|\widehat{\theta}^{(2)}_{n,k} - \theta_0\| \geqslant 2^M \alpha_{n,k}\Big) = 0,$$

we begin by observing that the inequality $\|\widehat{\theta}^{(2)}_{n,k} - \theta_0\| \geqslant 2^M \alpha_{n,k}$ implies that $\sup_{\theta\in S_{N,j}} \widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta) \geqslant \widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta_0)$ for some $j > M$. If $2^j \alpha_{n,k} \leqslant \eta_3$ for sufficiently small constant $\eta_3 > 0$, we see that it further entails the inequality

$$\begin{aligned}
&\sup_{\theta\in S_{N,j}} \Big(\widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta) - L(\widehat{\theta}^{(1)}_{n,k}, \theta) - \widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta_0) + L(\widehat{\theta}^{(1)}_{n,k}, \theta_0)\Big) \\
&\qquad \geqslant - \sup_{\theta\in S_{N,j}} L(\widehat{\theta}^{(1)}_{n,k}, \theta) + L(\widehat{\theta}^{(1)}_{n,k}, \theta_0) = \inf_{\theta\in S_{N,j}} L(\theta, \widehat{\theta}^{(1)}_{n,k}) + L(\widehat{\theta}^{(1)}_{n,k}, \theta_0) \\
&\qquad\qquad = \inf_{\theta\in S_{N,j}} L(\theta, \theta_0) \geqslant c\, 2^{2j-2} \alpha^2_{n,k}.
\end{aligned}$$

We deduce from the display above that

$$\begin{aligned}
\mathbb{P}\Big(\|\widehat{\theta}^{(2)}_{n,k} - \theta_0\| \geqslant 2^M \alpha_{n,k}\Big) &\leqslant \mathbb{P}\Big(\|\widehat{\theta}^{(2)}_{n,k} - \theta_0\| \geqslant \eta_3\Big) + \mathbb{P}\Big(\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| \geqslant 2^M \alpha^2_{n,k}\Big) \\
&+ \mathbb{P}\left(\bigcup_{j: j\geqslant M+1,\ 2^j \alpha_{n,k} \leqslant \eta_3} \sup_{\theta\in S_{N,j}, \theta'\in \bar{S}_{N,M/2}} \Big|\widehat{L}(\theta', \theta) - L(\theta', \theta)\Big| \geqslant c_1\, 2^{2j-2} \alpha^2_{n,k}\right) \\
&\qquad + \mathbb{P}\left(\sup_{\theta\in \bar{S}_{N,M/2}} \Big|\widehat{L}(\theta, \theta_0) - L(\theta, \theta_0)\Big| \geqslant c_1\, 2^{2M} \alpha_{n,k}\right).
\end{aligned}$$

We have shown before that the first and second term on the right side of the previous display converge to 0 as M, n and k tend to infinity, while the last term converges to 0 in view of argument presented previously in detail

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0268

(see representation (3.4) and the bounds that follow). It remains to estimate $\mathbb{P}\Big(\bigcup_{j:j\geqslant M+1,\ 2^j\alpha_{n,k}\leqslant\eta_3}\sup_{\theta\in S_{N,j},\theta'\in\bar{S}_{N,M/2}}\Big|\widehat{L}(\theta',\theta)-L(\theta',\theta)\Big|\geqslant c_1\,2^{2j-2}\alpha_{n,k}^2\Big)$. To this end, we again invoke Lemma 3 applied to the class

$$\left\{\ell(\theta_1,\cdot)-\ell(\theta_2,\cdot),\ \theta_1\in\bar{S}_{N,M/2},\theta_2\in\bar{S}_{N,j}\right\}.$$

Here, the "reference point" is (θ_0,θ_0). Since

$$|\ell(\theta,x)-\ell(\theta',x)|\leqslant V(x;r(\theta_0))(2^j+2^{M/2})\alpha_{n,k},$$

it is easy to see that $\sigma^2(\delta)\leqslant\mathbb{E}M^2_{\theta_0}(X)\left(2^{2j}+2^M\right)\alpha_{n,k}^2\leqslant C(\theta_0)2^{2j}\alpha^2_{n,k}$, and to deduce that

$$\begin{aligned}&\sqrt{N}\left(\widehat{L}(\theta',\theta)-L(\theta',\theta)\right)\\&=\frac{\Delta_n}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta')-\bar{L}_1(\theta)-L(\theta',\theta)\right)\right)}\frac{1}{\sqrt{k}}\sum_{i=1}^k\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_i(\theta')-\bar{L}_i(\theta)-L(\theta',\theta)\right)\right)\\&\qquad\qquad\qquad\qquad\qquad\qquad+\mathcal{R}_{n,k,j}(\theta',\theta),\end{aligned}$$

where

$$\sup_{\theta\in\bar{S}_{N,j},\theta'\in\bar{S}_{N,M/2}}|\mathcal{R}_{n,k,j}(\theta',\theta)|\leqslant C(d,\theta_0)\left(\frac{2^{2j}}{N}\frac{j^4}{\sqrt{k}}+\sqrt{k}\frac{2^{3j}}{N^{3/2}}+\frac{\mathcal{O}^2}{k^{3/2}}\right)$$

uniformly over all $M\leqslant j\leqslant J_{\max}$ with probability at least $1-\frac{C}{M}$. The remaining steps again closely mimic the argument outlined in detail after display (3.4) and yield that

$$\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ 2^j\alpha_{n,k}\leqslant\eta_3}\ \sup_{\theta\in S_{N,j},\theta'\in\bar{S}_{N,M/2}}\left|\widehat{L}(\theta',\theta)-L(\theta',\theta)\right|\geqslant c_1\,2^{2j-2}\alpha_{n,k}^2\right)$$
$$\leqslant\frac{C(d,\theta_0)}{\Delta_n}2^{-M+1}\to 0$$

as $M\to\infty$, therefore implying the last claim in the first part of the proof.

Step two. Now we are ready to establish the asymptotic normality of $\widehat{\theta}^{(1)}_{n,k}$ and $\widehat{\theta}^{(2)}_{n,k}$. To this end, observe that the first claim of the theorem holds with $\alpha_{n,k}=\frac{1}{\sqrt{nk}}=\frac{1}{\sqrt{N}}$, and consider the stochastic process $M_N(h,q)$ indexed by $h,q\in\mathbb{R}^d$ and defined via

$$M_N(h,q):=N\left(\widehat{L}(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N}))-L(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N})\right).$$

Below, we will show that $M_N(h,q)$ converges weakly to the Gaussian process $W(h,q):=W^T(h-q)$, $h,q\in\mathbb{R}^d$, where $W\sim N(0,\Sigma_W)$ and $\Sigma_W=\mathbb{E}\left[\partial_\theta\ell(\theta_0,X)\partial_\theta\ell(\theta_0,X)^T\right]$. Let us deduce the conclusion assuming that weak convergence has already been established. We have that

$$N\cdot\widehat{L}(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N}))=N\cdot L(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N})+M_N(h,q).$$

Note that, in view of Assumption 2 and the fact that θ_0 minimizes $L(\theta_0)$,

$$N\cdot L(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N})\to\frac{1}{2}h^T\partial^2_\theta L(\theta_0)h-\frac{1}{2}q^T\partial^2_\theta L(\theta_0)q \text{ as } N\to\infty,$$

therefore

$$N\cdot\widehat{L}(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N}))\xrightarrow{d}W^Th+\frac{1}{2}h^T\partial^2_\theta L(\theta_0)h-\left(W^Tq-\frac{1}{2}q^T\partial^2_\theta L(\theta_0)q\right).$$

It is easy to see that

$$\left(-\left[\partial_\theta^2 L(\theta_0)\right]^{-1} W, -\left[\partial_\theta^2 L(\theta_0)\right]^{-1} W\right)$$
$$= \operatorname*{argmin}_h \max_q W^T h + \frac{1}{2} h^T \partial_\theta^d L(\theta_0) h - \left(W^T q - \frac{1}{2} q^T \partial_\theta^d L(\theta_0) q\right),$$

where $-\left[\partial_\theta^2 L(\theta_0)\right]^{-1} W \sim N\left(0, \left[\partial_\theta^2 L(\theta_0)\right]^{-1} \Sigma_W \left[\partial_\theta^2 L(\theta_0)\right]^{-1}\right)$. Therefore, since

$$\left(\sqrt{N}\left(\widehat{\theta}^{(1)}_{n,k} - \theta_0\right), \sqrt{N}\left(\widehat{\theta}^{(2)}_{n,k} - \theta_0\right)\right) = \operatorname*{argmin}_h \max_q \widehat{L}(\theta_0 + h/\sqrt{N}, \theta_0 + q/\sqrt{N})),$$

continuous mapping theorem yields the desired conclusion. Next, we will establish the required weak convergence.

- **Establishing weak convergence.** To this end, we apply Lemma 3 to the class

$$\widetilde{\mathcal{L}}_N := \left\{ \widetilde{\ell}_N(h, q, \cdot) := \ell(\theta_0 + h/\sqrt{N}, \cdot) - \ell(\theta_0 + q/\sqrt{N}, \cdot), \left\| \begin{pmatrix} h \\ q \end{pmatrix} \right\| \leqslant R \right\}, \tag{3.6}$$

and note that $\left\| \begin{pmatrix} \theta_0 + h/\sqrt{N} \\ \theta_0 + q/\sqrt{N} \end{pmatrix} - \begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix} \right\| \leqslant \frac{R}{\sqrt{N}}$. We will also introduce the following notation for brevity (that will be used only in this part of the proof):

$$\bar{L}_j(h, q) := \frac{1}{n} \sum_{i \in G_j} \widetilde{\ell}_N(h, q, X_i), \quad \widetilde{L}(h, q) := \mathbb{E}\widetilde{\ell}_N(h, q, X). \tag{3.7}$$

The quantities δ and $\sigma^2(\delta)$ defined in Lemma 3 admit the bounds $\delta \leqslant \frac{R}{\sqrt{N}}$ and, in view of Assumption 3,

$$\sigma^2(\delta) := \sup_{\|(h,q)^T\|\leqslant R} \operatorname{Var}\left(\widetilde{\ell}_N(h,q,X)\right) \leqslant 2\mathbb{E}V^2(X;r(\theta_0))\frac{R^2}{N}, \tag{3.8}$$

hence Lemma 3 yields that

$$M_N(h,q) = \frac{\Delta_n}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(h,q)-\widetilde{L}(h,q)\right)\right)}\frac{\sqrt{N}}{\sqrt{k}}\sum_{j=1}^{k}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q)-\widetilde{L}(h,q)\right)\right) + o_P(1)$$

uniformly over $\left\|(h,q)^T\right\| \leqslant R$. In view of Assumption 1,

$$\mathbb{P}\left(\left|\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q)-\widetilde{L}(h,q)\right)\right| \leqslant 1\right) \leqslant \mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(h,q)-\widetilde{L}(h,q)\right)\right) \leqslant 1.$$

As $\sup_{\|(h,q)^T\|\leqslant R}\mathbb{P}\left(\left|\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q)-\widetilde{L}(h,q)\right)\right| \geqslant 1\right) \leqslant \sup_{\|(h,q)^T\|\leqslant R}\frac{\operatorname{Var}\left(\widetilde{\ell}(h,q,X)\right)}{\Delta_n^2} \to 0$ as $n,k\to\infty$, we deduce that $\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(h,q)-\widetilde{L}(h,q)\right)\right)\to 1$ and

$$M_N(h,q) = \Delta_n\frac{\sqrt{N}}{\sqrt{k}}\sum_{j=1}^{k}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q)-\widetilde{L}(h,q)\right)\right) + o_P(1). \tag{3.9}$$

It remains to establish convergence of the finite dimensional distributions as well as asymptotic equicontinuity. Convergence of finite dimensional distributions will be deduced from Lindeberg-Feller's central limit theorem. As $\rho'(x) = x$ for $|x|\leqslant 1$ by Assumption 1,

$$\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q)-\widetilde{L}(h,q)\right)\right) = \frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q)-\widetilde{L}(h,q)\right)$$

on the event $\mathcal{C}_j := \left\{\left|\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right)\right| \leqslant 1\right\}$. Chebyshev's inequality and Assumption 3 imply that

$$\begin{aligned}\mathbb{P}\left(\bar{\mathcal{C}}_j\right) \leqslant \mathrm{Var}&\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right)\right) \\ &\leqslant \frac{\mathbb{E}\widetilde{\ell}^2(h,q,X)}{\Delta_n^2} \leqslant \frac{\mathbb{E}\mathcal{V}^2(X;r(\theta_0))\|h-q\|^2}{\Delta_n^2 N},\end{aligned}$$

therefore, $\mathbb{P}\left(\bigcup_{j=1}^k \bar{\mathcal{C}}_j\right) \leqslant \frac{\mathbb{E}\mathcal{V}^2(X;r(\theta_0))\|h-q\|^2}{\Delta_n^2 n} \to 0$ as $n \to \infty$, and

$$\begin{aligned}M_N(h,q) &= \Delta_n \frac{\sqrt{N}}{\sqrt{k}} \sum_{j=1}^{k} \frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right) + o_P(1) \\ &= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \sqrt{N}\left(\widetilde{\ell}_N(h,q,X_j) - \widetilde{L}(h,q)\right) + o_P(1)\end{aligned}$$

on the event $\bigcap_{j=1}^k \mathcal{C}_j$. Hence, the limits of the finite dimensional distributions of the processes $M_N(h,q)$ and

$$\widehat{M}_N(h,q) := \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \sqrt{N}\left(\widetilde{\ell}_N(h,q,X_j) - \widetilde{L}(h,q)\right)$$

coincide. It is easy to conclude from the Lindeberg-Feller's theorem that the finite dimensional distributions of the process $(h,q) \mapsto \widehat{M}_N(h,q)$ are Gaussian, with covariance function

$$\begin{aligned}\lim_{N\to\infty} \mathrm{cov}&\left(\widehat{M}_N(h_1,q_1), \widehat{M}_N(h_2,q_2)\right) \\ &= (h_1 - q_1)^T\, \mathbb{E}\left[\partial_\theta \ell(\theta_0,X)\left(\partial_\theta \ell(\theta_0,X)\right)^T\right](h_2 - q_2)\,, \quad (3.10)\end{aligned}$$

Indeed, the aforementioned relation follows from the dominated convergence theorem, where pointwise convergence and the “domination” hold due to Assumption 3. Lindeberg's condition is also easily verified, as $\left(\sqrt{N}\widetilde{\ell}_N(h,q,X)\right)^2 \leqslant \mathcal{V}^2(X;r(\theta_0))\|h-q\|^2$, implying that the sequence $\left\{\left(\sqrt{N_j}\,\widetilde{\ell}_{N_j}(h,q,X)\right)^2\right\}_{j\geqslant 1}$ is uniformly integrable, where $N_j = n_j \cdot k_j$.

Finally, we will establish the asymptotic equicontinuity of the process $M_N(h,q)$. To this end, it suffices to prove that for any $\varepsilon > 0$,

$$\lim_{\delta\to 0}\limsup_{n,k\to\infty}\mathbb{P}\left(\sup_{\|(h_1,q_1)^T-(h_2,q_2)^T\|\leqslant\delta}|M_N(h_1,q_1)-M_N(h_2,q_2)|\geqslant\varepsilon\right)\to 0,$$

which would follow, in view of Lemma 3, from the relation

$$\lim_{\delta\to 0}\limsup_{n,k\to\infty}\mathbb{E}\sup_{\|(h_1,q_1)^T-(h_2,q_2)^T\|\leqslant\delta}\left|\Delta_n\frac{\sqrt{N}}{\sqrt{k}}\sum_{j=1}^{k}\left(\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h_1,q_1)-\widetilde{L}(h_1,q_1)\right)\right)\right.\right.$$
$$\left.\left. - \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h_2,q_2)-\widetilde{L}(h_2,q_2)\right)\right)\right)\right| = 0. \quad (3.11)$$

To estimate the expected supremum in (3.11), we first observe that for any h,q,

$$\sqrt{Nk}\left|\mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(h,q)-\widetilde{L}(h,q)\right)\right)\right| = o(1) \qquad (3.12)$$

as $k,n\to\infty$ by Lemma 1 and inequality (3.8). Therefore, we only need to show that

$$\limsup_{n,k\to\infty}\mathbb{E}\sup_{\|(h_1,q_1)^T-(h_2,q_2)^T\|\leqslant\delta}|M_N(h_1,q_1)-M_N(h_2,q_2)-$$
$$(\mathbb{E}M_N(h_1,q_1)-\mathbb{E}M_N(h_2,q_2))| \xrightarrow{\delta\to 0} 0.$$

Next, we will apply symmetrization inequality with Gaussian weights (van der Vaart and Wellner, 1996). Specifically, let $g_1, \ldots, g_k$ be i.i.d. $N(0,1)$ random variables independent of the data $X_1, \ldots, X_N$. Then, setting $B(\delta) := \left\{(h_1, q_1), (h_2, q_2) : \ \|(h_1, q_1)^T - (h_2, q_2)^T\| \leqslant \delta\right\}$, we have that

$$\mathbb{E} \sup_{B(\delta)} \left| M_N(h_1, q_1) - M_N(h_2, q_2) - (\mathbb{E} M_N(h_1, q_1) - \mathbb{E} M_N(h_2, q_2)) \right| \leqslant \\ C(\rho) \Delta_n \mathbb{E} \sup_{B(\delta)} \left| \frac{\sqrt{N}}{\sqrt{k}} \sum_{j=1}^{k} g_j \left(\rho' \left(\frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(h_1, q_1) - \widetilde{L}(h_1, q_1) \right) \right) \right.\right. \\ \left.\left. - \rho' \left(\frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(h_2, q_2) - \widetilde{L}(h_2, q_2) \right) \right) \right) \right|.$$

Let us condition everything on $X_1, \ldots, X_N$; we will write $\mathbb{E}_g$ to denote the expectation with respect to $g_1, \ldots, g_k$ only. Consider the Gaussian process $Y_{n,k}(t)$ defined via $\mathbb{R}^k \ni t \mapsto Y_{n,k}(t) := \frac{1}{\sqrt{k}} \sum_{j=1}^{k} g_j \sqrt{N} \rho'(t_j)$, where

$$t_j := t_j(h, q) = \frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(h, q) - \widetilde{L}(h, q) \right), \ j = 1, \ldots, k.$$

In what follows, we will rely on the ideas behind the proof of Theorem 2.10.6 in van der Vaart and Wellner (1996). Let us partition the set $\{(h, q) : \|(h, q)\| \leqslant R\}$ into the subsets $S_j, \ j = 1, \ldots, N(\delta)$ of diameter at most δ with respect to the Euclidean distance $\|\cdot\|$, and let $t^{(j)} := t^{(j)}(h^{(j)}, q^{(j)}) \in S_j \ j = 1, \ldots, N(\delta)$ be arbitrary points; we also note that $N(\delta) \leqslant \left(\frac{6R}{\delta}\right)^{2d}$. Next, set $T^{(j)} := \{t(h, q) : \ (h, q) \in S_j\}$. Our goal will be to show that

$$\limsup_{n,k \to \infty} \mathbb{E} \max_{j=1,\ldots,N(\delta)} \sup_{t \in T^{(j)}} \left| Y_{n,k}(t) - Y_{n,k}(t^{(j)}) \right| \to 0 \text{ as } \delta \to 0,$$

whence the desired conclusion would follow from Theorem 1.5.6 in van der Vaart and Wellner (1996). By Lemma 2.10.16 in van der Vaart and Wellner (1996),

$$\begin{aligned}\mathbb{E}_g \max_{j=1,\dots,N(\delta)} \sup_{t\in T^{(j)}} \left|Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right| \\ \leqslant C\Bigg(\max_{j=1,\dots,N(\delta)} \mathbb{E}_g \sup_{t\in T^{(j)}} \left|Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right| \\ + \sqrt{\log N(\delta)} \max_{1\leqslant j\leqslant N(\delta)} \sup_{t\in T^{(j)}} \mathrm{Var}_g^{1/2}\left(Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right)\Bigg). \quad (3.13)\end{aligned}$$

Observe that $\mathrm{Var}_g\left(Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right) = \frac{N}{k}\sum_{i=1}^k \left(\rho'(t_i) - \rho'(t_i^{(j)})\right)^2$, hence

$$\begin{aligned}\mathbb{E} \max_{1\leqslant j\leqslant N(\delta)} \sup_{t\in T^{(j)}} \mathrm{Var}_g^{1/2}\left(Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right) &\leqslant \mathbb{E}^{1/2} \sup_{t^{(1)},t^{(2)}} \frac{N}{k}\sum_{i=1}^k \left(\rho'(t_i^{(1)}) - \rho'(t_i^{(2)})\right)^2 \\ &\leqslant \sqrt{N} L(\rho')\mathbb{E}^{1/2} \sup_{t^{(1)},t^{(2)}} \left(t_1^{(1)} - t_1^{(2)}\right)^2 \\ = L(\rho')\mathbb{E}^{1/2} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} &\Bigg(\frac{\sqrt{nN}}{\Delta_n}\Big(\bar{L}_1(h_1,q_1) - \bar{L}_1(h_2,q_2) \\ &\qquad - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2)\Big)\Bigg)^2,\end{aligned}$$

where the supremum is taken over all $t^{(1)}(h_1,q_1)$, $t^{(2)}(h_2,q_2)$ such that $\|(h_1,q_1)-(h_2,q_2)\| \leqslant \delta$. To estimate the last expected supremum, we invoke Lemma 2 with $f_{h,q}(X) := \ell(\theta_0 + h/\sqrt{N}, X) - \ell(\theta_0 + q/\sqrt{N}, X)$, noting that, in view of Assumption 3,

$$\sqrt{N}|f_{h_1,q_1}(X) - f_{h_2,q_2}(X)| \leqslant \mathcal{V}(X; r(\theta_0))\left(\|h_1 - h_2\| + \|q_1 - q_2\|\right)$$

$$\leqslant 2\mathcal{V}(X;r(\theta_0))\left\|(h_1,q_1)-(h_2,q_2)\right\|. \quad (3.14)$$

Therefore,

$$\mathbb{E}^{1/2}\sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta}\left(\frac{\sqrt{nN}}{\Delta_n}\left(\bar{L}_1(h_1,q_1)-\bar{L}_1(h_2,q_2)-(\widetilde{L}(h_1,q_1)-\widetilde{L}(h_2,q_2)\right)\right)^2$$
$$\leqslant C\sqrt{d}\mathbb{E}^{1/2}\mathcal{V}^2(X;r(\theta_0))\cdot\delta,$$

yielding that the second term on the right side of (3.13) converges in probability to 0 as $\delta \to 0$. It remains to show that the first term

$$\max_{j=1,\dots,N(\delta)}\mathbb{E}_g\sup_{t\in T^{(j)}}\left|Y_{n,k}(t)-Y_{n,k}(t^{(j)})\right|$$

converges to 0 in probability. As ρ' is Lipschitz continuous, the covariance function of $Y_{n,k}(t)$ satisfies

$$\mathbb{E}\left(Y_{n,k}(t^{(1)})-Y_{n,k}(t^{(2)})\right)^2\leqslant L^2(\rho')\frac{N}{k}\sum_{j=1}^{k}\left(t_j^{(1)}-t_j^{(2)}\right)^2,$$

where the right side corresponds to the variance of increments of the process

$$Z_{n,k}(t)=\frac{L(\rho')}{\sqrt{k}}\sum_{j=1}^{k}g_j\sqrt{N}t_j.$$

Therefore, Slepian's lemma (Ledoux and Talagrand, 1991) implies that for any j,

$$\mathbb{E}_g\sup_{t\in T^{(j)}}\left|Y_{n,k}(t)-Y_{n,k}(t^{(j)})\right|$$

$$\leqslant \mathbb{E}_g \sup_{(h,q)\in S_j} \frac{1}{\sqrt{k}} \left| \frac{\sqrt{Nn}}{\Delta_n} \sum_{i=1}^{k} g_j \left(\bar{L}_i(h,q) - \bar{L}_i(h^{(j)},q^{(j)}) - (\widetilde{L}(h,q) - L(h^{(j)},q^{(j)})) \right) \right|.$$

In turn, it yields the inequality

$$\begin{aligned} &\mathbb{E} \max_{j=1,\dots,N(\delta)} \mathbb{E}_g \sup_{t\in T^{(j)}} \left| Y_{n,k}(t) - Y_{n,k}(t^{(j)}) \right| \\ &\qquad \leqslant \mathbb{E} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} \frac{1}{\sqrt{k}} \left| \frac{\sqrt{Nn}}{\Delta_n} \sum_{i=1}^{k} g_j \Big(\bar{L}_i(h_1,q_1) - \bar{L}_i(h_2,q_2) \right. \\ &\qquad\qquad \left. - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2)) \Big) \right|. \end{aligned}$$

To complete the proof, we will apply the multiplier inequality (Lemma 2.9.1 in van der Vaart and Wellner, 1996) to deduce that the last display is bounded, up to a multiplicative constant, by

$$\begin{aligned} &\max_{m=1,\dots,k} \mathbb{E} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} \frac{1}{\sqrt{m}} \left| \frac{\sqrt{Nn}}{\Delta_n} \sum_{i=1}^{m} \varepsilon_j \Big(\bar{L}_i(h_1,q_1) - \bar{L}_i(h_2,q_2) \right. \\ &\qquad\qquad \left. - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2)) \Big) \right| \end{aligned}$$

where $\varepsilon_1, \dots, \varepsilon_k$ are i.i.d. Rademacher random variables. Next, desymmetrization inequality (Lemma 2.3.6 in van der Vaart and Wellner, 1996) implies that for any $m = 1, \dots, k$,

$$\begin{aligned} &\mathbb{E} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} \frac{1}{\sqrt{m}} \left| \frac{\sqrt{Nn}}{\Delta_n} \sum_{i=1}^{m} \varepsilon_j \Big(\bar{L}_i(h_1,q_1) - \bar{L}_i(h_2,q_2) \right. \\ &\qquad\qquad \left. - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2)) \Big) \right| \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0268

$$\leqslant 2\mathbb{E} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} \frac{1}{\sqrt{mn}} \left| \frac{\sqrt{N}}{\Delta_n} \sum_{i=1}^{mn} \Big(\widetilde{\ell}_N(h_1,q_1,X_i) - \widetilde{\ell}_N(h_2,q_2,X_i) \right.$$
$$\left. - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2)) \Big) \right|$$

where $\widetilde{\ell}_N(h,q,X)$ and $\widetilde{L}(h,q)$ were defined in (3.6) and (3.7) respectively. It remains to apply Lemma 2 in exactly the same way as before (see (3.14)) to deduce that the last display is bounded from above by $C\sqrt{d}\mathbb{E}^{1/2}\mathcal{V}^2(X;r(\theta_0))\cdot \delta \to 0$ as $\delta \to 0$. This completes the proof of asymptotic equicontinuity, and therefore weak convergence, of the sequence of processes $M_N(h,q)$.

Supplementary Material

The online supplementary material includes the proof of Theorem 1, the proofs of technical results and description of numerical simulation.

Acknowledgments

This research was partially supported by the National Science Foundation under grants DMS CAREER-2045068 and CCF-1908905. The author would like to thank Timothée Mathieu for sharing his code used in the simulation results (the code is available at https://github.com/TimotheeMathieu/Excess-risk-bounds-in-robust-empirical-risk-minimization/).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0268

References

Alistarh, D., Z. Allen-Zhu, and J. Li (2018). Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 4613–4623.

Alon, N., Y. Matias, and M. Szegedy (1996). The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 20–29. ACM.

Audibert, J.-Y., O. Catoni, et al. (2011). Robust linear least squares regression. *The Annals of Statistics 39*(5), 2766–2794.

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

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

Chen, Y., L. Su, and J. Xu (2017). Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems 1*(2), 1–25.

Cherapanamjeri, Y., S. B. Hopkins, T. Kathuria, P. Raghavendra, and N. Tripuraneni (2019). Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. *arXiv preprint arXiv:1912.11071*.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0268

Devroye, L., M. Lerasle, G. Lugosi, and R. I. Oliveira (2016). Sub-Gaussian mean estimators. *The Annals of Statistics 44*(6), 2695–2725.

Feller, W. (1968). On the Berry-Esseen theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 10*(3), 261–268.

Holland, M. J. and K. Ikeda (2017). Robust regression using biased objectives. *Machine Learning 106*(9-10), 1643–1679.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics 35*(1), 73–101.

Ibragimov, R. and S. Sharakhmetov (2001). The best constant in the Rosenthal inequality for nonnegative random variables. *Statistics & probability letters 55*(4), 367–376.

Lecué, G. and M. Lerasle (2020). Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics 48*(2), 906–931.

Lecué, G., M. Lerasle, and T. Mathieu (2020). Robust classification via MOM minimization. *Machine learning 109*, 1635–1665.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: isoperimetry and processes.* Berlin: Springer-Verlag.

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

Li, K., H. Bao, and L. Zhang (2022). Robust covariance estimation for distributed principal

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0268

component analysis. *Metrika 85*(6), 707–732.

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

Lugosi, G. and S. Mendelson (2019b). Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society 22*(3), 925–965.

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.

Minsker, S. (2019). Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics 13*(2), 5213–5252.

Minsker, S. (2025). Uniform bounds for robust mean estimators. *To appear in Stochastic Processes and Their Applications*.

Minsker, S. and S. Yao (2025). Generalized median of means principle for Bayesian inference. *Machine Learning 114*(4).

Nemirovski, A. and D. Yudin (1983). *Problem complexity and method efficiency in optimization.* John Wiley & Sons Inc.

O'Donnell, R. (2014). *Analysis of boolean functions.* Cambridge University Press.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*(Oct), 2825–2830.

Prasad, A., A. S. Suggala, S. Balakrishnan, P. Ravikumar, et al. (2020). Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B 82*(3), 601–627.

Talagrand, M. (2005). *The generic chaining.* Springer.

van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes.* Springer Series in Statistics. New York: Springer-Verlag.

Department of Mathematics, University of Southern California

E-mail: minsker@usc.edu