Information and Inference: A Journal of the IMA (2021) Page 1 of 62
doi:10.1093/imaiai/drn000

Excess risk bounds in robust empirical risk minimization

TIMOTHÉE MATHIEU
Laboratoire de Mathematiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France
timothee.mathieu@u-psud.fr
STANISLAV MINSKER*,
Department of Mathematics, University of Southern California, Los Angeles, CA 90089.
*Corresponding author: minsker@usc.edu

[Received on 26 January 2021]

This paper investigates robust versions of the general empirical risk minimization algorithm, one of the core techniques underlying modern statistical methods. Success of the empirical risk minimization is based on the fact that for a "well-behaved" stochastic process $\{f(X),\ f \in \mathscr{F}\}$ indexed by a class of functions $f \in \mathscr{F}$, averages $\frac{1}{N}\sum_{j=1}^{N} f(X_j)$ evaluated over a sample $X_1, \dots, X_N$ of i.i.d. copies of X provide good approximation to the expectations $\mathbb{E}f(X)$, uniformly over large classes $f \in \mathscr{F}$. However, this might no longer be true if the marginal distributions of the process are heavy-tailed or if the sample contains outliers. We propose a version of empirical risk minimization based on the idea of replacing sample averages by robust proxies of the expectations, and obtain high-confidence bounds for the excess risk of resulting estimators. In particular, we show that the excess risk of robust estimators can converge to 0 at fast rates with respect to the sample size N, referring to the rates faster than $N^{-1/2}$. We discuss implications of the main results to the linear and logistic regression problems, and evaluate the numerical performance of proposed methods on simulated and real data.

Keywords: Keywords: robust estimation, excess risk, median-of-means, regression, classification

1. Introduction

This work is devoted to robust algorithms in the framework of statistical learning. A recent Forbes article [53] states that "Machine learning algorithms are very dependent on accurate, clean, and well-labeled training data to learn from so that they can produce accurate results" and "According to a recent report from AI research and advisory firm Cognilytica, over 80% of the time spent in AI projects are spent dealing with and wrangling data." While some abnormal elements of the sample, or outliers, can be detected and filtered during the preprocessing steps, others are more difficult to detect: for instance, a sophisticated adversary might try to "poison" data to force a desired outcome [42]. Other seemingly abnormal observations could be inherent to the underlying data-generating process. An "ideal" learning method should not discard informative samples, while limiting the effect of individual observation on the output of the learning algorithm at the same time. We are interested in robust methods that are model-free, and require minimal assumptions on the underlying distribution. We study two types of robustness: robustness to heavy tails expressed in terms of the moment requirements, as well as robustness to (a variant of) adversarial contamination. Heavy tails can be used to model variation and randomness naturally occurring in the sample, while adversarial contamination is a convenient way to model outliers of unknown nature.

© The author 2021. Published by Oxford University Press on behalf of the Institute of Mathematics and its Applications. All rights reserved.

The statistical framework used throughout the paper is defined as follows. Let $(S,\mathscr{S})$ be a measurable space, and let $X \in S$ be a random variable with distribution P. Suppose that $X_1,\dots,X_N$ are i.i.d. copies of X. Moreover, assume that $\mathscr{F}$ is a class of measurable functions from S to $\mathbb{R}$ and $\ell : \mathbb{R} \to \mathbb{R}_+$, where $\mathbb{R}_+$ is a set of non-negative integers, is a loss function. Many problems in statistical learning theory can be formulated as risk minimization of the form

$$\mathbb{E}\,\ell(f(X)) \to \min_{f\in\mathscr{F}}.$$

We will frequently write $P\ell(f)$ or simply $\mathscr{L}(f)$ in place of the expected loss $\mathbb{E}\ell(f(X))$. Throughout the paper, we will also assume that the minimum above is attained for some (unique) $f_* \in \mathscr{F}$ (however, f_* does not necessarily coincide with the global minimizer of $\mathscr{L}(f)$ over all measurable functions that might not belong to $\mathscr{F}$). For example, in the context of regression, $X=(Z,Y) \in \mathbb{R}^d \times \mathbb{R}$, $f(Z,Y) = Y - g(Z)$ for some g in a class $\mathscr{G}$ (such as the class of linear functions), $\ell(x) = x^2$, and $f_*(z,y) = y - g_*(z)$, where $g_* = \operatorname{argmin}_{g\in\mathscr{G}} \mathbb{E}(Y-g(Z))^2$. As the true distribution P is usually unknown, a proxy of f_* is obtained via *empirical risk minimization* (ERM), namely

$$\tilde{f}_N := \operatorname*{argmin}_{f\in\mathscr{F}} \mathscr{L}_N(f), \tag{1.1}$$

where P_N is the empirical distribution based on the sample $X_1,\dots,X_N$ and

$$\mathscr{L}_N(f) := P_N\,\ell(f) = \frac{1}{N}\sum_{j=1}^{N} \ell\left(f(X_j)\right).$$

Performance of any $f \in \mathscr{F}$ (in particular, $\tilde{f}_N$) is measured via the excess risk $\mathscr{E}(f) := P\ell(f) - P\ell(f_*)$. The excess risk of $\tilde{f}_N$ is a random variable defined as

$$\mathscr{E}(\tilde{f}_N) := P\ell\big(\tilde{f}_N\big) - P\ell(f_*) = \mathbb{E}\left[\ell\big(\tilde{f}(X)\big)|X_1,\dots,X_N\right] - \mathbb{E}\ell\left(f_*(X)\right).$$

General bounds for the excess risk have been extensively studied; a small subsample of the relevant works includes the papers [57, 58, 31, 4, 7, 55] and references therein. However, until recently sharp estimates were known only in the situation when the functions in the class $\ell(\mathscr{F}) := \{\ell(f),\ f\in\mathscr{F}\}$ are uniformly bounded, or when the envelope $F_\ell(x) := \sup_{f\in\mathscr{F}} |\ell(f(x))|$ of the class $\ell(\mathscr{F})$ possesses finite exponential moments. Our focus is on the situation when marginal distributions of the process $\{\ell(f(X)),\ f\in\mathscr{F}\}$ indexed by $\mathscr{F}$ are allowed to be heavy-tailed, meaning that they possess finite moments of low order only (in this paper, "low order" usually means between 2 to 4). In such cases, the tail probabilities of the random variables $\left\{\frac{1}{\sqrt{N}}\sum_{j=1}^N \left(\ell(f(X_j)) - \mathbb{E}\ell(f(X))\right),\ f\in\mathscr{F}\right\}$ decay polynomially, thus rendering many existing techniques ineffective. Moreover, we consider a challenging framework of *adversarial contamination* where the initial dataset of cardinality N is merged with a set of $\mathscr{O} < N$ outliers which are generated by an adversary who has an opportunity to inspect the data, and the combined dataset of cardinality $N^\circ = N + \mathscr{O}$ is presented to an algorithm; in this paper, we assume that the proportion of contamination $\frac{\mathscr{O}}{N}$ (or its upper bound) is known.

The approach that we propose is based on replacing the sample mean at the core of ERM by a more "robust" estimator of $\mathbb{E}\ell(f(X))$ that exhibits tight concentration under minimal moment assumptions. Well known examples of such estimators include the median-of-means estimator [48, 2, 38] and Catoni's estimator [14]. Both the median-of-means and Catoni's estimators gain robustness at the cost of being biased. The ways that the bias of these estimators is controlled is based on different principles however.

Informally speaking, Catoni's estimator relies on delicate "truncation" of the data, while the median-of-means (MOM) estimator exploits the fact that the median and the mean of a symmetric distribution both coincide with its center of symmetry. In this paper, we will use "hybrid" estimators that take advantage of both symmetry and truncation. This family of estimators has been introduced and studied in [46, 47], and we review the construction below.

1.1 *Organization of the paper.*

The main ideas behind the proposed estimators are explained in Section 1.3, followed by the high-level overview of the main theoretical results and comparison to existing literature in Section 1.4. The complete statements of the key results are given in Section 2, and in Section 3 we deduce the corollaries of these results for specific examples. Finally, the main ideas and key inequalities necessary for the proofs is explained in Section 4. The remaining technical arguments are contained in the Supplementary material [41]. Finally, in Section C of the Supplement we discuss practical implementation and numerical performance of our methods on synthetic and real data.

1.2 *Notation.*

For two sequences $\{a_j\}_{j\geqslant 1} \subset \mathbb{R}$ and $\{b_j\}_{j\geqslant 1} \subset \mathbb{R}$ for $j \in \mathbb{N}$, the expression $a_j \lesssim b_j$ means that there exists a constant $c > 0$ such that $a_j \leqslant c b_j$ for all $j \in \mathbb{N}$; $a_j \asymp b_j$ means that $a_j \lesssim b_j$ and $b_j \lesssim a_j$. Absolute constants will be denoted c, c_1, C, C', etc, and may take different values in different parts of the paper. For a function $h : \mathbb{R}^d \mapsto \mathbb{R}$, we define

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

and $\|h\|_\infty := \operatorname{ess\,sup}\{|h(y)| : y \in \mathbb{R}^d\}$. Moreover, $L(h)$ will stand for a Lipschitz constant of h. For $f \in \mathscr{F}$, let $\sigma^2(\ell, f) = \operatorname{Var}(\ell(f(X)))$ and for any subset $\mathscr{F}' \subseteq \mathscr{F}$, denote $\sigma^2(\ell, \mathscr{F}') = \sup_{f\in\mathscr{F}'} \sigma^2(\ell, f)$. Additional notation and auxiliary results are introduced on demand.

1.3 *Robust mean estimators.*

Let $k \leqslant N$ be an integer, and assume that $G_1, \dots, G_k$ are disjoint subsets of the index set $\{1, \dots, N\}$ of cardinality $|G_j| = n \geqslant \lfloor N/k \rfloor$ each. Given $f \in \mathscr{F}$, let

$$\overline{\mathscr{L}}_j(f) := \frac{1}{n} \sum_{i\in G_j} \ell(f(X_i))$$

be the empirical mean evaluated over the subsample indexed by G_j. Given a convex, even function $\rho : \mathbb{R} \mapsto \mathbb{R}_+$ and $\Delta > 0$, set

$$\widehat{\mathscr{L}}^{(k)}(f) := \operatorname*{argmin}_{y\in\mathbb{R}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\,\frac{\overline{\mathscr{L}}_j(f) - y}{\Delta}\right). \tag{1.2}$$

Clearly, if $\rho(x) = x^2$, $\widehat{\mathscr{L}}^{(k)}(f)$ is equal to the sample mean. If $\rho(x) = |x|$, then $\widehat{\mathscr{L}}^{(k)}(f)$ is the median-of-means estimator [48, 2, 19]. We will be interested in the situation when ρ is smooth and "shaped" like Huber's loss, in particular, that ρ' is bounded and Lipschitz continuous (exact conditions imposed

on ρ are specified in Assumption 1 below). Note that (1.2) defines a whole family of estimators for different values of k and n. It is instructive to consider two cases: first, when $k = N$ (so that $n = 1$) and the scaling factor $\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))}\sqrt{N}$, $\widehat{\mathscr{L}}^{(k)}(f)$ is akin to Catoni's estimator [14], and when n is large (e.g. $\sqrt{N} \ll n \ll N$ and $\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))}$), $\widehat{\mathscr{L}}^{(k)}(f)$ is the "median-of-means type" estimator. Let us elaborate on these two cases further. Generally speaking, the estimator $\widehat{\mathscr{L}}^{(k)}(f)$ is biased, and, as we already mentioned in the introduction, one way to understand the difference between Catoni's and median-of-means type estimators is via the difference in mechanisms used to control the bias. In the case of Catoni's estimator, this mechanism is based on truncating each observation at the level of order $\sqrt{N}$ [1] encoded in the choice of $\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))}\sqrt{N}$, while in the case of the median-of-means estimator it relies on the approximate symmetry, implied by the Central Limit Theorem, of the distribution of the empirical averages $\overline{\mathscr{L}}_j(f)$, and in particular the fact that any reasonable estimator of location for this distribution will be close to the mean $\mathscr{L}(f)$ when $|G_j|$ is large. In Section 2.1, we formally introduce the key quantities that allow us to control the bias under various moment assumptions on the underlying classes.

We also construct a permutation-invariant version of the estimator $\widehat{\mathscr{L}}^{(k)}(f)$ that does not depend on the specific choice of the subgroups $G_1, \dots, G_k$. We conjecture that this estimator is more efficient than $\widehat{\mathscr{L}}^{(k)}(f)$; see remark 1.1 below for more details. Next, let

$$\mathscr{A}_N^{(n)} := \{J : J \subseteq \{1, \dots, N\},\ |J| = n\}.$$

Let h be a measurable, permutation-invariant function of n variables. Recall that a U-statistic of order n with kernel h based on an i.i.d. sample $X_1, \dots, X_N$ is defined as [25]

$$U_{N,n} = \frac{1}{\binom{N}{n}} \sum_{J \in \mathscr{A}_N^{(n)}} h(\{X_j\}_{j \in J}). \tag{1.3}$$

Given $J \in \mathscr{A}_N^{(n)}$, let $\overline{\mathscr{L}}(f;J) := \frac{1}{n}\sum_{i \in J} f(X_i)$. Consider U-statistics of the form

$$U_{N,n}(z;f) = \sum_{J \in \mathscr{A}_N^{(n)}} \rho\left(\sqrt{n}\,\frac{\overline{\mathscr{L}}(f;J) - z}{\Delta}\right).$$

Then the permutation-invariant version of $\widehat{\mathscr{L}}^{(k)}(f)$ is defined as

$$\widehat{\mathscr{L}}_U^{(k)}(f) := \operatorname*{argmin}_{z \in \mathbb{R}} U_{N,n}(z;f).$$

Finally, assuming that $\widehat{\mathscr{L}}^{(k)}(f)$ provides good approximation of the expected loss $\mathscr{L}(f)$ of each individual $f \in \mathscr{F}$, it is natural to consider

$$\widehat{f}_N := \operatorname*{argmin}_{f \in \mathscr{F}} \widehat{\mathscr{L}}^{(k)}(f), \tag{1.4}$$

as well as its permutation-invariant analogue

$$\widehat{f}_N^U := \operatorname*{argmin}_{f \in \mathscr{F}} \widehat{\mathscr{L}}_U^{(k)}(f) \tag{1.5}$$

[1] Reference to truncation can be made explicit by setting $\rho(x) = \min\left(x^2/2, |x| - 1/2\right)$ to be Huber's loss and considering the gradient descent iteration for the optimization problem (1.2).

as an alternative to standard empirical risk minimization (1.1). The main goal of this paper is to obtain general bounds for the excess risk of the estimators $\widehat{f}_N$ and $\widehat{f}_N^U$ under minimal assumptions on the stochastic process $\{\ell(f(X)),\ f\in\mathscr{F}\}$. More specifically, we are interested in scenarios when the excess risk converges to 0 at fast, or "optimistic" rates, referring to the rates faster than $N^{-1/2}$. Rate of order $N^{-1/2}$ ("slow rates") are easier to establish: in particular, results of this type follow from bounds on the uniform deviations $\sup_{f\in\mathscr{F}}\left|\widehat{\mathscr{L}}^{(k)}(f)-\mathscr{L}(f)\right|$ that have been investigated in [47]. Proving fast rates is a more technically challenging task: to achieve the goal, we develop Bahadur-type representations [6] of the estimators $\widehat{\mathscr{L}}^{(k)}(f)$ and $\widehat{\mathscr{L}}_U^{(k)}(f)$ that provide linear, in $\ell(f)$, approximations of these nonlinear statistics that are easier to study, and carefully analyze the remainder terms. Introduction of such representations in the framework of median-of-means estimation is one of the main technical novelties of the paper; the tools we develop could prove useful in other related problems, such as study of the asymptotic distributions of the robust estimators $\widehat{f}_N$ and $\widehat{f}_N^U$.

REMARK 1.1 The main reason we introduce the permutation-invariant estimator $\widehat{f}_N^U$ is our conjecture that it has superior, compared to $\widehat{f}_N$, performance. We were able to confirm this fact numerically in our experiments; however, complete theoretical confirmation is not yet available, and requires new technical tools beyond those developed in the present work. Specifically, we conjecture that $\widehat{f}_N^U$ is more *efficient* than $\widehat{f}_N$: when $\mathscr{F}$ is finite dimensional, this means, informally, that the asymptotic distribution of $\sqrt{N}(\widehat{f}_N^U-f_*)$ has smaller variance than the asymptotic distribution of $\sqrt{N}(\widehat{f}_N-f_*)$. In other words, the conjectured difference in performance is about the constant factors rather than the rates. Such improvements are too subtle to be captured by the non-asymptotic bounds for the excess risk that are being pursued in this work, nevertheless they are clearly noticeable in the simulations.

It should also be acknowledged that exact evaluation of the U-statistics-based estimators $\widehat{\mathscr{L}}_U^{(k)}(f)$ and $\widehat{f}_N^U$ is not feasible due to the number of summands $\binom{N}{n}$ being very large even for small values of n. However, exact computation is typically not required, and throughout our detailed simulation studies, gradient descent methods proved to be very efficient for the problem (1.5) in scenarios like least-squares and logistic regression. These points, as well as comparison of the numerical performance of the estimators $\widehat{f}_N^U$ and $\widehat{f}_N$, are further discussed in Section C of the Supplementary material [41].

1.4 *Overview of the main results and comparison to existing bounds.*

Our main contribution is the proof of high-confidence bounds for the excess risk of the estimators $\widehat{f}_N$ and $\widehat{f}_N^U$. First, we show (see Theorem 2.1 and (2.4)) that the excess risk is bounded from above by the quantity of order $N^{-1/2}$ (referred to as "slow rates") with exponentially high probability if

$$\sigma^2(\ell,\mathscr{F})=\sup_{f\in\mathscr{F}}\sigma^2(\ell,f)<\infty \text{ and } \mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{\sqrt{N}}\sum_{j=1}^N\left(\ell(f(X_j))-\mathbb{E}\ell(f(X))\right)<\infty.$$

The latter is true if the class $\{\ell(f),\ f\in\mathscr{F}\}$ is P-Donsker [24], in other words, if the empirical process $f\mapsto\frac{1}{\sqrt{N}}\sum_{j=1}^N\left(\ell(f(X_j))-\mathbb{E}\ell(f(X))\right)$ converges weakly to a Gaussian limit. This result is analogous to its counterpart in the standard empirical risk minimization framework. Moreover, it is known [43, 36] that in general, the $N^{-1/2}$ rate for the excess risk of the empirical risk minimizers can not be improved. Our main contribution is the proof of the fact that that under additional assumption requiring that any $f\in\mathscr{F}$ with small excess risk is itself close to f_* (that minimizes the expected loss), $\widehat{f}_N$ and $\widehat{f}_N^U$ attain *fast* rates. This fact is well-known in the usual empirical risk minimization framework [10, 31] but is

new for the type of robust estimators considered here. We state the bounds below only for $\widehat{f}_N$ while the results for the U-statistics based $\widehat{f}_N^U$ are similar, up to the change in absolute constants. In order to avoid excessive technical details at this stage, we will first illustrate our general results by stating corollaries for the popular frameworks of logistic regression and regression with quadratic loss, while the most general versions of the theorems and additional examples will be stated afterwards.

Binary classification and logistic regression. Assume that $(Z,Y) \in S \times \{\pm 1\}$ is a random couple where Z is an instance and Y is a binary label, and let $g_*(z) := \mathbb{E}[Y|Z=z]$ be the regression function. It is well-known that the binary classifier $b_*(z) := \operatorname{sign}(g_*(z))$ achieves smallest possible misclassification error defined as $P(Y \neq g(Z))$. Let $\mathscr{F}$ be a given convex class of functions mapping S to $\mathbb{R}$, $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ a convex, nondecreasing, Lipschitz loss function, and let

$$h_* = \operatorname*{argmin}_{\text{all measurable } f} \mathbb{E}\ell(Yf(Z)).$$

The loss ℓ is classification-calibrated if $\operatorname{sign}(h_*(z)) = b_*(z)$ P-almost surely; we refer the reader to [9] for a detailed exposition. In the case of logistic regression considered below, $S = \mathbb{R}^d$,

$$\ell(y, f(z)) = \ell(yf(z)) := \log\left(1 + e^{-yf(z)}\right)$$

is the classification-calibrated loss and $\mathscr{F} = \left\{f_\beta(\cdot) = \langle \cdot, v\rangle,\ v \in \mathbb{R}^d,\ \|v\|_2 \leqslant R\right\}$. Note that results stated below hold without assuming that $h_* \in \mathscr{F}$.

Regression with quadratic loss. Let $(Z,Y) \in S \times \mathbb{R}$ be a random couple satisfying $Y = f_*(Z) + \eta$ where the noise variable η is independent of Z and $f_*(z) = \mathbb{E}[Y|Z=z]$ is the regression function. Linear regression with quadratic loss corresponds to $S = \mathbb{R}^d$,

$$\ell(y, f(z)) = \ell(y - f(z)) := (y - f(z))^2$$

and $\mathscr{F} = \left\{f_\beta(\cdot) = \langle \cdot, v\rangle,\ \ v \in \mathbb{R}^d, \|v\|_2 \leqslant R\right\}$. In this case, we will assume that $f_* \in \mathscr{F}$; it is possible to avoid this assumption at the cost of additional technicalities and taking advantage of the deep results of S. Mendelson [45] on the multiplier inequalities.

In the statements below, we will assume that we are given an i.i.d. sample $(Z_1,Y_1),\dots,(Z_N,Y_N)$ having the same distribution as (Z,Y) where the marginal distribution of Z is supported on a compact set. Moreover, suppose that $\mathbb{E}|\eta|^8 < \infty$ in the case of regression with quadratic loss; Section 3 contains other examples covering a wider class of distributions and classes $\mathscr{F}$.

THEOREM 1.1 (Informal) Assume the framework of either logistic regression or linear regression with quadratic loss. Then, for appropriately chosen k and Δ,

$$\mathscr{E}(\widehat{f}_N) \leqslant C(R,P,\rho)\left(\frac{d}{N} + \frac{s}{N^{3/4}} + \left(\frac{\mathscr{O}}{N}\right)^{3/4}\right)$$

with probability at least $1 - e^{-s}$ for all $s \lesssim k$.

Moreover, we construct a two-step estimator $\widehat{f}_N''$ based on $\widehat{f}_N$ that is capable of achieving further improved rates.

THEOREM 1.2 (Informal) Assume the framework of either logistic regression or linear regression with quadratic loss. There exists an estimator $\widehat{f}_N''$, defined later in the paper, such that

$$\mathscr{E}\left(\widehat{f}_N''\right) \leqslant C(R,P,\rho)\left(\frac{d}{N} + \frac{s}{N} + \frac{\mathscr{O}}{N}\right)$$

with probability at least $1-e^{-s}$ for all $1 \leqslant s \leqslant s_{\max}$ where $s_{\max} := s_{\max}(N) \to \infty$ as $N \to \infty$.

The estimator $\widehat{f}_N''$ mentioned in Theorem 1.2 is based on a two-step procedure, where $\widehat{f}_N$ serves as an initial approximation that is refined on the second step via risk minimization restricted to a "small neighborhood" of $\widehat{f}_N$. All of the bounds in this paper have the form $\mathscr{E}(\widehat{f}_N) \leqslant \overline{\delta} + C(\mathscr{F},P)\left(\frac{s}{N^\gamma} + \left(\frac{\mathscr{O}}{N}\right)^\gamma\right)$, where $\frac{1}{2} \leqslant \gamma \leqslant 1$ and $\overline{\delta}$ is the quantity (formally defined in (2.5)) that often coincides, up to log-factors, with the optimal rate for the excess risk [3, 40] – for instance, $\overline{\delta} \asymp \frac{d}{N}$ in the examples above. In the standard empirical risk minimization, the excess risk bounds in the linear and logistic regression admit the bounds of order $\frac{d}{N} + \frac{s}{N}$, albeit under more restrictive assumptions and in the corruption-free framework. Therefore, the bound of Theorem 1.1 is suboptimal in these cases due to the "remainder terms" being of order $N^{-3/4}$, and the improvement achieved by the two-step estimator $\widehat{f}_N''$, as described in Theorem 1.2, becomes important.

Next, we provide a brief overview of the literature on the topic and compare our results to the state of the art. Robustness of statistical learning algorithms has been studied extensively in recent years. Existing research has mainly focused on addressing robustness to heavy tails as well as adversarial contamination. One line of work investigated robust versions of the gradient descent method for the optimization problem (1.1) based on variants of the multivariate median-of-means technique [51, 16, 59, 1], as well as Catoni's estimator [28]. The line works initiated in the theoretical computer science community [33, 20, 22, also see the survey paper [23]] tackled the problem of optimal mean estimation in the adversarial contamination framework by establishing deep connections between the mean and covariance estimation problems that culminated in the family of powerful filtering algorithms; these algorithms can also be used as subroutines in robust gradient descent-type methods [21, 17]. While these algorithms admit strong theoretical guarantees, they require robustly estimating the gradient vector at every step (with the exception of [21] that offers a more efficient approach) hence are computationally demanding; moreover, results are weaker for losses that are not strongly convex (for instance, the hinge loss). The line of research that is closest in spirit to the approach of this paper includes the works that employ robust risk estimators based on Catoni's idea [5, 13, 27] and the median-of-means technique, such as "tournaments" and the "min-max median-of-means" [40, 39, 34, 35, 18], also see [17, 29] for the computationally efficient algorithms related to the tournament-type procedures. As it was mentioned in the introduction, the core of our methods can be viewed as a "hybrid" between Catoni's and the median-of-means estimators. We provide a more detailed comparison to the results of the aforementioned papers:

1. We show that risk minimization based on a version of Catoni's estimator is capable of achieving fast rates, thus improving the results and weakening the assumptions stated in [13] that only allowed the slow rates to be established;
2. We develop new tools and techniques to analyze proposed estimators. In particular, we do not rely on the "small ball" method [32, 44] and the standard "majority vote-based" analysis [34, 40] of the median-of-means estimators. Instead, we provide accurate bounds for the bias and investigate the remainder terms for the Bahadur-type linear approximations of the estimators defined in (1.2). In particular, we demonstrate that the order of typical deviations of the estimator $\widehat{\mathscr{L}}^{(k)}(f)$ around $\mathscr{L}(f)$ are significantly smaller than the deviations of the subsample averages $\overline{\mathscr{L}}_j(f)$, which is not easy to do using the majority vote-based proof techniques; consequently, this fact allows us to "decouple" the confidence parameter s that controls the deviation probabilities from parameters k and $\mathscr{O}$ responsible for the number of subsamples and the degree of contamination respectively. Unlike the tournaments-based estimators, in some regimes our algorithms admit a

"universal" choice of k that is independent of the parameter $\overline{\delta}$ controlling the optimal rate. In the previous works, parameter k was often overloaded as it controlled the deviation probabilities while depending on $\overline{\delta}$ (or a closely related quantity) at the same time. Finally, our techniques allow us to establish bounds that are uniform over a certain range of confidence parameter s while the previously existing deviation results were only available for $s \asymp k$.

3. We are able to simultaneously treat the case of Lipschitz as well as non-Lipschitz (e.g., quadratic) loss functions ℓ. At the same time, in some situations (e.g. linear regression with quadratic loss), the required assumptions are stronger compared to the best results in the literature tailored specifically to the task, e.g. [34, 40] that treat the case of regression with quadratic loss.
4. Existing approaches based on the median-of-means estimators are either computationally intractable [40], or outputs of practically efficient algorithms do not admit strong theoretical guarantees [34, 35, 18]. We design numerical algorithms specifically for the estimators $\widehat{f}_N$ and $\widehat{f}_N^U$ defined via (1.4) and (1.5), and show that they enjoy good performance in numerical experiments as well as strong theoretical guarantees.

2. Theoretical guarantees for the excess risk.

In this section, we give complete statements of the main results and explain the high-level ideas behind their proofs.

2.1 *Preliminaries.*

We start by introducing the main quantities that appear in our results, and state the key assumptions. Recall that $\sigma^2(\ell, \mathscr{F}')$ stands for $\sup_{f\in\mathscr{F}'}\sigma^2(\ell, f)$, where $\mathscr{F}' \subseteq \mathscr{F}$. The loss functions ρ that will be of interest to us satisfy the following assumption.

Assumption 1 Suppose that the function $\rho : \mathbb{R} \mapsto \mathbb{R}$ is convex, even, 5 times continuously differentiable and such that

(i) $\rho'(z) = z$ for $|z| \leqslant 1$ and $\rho'(z) = \text{const}$ for $z \geqslant 2$,
(ii) $z - \rho'(z)$ is nondecreasing.

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 ϕ be the "bump function" $\phi(x) = C\exp\left(-\frac{4}{1-4x^2}\right) I\left\{|x| \leqslant \frac{1}{2}\right\}$ where C is chosen so that $\int_{\mathbb{R}} \phi(x)\mathrm{d}x = 1$. Then ρ given by the convolution $\rho(x) = (h * \phi)(x)$ satisfies Assumption 1.

REMARK 2.1 (a) The requirements that ρ is 5 times continuously differentiable is of the technical nature and is likely not necessary. It appears due to the fact that we need to control higher order terms in the Bahadur-Kiefer type representations of the estimator $\widehat{\mathscr{L}^{(k)}}(f)$, as well as rely on the Lindeberg replacement-type arguments in our proofs.

(b) The derivative ρ' has a natural interpretation of being a smooth version of the truncation function. Moreover, observe that $\rho'(2) - 2 \leqslant \rho'(1) - 1 = 0$ by (ii), hence $\|\rho'\|_\infty \leqslant 2$. It is also easy to see

that for any $x > y$, $\rho'(x) - \rho'(y) = y - \rho'(y) - (x - \rho'(x)) + x - y \leqslant x - y$, hence ρ' is Lipschitz continuous with Lipschitz constant $L(\rho') = 1$.

In Section 1.3, we have briefly discussed the bias of robust mean estimators and various ways that it can be controlled. Now we will introduce the key quantities necessary to make the bounds precise. Everywhere below, $\Phi(\cdot)$ stands for the cumulative distribution function of the standard normal random variable and $W(f)$ denotes a random variable with distribution $N\big(0, \sigma^2(f)\big)$. For $f \in \mathscr{F}$ such that $\sigma(f) > 0$, $n \in \mathbb{N}$ and $t > 0$, define

$$\mathscr{M}_f(t,n) := \left| \Pr\left(\frac{\sum_{j=1}^n (f(X_j) - Pf)}{\sigma(f)\sqrt{n}} \leqslant t \right) - \Phi(t) \right|,$$

where $Pf := \mathbb{E} f(X)$. In other words, $\mathscr{M}_f(t,n)$ controls the rate of convergence in the central limit theorem. It follows from the results of L. Chen and Q.-M. Shao (Theorem 2.2 in [15]) that

$$\mathscr{M}_f(t,n) \leqslant g_f(t,n) := C\left(\frac{\mathbb{E}(f(X) - \mathbb{E}f(X))^2 I\left\{ \frac{|f(X) - \mathbb{E}f(X)|}{\sigma(f)\sqrt{n}} > 1 + \left| \frac{t}{\sigma(f)} \right| \right\}}{\sigma^2(f)\left(1 + \left| \frac{t}{\sigma(f)} \right|\right)^2} + \frac{1}{\sqrt{n}} \frac{\mathbb{E}|f(X) - \mathbb{E}f(X)|^3 I\left\{ \frac{|f(X) - \mathbb{E}f(X)|}{\sigma(f)\sqrt{n}} \leqslant 1 + \left| \frac{t}{\sigma(f)} \right| \right\}}{\sigma^3(f)\left(1 + \left| \frac{t}{\sigma(f)} \right|\right)^3} \right)$$

given that the absolute constant C is large enough. Note that, crucially, the control of the rate in terms of $g_f(t,n)$ is non-uniform, since $g_f(t,n)$ is a decreasing function of t. Moreover, let

$$G_f(n,\Delta) := \int_0^\infty g_f\left(\Delta\left(\frac{1}{2} + t \right), n \right) \mathrm{d}t.$$

The quantity $\frac{G_f(n,\Delta)}{\sqrt{n}}$ plays the key role in controlling the bias of the estimator $\widehat{\mathscr{L}}^{(k)}(f)$: it decreases both as Δ get large and as the subsample size n increases, referring to different bias-controlling mechanisms of Catoni's and the median-of-means type estimators, see discussion after (1.2). The following statement provides simple upper bounds for $g_f(t,n)$ and $G_f(n,\Delta)$ that depend on the tail properties of $f(X)$; its proof can be found in [47, Section 4.4].

LEMMA 2.1 Let $X_1, \dots, X_n$ be i.i.d. copies of X, and assume that $\mathrm{Var}(f(X)) < \infty$. Then $g_f(t,n) \to 0$ as $|t| \to \infty$ and $g_f(t,n) \to 0$ as $n \to \infty$, with the convergence being monotone. Moreover, if $\mathbb{E}|f(X) - \mathbb{E}f(X)|^{2+\delta} < \infty$ for some $\delta \in [0,1]$, then for all $t > 0$

$$g_f(t,n) \leqslant C' \frac{\mathbb{E}\big|f(X) - \mathbb{E}f(X)\big|^{2+\delta}}{n^{\delta/2}\left(\sigma(f) + |t|\right)^{2+\delta}} \leqslant C' \frac{\mathbb{E}\big|f(X) - \mathbb{E}f(X)\big|^{2+\delta}}{n^{\delta/2}|t|^{2+\delta}}, \tag{2.1}$$

$$G_f(n,\Delta) \leqslant C'' \frac{\mathbb{E}\big|f(X) - \mathbb{E}f(X)\big|^{2+\delta}}{\Delta^{2+\delta} n^{\delta/2}},$$

where $C', C'' > 0$ are absolute constants.

We can rewrite the bound for $\sup_{f\in\mathscr{F}} G_f(n,\Delta)$ as $\sup_{f\in\mathscr{F}} G_f(n,\Delta) \leqslant C'' \frac{\sup_{f\in\mathscr{F}} \mathbb{E}(|f(X) - \mathbb{E}f(X)|/\sigma(\ell,\mathscr{F}))^{2+\delta}}{(\Delta/\sigma(\ell,\mathscr{F}))^{2+\delta} n^{\delta/2}}$, where the numerator $\sup_{f\in\mathscr{F}} \mathbb{E}\left(|f(X) - \mathbb{E}f(X)|/\sigma(\ell,\mathscr{F})\right)^{2+\delta}$ is the quantity akin the kurtosis while the

ratio $M_\Delta := \frac{\Delta}{\sigma(\ell,\mathscr{F})}$ appearing in the denominator can be interpreted as a truncation level expressed in the "units" of $\sigma(\ell,\mathscr{F})$. This "truncation level," along with the subgroup size n, are the two main quantities controlling the bias of the estimators $\widehat{\mathscr{L}}^{(k)}(f),\ f\in\mathscr{F}$.

2.2 *Slow rates for the excess risk.*

Let

$$\widehat{\delta}_N := \mathscr{E}(\widehat{f}_N) = \mathscr{L}(\widehat{f}_N) - \mathscr{L}(f_*),$$
$$\widehat{\delta}_N^U := \mathscr{E}(\widehat{f}_N^U) = \mathscr{L}(\widehat{f}_N^U) - \mathscr{L}(f_*)$$

be the excess risk of $\widehat{f}_N$ and its permutation-invariant analogue $\widehat{f}_N^U$ which are the main objects of our interest. The following bound for the excess risk is well known in the empirical risk minimization literature [31], and it easily leads to control of the excess risk in terms of the uniform deviations of robust mean estimators.

$$\begin{aligned}\mathscr{E}(\widehat{f}_N) &= \mathscr{L}(\widehat{f}_N) - \mathscr{L}(f_*)\\ &= \mathscr{L}(\widehat{f}_N) + \widehat{\mathscr{L}}^{(k)}(\widehat{f}_N) - \widehat{\mathscr{L}}^{(k)}(\widehat{f}_N) + \widehat{\mathscr{L}}^{(k)}(f_*) - \widehat{\mathscr{L}}^{(k)}(f_*) - \mathscr{L}(f_*)\\ &= \Big(\mathscr{L}(\widehat{f}_N) - \widehat{\mathscr{L}}^{(k)}(\widehat{f}_N)\Big) - \Big(\mathscr{L}(f_*) - \widehat{\mathscr{L}}^{(k)}(f_*)\Big) + \underbrace{\widehat{\mathscr{L}}^{(k)}(\widehat{f}_N) - \widehat{\mathscr{L}}^{(k)}(f_*)}_{\leqslant 0}\\ &\leqslant 2\sup_{f\in\mathscr{F}}\Big|\widehat{\mathscr{L}}^{(k)}(f) - \mathscr{L}(f)\Big|. \qquad (2.2)\end{aligned}$$

The first result, Theorem 2.1 below, together with the inequality (2.2) immediately implies the "slow rate bound" (meaning rate not faster than $N^{-1/2}$) for the excess risk. This result has been previously established in [47]. Define

$$\widetilde{\Delta} := \max\left(\Delta, \sigma(\ell,\mathscr{F})\right).$$

THEOREM 2.1 There exist absolute constants $c, C>0$ such that for all $s>0$, n and k satisfying

$$\frac{1}{\Delta}\left(\frac{1}{\sqrt{k}}\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{\sqrt{N}}\sum_{j=1}^N\left(\ell(f(X_j)) - P\ell(f)\right) + \sigma(\ell,\mathscr{F})\sqrt{\frac{s}{k}}\right) + \sup_{f\in\mathscr{F}} G_f(n,\Delta) + \frac{s}{k} + \frac{\mathscr{O}}{k} \leqslant c, \qquad (2.3)$$

the following inequality holds with probability at least $1-2e^{-s}$:

$$\begin{aligned}\sup_{f\in\mathscr{F}}\Big|\widehat{\mathscr{L}}^{(k)}(f) - \mathscr{L}(f)\Big| \leqslant C\Bigg[\frac{\widetilde{\Delta}}{\Delta}\left(\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{N}\sum_{j=1}^N\left(\ell(f(X_j)) - P\ell(f)\right) + \sigma(\ell,\mathscr{F})\sqrt{\frac{s}{N}}\right)\\ + \widetilde{\Delta}\left(\sqrt{n}\frac{s}{N} + \frac{\sup_{f\in\mathscr{F}} G_f(n,\Delta)}{\sqrt{n}} + \frac{\mathscr{O}}{k\sqrt{n}}\right)\Bigg].\end{aligned}$$

Moreover, the same bounds hold for the permutation-invariant estimators $\widehat{\mathscr{L}}_U^{(k)}(f)$, up to the change in absolute constants.

An immediate corollary is the bound for the excess risk

$$\mathscr{E}(\widehat{f}_N) \leqslant C\left[\frac{\widetilde{\Delta}}{\Delta}\left(\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{N}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))+\sigma(\ell,\mathscr{F})\sqrt{\frac{s}{N}}\right)\right.$$
$$\left.+\widetilde{\Delta}\sqrt{n}\left(\frac{s}{N}+\frac{\sup_{f\in\mathscr{F}}G_f(n,\Delta)}{n}+\frac{\mathscr{O}}{N}\right)\right] \quad (2.4)$$

that holds under the assumptions of Theorem 2.1 with probability at least $1-2e^{-s}$. When the class $\{\ell(f),\ f\in\mathscr{F}\}$ is P-Donsker [24], $\limsup_{N\to\infty}\left|\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))\right|$ is bounded, hence condition (2.3) holds for N large enough whenever s is not too big and Δ and k are not too small, namely, $s\leqslant c'k$ and $\Delta\sqrt{k}\geqslant c''\sigma(\mathscr{F})$. The bound of Theorem 2.1 also suggests that the natural "unit" to measure the magnitude of the parameter Δ is $\sigma(\ell,\mathscr{F})$.

To put these results in perspective, let us consider two examples. First, assume that $n=1$, $k=N$ and set $\Delta=\Delta(s):=\sigma(\mathscr{F})\sqrt{\frac{N}{s}}$ for $s\leqslant c'N$. Using Lemma 2.1 with $\delta=0$ to estimate $G_f(n,\Delta)$, we deduce that

$$\mathscr{E}(\widehat{f}_N)\leqslant C\left[\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{N}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))+\sigma(\ell,\mathscr{F})\left(\sqrt{\frac{s}{N}}+\frac{\mathscr{O}}{\sqrt{N}}\right)\right]$$

with probability at least $1-2e^{-s}$. This inequality improves upon excess risk bounds obtained for Catoni-type estimators in [13], as it does not require functions in $\mathscr{F}$ to be uniformly bounded.

The second case we consider is when $N\gg n\geqslant 2$. For the choice of $\Delta\asymp\sigma(\ell,\mathscr{F})$, the estimator $\widehat{\mathscr{L}^{(k)}}(f)$ most closely resembles the median-of-means estimator, as we have explained in Section 1.3. In this case, Theorem 2.1 yields the excess risk bound of the form

$$\mathscr{E}(\widehat{f}_N)\leqslant C\left[\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{N}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))+\sigma(\ell,\mathscr{F})\left(\sqrt{\frac{s}{N}}+\sqrt{\frac{k}{N}}\sup_{f\in\mathscr{F}}G_f(n,\sigma(\mathscr{F}))+\frac{\mathscr{O}}{k}\sqrt{\frac{k}{N}}\right)\right]$$

that holds with probability $\geqslant 1-2e^{-s}$ for all $s\leqslant c'k$. As $\sup_{f\in\mathscr{F}}G_f(n,\Delta)$ is small for large n and $\frac{\mathscr{O}}{k}\sqrt{\frac{k}{N}}\leqslant\sqrt{\frac{\mathscr{O}}{N}}$ whenever $\mathscr{O}\leqslant k$, this bound improves upon Theorem 2 in [35] that provides bounds for the excess risk for robust classifiers based on the the median-of-means estimators.

2.3 *Towards fast rates for the excess risk.*

It is well known that in regression and binary classification problems, the excess risk often converges to 0 at a rate faster than $N^{-1/2}$, and could be as fast as N^{-1}. Such rates are often referred to as "fast" or "optimistic" rates. In particular, this is the case when there exists a "link" between the excess risk and the variance of the loss class, namely, if for some convex nondecreasing and nonnegative function ϕ such that $\phi(0)=0$,

$$\mathscr{E}(f)=P\ell(f)-P\ell(f_*)\geqslant\phi\left(\sqrt{\mathrm{Var}\left(\ell(f(X))-\ell(f_*(X))\right)}\right).$$

It is thus natural to ask if fast rates can be attained by estimators produced by the robust algorithms proposed above. Results presented in this section give an affirmative answer to this question. Let us

introduce the main quantities that commonly appear in the excess risk bounds [31, 40]. For $\delta > 0$, let

$$\mathscr{F}(\delta) := \{\ell(f) : f \in \mathscr{F}, \mathscr{E}(f) \leqslant \delta\},$$
$$\nu(\delta) := \sup_{\ell(f)\in\mathscr{F}(\delta)} \sqrt{\mathrm{Var}\left(\ell(f(X)) - \ell(f_*(X))\right)},$$
$$\omega(\delta) := \mathbb{E} \sup_{\ell(f)\in\mathscr{F}(\delta)} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \Big((\ell(f) - \ell(f_*))(X_j) - P(\ell(f) - \ell(f_*)) \Big) \right|.$$

Moreover, define

$$\mathfrak{B}(\ell, \mathscr{F}) := \frac{\sup_{f\in\mathscr{F}} \mathbb{E}^{1/4} \left(\ell(f(X)) - \mathbb{E}\ell(f(X))\right)^4}{\sigma(\ell, \mathscr{F})}.$$

The following condition, known as *Bernstein's condition* following [10], plays the crucial role in the analysis of excess risk bounds.

Assumption 2 There exist constants $D > 0$, $\delta_B > 0$ such that

$$\mathrm{Var}\left(\ell(f(X)) - \ell(f_*(X))\right) \leqslant D^2 \mathscr{E}(f)$$

whenever $\mathscr{E}(f) \leqslant \delta_B$.

Informally speaking, Assumption 2 postulates that any $f \in \mathscr{F}$ (more precisely, the loss $\ell(f)$ induced by it) with small excess risk is itself close to f_*. If this is true, it turns out that one can avoid global bounds on the expected supremum of the empirical process used to obtain "slow" rates, and instead rely on the modulus of continuity $\omega(\delta)$ of the empirical process locally in the neighborhood of $\ell(f_*)$ in order to get better upper bounds on the excess risk. The basics of this approach in the classical empirical risk minimization frameworks are clearly explained in [31, Chapter 1.2], and we rely on similar ideas below.

Assumption 2 is known to hold in many concrete cases of prediction and classification tasks, and we provide examples and references in Section 3 below. More general versions of the Bernstein's condition are often considered in the literature: for instance, it can be replaced by assumption requiring that $\mathrm{Var}\left(\ell(f(X)) - \ell(f_*(X))\right) \leqslant D^2 \left(\mathscr{E}(f)\right)^{\tau}$ for some $\tau \in (0, 1]$, as was done in [10]; clearly, our assumption corresponds to $\tau = 1$. Results of this paper admit straightforward extensions to the slightly less restrictive scenario when $\tau < 1$; we omit the details to reduce the level of technical burden on the statements of our results.

Following [31, Chapter 4], we will say that the function $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is of concave type if it is nondecreasing and $x \mapsto \frac{\psi(x)}{x}$ is decreasing. Moreover, if for some $\gamma \in (0, 1)$ $x \mapsto \frac{\psi(x)}{x^\gamma}$ is decreasing, we will say that ψ is of strictly concave type with exponent γ. We will assume that $\omega(\delta)$ admits an upper bound $\widetilde{\omega}(\delta)$ of strictly concave type (with some exponent γ), and that $\nu(\delta)$ admits an upper bound $\widetilde{\nu}(\delta)$ of concave type. For instance, when Assumption 2 holds, $\nu(\delta) \leqslant D\sqrt{\delta}$ for $\delta \leqslant \delta_B$, implying that $\widetilde{\nu}(\delta) = D\sqrt{\delta}$ is an upper bound for $\nu(\delta)$ of strictly concave type with $\gamma = \frac{1}{2}$.[2] Moreover, the function $\omega(\delta)$ often admits an upper bound of the form $\widetilde{\omega}(\delta) = R_1 + \sqrt{\delta} R_2$ where R_1 and R_2 do not depend on δ; such an upper bound is also of concave type. Next, set

$$\overline{\delta} := \min\left\{ \delta > 0 : C_1(\rho) \frac{1}{\sqrt{N}} \frac{\widetilde{\Delta}}{\Delta} \frac{\widetilde{\omega}(\delta)}{\delta} \leqslant \frac{1}{7} \right\}, \tag{2.5}$$

[2] This is only true in some neighborhood of 0, but is sufficient for our purposes.

where $C_1(\rho)$ is a sufficiently large positive constant that depends only on ρ. The quantity $\overline{\delta}$ often coincides with the optimal rates for the excess risk in the classical empirical risk minimization framework: for example, it is of order $\frac{d}{N}$ up to logarithmic factors in linear regression with quadratic loss and in logistic regression when Bernstein's condition is satisfied; in general, the order of $\overline{\delta}$ ranges between the pessimistic $N^{-1/2}$ in "hard" problems and "optimistic" N^{-1} where the rates between correspond to weaker versions of Assumptions 2, for instance, see [9]. The theorems below provide estimates for the excess risk of robust risk minimizers under various conditions on the tails of the random variables $\{f(X),\ f\in\mathscr{F}\}$. All these bounds have the same structure that includes the term $\overline{\delta}$ as well as the "remainder terms" that account for the bias of the robust risk estimators $\widehat{\mathscr{L}^{(k)}}(f)$ as well as the outlier contamination proportion $\frac{\mathscr{O}}{N}$; naturally, stricter moment conditions result in better remainder terms.

THEOREM 2.2 Assume that conditions of Theorem 2.1 hold. Additionally, suppose that $M_\Delta := \frac{\Delta}{\sigma(\ell,\mathscr{F})} \geqslant 1$. Then

$$\widehat{\delta}_N \leqslant \overline{\delta} + C(\rho)\left(D^2\left(\frac{1}{M_\Delta^2 n}+\frac{s+\mathscr{O}}{N}\right)+\sigma(\ell,\mathscr{F})\sqrt{n}M_\Delta\left(\frac{1}{M_\Delta^4 n}+\frac{s+\mathscr{O}}{N}\right)\right).$$

with probability at least $1-10e^{-s}$, where the constant $C(\rho)$ depends on ρ only and D is a constant appearing in Assumption 2.

Under stronger moment assumptions, the excess risk bound can be strengthened and take the following form.

THEOREM 2.3 Assume that conditions of Theorem 2.1 hold. Additionally, suppose that

$$\sup_{f\in\mathscr{F}}\mathbb{E}^{1/4}\left(\ell(f(X))-\mathbb{E}\ell(f(X))\right)^4<\infty$$

and that $M_\Delta := \frac{\Delta}{\sigma(\ell,\mathscr{F})} \geqslant 1$. Then

$$\widehat{\delta}_N \leqslant \overline{\delta} + C(\rho)\left(D^2+\sigma(\ell,\mathscr{F})\sqrt{n}M_\Delta\right)\left(\frac{\mathfrak{B}^6(\ell,\mathscr{F})}{M_\Delta^4 n^2}+\frac{s+\mathscr{O}}{N}\right).$$

with probability at least $1-10e^{-s}$, where the constant $C(\rho)$ depends on ρ only and D is a constant appearing in Assumption 2.

The main ideas behind the proofs of Theorems 2.2 and 2.3 are explained in the beginning of Section 4.

REMARK 2.2
1. The bounds of Theorems 2.2 and 2.3 hold for the excess risk $\widehat{\delta}_N^U$ of the permutation-invariant estimator $\widehat{f}_N^U$, up to a change in absolute constants.
2. It is evident that whenever $\mathscr{O}=0$, the best possible rates implied by Theorem 2.2 are of order $N^{-2/3}$ (indeed, this is the case whenever $M_\Delta\sqrt{n}\asymp N^{1/3}$ and $\overline{\delta}\lesssim N^{-2/3}$), while the best possible rates attained by Theorem 2.3 are of order $N^{-3/4}$ (when $M_\Delta\sqrt{n}\asymp N^{1/4}$ and $\overline{\delta}\lesssim N^{-3/4}$); in particular, in this case the choice of M_Δ and n is independent of $\overline{\delta}$. In general, if $\mathscr{O}=\varepsilon N$ for $\varepsilon>0$, the best rates implied by Theorems 2.2 and 2.3 are $\overline{\delta}+C(\mathscr{F},\rho,P)\varepsilon^{2/3}$ and $\overline{\delta}+C(\mathscr{F},\rho,P)\varepsilon^{3/4}$ respectively.
3. Assumption requiring that $M_\Delta\geqslant 1$ is introduced for convenience: without it, extra powers of the ratio $\frac{\max(\Delta,\sigma(\ell,\mathscr{F}))}{\Delta}$ appear in the bounds.

Our next goal is to describe an estimator that is capable of achieving excess risk rates up to N^{-1}. The approach that we follow is similar in spirit to the "minmax" estimators studied in [5, 38, 34], among

others, as well as the "median-of-means tournaments" introduced in [40]; all these methods focus on estimating the differences $\mathscr{L}(f_1)-\mathscr{L}(f_2)$ for all $f_1,f_2\in\mathscr{F}$. Recall that $f_*=\operatorname{argmin}_{f\in\mathscr{F}}P\ell(f)$, and observe that for any fixed $f'\in\mathscr{F}$, f_* can be equivalently defined via

$$f_*=\operatorname*{argmin}_{f\in\mathscr{F}}P\left(\ell(f)-\ell(f')\right).$$

A version of the robust empirical risk minimizer (1.4) corresponding to this problem can be defined as

$$\widehat{\mathscr{L}}^{(k)}(f-f'):=\operatorname*{argmin}_{y\in\mathbb{R}}\frac{1}{\sqrt{N}}\sum_{j=1}^{k}\rho\left(\sqrt{n}\,\frac{\left(\overline{\mathscr{L}}_j(f)-\overline{\mathscr{L}}_j(f')\right)-y}{\Delta}\right)$$

for appropriately chosen $\Delta>0$, and

$$\widehat{f}_N:=\operatorname*{argmin}_{f\in\mathscr{F}}\widehat{\mathscr{L}}^{(k)}(f-f').$$

Moreover, if $f'\in\mathscr{F}$ is a priori known to be "close" to f_*, then it suffices to search for the minimizer in a neighborhood $\mathscr{F}'$ of f' that contains f_* instead of all $f\in\mathscr{F}$:

$$\widehat{f}'_N:=\operatorname*{argmin}_{f\in\mathscr{F}'}\widehat{\mathscr{L}}^{(k)}(f-f').$$

The advantage gained by this procedure is expressed by the fact that $\sup_{f\in\mathscr{F}'}\operatorname{Var}\left(\ell(f(X))-\ell(f'(X))\right)$ can be much smaller than $\sigma(\ell,\mathscr{F})$.

We will now formalize this argument and provide performance guarantees; we use the framework of Theorem 2.3 which leads to the bounds that are easier to state and interpret. However, similar reasoning applies to the setting of Theorem 2.2 as well. The presented algorithms also admit straightforward permutation-invariant modifications that we omit. Let

$$\widehat{\mathscr{E}}_N(f):=\widehat{\mathscr{L}}^{(k)}(f)-\widehat{\mathscr{L}}^{(k)}(\widehat{f}_N)$$

be the "empirical excess risk" of f. Indeed, this is a meaningful notion as $\widehat{f}_N$ is the minimizer of $\widehat{\mathscr{L}}^{(k)}(f)$ over $f\in\mathscr{F}$. Assume that the initial sample of size N is split into two disjoint parts S_1 and S_2 of cardinalities that differ by at most 1: $(X_1,Y_1),\dots,(X_N,Y_N)=S_1\cup S_2$. The algorithm proceeds in the following way:

1. Let $\widehat{f}_{|S_1|}$ be the estimator (1.4) evaluated over subsample S_1 of cardinality $|S_1|\geqslant\lfloor N/2\rfloor$, with the scale parameter Δ_1 and the partition parameter k_1 corresponding the group size $n_1=\lfloor|S_1|/k_1\rfloor$;

2. Let $\delta'=\overline{\delta}+C(\rho)\left(D^2+\sigma(\ell,\mathscr{F})\sqrt{n}M_{\Delta_1}\right)\left(\frac{\mathfrak{B}^6(\ell,\mathscr{F})}{M_{\Delta_1}^4\,n_1^2}+\frac{s+\mathscr{O}}{N}\right)$ be a known upper bound on the excess risk in Theorem 2.3 (while this condition is restrictive, it is similar to the requirements of existing approaches [13, 40]; discussion of adaptation issues is beyond the scope of this paper and will be addressed elsewhere). Set

$$\widehat{\mathscr{F}}(\delta'):=\left\{f\in\mathscr{F}:\ \widehat{\mathscr{E}}_N(f)\leqslant\delta'\right\}.$$

3. Define $\widehat{f}''_N := \operatorname{argmin}_{f\in\widehat{\mathscr{F}}(\delta')} \widehat{\mathscr{L}^{(k)}}(f-\widehat{f}_{|S_1|})$ where

$$\widehat{\mathscr{L}^{(k)}}\left(f-\widehat{f}_{|S_1|}\right) = \operatorname*{argmin}_{y\in\mathbb{R}} \sum_{j=1}^{k_2} \rho\left(\sqrt{n}\,\frac{\left(\overline{\mathscr{L}}_j(f)-\overline{\mathscr{L}}_j(\widehat{f}_{|S_1|})\right)-y}{\Delta_2}\right)$$

is based on the subsample S_2 of cardinality $|S_2|\geqslant\lfloor N/2\rfloor$, a scale parameter Δ_2 and the partition parameter k_2 corresponding the group size $n_2=\lfloor|S_2|/k_2\rfloor$.

It will be demonstrated in the course of the proofs that on an event of high probability, $\widehat{\mathscr{F}}(\delta')\subseteq\mathscr{F}(c\delta')$ for an absolute constant $c\leqslant 7$. Hence, on this event $\sup_{f\in\widehat{\mathscr{F}}(\delta')}\operatorname{Var}\left(\ell(f(X))-\ell(f_*(X))\right)\leqslant\nu^2(c\delta')\leqslant cD^2\delta'$ by the definition of $\nu(\delta)$ and Assumption 2, thus $\Delta_2=DM_{\Delta_2}\sqrt{c\delta'}$ with $M_{\Delta_2}\geqslant 1$ often leads to an estimator with improved performance.

THEOREM 2.4 Suppose that

$$\sup_{f\in\mathscr{F}}\mathbb{E}^{1/4}\left(\ell(f(X))-\mathbb{E}\ell(f(X))\right)^4<\infty$$

and that Δ_1, Δ_2 satisfy $M_{\Delta_1}:=\frac{\Delta_1}{\sigma(\ell,\mathscr{F})}\geqslant 1$ and $M_{\Delta_2}:=\frac{\Delta_2}{D\sqrt{7\delta'}}\geqslant 1$. Moreover, assume that for a sufficiently small absolute constant $c'>0$, $\sup_{f\in\mathscr{F}}\max\left(G_f(n_1,\Delta_1),G_f(n_2,\Delta_2)\right)\leqslant c'$ and $\frac{s+\mathscr{O}}{\min(k_1,k_2)}\leqslant c'$. Finally, we require that

$$\sqrt{k_1}M_{\Delta_1}\geqslant\frac{c'}{\sigma(\ell,\mathscr{F})}\,\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{\sqrt{|S_1|}}\sum_{j=1}^{|S_1|}\left(\ell(f(X_j))-P\ell(f)\right)\text{ and}\qquad(2.6)$$

$$\sqrt{k_2}M_{\Delta_2}\geqslant c'\frac{\sqrt{N\delta'}}{D}.$$

Then

$$\mathscr{E}\left(\widehat{f}''_N\right)\leqslant\overline{\delta}+C(\rho)\left(D^2+D\sqrt{\delta'}\sqrt{n}M_{\Delta_2}\right)\left(\frac{\mathfrak{B}^6(\ell,\mathscr{F})}{M_{\Delta_2}^4n^2}+\frac{s+\mathscr{O}}{N}\right)$$

with probability at least $1-20e^{-s}$, where $C(\rho)$ depends on ρ only and D is the constant appearing in Assumption 2.

The statement of Theorem 2.4 is technical, so let us try to distill the main ideas. The key difference between Theorem 2.3 and Theorem 2.4 is that the "remainder term"

$$\sigma(\ell,\mathscr{F})\sqrt{n}M_\Delta\left(\frac{\mathfrak{B}^6(\ell,\mathscr{F})}{M_\Delta^4n^2}+\frac{s+\mathscr{O}}{N}\right)$$

is replaced by a potentially much smaller quantity $\sqrt{\delta'}\sqrt{n}M_\Delta\left(\frac{\mathfrak{B}^6(\ell,\mathscr{F})}{M_\Delta^4n^2}+\frac{s+\mathscr{O}}{N}\right)$. In particular, if $\delta'\ll\left(nM_\Delta^2\right)^{-1}$, this term often becomes negligible. To be more specific, assume that $\bar{\delta}=\frac{C(\mathscr{F})}{\sqrt{N}}\cdot h(N)$ where $h(N)\to 0$ as $N\to\infty$ (meaning that fast rates are achievable) and that $\mathscr{O}=\varepsilon N$ for $\varepsilon\geqslant\frac{1}{N}$. Moreover, suppose that $\mathfrak{B}(\ell,\mathscr{F})$ is bounded above by a constant. If Δ_1 is chosen such that $\Delta_1\asymp\sigma(\ell,\mathscr{F})$, then

$\delta' = C\left(\overline{\delta} + \sigma(\ell, \mathscr{F})\left(\left(\frac{k}{N}\right)^{3/2} + \frac{s+\mathscr{O}}{\sqrt{kN}}\right)\right)$. Hence, if $\max\left(h(N)\sqrt{N}, N\varepsilon^{2/3}\right) \ll k_j \leqslant CN\sqrt{\varepsilon}$ for $j = 1,2$ and $\Delta_2 \asymp \sqrt{\delta'}$, then

$$\delta' \cdot nM_{\Delta_2}^2 = O(1),$$

and the excess risk of $\widehat{f}_N''$ admits the bound

$$\mathscr{E}\left(\widehat{f}_N''\right) \leqslant \overline{\delta} + C(\rho, D)\left(\varepsilon + \frac{s}{N}\right)$$

that holds with probability at least $1 - Ce^{-s}$. A possible choice satisfying all the required conditions is $k_j \asymp N\sqrt{\varepsilon},\ j = 1,2$ (indeed, it this case it is straightforward to check that conditions (2.6) hold for sufficiently large N as $k_j \gtrsim \sqrt{N},\ j = 1,2$). Analysis of the case when $\mathscr{O} = 0$ follows similar steps, with several simplifications.

3. Examples.

In this section, we consider two common prediction problems, regression and binary classification, and discuss the implications of our main results for these problems in detail.

3.1 *Binary classification with convex surrogate loss.*

The key elements of the binary classification framework were outlined in Section 1.4. Here, we recall few popular examples of classification-calibrated losses and present conditions that are sufficient for the Assumption 2 to hold.

Logistic loss $\ell(yf(z)) = \log\left(1 + e^{-yf(z)}\right)$. Consider two scenarios:

1. Uniformly bounded classes, meaning that for all $f \in \mathscr{F}$, $\sup_{z\in S}|f(z)| \leqslant B$. In this case, Assumption 2 holds with $D = 2e^B$ for all $f \in \mathscr{F}$. See [8] and Proposition 6.1 in [3].
2. Linear separators and Gaussian design: in this case, we assume that $S = \mathbb{R}^d$, $Z \sim N(0, I)$ is Gaussian, and $\mathscr{F} = \{\langle\cdot, v\rangle :\ \|v\|_2 \leqslant R\}$ is a class of linear functions. In this case, according to the Proposition 6.2 in [3], Bernstein's assumption is satisfied with $D = cR^{3/2}$ for some absolute constant $c > 0$.

Hinge loss $\ell(yf(z)) = \max(0, 1 - yf(z))$. In this case, sufficient condition for Assumption 2 to hold is the following: there exists $\tau > 0$ such that $|g_*(Z)| \geqslant \tau$ almost surely, where $g_*(z) = \mathbb{E}[Y|Z = z]$. It follows from Theorem 7 in [8] (see also [55]) that Assumption 2 holds with $D = \frac{1}{\sqrt{2\tau}}$ in this case.

Bound for $\overline{\delta}$. Let Π stand for the marginal distribution of Z and recall that

$$\omega(\delta) := \mathbb{E}\sup_{\ell(f)\in\mathscr{F}(\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left((\ell(Y_j f(Z_j)) - \ell(Y_j f_*(Z_j))) - \mathbb{E}(\ell(Yf(Z)) - \ell(Yf_*(Z)))\right)\right|.$$

Since ℓ is Lipschitz continuous by assumption (with Lipschitz constant denoted $L(\ell)$), consequent application of symmetrization and Talagrand's contraction inequalities [37, 56] yields that

$$\omega(\delta) \leqslant 4L(\ell)\,\mathbb{E}\sup_{\|f-f_*\|_{L_2(\Pi)}\leqslant D\sqrt{\delta}}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\varepsilon_j(f - f_*)(Z_j)\right|$$

where $\varepsilon_1,\dots,\varepsilon_N$ are i.i.d. random signs independent from Y_j's and Z_j's. The latter quantity is the modulus of continuity of a Rademacher process, and various upper bounds for it are well known. For instance, if $\mathscr{F}$ is a subset of a linear space of dimension d, then, according to Proposition 3.2 in [31], $\mathbb{E}\sup_{\|f-f_*\|_{L_2(\Pi)}\leqslant D\sqrt{\delta}}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N \varepsilon_j(f-f_*)(Z_j)\right| \leqslant D\sqrt{\delta}\sqrt{d}$, whence $\widetilde{\omega}(\delta) := 4DL(\ell)\sqrt{\delta d}$ is an upper bound for $\omega(\delta)$ and is of concave type, implying that

$$\overline{\delta} \leqslant C(\rho,\ell)D^2\,\frac{d}{N}.$$

More generally, assume that the class $\mathscr{F}$ has a measurable envelope $F(z) := \sup_{f\in\mathscr{F}}|f(z)|$ that satisfies $\|F(Z)\|_{\psi_2} < \infty$, where $\|\xi\|_{\psi_2} := \inf\left\{C>0:\ \mathbb{E}\exp\left(|\xi/C|^2\right)\leqslant 2\right\}$ is the ψ_2 (Orlicz) norm. Moreover, suppose that the covering numbers $N(\mathscr{F},Q,\varepsilon)$ of the class $\mathscr{F}$ with respect to the norm $L_2(Q)$ [3] satisfy the bound

$$N(\mathscr{F},Q,\varepsilon) \leqslant \left(\frac{A\|F\|_{L_2(Q)}}{\varepsilon}\right)^V \tag{3.1}$$

for some constants $A\geqslant 1$, $V\geqslant 1$, all $0<\varepsilon\leqslant 2\|F\|_{L_2(Q)}$ and all probability measures Q. For instance, VC-subgraph classes are known to satisfy this bound with V being the VC dimension of $\mathscr{F}$ [58, 31]. In this case, it is not difficult to show (see for example the proof of Lemma 3.1 in the Supplementary material [41]) that

$$\mathbb{E}\sup_{\|f-f_*\|_{L_2(\Pi)}\leqslant D\sqrt{\delta}}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N \varepsilon_j(f-f_*)(Z_j)\right| \leqslant \widetilde{\omega}(\delta) := C\sqrt{V\log(e^2A^2N)}\left(\sqrt{\delta}+\sqrt{\frac{V}{N}}\log(A^2N)\|F\|_{\psi_2}\right),$$

hence it is easy to check that in this case

$$\overline{\delta} \leqslant C(\rho)\frac{V\log^{3/2}(e^2A^2N)\|F\|_{\psi_2}}{N}.$$

It immediately follows from the discussion following Theorem 2.4 that the excess risk of the estimator $\widehat{f}_N''$ satisfies

$$\mathscr{E}\left(\widehat{f}_N''\right) \leqslant C(\rho,D)\left(\frac{\mathscr{O}}{N}+\frac{V\log^{3/2}(e^2A^2N)\|F\|_{\psi_2}+s}{N}\right)$$

with probability at least $1-20e^{-s}$. Note that we did not need to assume that $h_* := \operatorname*{argmin}_{\text{all measurable } f}\mathbb{E}\ell(Yf(Z))$ belongs to $\mathscr{F}$. Similar results hold for regression problems with Lipschitz losses, such as Huber's loss or quantile loss [3].

3.2 *Regression with quadratic loss.*

Let $X=(Z,Y)\in S\times\mathbb{R}$ be a random couple with distribution P satisfying $Y=f_*(Z)+\eta$ where the noise variable η is independent of Z and $f_*(z)=\mathbb{E}[Y|Z=z]$ is the regression function. Let $\|\eta\|_{2,1} := \int_0^\infty\sqrt{\Pr(|\eta|>t)}\mathrm{d}t$, and observe that $\|\eta\|_{2,1}<\infty$ as $\sup_{f\in\mathscr{F}}\mathbb{E}(Y-f(Z))^4<\infty$ by assumption. As

[3] Definition: the covering number $N(\mathscr{F},Q,\varepsilon)$ is the smallest integer $k\geqslant 1$ such that there exist $f_1,\dots,f_k\in L_2(Q)$ satisfying $\bigcup_{j=1}^k B(f_j,\varepsilon)\supseteq\mathscr{F}$, where $B(f_j,\varepsilon)$ is the $L_2(Q)$ ball of radius ε centered at f_j.

before, Π will stand for the marginal distribution of Z. Let $\mathscr{F}$ be a given convex class of functions mapping S to $\mathbb{R}$ and such that the regression function f_* belongs to $\mathscr{F}$, so that

$$f_* = \operatorname*{argmin}_{f\in\mathscr{F}} \mathbb{E}\,(Y - f(Z))^2.$$

In this case, the natural choice for the loss function is the quadratic loss $\ell(x) = x^2$ which is not Lipschitz continuous on unbounded domains. Assume that the class $\mathscr{F}$ has a measurable envelope $F(z) := \sup_{f\in\mathscr{F}} |f(z)|$ that satisfies $\|F(Z)\|_{\psi_2} < \infty$. Moreover, suppose that the covering numbers $N(\mathscr{F}, Q, \varepsilon)$ of the class $\mathscr{F}$ with respect to the norm $L_2(Q)$ satisfy the bound

$$N(\mathscr{F}, Q, \varepsilon) \leqslant \left(\frac{A\|F\|_{L_2(Q)}}{\varepsilon}\right)^V$$

for some constants $A \geqslant 1$, $V \geqslant 1$, all $0 < \varepsilon \leqslant 2\|F\|_{L_2(Q)}$, and all probability measures Q (see remark about VC-subgraph classes following display (3.1)).

Verification of Bernstein's assumption. It follows from Lemma 5.1 in [31] that

$$\mathscr{F}(\delta) \subseteq \left\{(y - f(z))^2 : \ f \in \mathscr{F},\ \mathbb{E}(f(Z) - f_*(Z))^2 \leqslant 2\delta\right\},$$

hence $v(\delta) \leqslant \sqrt{2\delta}$ so D can be taken to be $\sqrt{2}$ in Assumption 2.

Bound for $\bar{\delta}$. Required estimates follow from the following lemma:

LEMMA 3.1 Under the assumptions made in this section and for $\Delta \geqslant \sigma(\ell, \mathscr{F})$,

$$\bar{\delta} \leqslant C(\rho)\frac{V\log^2(A^2N)(\|F\|_{\psi_2}^2 + \|\eta\|_{2,1}^2)}{N}.$$

Moreover, if the functions if $\mathscr{F}$ are uniformly bounded, the $\log^2(A^2N)$ can be removed.

The proof is given in Section A.9 of the Supplementary material [41]. An immediate corollary of the lemma, according to the discussion following Theorem 2.4, is that the excess risk of the estimator $\widehat{f}_N'$ satisfies the inequality

$$\mathscr{E}\left(\widehat{f}_N'\right) \leqslant C(\rho)\left(\frac{\mathscr{O}}{N} + \frac{V\log^2(A^2N)(\|F\|_{\psi_2}^2 + \|\eta\|_{2,1}^2) + s}{N}\right)$$

with probability at least $1 - 20e^{-s}$, for $0 < s \leqslant cN^{1/4}$.

4. Proofs of the main results.

In the proofs of the main results, we will rely on the following convenient change of variables. Denote

$$\widehat{G}_k(z; f) = \frac{1}{\sqrt{k}}\sum_{j=1}^{k} \rho'\left(\sqrt{n}\,\frac{(\overline{\mathscr{L}}_j(f) - \mathscr{L}(f)) - z}{\Delta}\right),$$

$$G_k(z; f) = \sqrt{k}\,\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\overline{\mathscr{L}}_1(f) - \mathscr{L}(f)) - z}{\Delta}\right).$$

In particular, when $\mathscr{O}=0$, $G_k(z;f)=\mathbb{E}\widehat{G}_k(z;f)$. Let $\widehat{e}^{(k)}(f)$ and $e^{(k)}(f)$ be defined by the equations

$$\widehat{G}_k\left(\widehat{e}^{(k)}(f);f\right)=0,$$
$$G_k\left(e^{(k)}(f);f\right)=0.$$

Comparing this to the definition of $\widehat{\mathscr{L}^{(k)}}(f)$ (1.2), it is easy to see that $\widehat{e}^{(k)}(f)=\widehat{\mathscr{L}^{(k)}}(f)-\mathscr{L}(f)$.

Let us explain the main high-level ideas behind the proof. In the classical empirical risk minimization framework, $\widehat{\mathscr{L}^{(k)}}(f)$ is replaced by the empirical mean $P_N\ell(f)=\frac{1}{N}\sum_{j=1}^N\ell(f(X_j))$; in particular, it is linear in $\ell(f)$, meaning that $P_N(\ell(f_1)-\ell(f_2))=P_N\ell(f_1)-P_N\ell(f_2)$, while $\widehat{\mathscr{L}^{(k)}}(f)$ lacks this property. Imagine that $\widehat{\mathscr{L}^{(k)}}(f)$ was linear in $\ell(f)$. Then, setting $\widehat{\delta}_N=\mathscr{L}(\widehat{f}_N)-\mathscr{L}(f_*)$, we would be able to write that

$$\widehat{\delta}_N=\mathscr{L}(\widehat{f}_N)-\mathscr{L}(f_*)=(\mathscr{L}(\widehat{f}_N)-\widehat{\mathscr{L}^{(k)}}(\widehat{f}_N))-(\mathscr{L}(f_*)-\widehat{\mathscr{L}^{(k)}}(f_*))+\underbrace{\widehat{\mathscr{L}^{(k)}}(\widehat{f}_N)-\widehat{\mathscr{L}^{(k)}}(f_*)}_{\leqslant 0}$$
$$\leqslant\sup_{f:\mathscr{E}(f)\leqslant\widehat{\delta}_N}\left|\widehat{\mathscr{L}^{(k)}}(f-f_*)-\mathscr{L}(f-f_*)\right|. \quad (4.1)$$

It would then suffice to find a good upper bound for the supremum on the right side of (4.1) and solve the resulting inequality to get an upper bound for $\widehat{\delta}_N$. However, this argument does not work directly due to the lack of linearity. Instead, we use Bahadur-type representation of the $\widehat{e}^{(k)}(f)$ to introduce linearity into the problem. Specifically, we will show that $\widehat{e}^{(k)}(f)=-\frac{\widehat{G}_k(0;f)}{\partial_z G_k(0;f)}+r_N(f)$ where $r_N(f)$ is a small remainder term and $\partial_z G_k(0;f)$ is the partial derivative of $G_k(z;f)$ with respect to z evaluated in $z=0$. The process $\widehat{G}_k(0;f)$ is "almost" linear in $\ell(f)$, the only obstacle being the nonlinearity due to ρ'. Mimicking (4.1), we can write that

$$\widehat{\delta}_N=\widehat{e}^{(k)}(\widehat{f}_N)-\widehat{e}^{(k)}(f_*)+\underbrace{\widehat{\mathscr{L}^{(k)}}(\widehat{f}_N)-\widehat{\mathscr{L}^{(k)}}(f_*)}_{\leqslant 0}\leqslant\widehat{e}^{(k)}(\widehat{f}_N)-\widehat{e}^{(k)}(f_*)$$
$$=\left|\frac{\widehat{G}_k\left(0;\widehat{f}_N\right)}{\partial_z G\left(0;\widehat{f}_N\right)}-\frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}\right|+r'_N(\widehat{f}_N,f_*)\leqslant\sup_{f:\mathscr{E}(f)\leqslant\widehat{\delta}_N}\left(\left|\frac{\widehat{G}_k(0;f)}{\partial_z G(0;f)}-\frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}\right|+r'_N(f,f_*)\right)$$

for appropriately defined $r'_N(\cdot,\cdot)$. The difference $\frac{\widehat{G}_k(0;f)}{\partial_z G(0;f)}-\frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}$ can be tackled with the techniques commonly used to estimate suprema of the empirical processes; in particular, symmetrization and contraction inequalities for Rademacher sums [37] are used to remove the additional nonlinearity in the definition of $\widehat{G}_k(z,f)$ introduced by ρ'. At that point, one only needs to carefully estimate the remainder term r'_N.

4.1 *Technical tools.*

We summarize the key results that our proofs rely on.

LEMMA 4.1 Let ρ satisfy Assumption 1. Then for any random variable Y with $\mathbb{E}Y^2<\infty$,

$$\operatorname{Var}\left(\rho'(Y)\right)\leqslant\operatorname{Var}(Y).$$

Proof. See Lemma 5.3 in [47]. □

LEMMA 4.2 For any function h of with bounded third derivative and a sequence of i.i.d. random variables $\xi_1,\dots,\xi_n$ such that $\mathbb{E}\xi_1 = 0$ and $\mathbb{E}|\xi_1|^3 < \infty$,

$$\left|\mathbb{E}h\left(\sum_{j=1}^n \xi_j\right) - \mathbb{E}h\left(\sum_{j=1}^n Z_j\right)\right| \leqslant Cn\|h'''\|_\infty \mathbb{E}|\xi_1|^3,$$

where $C>0$ is an absolute constant and $Z_1,\dots,Z_n$ are i.i.d. centered normal random variables such that $\mathrm{Var}(Z_1) = \mathrm{Var}(\xi_1)$.

Proof. This bound follows from a standard application of Lindeberg's replacement method; see chapter 11 in [49]. □

LEMMA 4.3 Assume that $\mathbb{E}|f(X)-\mathbb{E}f(X)|^2 < \infty$ for all $f\in\mathscr{F}$ and that ρ satisfies Assumption 1. Then for all $f\in\mathscr{F}$ and $z\in\mathbb{R}$ satisfying $|z| \leqslant \frac{\Delta}{\sqrt{n}}\frac{1}{2}$,

$$\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{(\bar\theta_j(f)-Pf)-z}{\Delta}\right) - \mathbb{E}\rho'\left(\frac{W(f)-\sqrt{n}z}{\Delta}\right)\right| \leqslant 2G_f(n,\Delta).$$

Proof. See Lemma 4.2 in [47]. □

Given N i.i.d. random variables $X_1,\dots,X_N\in\mathscr{S}$, let $\|f-g\|_{L_\infty(\Pi_N)} := \max_{1\leqslant j\leqslant N}|f(X_j)-g(X_j)|$. Moreover, define

$$\Gamma_{n,\infty}(\mathscr{F}) := \mathbb{E}\gamma_2^2(\mathscr{F};L_\infty(\Pi_N)),$$

where $\gamma_2(\mathscr{F},L_\infty(\Pi_N))$ is Talagrand's generic chaining complexity [54].

LEMMA 4.4 Let $\sigma^2 := \sup_{f\in\mathscr{G}}\mathbb{E}f^2(X)$. Then there exists a universal constant $C>0$ such that

$$\mathbb{E}\sup_{f\in\mathscr{F}}\left|\frac{1}{N}\sum_{j=1}^N f^2(X_j)-\mathbb{E}f^2(X)\right| \leqslant C\left(\sigma\sqrt{\frac{\Gamma_{N,\infty}(\mathscr{F})}{N}}\bigvee\frac{\Gamma_{N,\infty}(\mathscr{F})}{N}\right).$$

Proof. See Theorem 3.16 in [31]. □

The following form of Talagrand's concentration inequality is due to Klein and Rio (see Section 12.5 in [12]).

LEMMA 4.5 Let $\{Z_j(f),\ f\in\mathscr{F}\},\ j=1,\dots,N$ be independent (not necessarily identically distributed) separable stochastic processes indexed by class $\mathscr{F}$ and such that $|Z_j(f)-\mathbb{E}Z_j(f)| \leqslant M$ a.s. for all $1\leqslant j\leqslant N$ and $f\in\mathscr{F}$. Then the following inequality holds with probability at least $1-e^{-s}$:

$$\sup_{f\in\mathscr{F}}\left(\sum_{j=1}^N (Z_j(f)-\mathbb{E}Z_j(f))\right) \leqslant 2\mathbb{E}\sup_{f\in\mathscr{F}}\left(\sum_{j=1}^N (Z_j(f)-\mathbb{E}Z_j(f))\right) + V(\mathscr{F})\sqrt{2s} + \frac{4Ms}{3}, \qquad (4.2)$$

where $V^2(\mathscr{F}) = \sup_{f\in\mathscr{F}}\sum_{j=1}^N \mathrm{Var}\,(Z_j(f))$.

It is easy to see, applying (4.2) to processes $\{-Z_j(f),\ f\in\mathscr{F}\}$, that

$$\inf_{f\in\mathscr{F}}\left(\sum_{j=1}^N (Z_j(f)-\mathbb{E}Z_j(f))\right) \geqslant -2\mathbb{E}\sup_{f\in\mathscr{F}}\left(\sum_{j=1}^N (\mathbb{E}Z_j(f)-Z_j(f))\right) - V(\mathscr{F})\sqrt{2s} - \frac{4Ms}{3}$$

with probability at least $1-e^{-s}$.

4.2 *Proof of Theorems 2.2 and 2.3.*

We will provide detailed proofs for the estimator $\widehat{f}_N$ that is based on disjoint subsamples indexed by $G_1,\dots,G_k$. The bounds for its permutation-invariant version $\widehat{f}_N^U$ follow exactly the same steps where all applications of the Talagrand's concentration inequality (Lemma 4.5) should be replaced by its version (B.3) for nondegenerate U-statistics stated in Section B of the Supplementary material [41].

Let $J \subset \{1,\dots,k\}$ of cardinality $|J| \geqslant k - \mathscr{O}$ be the set containing all j such that the subsample $\{X_i,\ i \in G_j\}$ does not include outliers. Clearly, $\{X_i :\ i \in G_j,\ j \in J\}$ are still conditionally i.i.d. as the partitioning scheme is independent of the data. Moreover, set $N_J := \sum_{j\in J}|G_j|$, and note that, since $\mathscr{O} < k/2$,

$$N_J \geqslant n|J| \geqslant \frac{N}{2}.$$

Consider stochastic process $R_N(f)$ defined as

$$R_N(f) = \widehat{G}_k(0;f) + \partial_z G_k(0;f)\cdot \widehat{e}^{(k)}(f), \tag{4.3}$$

where $\partial_z G_k(0;f) := \partial_z G_k(z;f)_{|_{z=0}}$. Whenever $\partial_z G_k(0;f) \neq 0$ (this assumption will be justified by Lemma 4.6 below), we can solve (4.3) for $\widehat{e}^{(k)}(f)$ to obtain

$$\widehat{e}^{(k)}(f) = -\frac{\widehat{G}_k(0;f)}{\partial_z G_k(0;f)} + \frac{R_N(f)}{\partial_z G_k(0;f)}, \tag{4.4}$$

which can be viewed as a Bahadur-type representation of $\widehat{e}^{(k)}(f)$. Setting $f := \widehat{f}_N$ and recalling that $\widehat{e}^{(k)}(f) = \widehat{\mathscr{L}}^{(k)}(f) - \mathscr{L}(f)$, we deduce that

$$\widehat{\mathscr{L}}^{(k)}(\widehat{f}_N) = \mathscr{L}(\widehat{f}_N) - \frac{\widehat{G}_k\left(0;\widehat{f}_N\right)}{\partial_z G_k\left(0;\widehat{f}_N\right)} + \frac{R_N(\widehat{f}_N)}{\partial_z G_k\left(0;\widehat{f}_N\right)}.$$

By the definition (1.4) of $\widehat{f}_N$, $\widehat{\mathscr{L}}^{(k)}(\widehat{f}_N) \leqslant \widehat{\mathscr{L}}^{(k)}(f_*)$, hence

$$\mathscr{L}(\widehat{f}_N) - \frac{\widehat{G}_k\left(0;\widehat{f}_N\right)}{\partial_z G_k\left(0;\widehat{f}_N\right)} + \frac{R_N(\widehat{f}_N)}{\partial_z G_k\left(0;\widehat{f}_N\right)} \leqslant \mathscr{L}(f_*) - \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)} + \frac{R_N(f_*)}{\partial_z G_k(0;f_*)}.$$

Rearranging the terms, it is easy to see that

$$\widehat{\delta}_N = \mathscr{L}(\widehat{f}_N) - \mathscr{L}(f_*) \leqslant \left|\frac{\widehat{G}_k\left(0;\widehat{f}_N\right)}{\partial_z G\left(0;\widehat{f}_N\right)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}\right| + 2\sup_{f\in\mathscr{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|. \tag{4.5}$$

REMARK 4.1 Similar argument also implies, in view of the inequality $\mathscr{L}(f_*) \leqslant \mathscr{L}(\widehat{f}_N)$, that

$$\widehat{\mathscr{L}}^{(k)}(f_*) + \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)} - \frac{R_N(f_*)}{\partial_z G_k(0;f_*)} \leqslant \widehat{\mathscr{L}}^{(k)}(\widehat{f}_N) + \frac{\widehat{G}_k\left(0;\widehat{f}_N\right)}{\partial_z G_k\left(0;\widehat{f}_N\right)} - \frac{R_N(\widehat{f}_N)}{\partial_z G_k\left(0;\widehat{f}_N\right)},$$

hence

$$\widehat{\mathscr{L}}^{(k)}(f_*)-\widehat{\mathscr{L}}^{(k)}(\widehat{f}_N)\leqslant\left|\frac{\widehat{G}_k\left(0;\widehat{f}_N\right)}{\partial_z G\left(0;\widehat{f}_N\right)}-\frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}\right|+2\sup_{f\in\mathscr{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|.$$

It follows from (4.5) that in order to estimate the excess risk of $\widehat{f}_N$, it suffices to obtain the upper bounds for

$$A_1:=\left|\frac{\widehat{G}_k\left(0;\widehat{f}_N\right)}{\partial_z G_k\left(0;\widehat{f}_N\right)}-\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)}\right| \text{ and } A_2:=\sup_{f\in\mathscr{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|.$$

Observe that

$$\begin{aligned}&\frac{\widehat{G}_k\left(0;\widehat{f}_N\right)}{\partial_z G_k\left(0;\widehat{f}_N\right)}-\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)}\\&\qquad=\frac{\widehat{G}_k\left(0;\widehat{f}_N\right)-\widehat{G}_k(0;f_*)}{\partial_z G_k\left(0;\widehat{f}_N\right)}+\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\,\partial_z G_k\left(0;\widehat{f}_N\right)}\left(\partial_z G_k(0;f_*)-\partial_z G_k\left(0;\widehat{f}_N\right)\right).\end{aligned}$$

Since ρ'' is Lipschitz continuous by assumption,

$$\begin{aligned}&\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\,\partial_z G_k\left(0;\widehat{f}_N\right)}\left(\partial_z G_k(0;f_*)-\partial_z G_k\left(0;\widehat{f}_N\right)\right)\right|\\&=\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\,\partial_z G_k\left(0;\widehat{f}_N\right)}\frac{\sqrt{nk}}{\Delta}\mathbb{E}\left(\rho''\left(\sqrt{n}\frac{\overline{\mathscr{L}}_1(f_*)-\mathscr{L}(f_*)}{\Delta}\right)-\rho''\left(\sqrt{n}\frac{\overline{\mathscr{L}}_1(\widehat{f}_N)-\mathscr{L}(\widehat{f}_N)}{\Delta}\right)\right)\right|\\&\leqslant L(\rho'')\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\,\partial_z G_k\left(0;\widehat{f}_N\right)}\frac{\sqrt{nk}}{\Delta^2}\mathrm{Var}^{1/2}\left(\ell(\widehat{f}_N(X))-\ell(f_*(X))\right)\right|\\&=C(\rho)\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\,\partial_z G_k\left(0;\widehat{f}_N\right)}\right|\frac{\sqrt{nk}}{\Delta^2}\nu(\widehat{\delta}_N).\end{aligned}\tag{4.6}$$

The following two lemmas are required to proceed.

LEMMA 4.6 There exist $C(\rho)>0$ such that for any $f\in\mathscr{F}$,

$$|\partial_z G_k(0;f)|\geqslant\frac{\sqrt{kn}}{2\sqrt{2\pi}\Delta}\left(\min\left(\frac{\Delta}{\sqrt{\mathrm{Var}\left(\ell(f(X))\right)}},2\sqrt{\log 2}\right)-\frac{C(\rho)}{\sqrt{n}}\mathbb{E}\left|\frac{\ell(f(X))-P\ell(f)}{\Delta}\right|^3\right).$$

Proof. See Section A.1. □

In particular, the bound of Lemma 4.6 implies that for n large enough,

$$\inf_{f\in\mathscr{F}}|\partial_z G_k(0;f)|\geqslant\frac{1}{4\sqrt{2\pi}}\frac{\sqrt{kn}}{\max\left(\Delta,\sigma(\ell,\mathscr{F})\right)}=\frac{1}{4\sqrt{2\pi}}\frac{\sqrt{kn}}{\widetilde{\Delta}}.\tag{4.7}$$

It is also easy to deduce from the proof of Lemma 4.6 that for small n and $\Delta > \sigma(\ell, \mathscr{F})$, $\inf_{f\in\mathscr{F}} |\partial_z G_k(0;f)| \geqslant c(\rho)\frac{\sqrt{kn}}{\Delta}$ for some positive $c(\rho)$.

LEMMA 4.7 For any $f \in \mathscr{F}$,

$$\widehat{G}_k(0;f) \leqslant 2\left(\sqrt{k}G_f(n,\Delta) + \frac{\sigma(\ell,f)}{\Delta}\sqrt{s} + \frac{2s}{\sqrt{k}} + \frac{\mathscr{O}}{\sqrt{k}}\right)$$

with probability at least $1-2e^{-s}$, where $C>0$ is an absolute constant.

Proof.

See Section A.2. □

Lemma 4.7 and (4.7) imply, together with (4.6), that

$$\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\,\partial_z G_k\left(0;\widehat{f}_N\right)}\left(\partial_z G_k(0;f_*) - \partial_z G_k\left(0;\widehat{f}_N\right)\right)\right| \leqslant C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{\frac{s}{N}} + \frac{G_{f_*}(n,\Delta)}{\sqrt{n}} + \sqrt{n}\frac{s}{N} + \sqrt{n}\frac{\mathscr{O}}{N}\right)\nu(\widehat{\delta}_N)$$

on event Θ_1 of probability at least $1-2e^{-s}$. As $\widetilde{\Delta} \geqslant \sigma(\ell,\mathscr{F})$ by assumption, we deduce that

$$\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\,\partial_z G_k\left(0;\widehat{f}_N\right)}\left(\partial_z G_k(0;f_*) - \partial_z G_k\left(0;\widehat{f}_N\right)\right)\right| \leqslant C(\rho)\nu(\widehat{\delta}_N)\left(\sqrt{\frac{s}{N}} + \frac{G_{f_*}(n,\Delta)}{\sqrt{n}} + \sqrt{n}\frac{s}{N} + \sqrt{n}\frac{\mathscr{O}}{N}\right).$$

Define

$$\overline{\delta}_1 := \min\left\{\delta > 0 :\ C_1(\rho)\left(\sqrt{\frac{s}{N}} + \frac{G_{f_*}(n,\Delta)}{\sqrt{n}} + \sqrt{n}\frac{s+\mathscr{O}}{N}\right)\frac{\widetilde{\nu}(\delta)}{\delta} \leqslant \frac{1}{7}\right\}, \tag{4.8}$$

where $C_1(\rho)$ is sufficiently large. It is easy to see that on event $\Theta_1 \cap \{\widehat{\delta}_N > \overline{\delta}_1\}$,

$$\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\,\partial_z G_k\left(0;\widehat{f}_N\right)}\left(\partial_z G_k(0;f_*) - \partial_z G_k\left(0;\widehat{f}_N\right)\right)\right| \leqslant \frac{\widehat{\delta}_N}{7}, \tag{4.9}$$

for appropriately chosen $C_1(\rho)$.

Our next goal is to obtain an upper bound for $\left|\frac{\widehat{G}_k(0;\widehat{f}_N)-\widehat{G}_k(0;f_*)}{\partial_z G_k(0;\widehat{f}_N)}\right|$. To this end, we will need to control the local oscillations of the process $\widehat{G}_k(0;f)$. Specifically, we are interested in the bounds on the random variable $\sup_{f\in\mathscr{F}(\delta)}\left|\widehat{G}_k(0;f) - \widehat{G}_k(0;f_*)\right|$. The following technical lemma is important for the analysis.

LEMMA 4.8 Let $(\xi_1,\eta_1),\dots,(\xi_n,\eta_n)$ be a sequence of independent identically distributed random couples such that $\mathbb{E}\xi_1 = 0$, $\mathbb{E}\eta_1 = 0$, and $\mathbb{E}|\xi_1|^2 + \mathbb{E}|\eta_1|^2 < \infty$. Let F be an odd, smooth function with

bounded derivatives up to fourth order. Then

$$\left|\mathbb{E}F\left(\sum_{j=1}^n \xi_j\right) - \mathbb{E}F\left(\sum_{j=1}^n \eta_j\right)\right| \leqslant \max_{\alpha\in[0,1]} \sqrt{n}\,\mathrm{Var}^{1/2}(\xi_1-\eta_1)\left(\mathbb{E}\left|F'\left(S_n^\eta + \alpha\left(S_n^\xi - S_n^\eta\right)\right)\right|^2\right)^{1/2}.$$

Moreover, if $\mathbb{E}|\xi_1|^4 + \mathbb{E}|\eta_1|^4 < \infty$, then

$$\left|\mathbb{E}F\left(\sum_{j=1}^n \xi_j\right) - \mathbb{E}F\left(\sum_{j=1}^n \eta_j\right)\right| \leqslant C(F)\cdot n\Bigg(\mathrm{Var}^{1/2}(\xi_1-\eta_1)\left(R_4^2 + \sqrt{n-1}R_4^3\right) + \left(\mathbb{E}|\xi_1-\eta_1|^4\right)^{1/4} R_4^3\Bigg),$$

where $R_4 = \left(\max\left(\mathbb{E}|\xi_1|^4, \mathbb{E}|\eta_1|^4\right)\right)^{1/4}$ and $C(F) > 0$ is a constant that depends only on F.

Proof. See Section A.3. □

Now we are ready to state the bound for the local oscillations of the process $\widehat{G}_k(0;f)$. Let

$$U(\delta,s) := \frac{2}{\Delta}\left(8\sqrt{2}\omega(\delta) + \nu(\delta)\sqrt{\frac{s}{2}}\right) + \frac{32s}{3\sqrt{k}}.$$

Moreover, if $\widetilde{\omega}(\delta)$ and $\widetilde{\nu}(\delta)$ are upper bounds for $\omega(\delta)$ and $\nu(\delta)$ and are of concave type, then

$$\widetilde{U}(\delta,s) := \frac{2}{\Delta}\left(c(\gamma)\,\widetilde{\omega}(\delta) + \widetilde{\nu}(\delta)\sqrt{\frac{s}{2}}\right) + \frac{32s}{\sqrt{k}}, \tag{4.10}$$

where $c(\gamma) > 0$ depends only on γ, is also an upper bound for $U(\delta,s)$ of strictly concave type. Moreover, define

$$R_4(\ell,\mathscr{F}) := \sup_{f\in\mathscr{F}} \mathbb{E}^{1/4}\Big(\ell(f(X)) - \mathbb{E}\ell(f(X))\Big)^4,$$

$$\nu_4(\delta) := \sup_{f\in\mathscr{F}(\delta)} \mathbb{E}^{1/4}\Big(\ell(f(X)) - \ell(f_*(X)) - \mathbb{E}\left(\ell(f(X)) - \ell(f_*(X))\right)\Big)^4,$$

$$\mathfrak{B}(\ell,\mathscr{F}) := \frac{R_4(\ell,\mathscr{F})}{\sigma(\ell,\mathscr{F})},$$

$$\widetilde{B}(\delta) := \begin{cases} \frac{\widetilde{\nu}(\delta)}{\Delta}\frac{1}{M_\Delta}, & R_4(\ell,\mathscr{F}) = \infty, \\ \frac{\mathfrak{B}^3(\ell,\mathscr{F})}{\sqrt{n}}\left(\frac{\widetilde{\nu}(\delta)}{\Delta}\frac{1}{M_\Delta^2} + \frac{\widetilde{\nu}_4(\delta)}{\Delta}\frac{1}{M_\Delta^3\sqrt{n}}\right), & R_4(\ell,\mathscr{F}) < \infty, \end{cases}$$

where $\widetilde{\nu}_4(\delta)$ upper bounds $\nu_4(\delta)$ and is of concave type. Below, we will use a crude bound $\nu_4(\delta) \leqslant 2R_4(\ell,\mathscr{F})$, but additional improvements are possible if better estimates of $\nu_4(\delta)$ are available.

LEMMA 4.9 With probability at least $1 - e^{-2s}$,

$$\sup_{f\in\mathscr{F}(\delta)} \left|\widehat{G}_k(0;f) - \widehat{G}_k(0;f_*)\right| \leqslant U(\delta,s) + C(\rho)\sqrt{k}\widetilde{B}(\delta) + 4\frac{\mathscr{O}}{\sqrt{k}},$$

where $C(\rho) > 0$ is constant that depends only on ρ.

Proof. See Section A.4. □

Next, we state the "uniform version" of Lemma 4.9.

LEMMA 4.10 With probability at least $1-e^{-s}$, for all $\delta \geqslant \delta_{\min}$ simultaneously,

$$\sup_{f\in\mathscr{F}(\delta)}\left|\widehat{G}_k(0;f)-\widehat{G}_k(0;f_*)\right| \leqslant C(\rho)\delta\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}+\sqrt{k}\frac{\widetilde{B}(\delta_{\min})}{\delta_{\min}}\right)+4\frac{\mathscr{O}}{\sqrt{k}},$$

where $C(\rho)>0$ is constant that depends only on ρ.

Proof. See Section A.5. □

It follows from Lemma 4.10 and inequality (4.7) that on event Θ_2 of probability at least $1-e^{-s}$, for all $\delta \geqslant \delta_{\min}$ simultaneously,

$$\sup_{f\in\mathscr{F}(\delta)}\left|\frac{\widehat{G}_k(0;f)-\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f)}\right| \leqslant C(\rho)\delta\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}+\frac{\widetilde{\Delta}}{\sqrt{n}}\frac{\widetilde{B}(\delta_{\min})}{\delta_{\min}}\right)+4\widetilde{\Delta}\sqrt{n}\frac{\mathscr{O}}{N}.$$

Define

$$\overline{\delta}_2 := \min\left\{\delta>0:\ C_2(\rho)\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta,s)}{\delta}\leqslant\frac{1}{7}\right\},$$
$$\overline{\delta}_3 := \min\left\{\delta>0:\ C_3(\rho)\frac{\widetilde{\Delta}}{\sqrt{n}}\frac{\widetilde{B}(\delta)}{\delta}\leqslant\frac{1}{7}\right\},$$

where $C_2(\rho)$, $C_3(\rho)$ are sufficiently large constants. Then, on event $\Theta_2\cap\left\{\widehat{\delta}_N>\max(\overline{\delta}_2,\overline{\delta}_3)\right\}$,

$$\sup_{f\in\mathscr{F}(\widehat{\delta}_N)}\left|\frac{\widehat{G}_k(0;f)-\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f)}\right| \leqslant \frac{2\widehat{\delta}_N}{7}+4\widetilde{\Delta}\sqrt{n}\frac{\mathscr{O}}{N} \tag{4.11}$$

for appropriately chosen $C_2(\rho), C_3(\rho)$.
Finally, we provide an upper bound for the process $R_N(f)$ defined via

$$R_N(f)=\widehat{G}_k(0;f)+\partial_z G_k(0;f)\cdot\widehat{e}^{(k)}(f).$$

LEMMA 4.11 Assume that conditions of Theorem 2.1 hold, and let $\delta_{\min}>0$ be fixed. Then for all $s>0$, $\delta\geqslant\delta_{\min}$, positive integers n and k such that

$$\delta\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}\sqrt{k}}+\sup_{f\in\mathscr{F}}G_f(n,\Delta)+\frac{s+\mathscr{O}}{k}\leqslant c(\rho), \tag{4.12}$$

the following inequality holds with probability at least $1-7e^{-s}$, uniformly over all δ satisfying (4.12):

$$\sup_{f\in\mathscr{F}(\delta)}|R_N(f)|\leqslant C(\rho)\sqrt{N}\frac{\widetilde{\Delta}^2}{\Delta^2}\left(n^{1/2}\delta^2\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}\sqrt{N}}\right)^2\bigvee\frac{\sigma^2(\ell,f_*)}{\Delta^2}\frac{n^{1/2}s}{N}\right.$$
$$\left.\bigvee n^{1/2}\left(\sup_{f\in\mathscr{F}}\frac{G_f(n,\Delta)}{\sqrt{n}}\right)^2\bigvee n^{3/2}\frac{s^2}{N^2}\bigvee n^{3/2}\frac{\mathscr{O}^2}{N^2}\right).$$

Moreover, the bound of Theorem 2.1 holds on the same event.

Proof. See Section A.6. □

Recall that

$$\overline{\delta}_2 = \min\left\{\delta > 0 : \ C_2(\rho)\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta,s)}{\delta} \leqslant \frac{1}{7}\right\},$$

where $C_2(\rho)$ is a large enough constant. Let Θ_3 be the event of probability at least $1-7e^{-s}$ on which Lemma 4.11 holds with $\delta_{\min} = \overline{\delta}_2$, and consider the event $\Theta_3 \cap \{\widehat{\delta}_N > \overline{\delta}_2\}$. We will now show that on this event, Lemma 4.11 applies with $\delta = \widehat{\delta}_N$. Indeed, the bound of Theorem 2.1 is valid on Θ_3, hence the inequality (2.4) implies that on Θ_3, $\widehat{\delta}_N \leqslant C(\rho)\frac{\widetilde{\Delta}}{\sqrt{n}}$, and it is straightforward to check that condition (4.12) of Lemma 4.11 holds with $\delta_{\min} = \overline{\delta}_2$ and $\delta = \widehat{\delta}_N$. It follows from inequality (4.7) that on event $\Theta_3 \cap \{\widehat{\delta}_N \geqslant \overline{\delta}_2\}$,

$$\sup_{f\in\mathscr{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right| \leqslant C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{n^{1/2}}{\widetilde{\Delta}}\widehat{\delta}_N^2\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta_2,s)}{\delta_2}\right)^2 \bigvee \widetilde{\Delta}\frac{\sigma^2(\ell,f_*)}{\Delta^2}\frac{n^{1/2}s}{N}\right.$$
$$\left. \bigvee n^{1/2}\widetilde{\Delta}\left(\sup_{f\in\mathscr{F}}\frac{G_f(n,\Delta)}{\sqrt{n}}\right)^2 \bigvee n^{3/2}\widetilde{\Delta}\frac{s^2+\mathscr{O}^2}{N^2}\right).$$

Consider the expression

$$C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\frac{n^{1/2}}{\widetilde{\Delta}}\widehat{\delta}_N^2\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta_2,s)}{\delta_2}\right)^2 = C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta_2,s)}{\delta_2}\right)^2\widehat{\delta}_N\cdot\frac{n^{1/2}\widehat{\delta}_N}{\widetilde{\Delta}},$$

and observe that whenever Theorem 2.1 holds, $\frac{n^{1/2}\widehat{\delta}_N}{\widetilde{\Delta}} \leqslant c(\rho)$, hence the latter is bounded from above by

$$\widehat{\delta}_N\cdot C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\overline{\delta}_2,s)}{\overline{\delta}_2}\right)^2 \leqslant \frac{\widehat{\delta}_N}{7}$$

whenever $\Delta \geqslant \sigma(\ell,\mathscr{F})$ (so that $\widetilde{\Delta} = \Delta$) and $C_2(\rho)$ in the definition of $\overline{\delta}_2$ is large enough. Moreover,

$$C(\rho)\frac{\widetilde{\Delta}^3}{\Delta^3}\frac{\sigma^2(\ell,f_*)}{\Delta}\frac{n^{1/2}s}{N} \leqslant C'(\rho)\cdot\sigma(\ell,f_*)\sqrt{n}\frac{s}{N} \leqslant C'(\rho)\widetilde{\Delta}\sqrt{n}\frac{s}{N}$$

if $\widetilde{\Delta} \geqslant \sigma(\ell,f_*)$. As $\frac{s+\mathscr{O}}{k} \leqslant c$ under the conditions of Theorem 2.1, $n^{3/2}\widetilde{\Delta}\frac{s^2+\mathscr{O}^2}{N^2} \leqslant C\widetilde{\Delta}\sqrt{n}\frac{s+\mathscr{O}}{N}$. Combining the inequalities obtained above, we deduce on event $\Theta_3 \cap \{\widehat{\delta}_N \geqslant \overline{\delta}_2\}$,

$$2\sup_{f\in\mathscr{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right| \leqslant \frac{2\widehat{\delta}_N}{7} + C(\rho)\widetilde{\Delta}\left(\sqrt{n}\frac{s+\mathscr{O}}{N}\bigvee\frac{\sup_{f\in\mathscr{F}}\left(G_f(n,\Delta)\right)^2}{\sqrt{n}}\right)$$

whenever $\widetilde{\Delta} \geqslant \sigma(\ell,\mathscr{F})$. Finally, define

$$\overline{\delta}_4 := C_4(\rho)\widetilde{\Delta}\left(\sqrt{n}\frac{s+\mathscr{O}}{N}\bigvee\frac{\sup_{f\in\mathscr{F}}\left(G_f(n,\Delta)\right)^2}{\sqrt{n}}\right),$$

where $C_4(\rho)$ is sufficiently large. Then on event $\Theta_3 \cap \left\{\widehat{\delta}_N \geqslant \max\left(\overline{\delta}_2, 7\overline{\delta}_4\right)\right\}$,

$$2 \sup_{f\in\mathscr{F}(\widehat{\delta}_N)} \left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right| + 4\widetilde{\Delta}\sqrt{n}\frac{\mathscr{O}}{N} \leqslant \frac{2\widehat{\delta}_N}{7} + \frac{\widehat{\delta}_N}{7} = \frac{3\widehat{\delta}_N}{7}. \tag{4.13}$$

Note that the expression above takes care of the term $4\widetilde{\Delta}\sqrt{n}\frac{\mathscr{O}}{N}$ that appeared in (4.11). Combining (4.9), (4.11), (4.13), we deduce that on event $\Theta_1 \cap \Theta_2 \cap \Theta_3 \cap \left\{\widehat{\delta}_N \geqslant \max\left(\overline{\delta}_1, \overline{\delta}_2, \overline{\delta}_3, 7\overline{\delta}_4\right)\right\}$,

$$\widehat{\delta}_N \leqslant \frac{6}{7}\widehat{\delta}_N,$$

leading to a contradiction, hence on event $\Theta_1 \cap \Theta_2 \cap \Theta_3$ of probability at least $1 - 10e^{-s}$,

$$\widehat{\delta}_N \leqslant \max\left(\overline{\delta}_1, \overline{\delta}_2, \overline{\delta}_3, 7\overline{\delta}_4\right). \tag{4.14}$$

Recall the definition (4.8) of $\overline{\delta}_1$. If condition 2 ("Bernstein condition") holds, then $\widetilde{\nu}(\delta) \leqslant D\sqrt{\delta}$ for small enough δ, in which case

$$\overline{\delta}_1 \leqslant C(\rho)D^2\left(\frac{s+\mathscr{O}}{N} + \frac{G^2_{f_*}(n,\Delta)}{n}\right),$$

where we used the fact that $\frac{s}{k} \leqslant c$ by assumption. Together with the bound (2.1) for $G_{f_*}(n,\Delta)$, we deduce that, under the assumption that $R_4(\ell,\mathscr{F}) < \infty$,

$$\overline{\delta}_1 \leqslant C(\rho)D^2\left(\frac{s+\mathscr{O}}{N} + \frac{\left(\mathbb{E}\left|f_*(X) - \mathbb{E}f_*(X)\right|^3\right)^2}{\Delta^6 n^2}\right).$$

Since $\Delta = \sigma(\ell,\mathscr{F})M_\Delta$, $\frac{\mathbb{E}\left|f_*(X)-\mathbb{E}f_*(X)\right|^3}{\Delta^3} \leqslant \frac{\sup_{f\in\mathscr{F}}\mathbb{E}\left|f(X)-\mathbb{E}f(X)\right|^3}{\sigma^3(\ell,\mathscr{F})M_\Delta^3} \leqslant \frac{\mathfrak{B}^3(\ell,\mathscr{F})}{M_\Delta^3}$, where

$$\mathfrak{B}(\ell,\mathscr{F}) = \frac{\sup_{f\in\mathscr{F}}\mathbb{E}^{1/4}\left(\ell(f(X)) - \mathbb{E}\ell(f(X))\right)^4}{\sigma(\ell,\mathscr{F})},$$

hence

$$\overline{\delta}_1 \leqslant C(\rho)D^2\left(\frac{s+\mathscr{O}}{N} + \frac{\mathfrak{B}^6(\ell,\mathscr{F})}{n^2 M_\Delta^6}\right). \tag{4.15}$$

At the same time, if only $\sigma(\ell,\mathscr{F}) < \infty$, we similarly obtain that

$$\overline{\delta}_1 \leqslant C(\rho)D^2\left(\frac{s+\mathscr{O}}{N} + \frac{1}{M_\Delta^4 n}\right). \tag{4.16}$$

Next we will estimate $\overline{\delta}_3$. Recall that, when $R_4(\ell,\mathscr{F}) < \infty$,

$$\widetilde{B}(\delta) = \frac{\mathfrak{B}^3(\ell,\mathscr{F})}{\sqrt{n}}\left(\frac{\widetilde{\nu}(\delta)}{\Delta}\frac{1}{M_\Delta^2} + \frac{\widetilde{\nu}_4(\delta)}{\Delta}\frac{1}{M_\Delta^3\sqrt{n}}\right).$$

For sufficiently small δ (namely, for which condition 2 holds) and $\Delta \geqslant \sigma(\ell, \mathscr{F})$,

$$\frac{\widetilde{\Delta}}{\sqrt{n}}\widetilde{B}(\delta) \leqslant \frac{\mathfrak{B}^3(\ell,\mathscr{F})}{n}\left(\frac{\widetilde{\nu}(\delta)}{M_\Delta^2} + \frac{R_4(\ell,\mathscr{F})}{M_\Delta^3\sqrt{n}}\right) \leqslant \frac{\mathfrak{B}^3(\ell,\mathscr{F})}{n}\left(D\frac{\sqrt{\delta}}{M_\Delta^2} + \sigma(\ell,\mathscr{F})\frac{\mathfrak{B}(\ell,\mathscr{F})}{M_\Delta^3\sqrt{n}}\right)$$

and

$$\overline{\delta}_3 \leqslant C(\rho)\left(D^2\frac{\mathfrak{B}^6(\ell,\mathscr{F})}{n^2 M_\Delta^4} + \sigma(\ell,\mathscr{F})\frac{\mathfrak{B}^4(\ell,\mathscr{F})}{n^{3/2}M_\Delta^3}\right). \tag{4.17}$$

At the same time, if only the second moments are finite, $\widetilde{B}(\delta) = \frac{\widetilde{\nu}(\delta)}{\Delta}\frac{1}{M_\Delta}$, and it is easy to deduce that in this case,

$$\overline{\delta}_3 \leqslant C(\rho)\frac{D^2}{M_\Delta^2 n}. \tag{4.18}$$

Next, we obtain a simpler bound for $\overline{\delta}_4$: as $\Delta \geqslant \sigma(\ell,\mathscr{F})$ by assumption, $\widetilde{\Delta} = \Delta = \sigma(\ell,\mathscr{F})M_\Delta$, and the estimate (2.1) for $G_{f_*}(n,\Delta)$ implies (if $R_4(\ell,\mathscr{F}) < \infty$) that

$$\overline{\delta}_4 \leqslant C(\rho)\,\sigma(\ell,\mathscr{F})\left(\sqrt{n}M_\Delta\frac{s+\mathscr{O}}{N} + \frac{\mathfrak{B}^6(\ell,\mathscr{F})}{M_\Delta^5 n^{3/2}}\right). \tag{4.19}$$

If only $\sigma(\ell,\mathscr{F}) < \infty$, we similarly deduce from (2.1) that

$$\overline{\delta}_4 \leqslant C(\rho)\,\sigma(\ell,\mathscr{F})\left(\sqrt{n}M_\Delta\cdot\frac{s+\mathscr{O}}{N} + \frac{1}{M_\Delta^3\sqrt{n}}\right). \tag{4.20}$$

Finally, recall that $\widetilde{U}(\delta,s) = \frac{2}{\Delta}\left(c(\gamma)\,\widetilde{\omega}(\delta) + \widetilde{\nu}(\delta)\sqrt{\frac{s}{2}}\right) + \frac{32s}{\sqrt{k}}$ and $\overline{\delta}_2 = \min\left\{\delta > 0 :\ C_2(\rho)\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta,s)}{\delta} \leqslant \frac{1}{7}\right\}$, hence

$$\overline{\delta}_2 \leqslant \overline{\delta}\bigvee C(\rho)D^2\frac{s}{N}\bigvee C(\rho)\sigma(\ell,\mathscr{F})\frac{s\sqrt{n}M_\Delta}{N}, \tag{4.21}$$

where $\overline{\delta}$ was defined in (2.5). Combining inequalities (4.15), (4.21), (4.17), (4.19) and (4.14), we obtain the final form of the bound under the stronger assumption $R_4(\ell,\mathscr{F}) < \infty$. Similarly, the combination of (4.16), (4.21), (4.18), (4.20) and (4.14) yields the bound under the weaker assumption $\sigma(\ell,\mathscr{F}) < \infty$.

4.3 *Proof of Theorem 2.4.*

Recall that $\widehat{\mathscr{E}}_N(f_*) := \widehat{\mathscr{L}}^{(k)}(f_*) - \widehat{\mathscr{L}}^{(k)}(\widehat{f}'_N)$ is the "empirical excess risk" of f_*, and let $\widehat{\delta}_N := \mathscr{E}(\widehat{f}'_N)$. It follows from Remark 4.1 that (using the notation used in the proof of Theorems 2.2 and 2.3)

$$\widehat{\mathscr{E}}_N(f_*) \leqslant \left|\frac{\widehat{G}_k\left(0;\widehat{f}'_N\right)}{\partial_z G\left(0;\widehat{f}'_N\right)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}\right| + 2\sup_{f\in\mathscr{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|.$$

On the event of Theorem 2.3 of probability at least $1-10e^{-s}$,

$$\mathscr{E}(\widehat{f}'_N) \leqslant \delta' := \overline{\delta} + C(\rho)\left(D^2\sigma(\ell,\mathscr{F})\sqrt{n}M_\Delta\right)\left(\frac{\mathfrak{B}^6(\ell,\mathscr{F})}{M_\Delta^4 n^2} + \frac{s+\mathscr{O}}{N}\right),$$

hence on this event

$$\widehat{\mathscr{E}}_N(f_*) \leqslant \sup_{f\in\mathscr{F}(\delta')}\left|\frac{\widehat{G}_k(0;f)}{\partial_z G(0;f)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}\right| + 2\sup_{f\in\mathscr{F}(\delta')}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right| \leqslant \frac{6}{7}\delta',$$

where the last inequality again follows from main steps in the proof of Theorem 2.3; note that similar result holds if δ' is replaced by its analogue from Theorem 2.3. Consider the set $\widehat{\mathscr{F}}(\delta') = \left\{f\in\mathscr{F}:\ \widehat{\mathscr{E}}_N(f)\leqslant\delta'\right\}$. First, observe that on the event $\mathscr{E}_1$ of Theorem 2.3, $f_*\in\widehat{\mathscr{F}}(\delta')$ as implied by the previous display. We will next show that $\widehat{\mathscr{F}}(\delta')\subseteq\mathscr{F}(7\delta')$ on the event $\mathscr{E}_1$ of Theorem 2.3, meaning that for any $f\in\widehat{\mathscr{F}}(\delta')$, $\mathscr{E}(f)\leqslant 7\delta'$. Indeed, let $f\in\widehat{\mathscr{F}}(\delta')$ be such that $\mathscr{E}(f)=\sigma$. Then (4.4) implies that

$$\begin{aligned}\mathscr{L}(f)-\mathscr{L}(f_*) &\leqslant \widehat{\mathscr{L}}^{(k)}(f)-\widehat{\mathscr{L}}^{(k)}(f_*) + \left|\frac{\widehat{G}_k(0;f)}{\partial_z G_k(0;f)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)}\right| + \left|\frac{R_N(f)}{\partial_z G_k(0;f)} + \frac{R_N(f_*)}{\partial_z G_k(0;f_*)}\right| \\ &\leqslant \widehat{\mathscr{E}}_N(f) + \sup_{f\in\mathscr{F}(\sigma)}\left|\frac{\widehat{G}_k(0;f)}{\partial_z G_k(0;f)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)}\right| + 2\sup_{f\in\mathscr{F}(\sigma)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|.\end{aligned}$$

Again, it follows from the arguments used in proof of Theorem 2.3 that on event $\mathscr{E}_1$ of probability at least $1-10e^{-s}$,

$$\sup_{f\in\mathscr{F}(\sigma)}\left|\frac{\widehat{G}_k(0;f)}{\partial_z G_k(0;f)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)}\right| + 2\sup_{f\in\mathscr{F}(\sigma)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right| \leqslant \frac{6}{7}\max\left(\delta',\sigma\right).$$

Consequently, $\sigma\leqslant\delta'+\frac{6}{7}\max(\delta',\sigma)$ on this event, implying that $\sigma\leqslant 7\delta'$. Next, Assumption 2 yields that

$$\begin{aligned}&\sup_{f\in\widehat{\mathscr{F}}(\delta')}\operatorname{Var}\left(\ell(f(X))-\ell(\widehat{f}'_N)\right) \\ &\qquad\leqslant 2\left(\sup_{f\in\widehat{\mathscr{F}}(\delta')}\operatorname{Var}\left(\ell(f(X))-\ell(f_*(X))\right) + \operatorname{Var}\left(\ell(\widehat{f}'_N(X))-\ell(f_*(X))\right)\right) \leqslant 2D(\sqrt{7}+1)\delta'\end{aligned}$$

on $\mathscr{E}_1$. It remains to apply Theorem 2.3, conditionally on $\mathscr{E}_1$, to the class

$$\widehat{\mathscr{F}}(\delta')-\widehat{f}'_N := \left\{f-\widehat{f}'_N,\ f\in\widehat{\mathscr{F}}(\delta')\right\}.$$

To this end, we need to verify the assumption of Theorem 2.1 that translates into the requirement

$$c\Delta_2 \geqslant \frac{1}{\sqrt{k_2}}\,\mathbb{E}\sup_{f\in\mathscr{F}(7\delta')}\frac{1}{\sqrt{|S_2|}}\sum_{j=1}^{|S_2|}\left(\ell(f(X_j))-\ell(f_*(X_j))-P(\ell(f)-\ell(f_*))\right).$$

As $\delta'>\overline{\delta}$ and $|S_2|\geqslant\lfloor N/2\rfloor$, we have the inequality

$$\mathbb{E}\sup_{f\in\mathscr{F}(7\delta')}\frac{1}{\sqrt{|S_2|}}\sum_{j=1}^{|S_2|}\left(\ell(f(X_j))-\ell(f_*(X_j))-P(\ell(f)-\ell(f_*))\right) \leqslant C\delta'\sqrt{N},$$

hence it suffices to check that $\Delta_2 = DM_{\Delta_2}\sqrt{7\delta'} \geqslant C\delta'\sqrt{\frac{N}{k_2}}$. The latter is equivalent to $\delta'\leqslant CD^2M_{\Delta_2}^2\frac{k_2}{N}$ that holds by assumption. Result now follows easily as we assumed that the subsamples S_1 and S_2 used to construct $\widehat{f}'_N$ and $\widehat{f}''_N$ are disjoint.

Funding

Stanislav Minsker gratefully acknowledges support by the National Science Foundation [DMS-1712956 and CCF-1908905].

Data Availability Statement

The data underlying this article are available in UCI Machine Learning Repository [52] at https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized. Artificially generated data underlying this article can be generated using the code that is available on github at https://github.com/TimotheeMathieu/Excess-risk-bounds-in-robust-empirical-risk-minimization/.

REFERENCES

[1] ALISTARH, D., ALLEN-ZHU, Z. & LI, J. (2018) Byzantine stochastic gradient descent. in *Advances in Neural Information Processing Systems*, pp. 4613–4623.

[2] ALON, N., MATIAS, Y. & SZEGEDY, M. (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.

[3] ALQUIER, P., COTTET, V. & LECUÉ, G. (2019) Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *Annals of Statistics*, **47**(4), 2117–2144.

[4] ANTHONY, M. & BARTLETT, P. L. (2009) *Neural network learning: Theoretical foundations*. Cambridge University Press.

[5] AUDIBERT, J.-Y. & CATONI, O. (2011) Robust linear least squares regression. *The Annals of Statistics*, **39**(5), 2766–2794.

[6] BAHADUR, R. R. (1966) A note on quantiles in large samples. *The Annals of Mathematical Statistics*, **37**(3), 577–580.

[7] BARTLETT, P. L., BOUSQUET, O. & MENDELSON, S. (2005) Local Rademacher complexities. *The Annals of Statistics*, **33**(4), 1497–1537.

[8] BARTLETT, P. L., JORDAN, M. I. & MCAULIFFE, J. D. (2004) Large margin classifiers: convex loss, low noise, and convergence rates. in *Advances in Neural Information Processing Systems*, pp. 1173–1180.

[9] ——— (2006) Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, **101**(473), 138–156.

[10] BARTLETT, P. L. & MENDELSON, S. (2006) Empirical minimization. *Probability Theory and Related Fields*, **135**(3), 311–334.

[11] BARTLETT, P. L., MENDELSON, S. & NEEMAN, J. (2012) ℓ_1-regularized linear regression: persistence and oracle inequalities. *Probability theory and related fields*, **154**(1-2), 193–224.

[12] BOUCHERON, S., LUGOSI, G. & MASSART, P. (2013) *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

[13] BROWNLEES, C., JOLY, E. & LUGOSI, G. (2015) Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, **43**(6), 2507–2536.

[14] CATONI, O. (2012) Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, **48**(4), 1148–1185.

[15] CHEN, L. H. & SHAO, Q.-M. (2001) A non-uniform Berry–Esseen bound via Stein's method. *Probability theory and related fields*, **120**(2), 236–254.

[16] CHEN, Y., SU, L. & XU, J. (2017) Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, **1**(2), 44.

[17] CHERAPANAMJERI, Y., HOPKINS, S. B., KATHURIA, T., RAGHAVENDRA, P. & TRIPURANENI, N. (2020) Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 601–609.

[18] CHINOT, G., LECUÉ, G. & LERASLE, M. (2019) Robust statistical learning with Lipschitz and convex loss functions. *Probability Theory and related fields*, pp. 1–44.
[19] DEVROYE, L., LERASLE, M., LUGOSI, G. & OLIVEIRA, R. I. (2016) Sub-Gaussian mean estimators. *The Annals of Statistics*, **44**(6), 2695–2725.
[20] DIAKONIKOLAS, I., KAMATH, G., KANE, D., LI, J., MOITRA, A. & STEWART, A. (2019) Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, **48**(2), 742–864.
[21] DIAKONIKOLAS, I., KAMATH, G., KANE, D., LI, J., STEINHARDT, J. & STEWART, A. (2019) Sever: A robust meta-algorithm for stochastic optimization. *Proceedings of the International Conference on Machine Learning*, pp. 1596–1606.
[22] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. & STEWART, A. (2017) Being robust (in high dimensions) can be practical. *Proceedings of the International Conference on Machine Learning*, pp. 999–1008.
[23] DIAKONIKOLAS, I. & KANE, D. M. (2019) Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.
[24] DUDLEY, R. M. (2014) *Uniform central limit theorems*, vol. 142. Cambridge University Press.
[25] HOEFFDING, W. (1948) A Class of Statistics with Asymptotically Normal Distribution. *Annals of Mathematical Statististics*, **19**(3), 293–325.
[26] ——— (1963) Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, **58**(301), 13–30.
[27] HOLLAND, M. J. & IKEDA, K. (2017) Robust regression using biased objectives. *Machine Learning*, **106**(9-10), 1643–1679.
[28] ——— (2019) Efficient learning with robust gradient descent. *Machine Learning*, **108**(8-9), 1523–1560.
[29] HOPKINS, S. B. (2020) Mean estimation with sub-Gaussian rates in polynomial time. *Annals of Statistics*, **48**(2), 1193–1213.
[30] HUBER, P. J. & RONCHETTI, E. M. (2009) *Robust statistics; 2nd ed.*, Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ.
[31] KOLTCHINSKII, V. (2011) *Oracle inequalities in empirical risk minimization and sparse recovery problems*, vol. 2033 of *Lecture Notes in Mathematics*. Springer, École d'Été de Probabilités de Saint-Flour 2008.
[32] KOLTCHINSKII, V. & MENDELSON, S. (2015) Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, **2015**(23), 12991–13008.
[33] LAI, K. A., RAO, A. B. & VEMPALA, S. (2016) Agnostic estimation of mean and covariance. in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 665–674. IEEE.
[34] LECUÉ, G. & LERASLE, M. (2020) Robust machine learning by median-of-means: theory and practice. *Annals of Statistics*, **48**(2), 906–931.
[35] LECUÉ, G., LERASLE, M. & MATHIEU, T. (2020) Robust classification via MOM minimization. *Machine Learning*, **109**(8), 1635–1665.
[36] LECUÉ, G. & MENDELSON, S. (2010) Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli*, pp. 605–613.
[37] LEDOUX, M. & TALAGRAND, M. (1991) *Probability in Banach spaces: isoperimetry and processes*. Springer-Verlag, Berlin.
[38] LERASLE, M. & OLIVEIRA, R. I. (2011) Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.
[39] LUGOSI, G. & MENDELSON, S. (2019a) Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*, **25**(3), 2075–2106.
[40] ——— (2019b) Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, **22**(3), 925–965.
[41] MATHIEU, T. & MINSKER, S. (2021) Excess risk bounds in robust empirical risk minimization: Supplementary Material. *Information and Inference: A Journal of the IMA*.
[42] MAYZLIN, D., DOVER, Y. & CHEVALIER, J. (2014) Promotional reviews: An empirical investigation of

online review manipulation. *American Economic Review*, **104**(8), 2421–55.
[43] MENDELSON, S. (2008) Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, **54**(8), 3797–3803.
[44] ——— (2014) Learning without concentration. in *Conference on Learning Theory*, pp. 25–39.
[45] ——— (2016) Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications*, **126**(12), 3652–3680.
[46] MINSKER, S. (2019a) Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, **13**(2), 5213–5252.
[47] ——— (2019b) Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*.
[48] NEMIROVSKI, A. & YUDIN, D. (1983) *Problem complexity and method efficiency in optimization*. John Wiley & Sons.
[49] O'DONNELL, R. (2014) *Analysis of Boolean functions*. Cambridge University Press.
[50] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R. & DUBOURG, V. (2011) Scikit-learn: Machine learning in Python. *Journal of machine learning research*, **12**(Oct), 2825–2830.
[51] PRASAD, A., SUGGALA, A. S., BALAKRISHNAN, S. & RAVIKUMAR, P. (2020) Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B*, **82**(3), 601–627.
[52] REDMOND, M. (2011) Communities and Crime Unnormalized Data Set. Creating using 1990 US Census and other sources. Available at https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized.
[53] RON SCHMELZER, F. (2019) The Achilles' heel Of AI. https://www.forbes.com/sites/cognitiveworld/2019/03/07/the-achilles-heel-of-ai.
[54] TALAGRAND, M. (2014) *Upper and lower bounds for stochastic processes: modern methods and classical problems*, vol. 60. Springer Science & Business Media.
[55] TSYBAKOV, A. B. (2004) Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, **32**(1), 135–166.
[56] VAN DE GEER, S. (2016) Estimation and testing under sparsity. *Lecture Notes in Mathematics*, **2159**.
[57] VAN DE GEER, S. A. & VAN DE GEER, S. (2000) *Empirical Processes in M-estimation*, vol. 6. Cambridge University Press.
[58] VAN DER VAART, A. W. & WELLNER, J. A. (2000) *Weak convergence and empirical Processes: with applications to statistics*. Springer.
[59] YIN, D., CHEN, Y., KANNAN, R. & BARTLETT, P. (2018) Byzantine-robust distributed learning: towards optimal statistical rates. in *International Conference on Machine Learning*, pp. 5650–5659.