Excess risk bounds in robust empirical risk minimization

Timothée Mathieu, Stanislav Minsker

Excess risk bounds in robust empirical risk minimization

Timothée Mathieu [1,*] **and Stanislav Minsker**[2,**]

[1] *1. Laboratoire de Mathematiques d'Orsay, Univ. Paris-Sud, CNRS, Universit Paris-Saclay, 91405 Orsay, France and Inria Saclay - Ile-de-France, Bt. Turing, Campus de l'Ecole Polytechnique, 91120 Palaiseau, France.*
e-mail: *timothee.mathieu@u-psud.fr

[2] *Department of Mathematics, University of Southern California, Los Angeles, CA 90089. e-mail:* **minsker@usc.edu

Abstract: 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 $\{f(X),\ f \in \mathcal{F}\}$ indexed by a class of functions $f \in \mathcal{F}$, 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 $f \in \mathcal{F}$. 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 expectation, 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. 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 and phrases: 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 [41] 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 samples, 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 [33]. 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 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 statistical framework used throughout the paper is defined as follows. Let $(S, \mathcal{S})$ 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 $\mathcal{F}$ is a class of measurable functions from S to $\mathbb{R}$ and $\ell : \mathbb{R} \to \mathbb{R}_+$ is a loss function. Many problems in statistical learning theory can be formulated as risk minimization of the form

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

We will frequently write $P\ell(f)$ or simply $\mathcal{L}(f)$ in place of the expected loss $\mathbb{E}\ell\,(f(X))$. Throughout the paper, we will also assume that the minimum is attained for some (unique) $f_* \in \mathcal{F}$. 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 $\mathcal{G}$ (such as the class of linear functions), $\ell(x) = x^2$, and $f_*(z, y) = y - g_*(z)$, where $g_*(z) = \mathbb{E}\,[Y|Z = z]$ is the conditional expectation. As the true distribution P is usually unknown, a proxy of f_* is obtained via *empirical risk minimization* (ERM), namely

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

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

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

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

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

General bounds for the excess risk have been extensively studied; a small subsample of the relevant works includes the papers [45, 46, 24, 4, 10, 43] and references therein. However, until recently sharp estimates were known only in the situation when the functions in the class $\ell(\mathcal{F}) := \{\ell(f),\ f \in \mathcal{F}\}$ are uniformly bounded, or when the envelope $F_\ell(x) := \sup_{f\in\mathcal{F}} |\ell(f(x))|$ of the class $\ell(\mathcal{F})$ possesses finite exponential moments. Our focus is on the situation when marginal distributions of the process $\{\ell(f(X)),\ f \in \mathcal{F}\}$ indexed by $\mathcal{F}$ 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 $\left\{\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \ell(f(X_j)) - \mathbb{E}\ell(f(X)),\ f \in \mathcal{F}\right\}$ 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 $\mathcal{O} < N$ outliers which are generated by an adversary who has an opportunity to inspect the data, and the combined dataset of cardinality $N^\circ = N + \mathcal{O}$ is presented to an algorithm; in this paper, we assume that the proportion of contamination $\frac{\mathcal{O}}{N}$ (or its upper bound) is known.

The approach that we propose is based on replacing the sample mean that is 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 [37, 2, 30] and Catoni's estimator [13]. 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 [36, 35], 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. In Section 2, we discuss practical implementation and numerical performance of our methods for two problems, linear regression and binary classification. The complete statements of the key results are given in Section 3, and in Section 4 we deduce the corollaries of these results for specific examples. Finally, the architecture of the proofs is explained in Section 5, while the remaining technical arguments and additional numerical results are contained in the appendix.

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 $f \in \mathcal{F}$, let $\sigma^2(\ell, f) = \operatorname{Var}\left(\ell(f(X))\right)$ and for any subset $\mathcal{F}' \subseteq \mathcal{F}$, denote $\sigma^2(\ell, \mathcal{F}') = \sup_{f\in\mathcal{F}'} \sigma^2(\ell, f)$. 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 $f \in \mathcal{F}$, let

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

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

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

Clearly, if $\rho(x) = x^2$, $\widehat{\mathcal{L}}^{(k)}(f)$ is equal to the sample mean. If $\rho(x) = |x|$, then $\widehat{\mathcal{L}}^{(k)}(f)$ is the median-of-means estimator [37, 2, 17]. We will be interested in the situation when ρ is similar to Huber's loss, whence ρ' is bounded and Lipchitz continuous (exact conditions imposed on ρ are specified in Assumption 1 below). It is instructive to consider two cases: first, when $k = N$ (so that $n = 1$) and $\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))}\sqrt{N}$, $\widehat{\mathcal{L}}^{(k)}(f)$ is akin to Catoni's estimator [13], and when n is large and $\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))}$, we recover the "median-of-means type" estimator. [1]

We also construct a permutation-invariant version of the estimator $\widehat{\mathcal{L}}^{(k)}(f)$ that does not depend on the specific choice of the subgroups $G_1, \dots, G_k$. Define

$$\mathcal{A}_N^{(n)} := \{J : \ J \subseteq \{1, \dots, N\}, \mathrm{Card}(J) = n\}.$$

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 [19]

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

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

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

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

$$\widehat{\mathcal{L}}_U^{(k)}(f) := \operatorname*{argmin}_{z \in \mathbb{R}} U_{N,n}(z; f). \tag{1.4}$$

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

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

as well as its permutation-invariant analogue

$$\widehat{f}_N^U := \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{\mathcal{L}}_U^{(k)}(f) \tag{1.6}$$

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 $\{\ell(f(X)), \ f \in \mathcal{F}\}$. 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

[1] The "standard" median-of-means estimator corresponds to $\rho(x) = x$ and can be seen as a limit of $\widehat{\mathcal{L}}^{(k)}(f)$ when $\Delta \to 0$; this case is not covered by results of the paper, as we will require that ρ' is smooth and Δ is bounded from below.

$N^{-1/2}$ ("slow rates") are easier to establish: in particular, results of this type follow from bounds on the uniform deviations $\sup_{f\in\mathcal{F}}\left|\widehat{\mathcal{L}}^{(k)}(f)-\mathcal{L}(f)\right|$ that have been investigated in [35]. Proving fast rates is a more technically challenging task: to achieve the goal, we study remainder terms in Bahadur-type representations of the estimators $\widehat{\mathcal{L}}^{(k)}(f)$ and $\widehat{\mathcal{L}}_U^{(k)}(f)$ that provide linear (in $\ell(f)$) approximations of these nonlinear statistics and are easier to study.

Let us remark that exact evaluation of the U-statistics based estimators $\widehat{\mathcal{L}}_U^{(k)}(f)$ 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 (1.6) in scenarios like least-squares and logistic regression. Moreover, numerical performance of the permutation-invariant estimator $\widehat{f}_N^U$ is never worse than $\widehat{f}_N$, and often is significantly better; these points are further discussed in Section 2.

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 that rates of order $N^{-1/2}$ are achieved with exponentially high probability if $\sigma(\ell,\mathcal{F}) = \sup_{f\in\mathcal{F}}\sigma^2(\ell,f) < \infty$ and $\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{N}}\sum_{j=1}^N\left(\ell(f(X_j))-\mathbb{E}\ell(f(X))\right) < \infty$. The latter is true if the class $\{\ell(f),\ f\in\mathcal{F}\}$ is P-Donsker [18], 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. Next, we demonstrate that under additional assumption requiring that any $f\in\mathcal{F}$ with small excess risk must be close to f_* that minimizes the expected loss, $\widehat{f}_N$ and $\widehat{f}_N^U$ attain fast rates; we state the bounds only for $\widehat{f}_N$ while the results for $\widehat{f}_N^U$ are similar, up to the change in absolute constants.

Theorem 1.1 (Informal). *Assume that $\sigma(\ell,\mathcal{F})<\infty$. Then, for appropriately set k and Δ,*

$$\mathcal{E}(\widehat{f}_N) \leqslant \overline{\delta} + C(\mathcal{F},P)\left(\frac{s}{N^{2/3}}+\left(\frac{\mathcal{O}}{N}\right)^{2/3}\right)$$

with probability at least $1-e^{-s}$ for all $s\lesssim k$. Moreover, if $\sup_{f\in\mathcal{F}}\mathbb{E}^{1/4}\left(\ell(f(X))-\mathbb{E}\ell(f(X))\right)^4<\infty$, then

$$\mathcal{E}(\widehat{f}_N) \leqslant \overline{\delta} + C(\mathcal{F},P)\left(\frac{s}{N^{3/4}}+\left(\frac{\mathcal{O}}{N}\right)^{3/4}\right),$$

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

Here, $\overline{\delta}$ is the quantity (formally defined in (3.5) below) that often coincides with the optimal rate for the excess risk [3, 31]. Moreover, we design a two-step estimator based on $\widehat{f}_N$ that is capable of achieving faster rates whenever $\overline{\delta}\ll N^{-3/4}$.

Theorem 1.2 (Informal). *Assume that $\sup_{f\in\mathcal{F}}\mathbb{E}^{1/4}\left(\ell(f(X))-\mathbb{E}\ell(f(X))\right)^4<\infty$. There exists an estimator $\widehat{f}_N''$ such that*

$$\mathcal{E}\left(\widehat{f}_N''\right) \leqslant \overline{\delta} + C(\mathcal{F},P,\rho)\left(\frac{\mathcal{O}}{N}+\frac{s}{N}\right)$$

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

Estimator $\widehat{f}_N''$ is based on a two-step procedure, where $\widehat{f}_N$ serves as an initial approximation that is refined on the second step via the risk minimization restricted to a "small neighborhood" of $\widehat{f}_N$.

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 for the optimization problem (1.1) based on variants of the multivariate median-of-means technique [40, 15, 47, 1], as well as Catoni's estimator [21]. While these algorithms admits strong theoretical guarantees, they require robustly estimating the gradient vector at every step hence are computationally demanding; moreover, results are weaker for losses that are not strongly convex (for instance, the hinge loss).

The line of work that is closest in spirit to the approach of this paper has includes the works that employ robust risk estimators based on Catoni's approach [5, 12, 22] and the median-of-means technique, such as "tournaments" and the "min-max median-of-means" [31, 32, 27, 28, 16]. 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 Catoni's estimator is capable of achieving fast rates, thus improving the results and weakening the assumptions stated in [12];
2. Existing approaches based on the median-of-means estimators are either computationally intractable [31], or outputs of practically efficient algorithms do not admit strong theoretical guarantees [27, 28, 16]. Our algorithms are designed specifically for the estimators $\widehat{f}_N$ and $\widehat{f}_N^U$, and enjoy good performance in numerical experiments along with strong theoretical guarantees simultaneously.
3. We develop new tools and techniques to analyze proposed estimators. In particular, we do not rely on the "small ball" method [25, 34] and the standard "majority vote-based" analysis 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 (1.2). In particular, we demonstrate that the typical deviations of the estimator $\widehat{\mathcal{L}}^{(k)}(f)$ around $\mathcal{L}(f)$ are significantly smaller than the deviations of the subsample averages $\bar{\mathcal{L}}_j(f)$; consequently, this fact allows us to "decouple" the parameter k responsible for the cardinality of subsamples from the confidence parameter s that controls the deviation probabilities, and establish bounds that are uniform over a certain range of s instead of a fixed level $s \asymp k$. Moreover, in cases when adversarial contamination is insignificant (e.g. $\mathcal{O} = O(1)$), our algorithms, unlike existing results, admit a "universal" choice of k that is independent of the parameter $\bar{\delta}$ controlling the optimal rate.
   We are able to 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), our required assumptions are slightly stronger compared to the best results in the literature tailored specifically to the task [e.g. 31, 27].

2. Numerical algorithms and examples.

The main goal of this section is to discuss numerical algorithms used to approximate estimators $\widehat{f}_N$ and $\widehat{f}_N^U$, as well as assess the quality of resulting solutions. We will also compare our methods with the ones known previously, specifically, the median-of-means based approach proposed in [28]. Finally, we perform the numerical study of dependence of the solutions on the parameters Δ and k. All evaluations are performed for logistic regression in the framework of binary classification as well as linear regression with quadratic loss using simulated data, while applications to real data are shown in the appendix. Let us mention that the numerical methods for closely related approach in the special case of linear regression have been investigated in a recent work [22]. Here, we focus on general algorithms that can easily be adapted to other predictions tasks and loss functions. Let us first briefly recall the formulations of both the binary classification and the linear regression problems.

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 $\mathcal{F}$ 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

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

The loss ℓ is classification-calibrated if $\operatorname{sign}(\rho_*(z)) = b_*(z)$ P-almost surely; we refer the reader to [7] 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 a classification-calibrated loss and $\mathcal{F} = \left\{f_\beta(\cdot) = \langle \cdot, \beta\rangle,\ \beta \in \mathbb{R}^d\right\}$ (as usual, the intercept term can be included if the vector Z is replaced by $\tilde{Z} = (Z, 1)$).

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 $\mathcal{F} = \left\{ f_\beta(\cdot) = \langle \cdot, \beta \rangle, \quad \beta \in \mathbb{R}^d \right\}$.

In both examples, we will assume that we are given an i.i.d. sample $(Z_1, Y_1), \ldots, (Z_N, Y_N)$ having the same distribution as (Z, Y).

2.1. Gradient descent algorithms.

Optimization problems (1.5) and (1.6) are not convex, so we will focus our attention of the variants of the gradient descent method employed to find local minima. We will first derive the expression for $\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta)$, the gradient of $\widehat{\mathcal{L}}^{(k)}(\beta) := \widehat{\mathcal{L}}^{(k)}(f_\beta)$, for the problems corresponding to logistic regression and regression with quadratic loss. It follows from (1.2) that $\widehat{\mathcal{L}}^{(k)}(\beta)$ satisfies the equation

$$\sum_{j=1}^{k} \rho' \left(\sqrt{n} \frac{\bar{\mathcal{L}}_j(\beta) - \widehat{\mathcal{L}}^{(k)}(\beta)}{\Delta} \right) = 0. \tag{2.1}$$

Taking the derivative in (2.1) with respect to β, we retrieve $\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta)$:

$$\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta) = \frac{\sum_{j=1}^{k} \left(\frac{1}{n} \sum_{i \in G_j} Z_i \, \ell'(Y_i, f_\beta(Z_i)) \right) \rho'' \left(\sqrt{n} \frac{\bar{\mathcal{L}}_j(\beta) - \widehat{\mathcal{L}}^{(k)}(\beta)}{\Delta} \right)}{\sum_{j=1}^{k} \rho'' \left(\sqrt{n} \frac{\bar{\mathcal{L}}_j(\beta) - \widehat{\mathcal{L}}^{(k)}(\beta)}{\Delta} \right)}, \tag{2.2}$$

where $\ell'(Y_i, f_\beta(Z_i))$ stands for the partial derivative $\frac{\partial \ell(y,t)}{\partial t}$ with respect to the second argument t, so that $\ell'(Y_i, f_\beta(Z_i)) = -Y_i \frac{e^{-Y_i \langle \beta, Z_i \rangle}}{1 + e^{-Y_i \langle \beta, Z_i \rangle}}$ in the case of logistic regression and $\ell'(Y_i, f_\beta(Z_i)) = 2\left(\langle \beta, Z_i \rangle - Y_i \right)$ for regression with quadratic loss. In most of our numerical experiments, we choose ρ to be Huber's loss,

$$\rho(y) = \frac{y^2}{2} I\left\{ |y| \leqslant 1 \right\} + \left(|y| - \frac{1}{2} \right) I\left\{ |y| > 1 \right\}.$$

In this case, $\rho''(y) = I\{|y| \leqslant 1\}$ for all $y \in \mathbb{R}$, hence the expression for the gradient can be simplified to

$$\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta) = \frac{\sum_{j=1}^{k} \left(\frac{1}{n} \sum_{i \in G_j} Z_i \, \ell'(Y_i, f_\beta(Z_i)) \right) I\left\{ \left| \bar{\mathcal{L}}_j(\beta) - \widehat{\mathcal{L}}^{(k)}(\beta) \right| \leqslant \frac{\Delta}{\sqrt{n}} \right\}}{\# \left\{ j : \left| \bar{\mathcal{L}}_j(\beta) - \widehat{\mathcal{L}}^{(k)}(\beta) \right| \leqslant \frac{\Delta}{\sqrt{n}} \right\}}, \tag{2.3}$$

where we implicitly assume that Δ is chosen large enough so that the denominator is not equal to 0. To evaluate $\widehat{\mathcal{L}}^{(k)}(\beta)$, we use the "modified weights" algorithm due to Huber and Ronchetti [23, section 6.7]. Complete version of the gradient descent algorithm used to approximate $\widehat{\beta}_N$ (identified with the solution $\widehat{f}_N$ of the problem (1.5)) is presented in Figure 1.

Next, we discuss a variant of a stochastic gradient descent for approximating the "permutation-invariant" estimator $\widehat{f}_N^U$ used when the subgroup size $n > 1$; in our numerical experiments (see Section B.2 for the numerical comparison of two approaches), this method demonstrated consistently superior performance. Below, we will identify $\widehat{f}_N^U$ with the vector of corresponding coefficients $\widehat{\beta}_N^U$. Recall that $\mathcal{A}_N^{(n)} := \{J : \ J \subseteq \{1, \ldots, N\}, \ \mathrm{Card}(J) = n\}$, and that

$$\widehat{\mathcal{L}}_U^{(k)}(\beta) = \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{J \in \mathcal{A}_N^{(n)}} \rho \left(\sqrt{n} \, \frac{\bar{\mathcal{L}}(f_\beta; J) - z}{\Delta} \right). \tag{2.4}$$

Fig 1: Algorithm 1 – evaluation of $\widehat{\beta}_N$.

Input: the dataset $(Z_i, Y_i)_{1\leqslant i\leqslant N}$, number of blocks $k \in \mathbb{N}$, step size parameter $\eta > 0$, maximum number of iterations M, initial guess $\beta_0 \in \mathbb{R}^d$, tuning parameter $\Delta \in \mathbb{R}$.
Construct blocks $G_1, \ldots, G_k$;
for all $t = 0, \ldots, M$ **do**
Compute $\widehat{\mathcal{L}}^{(k)}(\beta_t)$ using the Modified Weights algorithm;
Compute $\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta_t)$ from equation 2.3;
Update

$$\beta_{t+1} = \beta_t - \eta \nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta_t).$$

end for
Output: β_{M+1}.

Similarly to the way that we derived the expression for $\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta)$ from (1.2), it follows from (2.4), with ρ again being the Huber's loss, that

$$\sum_{J\in\mathcal{A}_N^{(n)}} \rho'\left(\sqrt{n}\,\frac{\bar{\mathcal{L}}(f_\beta; J) - \widehat{\mathcal{L}}_U^{(k)}(\beta)}{\Delta}\right) = 0 \quad \text{and}$$

$$\nabla_\beta \widehat{\mathcal{L}}_U^{(k)}(\beta) = \frac{\sum_{J\in\mathcal{A}_N^{(n)}} \left(\frac{1}{n}\sum_{i\in J} Z_i\, \ell'(Y_i, f_\beta(Z_i))\right) I\left\{\left|\bar{\mathcal{L}}(\beta; J) - \widehat{\mathcal{L}}^{(k)}(\beta)\right| \leqslant \frac{\Delta}{\sqrt{n}}\right\}}{\#\left\{J \in \mathcal{A}_N^{(n)} : \left|\bar{\mathcal{L}}(\beta; J) - \widehat{\mathcal{L}}^{(k)}(\beta)\right| \leqslant \frac{\Delta}{\sqrt{n}}\right\}}. \tag{2.5}$$

Expressions in (2.5) are closely related to U-statistics, and it will be convenient to write them in a slightly different form. To this end, let π_N be the collection of all permutations $i : \{1, \ldots, N\} \mapsto \{1, \ldots, N\}$. Given $\tau = (i_1, \ldots, i_N) \in \pi_N$ and an arbitrary U-statistic $U_{N,n}$ defined in (1.3), let

$$T_{i_1,\ldots,i_N} := \frac{1}{k}\left(h\left(X_{i_1}, \ldots, X_{i_n}\right) + h\left(X_{i_{n+1}}, \ldots, X_{i_{2n}}\right) + \ldots + h\left(X_{i_{(k-1)n+1}}, \ldots, X_{i_{kn}}\right)\right).$$

Equivalently, for $\tau = (i_1, \ldots, i_N) \in \pi_N$, let

$$G_j(\tau) = \left(i_{(j-1)n+1}, \ldots, i_{jn}\right),\ j = 1, \ldots, k = \lfloor N/n \rfloor, \tag{2.6}$$

which gives a compact form

$$T_\tau = \frac{1}{k}\sum_{j=1}^{k} h\left(X_i, i \in G_j(\tau)\right).$$

It is well known (section 5 in [20]) that the following representation of the U-statistic holds:

$$U_{N,n} = \frac{1}{N!}\sum_{\tau\in\pi_N} T_\tau. \tag{2.7}$$

Applying representation (2.7) to (2.4), we deduce that

$$\widehat{\mathcal{L}}_U^{(k)}(\beta) = \operatorname*{argmin}_{z\in\mathbb{R}} \sum_{\tau\in\pi_N} \mathcal{R}_\tau(\beta, z), \tag{2.8}$$

with $\mathcal{R}_\tau(\beta, z) = \sum_{j=1}^k \rho\left(\sqrt{n}\,\frac{\bar{\mathcal{L}}(f_\beta; G_j(\tau)) - z}{\Delta}\right)$. Similarly, applying representation (2.7) to the numerator and the denominator in (2.5), we see that $\nabla_\beta \widehat{\mathcal{L}}_U^{(k)}(\beta)$ can be written as a weighted sum

$$\nabla_\beta \widehat{\mathcal{L}}_U^{(k)}(\beta) = \sum_{\tau\in\pi_N} \underbrace{\frac{\sum_{j=1}^k I\left\{\left|\bar{\mathcal{L}}(\beta; G_j(\tau)) - \widehat{\mathcal{L}}^{(k)}(\beta)\right| \leqslant \frac{\Delta}{\sqrt{n}}\right\}}{\sum_{\pi\in\pi_N}\sum_{j=1}^k I\left\{\left|\bar{\mathcal{L}}(\beta; G_j(\pi)) - \widehat{\mathcal{L}}^{(k)}(\beta)\right| \leqslant \frac{\Delta}{\sqrt{n}}\right\}}}_{=\omega_\tau,\ \text{weight corresponding to permutation } \tau} \cdot \widetilde{\Gamma}_\tau(\beta),$$

where

$$\widetilde{\Gamma}_\tau(\beta) := \frac{\sum_{j=1}^k \left(\frac{1}{n}\sum_{i\in G_j(\tau)} Z_i\, \ell'(Y_i, f_\beta(Z_i))\right) I\left\{\left|\bar{\mathcal{L}}(\beta; G_j(\tau)) - \widehat{\mathcal{L}}^{(k)}(\beta)\right| \leqslant \frac{\Delta}{\sqrt{n}}\right\}}{\sum_{j=1}^k I\left\{\left|\bar{\mathcal{L}}(\beta; G_j(\tau)) - \widehat{\mathcal{L}}^{(k)}(\beta)\right| \leqslant \frac{\Delta}{\sqrt{n}}\right\}} \tag{2.9}$$

is similar to the expression for the gradient of $\widehat{\mathcal{L}}^{(k)}(\beta)$ defined for a fixed partition $G_1(\tau),\dots,G_k(\tau)$, see equation (2.3). Representations in (2.8) and (2.9) can be simplified even further noting that permutations that do not alter the subgroups $G_1,\dots,G_k$ also do not change the values of $\mathcal{R}_\tau(\beta,z)$, ω_τ and $\widetilde{\Gamma}_\tau(\beta)$. To this end, let us say that $\tau_1,\tau_2\in\pi_N$ are equivalent if $G_j(\tau_1)=G_j(\tau_2)$ for all $j=1,\dots,k$. It is easy to see that there are $\frac{N!}{(n!)^k\cdot(N-nk)!}$ equivalence classes, and let $\pi_{N,n,k}$ be the set of permutations containing exactly one permutation from each equivalence class. We can thus write

$$\begin{aligned}\widehat{\mathcal{L}}_U^{(k)}(\beta)&=\operatorname*{argmin}_{z\in\mathbb{R}}Q(\beta,z):=\operatorname*{argmin}_{z\in\mathbb{R}}\sum_{\tau\in\pi_{N,n,k}}\mathcal{R}_\tau(\beta,z),\\ \nabla_\beta\widehat{\mathcal{L}}_U^{(k)}(\beta)&=\sum_{\tau\in\pi_{N,n,k}}\widetilde{\omega}_\tau\cdot\widetilde{\Gamma}_\tau(\beta),\end{aligned}\tag{2.10}$$

where $\widetilde{\omega}_\tau=(n!)^k\,(N-nk)!\cdot\omega_\tau$. Representation (2.10) suggests that in order to obtain an unbiased estimator of $\nabla_z Q(\beta,z)$, one can sample a permutation $\tau\in\pi_{N,n,k}$ uniformly at random, compute $\nabla_z\mathcal{R}_\tau(\beta,z)$ and use it as a descent direction. This yields a version of the stochastic gradient descent for evaluating $\widehat{\mathcal{L}}_U^{(k)}(\beta)$ presented in Figure 2. Once a method for computing $\widehat{\mathcal{L}}_U^{(k)}(\beta)$ is established, similar reasoning

Fig 2: Algorithm 2 – evaluation of $\widehat{\mathcal{L}}_U^{(k)}(\beta)$.

Input: the dataset $(Z_i,Y_i)_{1\leqslant i\leqslant N}$, number of blocks $k\in\mathbb{N}$, step size parameter $\eta>0$, maximum number of iterations M, initial guess $z_0\in\mathbb{R}$, tuning parameter $\Delta\in\mathbb{R}$.
for all $t=0,\dots,M$ **do**
Sample permutation τ uniformly at random from $\pi_{N,n,k}$, construct blocks $G_1(\tau),\dots,G_k(\tau)$ according to (2.6);
Compute $\nabla_z\mathcal{R}_\tau(\beta,z_t)=-\frac{\sqrt{n}}{\Delta}\sum_{j=1}^k\rho'\left(\sqrt{n}\frac{\mathcal{L}(f_\beta;G_j(\tau))-z_t}{\Delta}\right)$;
Update
$$z_{t+1}=z_t-\eta\nabla_z\mathcal{R}_\tau(\beta,z_t).$$
end for
Output: z_{M+1}.

leads to an algorithm for finding $\widehat{f}_N^U$. Indeed, using representation (2.10), it is easy to see that an unbiased estimator of $\nabla_\beta\widehat{\mathcal{L}}_U^{(k)}(\beta)$ can be obtained by first sampling a permutation $\tau\in\pi_{N,n,k}$ according to the probability distribution given by the weights $\{\widetilde{\omega}_\tau,\ \tau\in\pi_{N,n,k}\}$, then evaluating $\widetilde{\Gamma}_\tau(\beta)$ using formula (2.9), and using $\widetilde{\Gamma}_\tau(\beta)$ as a direction of descent. In most typical cases, the number M of the gradient descent iterations is much smaller than $\frac{N!}{(n!)^k\cdot(N-nk)!}$, whence it is unlikely that the same permutation will be repeated twice in the sampling process. This reasoning suggests the idea of replacing the weights $\widetilde{\omega}_\tau$ by the uniform distribution over $\pi_{N,n,k}$ that leads to a much faster practical implementation which is detailed in Figure 3. It is easy to see that presented gradient descent algorithms for evaluating $\widehat{f}_N$ and $\widehat{f}_N^U$ have

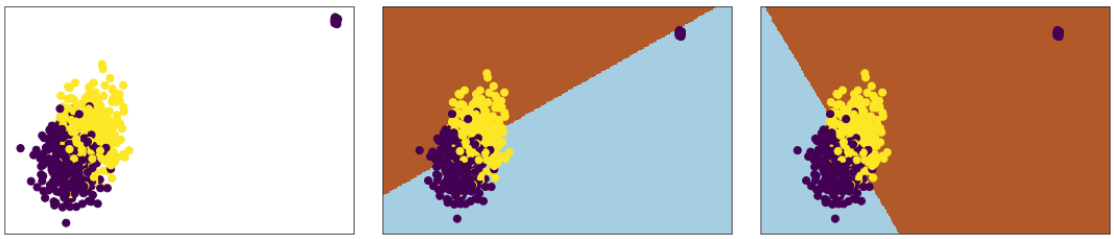
Fig 3: Algorithm 3 – evaluation of $\widehat{\beta}_N^U$.

Input: the dataset $(Z_i,Y_i)_{1\leqslant i\leqslant N}$, number of blocks $k\in\mathbb{N}$, step size parameter $\eta>0$, maximum number of iterations M, initial guess $\beta_0\in\mathbb{R}^d$, tuning parameter $\Delta\in\mathbb{R}$.
for all $t=0,\dots,M$ **do**
Sample permutation τ uniformly at random from $\pi_{N,n,k}$, construct blocks $G_1(\tau),\dots,G_k(\tau)$ according to (2.6);
Compute $\widehat{\mathcal{L}}_U^{(k)}(\beta_t)$ using Algorithm 2 in Figure 2;
Compute $\widetilde{\Gamma}_\tau(\beta_t)$ via equation 2.9;
Update
$$\beta_{t+1}=\beta_t-\eta\widetilde{\Gamma}_\tau(\beta_t).$$
end for
Output: β_{M+1}.

the same numerical complexity. The following subsections provide several "proof-of-concept" examples illustrating the performance of proposed methods, as well as comparison to the existing techniques.

2.2. Logistic regression.

The dataset consists of pairs $(Z_j, Y_j) \in \mathbb{R}^2 \times \{\pm 1\}$, where the marginal distribution of the labels is uniform and conditional distributions of Z are normal, namely, Law $(Z \,|\, Y = 1) = \mathcal{N}\left((-1,-1)^T, 1.4 I_2\right)$, Law $(Z \,|\, Y = -1) \sim \mathcal{N}\left((1,1), 1.4 I_2\right)$, and $\Pr(Y = 1) = \Pr(Y = -1) = 1/2$. The dataset includes outliers for which $Y \equiv 1$ and $Z \sim \mathcal{N}\left((24, 8), 0.1 I_2\right)$, where I_2 stands for the 2×2 identity matrix. We generated 600 "informative" observations along with 30 outliers, and compared performance or our robust method (based on evaluating $\widehat{\beta}_N^U$) with the standard logistic regression that is known to be sensitive to outliers in the sample (we used implementation available in the Scikit-learn package [39]). Results of the experiment are presented in Figure 4. Parameters k and Δ in our implementation were tuned via cross-validation.



(a) Training Dataset (b) Decision function – standard Logistic Regression (c) Decision function – Algorithm 3

Fig 4: Scatter plot of 630 samples from the training dataset (600 informative observations, 30 outliers), the color of the points correspond to their labels and the background color – to the predicted labels (brown region corresponds to "yellow" labels and blue – to "purple").

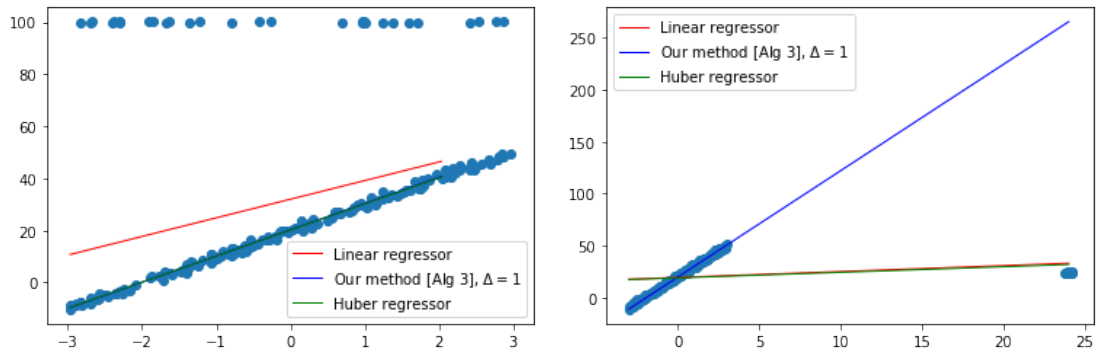
2.3. Linear regression.

In this section, we compare performance of our method (again based on evaluating $\widehat{\beta}_N^U$) with standard linear regression as well as with robust Huber's regression estimator [23, section 7]; linear regression and Huber's regression were implemented using 'LinearRregression' and 'HuberRegressor' functions in the Scikit-learn package [39]. As in the previous example, the dataset consists of informative observations and outliers. Informative data (Z_j, Y_j), $j = 1, \ldots, 570$ are i.i.d. and satisfy the linear model $Y_j = 10 Z_j + \varepsilon_j$ where $Z_j \sim \text{Unif}[-3, 3]$ and $\varepsilon_j \sim \mathcal{N}(0, 1)$. We consider two types of outliers: (a) outliers in the response variable Y only, and (b) outliers in the predictor Z. It is well-known that standard linear regression is not robust in any of these scenarios, Huber's regression estimator is robust to outliers in response Y only, while our approach is shown to be robust to corruption of both types. In both test scenarios, we generated 30 outliers. Given Z_j, the outliers Y_j of type (a) are sampled from a $\mathcal{N}(100, 0.01)$ distribution, while the outliers of type (b) are $Z_j \sim \mathcal{N}\left((24, 24)^T, 0.01\, I_2\right)$. Results are presented in Figure 5, and confirm the expected outcomes.

2.4. Choice of k and Δ.

In this subsection, we evaluate the effect of different choices of k and Δ in the linear regression setting of Section 2.3, again with 570 informative observations and 30 outliers of type (b) as described in section 2.3 above. Figure 6a shows the plot of the resulting mean square error (MSE) against the number of subgroups k. As expected, the error decreases significantly when k exceeds 60, twice the number of outliers. At the same time, the MSE remains stable as k grows up to $k \simeq 100$, which is a desirable property for practical applications. In this experiment, Δ was set using the "median absolute deviation" (MAD) estimator defined as follows. We start with Δ_0 being a small number (e.g., $\Delta_0 = 0.1$). Given a current approximate solution β_t, a permutation τ and the corresponding subgroups $G_1(\tau), \ldots, G_k(\tau)$, set $\widehat{M}(\beta_t) := \text{median}\left(\widehat{\mathcal{L}}^{(k)}(\beta_t; G_1(\tau), \ldots, \widehat{\mathcal{L}}^{(k)}(\beta_t; G_k(\tau)\right)$, and

$$\text{MAD}(\beta_t) = \text{median}\left(\left|\widehat{\mathcal{L}}^{(k)}(\beta_t; G_1(\tau) - \widehat{M}(\beta_t)\right|, \ldots, \left|\widehat{\mathcal{L}}^{(k)}(\beta_t; G_k(\tau) - \widehat{M}(\beta_t)\right|\right).$$

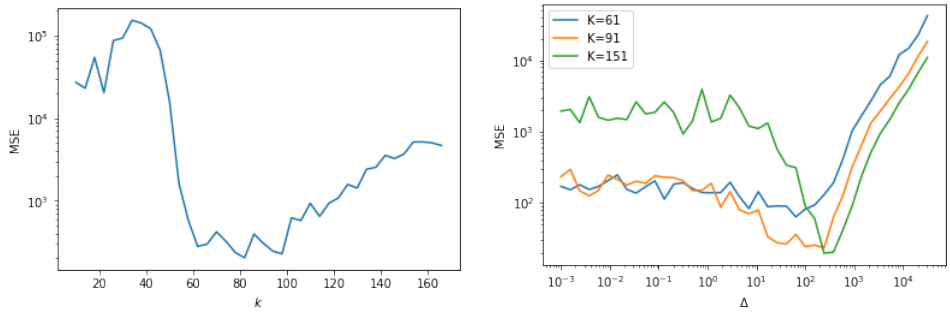


(a) Outliers in response variable (b) Outliers in predictors

Fig 5: Scatter plot of 600 training samples (570 informative data and 30 outliers) and the corresponding regression lines for our method, Huber's regression and regression with quadratic loss.

Finally, define $\widehat{\Delta}_{t+1} := \frac{\mathrm{MAD}(\beta_t)}{\Phi^{-1}(3/4)}$, where Φ is the distribution function of the standard normal law. After a small number m (e.g. $m = 10$) of "burn-in" iterations of Algorithm 3, Δ is fixed at the level $\widehat{\Delta}_m$ for all the remaining iterations.

Next, we study the effect of varying Δ for different but fixed values of k. To this end, we set $k \in \{61, 91, 151\}$, and evaluated the MSE as a function of Δ. Resulting plot is presented in Figure 6b. The MSE achieves its minimum for $\Delta \asymp 10^2$; for larger values of Δ, the effect of outliers becomes significant as the algorithm starts to resemble regression with quadratic loss (indeed, outliers in this specific example are at a distance ≈ 100 from the bulk of the data).



(a) MSE vs k (b) MSE vs Δ (log-log scale)

Fig 6: Plot of the tuning parameter (x-axis) against the MSE (y-axis) obtained with Algorithm 3. The MSE was evaluated via the Monte-Carlo approximation over 500 samples of the data.

2.4.1. Comparison with existing methods.

In this section, we compare performance of Algorithm 3 with a median-of-means-based robust gradient descent algorithm studied in [28]. The main difference of this method is in the way the descent direction is computed at every step. Specifically, $\widetilde{\Gamma}_\tau(\beta)$ employed in Algorithm 3 is replaced by $\nabla_\beta \mathcal{L}^\diamond(\beta)$ where $\mathcal{L}^\diamond(\beta) := \text{median}\big(\bar{\mathcal{L}}(\beta; G_1(\tau), \ldots, \bar{\mathcal{L}}(\beta; G_k(\tau)\big)$, see Figure 7 and [28] for the detailed description. Experiments were performed for the logistic regression problem based on the "two moons" pattern, one of the standard datasets in the Scikit-learn package [39] presented in Figure 8a. We performed two sets of experiments, one on the outlier-free dataset and one on the dataset consisting of 90% of informative observations and 10% of outliers, depicted as a yellow dot with coordinates $(0, 5)$ on the plot. In both

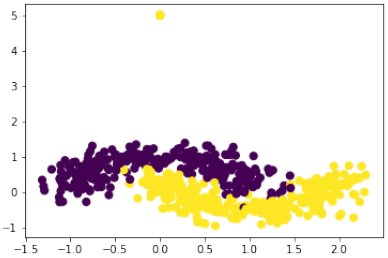
Fig 7: Algorithm 4.

Input: the dataset $(Z_i, Y_i)_{1 \leqslant i \leqslant N}$, number of blocks $k \in \mathbb{N}$, step size parameter $\eta > 0$, maximum number of iterations M, initial guess $\beta_0 \in \mathbb{R}^d$.
for all $t = 0, \dots, M$ **do**
Sample permutation τ uniformly at random from $\pi_{N,n,k}$, construct blocks $G_1(\tau), \dots, G_k(\tau)$ according to (2.6);
Compute $\nabla_\beta \mathcal{L}^\diamond(\beta)$;
Update

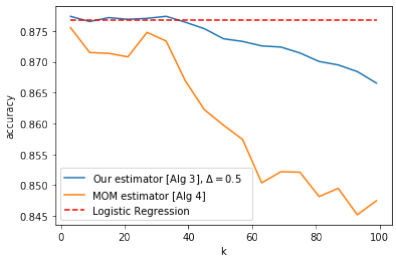
$$\beta_{t+1} = \beta_t - \eta \nabla_\beta \mathcal{L}^\diamond(\beta).$$

end for
Output: β_{M+1}.

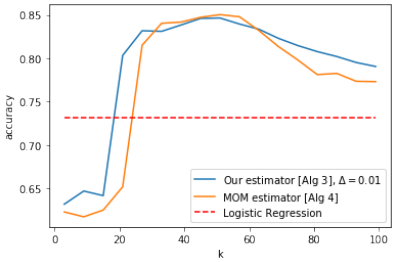
scenarios, we tested the "small" ($N = 100$) and "moderate" ($N = 1000$) sample size regimes. We used standard logistic regression trained on an outlier-free sample as a benchmark; its accuracy is shown as a dotted red line on the plots. In all the cases, parameter Δ was tuned via cross-validation. In the outlier-free setting, our method (based on Algorithm 3) performed nearly as good as logistic regression; notably, performance of the method was strong even for large values of k, while classification accuracy decreased noticeably for Algorithm 4 for large k. In the presence of outliers, our method performed similar to Algorithm 4, while both methods outperformed standard logistic regression; for large values of k, our method was again slightly better. At the same time, Algorithm 4 was consistently faster than Algorithm 3 across the experiments.



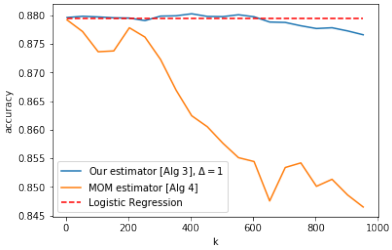
(a) "Two moons" dataset [39] with outliers.



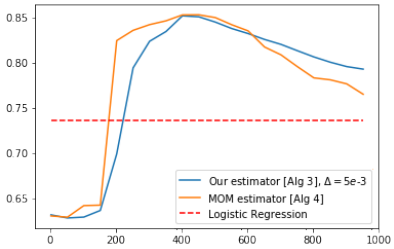
(b) $N = 100$, no outliers



(c) $N = 100$, with 10 outliers



(d) $N = 1000$, no outliers



(e) $N = 1000$, with 100 outliers

Fig 8: Comparison of Algorithm 3, Algorithm 4 and standard logistic regression. The accuracy was evaluated using Monte-Carlo simulation over 300 runs.

3. Theoretical guarantees for the excess risk.

3.1. Preliminaries.

In this section, we introduce the main quantities that appear in our results, and state the key assumptions. $\sigma^2(\ell, \mathcal{F}') = \sup_{f \in \mathcal{F}'} \sigma^2(\ell, f)$. 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, continuously differentiable 5 times 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" $\phi(x) = C \exp\left(-\frac{4}{1-4x^2}\right) \{|x| \leqslant \frac{1}{2}\}$ where C is chosen so that $\int_{\mathbb{R}} \phi(x)dx = 1$. Then ρ given by the convolution $\rho(x) = (h * \phi)(x)$ satisfies assumption 1.

Remark 3.1. *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 $L(\rho') = 1$.*

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 $N\left(0, \sigma^2(f)\right)$. For $f \in \mathcal{F}$ such that $\sigma(f) > 0$, $n \in \mathbb{N}$ and $t > 0$, define

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

where $Pf := \mathbb{E}f(X)$. In other words, $g_f(t, n)$ 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 14] that

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

given that the absolute constant C is large enough. Moreover, let

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

This quantity (more specifically, its scaled version $\frac{G_f(n,\Delta)}{\sqrt{n}}$ plays the key role in controlling the bias of the estimator $\widehat{\mathcal{L}}^{(k)}(f)$. The following statement provides simple upper bounds for $g_f(t, n)$ and $G_f(n, \Delta)$.

Lemma 3.1. *Let $X_1, \ldots, X_n$ be i.i.d. copies of X, and assume that $Var(f(X)) < \infty$. Then $g_f(t, n) \to 0$ as $|t| \to \infty$ and $g_f(t, n) \to 0$ as $n \to \infty$, with 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 $t > 0$*

$$g_f(t, n) \leqslant C' \frac{\mathbb{E}\left|f(X) - \mathbb{E}f(X)\right|^{2+\delta}}{n^{\delta/2}\left(\sigma(f) + |t|\right)^{2+\delta}} \leqslant C' \frac{\mathbb{E}\left|f(X) - \mathbb{E}f(X)\right|^{2+\delta}}{n^{\delta/2}|t|^{2+\delta}}, \tag{3.1}$$

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

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

3.2. *Slow rates for the excess risk.*

Let

$$\widehat{\delta}_N := \mathcal{E}(\widehat{f}_N) = \mathcal{L}\big(\widehat{f}_N\big) - \mathcal{L}(f_*),$$
$$\widehat{\delta}^U_N := \mathcal{E}(\widehat{f}^U_N) = \mathcal{L}\big(\widehat{f}^U_N\big) - \mathcal{L}(f_*)$$

be the excess risk of $\widehat{f}_N$ and its permutation-invariant analogue $\widehat{f}^U_N$ which are the main objects of our interest. The following bound for the excess risk is well known:

$$\begin{aligned}\mathcal{E}\big(\widehat{f}_N\big) &= \mathcal{L}\big(\widehat{f}_N\big) - \mathcal{L}(f_*)\\ &= \mathcal{L}\big(\widehat{f}_N\big) + \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) - \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) + \widehat{\mathcal{L}}^{(k)}(f_*) - \widehat{\mathcal{L}}^{(k)}(f_*) - \mathcal{L}(f_*)\\ &= \Big(\mathcal{L}\big(\widehat{f}_N\big) - \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N)\Big) - \Big(\mathcal{L}(f_*) - \widehat{\mathcal{L}}^{(k)}(f_*)\Big) + \underbrace{\widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) - \widehat{\mathcal{L}}^{(k)}(f_*)}_{\leqslant 0}\\ &\leqslant 2\sup_{f\in\mathcal{F}}\Big|\widehat{\mathcal{L}}^{(k)}(f) - \mathcal{L}(f)\Big|. \qquad (3.2)\end{aligned}$$

The first result, Theorem 3.1 below, together with the inequality (3.2) 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 [35]. Define

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

Theorem 3.1. *There exist absolute constants $c, C > 0$ such that for all $s > 0$, n and k satisfying*

$$\frac{1}{\Delta}\left(\frac{1}{\sqrt{k}}\,\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(\ell(f(X_j)) - P\,\ell(f)\right) + \sigma(\ell,\mathcal{F})\sqrt{\frac{s}{k}}\right) + \sup_{f\in\mathcal{F}} G_f(n,\Delta) + \frac{s}{k} + \frac{\mathcal{O}}{k} \leqslant c, \qquad (3.3)$$

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

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

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

An immediate corollary is the bound for the excess risk

$$\begin{aligned}\mathcal{E}(\widehat{f}_N) \leqslant C\Bigg[&\frac{\widetilde{\Delta}}{\Delta}\left(\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{j=1}^{N}\left(\ell(f(X_j)) - P\,\ell(f)\right) + \sigma(\ell,\mathcal{F})\sqrt{\frac{s}{N}}\right)\\ &+ \widetilde{\Delta}\sqrt{n}\left(\frac{s}{N} + \frac{\sup_{f\in\mathcal{F}} G_f(n,\Delta)}{n} + \frac{\mathcal{O}}{N}\right)\Bigg] \qquad (3.4)\end{aligned}$$

that holds under the assumptions of Theorem 3.1 with probability at least $1 - 2e^{-s}$. When the class $\{\ell(f),\ f\in\mathcal{F}\}$ is P-Donsker [18], $\limsup_{N\to\infty}\Big|\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(\ell(f(X_j)) - P\ell(f)\right)\Big|$ is bounded, hence condition (3.3) 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 $\Delta\sqrt{k} \geqslant c''\sigma(\mathcal{F})$. The bound of Theorem 3.1 also suggests that the natural "unit" to measure the magnitude of parameter Δ is $\sigma(\ell,\mathcal{F})$. We will often use the ratio $M_\Delta := \frac{\Delta}{\sigma(\ell,\mathcal{F})}$ that can be interpreted as a level of truncation expressed in the units of $\sigma(\ell,\mathcal{F})$, and is one of the two main quantities controlling the bias of the estimator $\widehat{\mathcal{L}}^{(k)}(f)$, the second one being the subgroup size n.

To put these results in perspective, let us consider two examples. First, assume that $n = 1$, $k = N$ and set $\Delta = \Delta(s) := \sigma(\mathcal{F})\sqrt{\frac{N}{s}}$ for $s \leqslant c'N$. Using Lemma 3.1 with $\delta = 0$ to estimate $G_f(n, \Delta)$, we deduce that

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

with probability at least $1-2e^{-s}$. This inequality improves upon excess risk bounds obtained for Catoni-type estimators in [12], as it does not require functions in $\mathcal{F}$ to be uniformly bounded.

The second case we consider is when $N \gg n \geqslant 2$. For the choice of $\Delta \asymp \sigma(\ell,\mathcal{F})$, the estimator $\widehat{\mathcal{L}}^{(k)}(f)$ most closely resembles the median-of-means estimator. In this case, Theorem 3.1 yields the excess risk bound of the form

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

that holds with probability $\geqslant 1 - 2e^{-s}$ for all $s \leqslant c'k$. As $\sup_{f\in\mathcal{F}}G_f(n,\Delta)$ is small for large n and $\frac{\mathcal{O}}{k}\sqrt{\frac{k}{N}} \leqslant \sqrt{\frac{\mathcal{O}}{N}}$ whenever $\mathcal{O} \leqslant k$, this bound is improves upon Theorem 2 in [28] that provides bounds for the excess risk for robust classifiers based on the the median-of-means estimators.

3.3. Towards fast rates for the excess risk.

It is well known that in regression and binary classification problems, 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 $\phi(0) = 0$,

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

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 appear in the excess risk bounds. For $\delta > 0$, let

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

Moreover, define

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

The following condition, known as *Bernstein's condition* following [8], plays the crucial role in the analysis of excess risk bounds.

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

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

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

Assumption 2 is known to hold in many concrete cases of prediction and classification tasks, and we provide examples and references in Section 4 below. Informally speaking, it postulates that any f with small excess risk must be "close" to f_*. More general versions of the Bernstein's condition are often considered in the literature: for instance, it can be replaced by assumption [8] requiring that

$\mathrm{Var}\left(\ell(f(X)) - \ell(f_*(X))\right) \leqslant D^2 \left(\mathcal{E}(f)\right)^\tau$ for some $\tau \in (0,1]$ (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 [24, Chapter 4], we will say the 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

$$\bar{\delta} := \min\left\{\delta > 0 : \; C_1(\rho)\frac{1}{\sqrt{N}}\frac{\widetilde{\Delta}}{\Delta}\frac{\widetilde{\omega}(\delta)}{\delta} \leqslant \frac{1}{7}\right\}, \tag{3.5}$$

where $C_1(\rho)$ is a sufficiently large positive constant that depends only on ρ. This quantity plays an important role in controlling the excess risk, as shown by the following theorems.

Theorem 3.2. *Assume that conditions of Theorem 3.1 hold. Additionally, suppose that* $M_\Delta := \frac{\Delta}{\sigma(\ell,\mathcal{F})} \geqslant 1$. *Then*

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

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. Moreover, same bound holds for* $\widehat{\delta}_N^U$, *up to a change in absolute constants.*

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

Theorem 3.3. *Assume that conditions of Theorem 3.1 hold. Additionally, suppose that*

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

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

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

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. Moreover, same bound holds for* $\widehat{\delta}_N^U$, *up to a change in absolute constants.*

Remark 3.2.
1. *It is evident that whenever* $\mathcal{O} = 0$, *the best possible rates implied by Theorem 3.2 are of order* $N^{-2/3}$ *(indeed, this is the case whenever* $M_\Delta\sqrt{n} \asymp N^{1/3}$ *and* $\bar{\delta} \lesssim N^{-2/3}$*), while the best possible rates attained by Theorem 3.3 are of order* $N^{-3/4}$ *(when* $M_\Delta\sqrt{n} \asymp N^{1/4}$ *and* $\bar{\delta} \lesssim N^{-3/4}$*); in particular, in this case the choice of* M_Δ *and* n *is independent of* $\bar{\delta}$. *In general, if* $\mathcal{O} = \varepsilon N$ *for* $\varepsilon > 0$, *the best rates implied by Theorems 3.2 and 3.3 are* $\bar{\delta} + C(\mathcal{F},\rho,P)\varepsilon^{-2/3}$ *and* $\bar{\delta} + C(\mathcal{F},\rho,P)\varepsilon^{-3/4}$ *respectively.*
2. *Assumption requiring that* $M_\Delta \geqslant 1$ *is introduced for convenience: without it, extra powers of the ratio* $\frac{\max(\Delta,\sigma(\ell,\mathcal{F}))}{\Delta}$ *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, 30, 27, among others], as well as the "median-of-means tournaments" introduced in [31]; all these methods focus on estimating the differences $\mathcal{L}(f_1) - \mathcal{L}(f_2)$ for all $f_1, f_2 \in \mathcal{F}$. Recall that $f_* = \mathrm{argmin}_{f\in\mathcal{F}} P\ell(f)$, and observe that for any fixed $f' \in \mathcal{F}$, f_* can be equivalently defined via

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

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

A version of the robust empirical risk minimizer (1.5) corresponding to this problem can be defined as

$$\widehat{\mathcal{L}}^{(k)}(f-f') := \operatorname*{argmin}_{y\in\mathbb{R}} \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\,\frac{\left(\bar{\mathcal{L}}_j(f)-\bar{\mathcal{L}}_j(f')\right)-y}{\Delta}\right) \tag{3.6}$$

for appropriately chose $\Delta > 0$, and

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

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

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

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

We will now formalize this argument and provide performance guarantees; we use the framework of Theorem 3.3 which leads to the bounds that are easier to state and interpret. However, similar reasoning applies to the setting of Theorem 3.2 as well. Presented algorithms also admit straightforward permutation-invariant modifications that we omit. Let

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

be the "empirical excess risk" of f. Indeed, this is a meaningful notion as $\widehat{f}_N$ is the minimizer of $\widehat{\mathcal{L}}^{(k)}(f)$ over $f \in \mathcal{F}$. Assume that the initial sample of size N is split into two disjoint parts S_1 and S_2 of cardinalities that differ at most by 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 (1.5) 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 $\delta' = \bar{\delta} + C(\rho)\left(D^2 + \sigma(\ell,\mathcal{F})\sqrt{n}M_{\Delta_1}\right)\left(\frac{\mathfrak{B}^6(\ell,\mathcal{F})}{M^4_{\Delta_1} n_1^2} + \frac{s+\mathcal{O}}{N}\right)$ be a known upper bound on the excess risk in Theorem 3.3 (while this condition is restrictive, it is similar to the requirements of existing approaches [12, 31]; discussion of adaptation issues is beyond the scope of this paper and will be addressed elsewhere). Set
$$\widehat{\mathcal{F}}(\delta') := \left\{ f \in \mathcal{F} :\ \widehat{\mathcal{E}}_N(f) \leqslant \delta' \right\}.$$
3. Define $\widehat{f}''_N := \operatorname{argmin}_{f\in\widehat{\mathcal{F}}(\delta')} \widehat{\mathcal{L}}^{(k)}(f - \widehat{f}_{|S_1|})$ where

$$\widehat{\mathcal{L}}^{(k)}\left(f-\widehat{f}_{|S_1|}\right) = \operatorname*{argmin}_{y\in\mathbb{R}} \sum_{j=1}^{k_2} \rho\left(\sqrt{n}\,\frac{\left(\bar{\mathcal{L}}_j(f)-\bar{\mathcal{L}}_j(\widehat{f}_{|S_1|})\right)-y}{\Delta_2}\right) \tag{3.7}$$

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

Theorem 3.4. *Suppose that*

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

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

Finally, we require that

$$\sqrt{k_1}M_{\Delta_1} \geqslant \frac{c'}{\sigma(\ell,\mathcal{F})}\,\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{|S_1|}}\sum_{j=1}^{|S_1|}\left(\ell(f(X_j))-P\,\ell(f)\right) \quad \textit{and} \tag{3.8}$$

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

Then

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

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 3.4 is technical, so let us try to distill the main ideas. The key difference between Theorem 3.3 and Theorem 3.4 is that the "remainder term"

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

is replaced by a potentially much smaller quantity $\sqrt{\delta'}\sqrt{n}M_\Delta\left(\frac{\mathfrak{B}^6(\ell,\mathcal{F})}{M_\Delta^4 n^2} + \frac{s+\mathcal{O}}{N}\right)$. In particular, if $\delta' \ll \left(nM_\Delta^2\right)^{-1}$, this term often becomes negligible. To be more specific, assume that $\bar{\delta} = \frac{C(\mathcal{F})}{\sqrt{N}}\cdot h(N)$ where $h(N)\to 0$ as $N\to\infty$ (meaning that fast rates are achievable) and that $\mathcal{O} = \varepsilon N$ for $\varepsilon \geqslant \frac{1}{N}$. Moreover, suppose that $\mathfrak{B}(\ell,\mathcal{F})$ is bounded above by a constant. If Δ_1 is chosen such that $\Delta_1 \asymp \sigma(\ell,\mathcal{F})$, then $\delta' = C\left(\bar{\delta} + \sigma(\ell,\mathcal{F})\left(\left(\frac{k}{N}\right)^{3/2} + \frac{s+\mathcal{O}}{\sqrt{kN}}\right)\right)$. 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

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

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 (3.8) hold for sufficiently large N as $k_j \gtrsim \sqrt{N},\ j=1,2$). Analysis of the case when $\mathcal{O}=0$ follows similar steps, with several simplifications.

4. Examples.

We consider two common prediction problems, regression and binary classification, and discuss the implications of our main results for these problems.

4.1. Binary classification with convex surrogate loss.

The key elements of the binary classification framework were outlined in Section 2. 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 $f\in\mathcal{F}$, $\sup_{z\in S}|f(z)| \leqslant B$. In this case, Assumption 2 holds with $D = 2e^B$ for all $f\in\mathcal{F}$. See [6] 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 $\mathcal{F} = \{\langle\cdot,v\rangle :\ \|v\|_2\leqslant R\}$ 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. It follows from Proposition 1 in [26] (see also [43]) that Assumption 2 holds with $D = \frac{1}{\sqrt{2\tau}}$ in this case.

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

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

Since ℓ is Lipchitz continuous by assumption (with Lipschitz constant denoted $L(\ell)$), consequent application of symmetrization and Talagrand's contraction inequalities [29, 44] 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 $\mathcal{F}$ is a subset of a linear space of dimension d, then, according to Proposition 3.2 in [24], $\mathbb{E} \sup_{\|f - f_*\|_{L_2(\Pi)} \leqslant D\sqrt{\delta}} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \varepsilon_j (f - f_*)(Z_j) \right| \leqslant D\sqrt{\delta}\sqrt{d}$, whence $\widetilde{\omega}(\delta) := 4D\, L(\ell)\sqrt{\delta d}$ is an upper bound for $\omega(\delta)$ and is of concave type, implying that

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

More generally, assume that the class $\mathcal{F}$ has a measurable envelope $F(z) := \sup_{f \in \mathcal{F}} |f(z)|$ 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 $N(\mathcal{F}, Q, \varepsilon)$ of the class $\mathcal{F}$ with respect to the norm $L_2(Q)$ satisfy the bound

$$N(\mathcal{F}, Q, \varepsilon) \leqslant \left(\frac{A \|F\|_{L_2(Q)}}{\varepsilon} \right)^V \tag{4.1}$$

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 $\mathcal{F}$ [46, 24]. In this case, it is not difficult to show (see for example the proof of Lemma 4.1 in the appendix) that

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

hence it is easy to check that in this case

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

It immediately follows from the discussion following Theorem 3.4 that the excess risk of the estimator $\widehat{f}''_N$ satisfies

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

with probability at least $1 - 20e^{-s}$. Similar results hold for regression problems with Lipschitz losses, such as Huber's loss or quantile loss [3].

4.2. 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 $\|\eta\|_{2,1} := \int_0^\infty \sqrt{\Pr(|\eta| > t)}dt$, and observe that $\|\eta\|_{2,1} < \infty$ as $\sup_{f\in\mathcal{F}} \mathbb{E}(Y - f(Z))^4 < \infty$ by assumption. As before, Π will stand for the marginal distribution of Z. Let $\mathcal{F}$ be a given convex class of functions mapping S to $\mathbb{R}$ and such that the regression function f_* belongs to $\mathcal{F}$, so that

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

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 $\mathcal{F}$ has a measurable envelope $F(z) := \sup_{f\in\mathcal{F}} |f(z)|$ that satisfies $\|F(Z)\|_{\psi_2} < \infty$. Moreover, suppose that the covering numbers [3] $N(\mathcal{F}, Q, \varepsilon)$ of the class $\mathcal{F}$ with respect to the norm $L_2(Q)$ satisfy the bound

$$N(\mathcal{F}, Q, \varepsilon) \leqslant \left(\frac{A\|F\|_{L_2(Q)}}{\varepsilon}\right)^V \tag{4.2}$$

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 $\mathcal{F}$ [46, 24].

Bernstein's assumption. It follows from Lemma 5.1 in [24] that

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

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

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

Lemma 4.1. *Under the assumptions made in this section and for $\Delta \geqslant \sigma(\ell, \mathcal{F})$,*

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

The proof is given in the appendix. An immediate corollary of the lemma, according to the discussion following Theorem 3.4, is that the excess risk of the estimator $\widehat{f}_N''$ satisfies the inequality

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

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

5. Proofs of the main results.

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

$$\widehat{G}_k(z; f) = \frac{1}{\sqrt{k}}\sum_{j=1}^k \rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f) - \mathcal{L}(f)) - z}{\Delta}\right),$$
$$G_k(z; f) = \sqrt{k}\,\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f) - \mathcal{L}(f)) - z}{\Delta}\right).$$

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

$$\widehat{G}_k\left(\widehat{e}^{(k)}(f); f\right) = 0, \tag{5.1}$$
$$G_k\left(e^{(k)}(f); f\right) = 0.$$

Comparing this to the definition of $\widehat{\mathcal{L}}^{(k)}(f)$ (1.2), it is easy to see that $\widehat{e}^{(k)}(f) = \widehat{\mathcal{L}}^{(k)}(f) - \mathcal{L}(f)$. Hence $e^{(k)}(f)$, the "population version" of $\widehat{e}^{(k)}(f)$, is a natural measure of bias of the estimator $\widehat{\mathcal{L}}^{(k)}(f)$.

[3] Definition..

5.1. Technical tools.

We summarize the key results that our proofs rely on.

Lemma 5.1. *Let ρ satisfy Assumption 1. Then for any random variable Y with $\mathbb{E}Y^2 < \infty$,*

$$Var\left(\rho'(Y)\right) \leqslant Var(Y)\,.$$

Proof. See Lemma 5.3 in [35]. □

Lemma 5.2. *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 $Var(Z_1) = Var(\xi_1)$.

Proof. This bound follows from a standard application of Lindeberg's replacement method; see [38, chapter 11]. □

Lemma 5.3. *Assume that $\mathbb{E}|f(X) - \mathbb{E}f(X)|^2 < \infty$ for all $f \in \mathcal{F}$ and that ρ satisfies Assumption 1. Then for all $f \in \mathcal{F}$ and $z \in \mathbb{R}$ satisfying $|z| \leqslant \frac{\Delta}{\sqrt{n}}\frac{1}{2}$,*

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

Proof. See Lemma 4.2 in [35]. □

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

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

where $\gamma_2(\mathcal{F}, L_\infty(\Pi_N))$ is Talagrand's generic chaining complexity [42].

Lemma 5.4. *Let $\sigma^2 := \sup_{f\in\mathcal{G}}\mathbb{E}f^2(X)$. Then there exists a universal constant $C > 0$ such that*

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

Proof. See Theorem 3.16 in [24]. □

The following form of Talagrand's concentration inequality is due to Klein and Rio (see section 12.5 in [11]).

Lemma 5.5. *Let $\{Z_j(f),\ f \in \mathcal{F}\},\ j = 1, \dots, N$ be independent (not necessarily identically distributed) separable stochastic processes indexed by class $\mathcal{F}$ and such that $|Z_j(f) - \mathbb{E}Z_j(f)| \leqslant M$ a.s. for all $1 \leqslant j \leqslant N$ and $f \in \mathcal{F}$. Then the following inequality holds with probability at least $1 - e^{-s}$:*

$$\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^{N}(Z_j(f) - \mathbb{E}Z_j(f))\right) \leqslant 2\mathbb{E}\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^{N}(Z_j(f) - \mathbb{E}Z_j(f))\right) + V(\mathcal{F})\sqrt{2s} + \frac{4Ms}{3}, \tag{5.2}$$

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

It is easy to see, applying (5.2) to processes $\{-Z_j(f),\ f \in \mathcal{F}\}$, that

$$\inf_{f\in\mathcal{F}}\left(\sum_{j=1}^{N}(Z_j(f) - \mathbb{E}Z_j(f))\right) \geqslant -2\mathbb{E}\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^{N}(\mathbb{E}Z_j(f) - Z_j(f))\right) - V(\mathcal{F})\sqrt{2s} - \frac{4Ms}{3} \tag{5.3}$$

with probability at least $1-e^{-s}$. Next, we describe the tools necessary to extend these concentration inequalities to nondegenerate U-statistics. Deviation inequality (5.2) is a corollary of the following bound for the moment generating function (section 12.5 in [11]):

$$\log \mathbb{E} e^{\lambda\left(\sum_{j=1}^{N}(Z_j(f)-\mathbb{E}Z_j(f))\right)} \leqslant \frac{e^{\lambda M}-\lambda M-1}{M^2}\left(V^2(\mathcal{F})+2M\,\mathbb{E}\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^{N}(Z_j(f)-\mathbb{E}Z_j(f))\right)\right) \tag{5.4}$$

that holds for all $\lambda>0$. We use this fact to demonstrate a straightforward extension of Lemma 5.5 to the case of U-statistics. Let π_N be the collection of all permutations $\pi:\{1,\dots,N\}\mapsto\{1,\dots,N\}$. Given $(i_1,\dots,i_N)\in\pi_N$ and a U-statistic $U_{N,n}$ with kernel h defined in (1.3), let

$$T_{i_1,\dots,i_N}:=\frac{1}{k}\left(h\left(X_{i_1},\dots,X_{i_n}\right)+h\left(X_{i_{n+1}},\dots,X_{i_{2n}}\right)+\dots+h\left(X_{i_{(k-1)n+1}},\dots,X_{i_{kn}}\right)\right).$$

It is well known (e.g., see section 5 in [20]) that the following representation holds:

$$U_{N,n}=\frac{1}{N!}\sum_{(i_1,\dots,i_N)\in\pi_N}T_{i_1,\dots,i_N}. \tag{5.5}$$

Let $U'_{N,n}(z;f)=\frac{1}{\binom{N}{n}}\sum_{J\in\mathcal{A}_N^{(n)}}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}(f;J)-\mathbb{E}\ell(f(X)))-z}{\Delta}\right)$. Applied to $U'_{N,n}(z;f)$, relation (5.5) yields that

$$U'_{N,n}(z;f)=\frac{1}{N!}\sum_{(i_1,\dots,i_N)\in\pi_N}T_{i_1,\dots,i_N}(z;f),$$

where

$$T_{i_1,\dots,i_N}(z;f)=\frac{1}{k}\Bigg(\rho'\left(\sqrt{n}\,\frac{\bar{\mathcal{L}}(f;\{i_1,\dots,i_n\})-\mathbb{E}\ell(f(X))-z}{\Delta}\right)+ \\ \dots+\rho'\left(\sqrt{n}\,\frac{\bar{\mathcal{L}}(f;\{i_{(k-1)n+1},\dots,i_{kn}\})-\mathbb{E}\ell(f(X))-z}{\Delta}\right)\Bigg).$$

Jensen's inequality implies that for any $\lambda>0$,

$$\mathbb{E}\exp\left(\frac{\lambda}{N!}\sum_{(i_1,\dots,i_N)\in\pi_N}(T_{i_1,\dots,i_N}(z;f)-\mathbb{E}T_{i_1,\dots,i_N}(z;f))\right) \\ \leqslant\frac{1}{N!}\sum_{(i_1,\dots,i_N)\in\pi_N}\mathbb{E}\exp\left(\lambda\left(T_{1,\dots,N}(z;f)-\mathbb{E}T_{1,\dots,N}(z;f)\right)\right),$$

hence bound (5.4) can be applied and yields that

$$\sup_{f\in\mathcal{F}}\left(U'_{N,n}(z;f)-\mathbb{E}U'_{N,n}(z;f)\right)\leqslant 2\mathbb{E}\sup_{f\in\mathcal{F}}\left(T_{1,\dots,N}(z;f)-\mathbb{E}T_{1,\dots,N}(z;f)\right) \\ +\sup_{f\in\mathcal{F}}\sqrt{\mathrm{Var}\left(\rho'\left(\sqrt{n}\,\frac{\bar{\theta}(f;\{1,\dots,n\})-Pf-z}{\Delta}\right)\right)}\sqrt{\frac{2s}{k}}+\frac{8s\|\rho'\|_\infty}{3k} \tag{5.6}$$

with probability at least $1-e^{-s}$. The expression can be further simplified by noticing that $\|\rho'\|_\infty\leqslant 2$ and that

$$\mathrm{Var}\left(\rho'\left(\sqrt{n}\,\frac{\bar{\theta}(f;\{1,\dots,n\})-Pf-z}{\Delta}\right)\right)\leqslant\frac{\sigma^2(f)}{\Delta^2}.$$

due to Lemma 5.1.

5.2. Proof of Theorems 3.2 and 3.3.

We will provide detailed proofs for the estimator $\widehat{f}_N$ that is based on disjoint groups $G_1, \dots, G_k$. 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 5.5) are replaced by its version for nondegenerate U-statistics (5.6).

Let $J \subset \{1, \dots, k\}$ of cardinality $|J| \geqslant k - \mathcal{O}$ 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 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 $\mathcal{O} < k/2$,

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

Consider stochastic process $R_N(f)$ defined as

$$R_N(f) = \widehat{G}_k\left(0; f\right) + \partial_z G_k\left(0; f\right) \cdot \widehat{e}^{(k)}(f), \tag{5.7}$$

where $\partial_z G_k\left(0; f\right) := \partial_z G_k\left(z; f\right)_{|_{z=0}}$. Whenever $\partial_z G_k\left(0; f\right) \neq 0$ (this assumption will be justified by Lemma 5.6 below), we can solve (5.7) for $\widehat{e}^{(k)}(f)$ to obtain

$$\widehat{e}^{(k)}(f) = -\frac{\widehat{G}_k\left(0; f\right)}{\partial_z G_k\left(0; f\right)} + \frac{R_N(f)}{\partial_z G_k\left(0; f\right)}, \tag{5.8}$$

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

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

By the definition (1.5) of $\widehat{f}_N$, $\widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) \leqslant \widehat{\mathcal{L}}^{(k)}(f_*)$, hence

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

Rearranging the terms, it is easy to see that

$$\widehat{\delta}_N = \mathcal{L}(\widehat{f}_N) - \mathcal{L}(f_*) \leqslant \left| \frac{\widehat{G}_k\left(0; \widehat{f}_N\right)}{\partial_z G\left(0; \widehat{f}_N\right)} - \frac{\widehat{G}_k\left(0; f_*\right)}{\partial_z G\left(0; f_*\right)} \right| + 2 \sup_{f \in \mathcal{F}(\widehat{\delta}_N)} \left| \frac{R_N(f)}{\partial_z G_k\left(0; f\right)} \right|. \tag{5.9}$$

Remark 5.1. *Similar argument also implies, in view of the inequality $\mathcal{L}(f_*) \leqslant \mathcal{L}(\widehat{f}_N)$, that*

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

hence

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

It follows from (5.9) that in order to estimate the excess risk of $\widehat{f}_N$, it suffices to obtain the upper bounds for

$$A_1 := \left| \frac{\widehat{G}_k\left(0; \widehat{f}_N\right)}{\partial_z G_k\left(0; \widehat{f}_N\right)} - \frac{\widehat{G}_k\left(0; f_*\right)}{\partial_z G_k\left(0; f_*\right)} \right| \tag{5.10}$$

and

$$A_2 := \sup_{f\in\mathcal{F}(\hat{\delta}_N)} \left| \frac{R_N(f)}{\partial_z G_k\left(0;f\right)} \right|. \tag{5.11}$$

Observe that

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

Since ρ'' is Lipschitz continuous by assumption,

$$\begin{aligned}
&\left| \frac{\widehat{G}_k\left(0;f_*\right)}{\partial_z G_k\left(0;f_*\right)\partial_z G_k\left(0;\widehat{f}_N\right)} \left(\partial_z G_k\left(0;f_*\right) - \partial_z G_k\left(0;\widehat{f}_N\right)\right) \right| \\
&\quad = \left| \frac{\widehat{G}_k\left(0;f_*\right)}{\partial_z G_k\left(0;f_*\right)\partial_z G_k\left(0;\widehat{f}_N\right)} \frac{\sqrt{nk}}{\Delta} \mathbb{E}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*) - \mathcal{L}(f_*)}{\Delta}\right) - \rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(\widehat{f}_N) - \mathcal{L}(\widehat{f}_N)}{\Delta}\right)\right)\right| \\
&\qquad \leqslant L(\rho'') \left| \frac{\widehat{G}_k\left(0;f_*\right)}{\partial_z G_k\left(0;f_*\right)\partial_z G_k\left(0;\widehat{f}_N\right)} \frac{\sqrt{nk}}{\Delta^2} \mathrm{Var}^{1/2}\left(\ell(\widehat{f}_N(X)) - \ell(f_*(X))\right)\right| \\
&\qquad = C(\rho) \left| \frac{\widehat{G}_k\left(0;f_*\right)}{\partial_z G_k\left(0;f_*\right)\partial_z G_k\left(0;\widehat{f}_N\right)} \right| \frac{\sqrt{nk}}{\Delta^2} \nu(\hat{\delta}_N).
\end{aligned} \tag{5.12}$$

We following two lemmas are required to proceed.

Lemma 5.6. *There exist $C(\rho) > 0$ such that for any $f \in \mathcal{F}$,*

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

Proof. See section A.1. □

In particular, the first bound of Lemma 5.6 implies that for n large enough,

$$\inf_{f\in\mathcal{F}} \left|\partial_z G_k\left(0;f\right)\right| \geqslant \frac{1}{2}\frac{\sqrt{kn}}{\max\left(\Delta, \sigma(\ell,\mathcal{F})\right)} = \frac{1}{2}\frac{\sqrt{kn}}{\widetilde{\Delta}}. \tag{5.13}$$

It is also easy to deduce from the proof of Lemma 5.6 that for small n and $\Delta > \sigma(\ell,\mathcal{F})$, $\inf_{f\in\mathcal{F}} \left|\partial_z G_k\left(0;f\right)\right| \geqslant c(\rho)\frac{\sqrt{kn}}{\Delta}$ for some positive $c(\rho)$.

Lemma 5.7. *For any $f \in \mathcal{F}$,*

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

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

Proof. See section A.2. □

Lemma 5.7 and (5.13) imply, together with (5.12), that

$$\begin{aligned}
&\left| \frac{\widehat{G}_k\left(0;f_*\right)}{\partial_z G_k\left(0;f_*\right)\partial_z G_k\left(0;\widehat{f}_N\right)} \left(\partial_z G_k\left(0;f_*\right) - \partial_z G_k\left(0;\widehat{f}_N\right)\right) \right| \\
&\qquad \leqslant C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{\frac{s}{N}} + \frac{G_{f_*}(n,\Delta)}{\sqrt{n}} + \sqrt{n}\frac{s}{N} + \sqrt{n}\frac{\mathcal{O}}{N}\right)\nu(\hat{\delta}_N)
\end{aligned} \tag{5.14}$$

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

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

Define

$$\overline{\delta}_1 := \min \left\{ \delta > 0 : \ C_1(\rho) \left(\sqrt{\frac{s}{N}} + \frac{G_{f_*}(n, \Delta)}{\sqrt{n}} + \sqrt{n} \frac{s + \mathcal{O}}{N} \right) \frac{\widetilde{\nu}(\delta)}{\delta} \leqslant \frac{1}{7} \right\} \tag{5.15}$$

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

$$\left| \frac{\widehat{G}_k\left(0; f_*\right)}{\partial_z G_k\left(0; f_*\right) \partial_z G_k\left(0; \widehat{f}_N\right)} \left(\partial_z G_k\left(0; f_*\right) - \partial_z G_k\left(0; \widehat{f}_N\right) \right) \right| \leqslant \frac{\widehat{\delta}_N}{7}, \tag{5.16}$$

for appropriately chosen $C_1(\rho)$.

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

Lemma 5.8. *Let $(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n)$ 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*

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

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

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

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

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

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

$$\widetilde{U}(\delta, s) := \frac{2}{\Delta} \left(c(\gamma) \, \widetilde{\omega}(\delta) + \widetilde{\nu}(\delta) \sqrt{\frac{s}{2}} \right) + \frac{32 s}{\sqrt{k}}, \tag{5.17}$$

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

define

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

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

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

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

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

Lemma 5.9. *With probability at least* $1 - e^{-2s}$,

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

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

Proof. See section A.4. □

Next, we state the "uniform version" of Lemma 5.9:

Lemma 5.10. *With probability at least* $1 - e^{-s}$, *for all* $\delta \geqslant \delta_{\min}$ *simultaneously,*

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

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

Proof. See section A.5. □

It follows from Lemma 5.10 and inequality (5.13) that on event Θ_2 of probability at least $1 - e^{-s}$, for all $\delta \geqslant \delta_{\min}$ simultaneously,

$$\sup_{f\in\mathcal{F}(\delta)} \left|\frac{\widehat{G}_k(0;f) - \widehat{G}_k(0;f_*)}{\partial_z G_k(0;f)}\right| \leqslant C(\rho)\delta\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}} + \frac{\widetilde{\Delta}}{\sqrt{n}}\frac{\widetilde{B}(\delta_{\min})}{\delta_{\min}}\right) + 4\widetilde{\Delta}\sqrt{n}\frac{\mathcal{O}}{N}. \tag{5.18}$$

Define

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

$$\bar{\delta}_3 := \min\left\{\delta > 0 :\ C_3(\rho)\frac{\widetilde{\Delta}}{\sqrt{n}}\frac{\widetilde{B}(\delta)}{\delta} \leqslant \frac{1}{7}\right\}$$

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

$$\sup_{f\in\mathcal{F}(\widehat{\delta}_N)} \left|\frac{\widehat{G}_k(0;f) - \widehat{G}_k(0;f_*)}{\partial_z G_k(0;f)}\right| \leqslant \frac{2\,\widehat{\delta}_N}{7} + 4\widetilde{\Delta}\sqrt{n}\frac{\mathcal{O}}{N} \tag{5.19}$$

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 5.11. *Assume that conditions of Theorem 3.1 hold, and let $\delta_{\min} > 0$ be fixed. Then for all $s > 0$, $\delta \geqslant \delta_{\min}$, positive integers n and k such that*

$$\delta \frac{\widetilde{U}(\delta_{\min}, s)}{\delta_{\min}\sqrt{k}} + \sup_{f\in\mathcal{F}} G_f(n,\Delta) + \frac{s+\mathcal{O}}{k} \leqslant c(\rho), \tag{5.20}$$

the following inequality holds with probability at least $1 - 7e^{-s}$, uniformly over all δ satisfying (5.20):

$$\sup_{f\in\mathcal{F}(\delta)} |R_N(f)| \leqslant C(\rho)\sqrt{N}\frac{\widetilde{\Delta}^2}{\Delta^2}\left(n^{1/2}\delta^2\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}\sqrt{N}}\right)^2 \bigvee \frac{\sigma^2(\ell,f_*)}{\Delta^2}\frac{n^{1/2}\,s}{N}\right.$$
$$\left. \bigvee n^{1/2}\left(\sup_{f\in\mathcal{F}}\frac{G_f(n,\Delta)}{\sqrt{n}}\right)^2 \bigvee n^{3/2}\frac{s^2}{N^2} \bigvee n^{3/2}\frac{\mathcal{O}^2}{N^2}\right). \tag{5.21}$$

Moreover, the bound of Theorem 3.1 holds on the same event.

Proof. See section A.6. □

Recall that

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

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 5.11 holds with $\delta_{\min} = \bar{\delta}_2$, and consider the event $\Theta_3 \cap \{\widehat{\delta}_N > \bar{\delta}_2\}$. We will now show that on this event, Lemma 5.11 applies with $\delta = \widehat{\delta}_N$. Indeed, the bound of Theorem 3.1 is valid on Θ_3, hence the inequality (3.4) implies that on Θ_3, $\widehat{\delta}_N \leqslant C(\rho)\frac{\widetilde{\Delta}}{\sqrt{n}}$, and it is straightforward to check that condition (5.20) of Lemma 5.11 holds with $\delta_{\min} = \bar{\delta}_2$ and $\delta = \widehat{\delta}_N$. It follows from inequality (5.13) that on event $\Theta_3 \cap \{\widehat{\delta}_N \geqslant \bar{\delta}_2\}$,

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

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

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

whenever $\Delta \geqslant \sigma(\ell,\mathcal{F})$ (so that $\widetilde{\Delta} = \Delta$) and $C_2(\rho)$ in the definition of $\bar{\delta}_2$ 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 $\frac{s+\mathcal{O}}{k} \leqslant c$ under the conditions of Theorem 3.1,

$$n^{3/2}\widetilde{\Delta}\frac{s^2+\mathcal{O}^2}{N^2} \leqslant C\widetilde{\Delta}\sqrt{n}\frac{s+\mathcal{O}}{N}.$$

Combining the inequalities obtained above, we deduce on event $\Theta_3 \cap \{\widehat{\delta}_N \geqslant \bar{\delta}_2\}$,

$$2 \sup_{f\in\mathcal{F}(\widehat{\delta}_N)} \left|\frac{R_N(f)}{\partial_z G_k\left(0;f\right)}\right| \leqslant \frac{2\widehat{\delta}_N}{7} + C(\rho)\widetilde{\Delta}\left(\sqrt{n}\frac{s+\mathcal{O}}{N} \bigvee \frac{\sup_{f\in\mathcal{F}}\left(G_f\left(n,\Delta\right)\right)^2}{\sqrt{n}}\right) \tag{5.22}$$

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

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

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

$$2 \sup_{f\in\mathcal{F}(\widehat{\delta}_N)} \left|\frac{R_N(f)}{\partial_z G_k\left(0;f\right)}\right| + 4\widetilde{\Delta}\sqrt{n}\frac{\mathcal{O}}{N} \leqslant \frac{2\widehat{\delta}_N}{7} + \frac{\widehat{\delta}_N}{7} = \frac{3\widehat{\delta}_N}{7}. \tag{5.23}$$

Note that the expression above takes care of the term $4\widetilde{\Delta}\sqrt{n}\frac{\mathcal{O}}{N}$ that appeared in (5.19). Combining (5.16),(5.19),(5.23), we deduce that on event $\Theta_1 \cap \Theta_2 \cap \Theta_3 \cap \left\{\widehat{\delta}_N \geqslant \max\left(\bar{\delta}_1, \bar{\delta}_2, \bar{\delta}_3, 7\,\bar{\delta}_4\right)\right\}$,

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

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

$$\widehat{\delta}_N \leqslant \max\left(\bar{\delta}_1, \bar{\delta}_2, \bar{\delta}_3, 7\bar{\delta}_4\right). \tag{5.24}$$

Recall the definition (5.15) of $\bar{\delta}_1$. If condition 2 ("Bernstein condition") holds, then $\widetilde{\nu}(\delta) \leqslant D\sqrt{\delta}$ for small enough δ, in which case

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

where we used the fact that $\frac{s}{k} \leqslant c$ by assumption. Together with the bound (3.1) for $G_{f_*}(n,\Delta)$, we deduce that, under the assumption that $R_4(\ell,\mathcal{F}) < \infty$,

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

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

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

hence

$$\bar{\delta}_1 \leqslant C(\rho)D^2\left(\frac{s+\mathcal{O}}{N} + \frac{\mathfrak{B}^6(\ell,\mathcal{F})}{n^2 M^6_\Delta}\right). \tag{5.25}$$

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

$$\bar{\delta}_1 \leqslant C(\rho)D^2\left(\frac{s+\mathcal{O}}{N} + \frac{1}{M^4_\Delta\, n}\right). \tag{5.26}$$

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

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

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

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

and

$$\bar{\delta}_3 \leqslant C(\rho)\left(D^2\frac{\mathfrak{B}^6(\ell,\mathcal{F})}{n^2 M_\Delta^4} + \sigma(\ell,\mathcal{F})\frac{\mathfrak{B}^4(\ell,\mathcal{F})}{n^{3/2}M_\Delta^3}\right). \tag{5.27}$$

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,

$$\bar{\delta}_3 \leqslant C(\rho)\frac{D^2}{M_\Delta^2 n}. \tag{5.28}$$

Next, we obtain a simpler bound for $\bar{\delta}_4$: as $\Delta \geqslant \sigma(\ell,\mathcal{F})$ by assumption, $\widetilde{\Delta} = \Delta = \sigma(\ell,\mathcal{F})M_\Delta$, and the estimate (3.1) for $G_{f_*}(n,\Delta)$ implies (if $R_4(\ell,\mathcal{F}) < \infty$) that

$$\bar{\delta}_4 \leqslant C(\rho)\,\sigma(\ell,\mathcal{F})\left(\sqrt{n}M_\Delta\frac{s+\mathcal{O}}{N} + \frac{\mathfrak{B}^6(\ell,\mathcal{F})}{M_\Delta^5 n^{3/2}}\right). \tag{5.29}$$

If only $\sigma(\ell,\mathcal{F}) < \infty$, we similarly deduce from (3.1) that

$$\bar{\delta}_4 \leqslant C(\rho)\,\sigma(\ell,\mathcal{F})\left(\sqrt{n}M_\Delta\cdot\frac{s+\mathcal{O}}{N} + \frac{1}{M_\Delta^3\sqrt{n}}\right). \tag{5.30}$$

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

$$\bar{\delta}_2 \leqslant \bar{\delta}\bigvee C(\rho)D^2\frac{s}{N}\bigvee C(\rho)\sigma(\ell,\mathcal{F})\frac{s\sqrt{n}M_\Delta}{N}, \tag{5.31}$$

where $\bar{\delta}$ was defined in (3.5). Combining inequalities (5.25), (5.31) (5.27), (5.29) and (5.24), we obtain the final form of the bound under the stronger assumption $R_4(\ell,\mathcal{F}) < \infty$. Similarly, the combination of (5.26), (5.31) (5.28), (5.30) and (5.24) yields the bound under the weaker assumption $\sigma(\ell,\mathcal{F}) < \infty$.

5.3. *Proof of Theorem 3.4.*

Recall that $\widehat{\mathcal{E}}_N(f_*) := \widehat{\mathcal{L}}^{(k)}(f_*) - \widehat{\mathcal{L}}^{(k)}(\widehat{f}'_N)$ is the "empirical excess risk" of f_*, and let $\widehat{\delta}_N := \mathcal{E}(\widehat{f}'_N)$. It follows from Remark 5.1 that (using the notation used in the proof of Theorems 3.2 and 3.3)

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

On the event of Theorem 3.3 of probability at least $1 - 10e^{-s}$,

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

hence on this event

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

where the last inequality again follows from main steps in the proof of Theorem 3.3. [4] Consider the set $\widehat{\mathcal{F}}(\delta') = \left\{f \in \mathcal{F} : \ \widehat{\mathcal{E}}_N(f) \leqslant \delta'\right\}$. First, observe that on the event $\mathcal{E}_1$ of Theorem 3.3, $f_* \in \widehat{\mathcal{F}}(\delta')$ as implied by the previous display. We will next show that $\widehat{\mathcal{F}}(\delta') \subseteq \mathcal{F}(7\delta')$ on the event $\mathcal{E}_1$ of Theorem 3.3, meaning that for any $f \in \widehat{\mathcal{F}}(\delta')$, $\mathcal{E}(f) \leqslant 7\delta'$. Indeed, let $f \in \widehat{\mathcal{F}}(\delta')$ be such that $\mathcal{E}(f) = \sigma$. Then (5.8) implies that

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

Again, it follows from the arguments used in proof of Theorem 3.3 that on event $\mathcal{E}_1$ of probability at least $1 - 10e^{-s}$,

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

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

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

on $\mathcal{E}_1$. It remains to apply Theorem 3.3, conditionally on $\mathcal{E}_1$, to the class

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

To this end, we need to verify the assumption of Theorem 3.1 that translates into the requirement

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

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

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

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.

Acknowledgements

Stanislav Minsker gratefully acknowledges support by the National Science Foundation grant DMS-1712956.

References

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

[4] Similar result holds if δ' is replaced by its analogue from Theorem 3.3.

[2] Alon, N., Matias, Y. and Szegedy, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing* 20–29. ACM.
[3] Alquier, P., Cottet, V. and Lecué, G. (2017). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *arXiv preprint arXiv:1702.01402*.
[4] Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.
[5] Audibert, J.-Y., Catoni, O. et al. (2011). Robust linear least squares regression. *The Annals of Statistics* **39** 2766–2794.
[6] Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2004). Large margin classifiers: convex loss, low noise, and convergence rates. In *Advances in Neural Information Processing Systems* 1173–1180.
[7] Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association* **101** 138–156.
[8] Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability Theory and Related Fields* **135** 311–334.
[9] Bartlett, P. L., Mendelson, S. and Neeman, J. (2012). ℓ_1-regularized linear regression: persistence and oracle inequalities. *Probability theory and related fields* **154** 193–224.
[10] Bartlett, P. L., Bousquet, O., Mendelson, S. et al. (2005). Local rademacher complexities. *The Annals of Statistics* **33** 1497–1537.
[11] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.
[12] Brownlees, C., Joly, E., Lugosi, G. et al. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics* **43** 2507–2536.
[13] Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **48** 1148–1185. Institut Henri Poincaré.
[14] Chen, L. H. and Shao, Q.-M. (2001). A non-uniform Berry–Esseen bound via Stein's method. *Probability theory and related fields* **120** 236–254.
[15] Chen, Y., Su, L. and 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** 44.
[16] Chinot, G., Guillaume, L. and Matthieu, L. (2018). Statistical learning with Lipschitz and convex loss functions. *arXiv preprint arXiv:1810.01090*.
[17] Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics* **44** 2695–2725.
[18] Dudley, R. M. (2014). *Uniform central limit theorems* **142**. Cambridge university press.
[19] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* 293–325.
[20] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* **58** 13–30.
[21] Holland, M. J. and Ikeda, K. (2017a). Efficient learning with robust gradient descent. *arXiv preprint arXiv:1706.00182*.
[22] Holland, M. J. and Ikeda, K. (2017b). Robust regression using biased objectives. *Machine Learning* **106** 1643–1679.
[23] Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics; 2nd ed. Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ.
[24] Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems. Lecture Notes in Mathematics* **2033**. Springer Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour.
[25] Koltchinskii, V. and Mendelson, S. (2015). Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices* **2015** 12991–13008.
[26] Lecué, G. et al. (2007). Optimal rates of aggregation in classification under low noise assumption. *Bernoulli* **13** 1000–1022.
[27] Lecué, G. and Lerasle, M. (2017). Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*.

[28] Lecué, G., Lerasle, M. and Mathieu, T. (2018). Robust classification via MOM minimization. *arXiv preprint arXiv:1808.03106*.
[29] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: isoperimetry and processes*. Springer-Verlag, Berlin.
[30] Lerasle, M. and Oliveira, R. I. (2011). Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.
[31] Lugosi, G. and Mendelson, S. (2016). Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*.
[32] Lugosi, G. and Mendelson, S. (2017). Regularization, sparse recovery, and median-of-means tournaments. *arXiv preprint arXiv:1701.04112*.
[33] Mayzlin, D., Dover, Y. and Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* **104** 2421–55.
[34] Mendelson, S. (2014). Learning without concentration. In *Conference on Learning Theory* 25–39.
[35] Minsker, S. (2018). Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*.
[36] Minsker, S. and Strawn, N. (2017). Distributed statistical estimation and rates of convergence in normal approximation. *arXiv preprint arXiv:1704.02658*.
[37] Nemirovski, A. and Yudin, D. (1983). *Problem complexity and method efficiency in optimization*. John Wiley & Sons Inc.
[38] O'Donnell, R. (2014). *Analysis of boolean functions*. Cambridge University Press.
[39] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12** 2825–2830.
[40] Prasad, A., Suggala, A. S., Balakrishnan, S. and Ravikumar, P. (2018). Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.
[41] Ron Schmelzer, F. (2019). The Achilles' Heel Of AI. https://www.forbes.com/sites/cognitiveworld/2019/03/07/the-achilles-heel-of-ai.
[42] Talagrand, M. (2014). *Upper and lower bounds for stochastic processes: modern methods and classical problems* **60**. Springer Science & Business Media.
[43] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics* **32** 135–166.
[44] Van de Geer, S. (2016). Estimation and testing under sparsity. *Lecture Notes in Mathematics* **2159**.
[45] Van de Geer, S. A. and van de Geer, S. (2000). *Empirical Processes in M-estimation* **6**. Cambridge university press.
[46] Wellner, J. and van der Vaart, A. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.
[47] Yin, D., Chen, Y., Ramchandran, K. and Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498*.

Appendix A: Remaining proofs.

A.1. Proof of Lemma 5.6.

As ρ is sufficiently smooth,

$$\partial_z G_k(0; f) = -\frac{\sqrt{kn}}{\Delta}\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f) - \mathcal{L}(f)}{\Delta}\right).$$

Let $W(\ell(f))$ denote a centered normal random variable variance equal to $\mathrm{Var}(\ell(f(X)))$. Lemma 5.2 implies that

$$\left|\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f) - \mathcal{L}(f)}{\Delta}\right) - \mathbb{E}\rho''\left(\frac{W(\ell(f))}{\Delta}\right)\right| \leqslant C\frac{\|\rho^{(5)}\|_\infty}{\Delta^3\sqrt{n}}\mathbb{E}\left|\ell(f(X)) - P\ell(f)\right|^3.$$

Next, as $\rho''(x) \geqslant I\{|x| \leqslant 1\}$ by assumption,

$$\mathbb{E}\rho''\left(\frac{W(\ell(f))}{\Delta}\right) \geqslant \Pr(|W(\ell(f))| \leqslant \Delta).$$

Gaussian tail bound implies that

$$\Pr(|W(\ell(f))| \leqslant \Delta) \geqslant 1 - 2\exp\left(-\frac{1}{2}\frac{\Delta^2}{\mathrm{Var}\left(\ell(f(X))\right)}\right) \geqslant \frac{1}{2}$$

whenever $\Delta^2 \geqslant 4\log(2)\mathrm{Var}\left(\ell(f(X))\right)$. On the other hand, if $\xi \sim N(0,1)$, then clearly $\Pr(Z \leqslant |t|) \geqslant \frac{2|t|}{\sqrt{2\pi}} e^{-t^2/2}$, hence

$$\Pr(|W(\ell(f))| \leqslant \Delta) \geqslant \frac{2\Delta}{\sqrt{2\pi\mathrm{Var}\left(\ell(f(X))\right)}} \exp\left(-\frac{1}{2}\frac{\Delta^2}{\mathrm{Var}\left(\ell(f(X))\right)}\right) \geqslant \frac{\Delta}{\sqrt{8\pi\mathrm{Var}\left(\ell(f(X))\right)}}$$

whenever $\Delta^2 < 4\log(2)\mathrm{Var}\left(\ell(f(X))\right)$. Combination of two bounds yields that

$$\Pr(|W(\ell(f))| \leqslant \Delta) \geqslant \frac{1}{2\sqrt{2\pi}} \min\left(\frac{\Delta}{\sqrt{\mathrm{Var}\left(\ell(f(X))\right)}}, 2\sqrt{\log 2}\right).$$

A.2. Proof of Lemma 5.7.

Observe that

$$\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) = \frac{1}{\sqrt{k}}\sum_{j\in J}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) + \frac{1}{\sqrt{k}}\sum_{j\notin J}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)$$
$$\leqslant \sqrt{\frac{|J|}{k}}\frac{1}{\sqrt{|J|}}\sum_{j\in J}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) + 2\frac{\mathcal{O}}{\sqrt{k}},$$

where we used the fact that $\|\rho'\|_\infty \leqslant 2$. Bernstein's inequality implies that

$$\left|\frac{1}{\sqrt{|J|}}\left(\sum_{j\in J}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) - \mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right)\right|$$
$$\leqslant 2\left(\mathrm{Var}^{1/2}\left(\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right)\sqrt{s} + \frac{2s}{\sqrt{|J|}}\right)$$

with probability at least $1-2e^{-s}$, where we again used the fact that $\|\rho'\|_\infty \leqslant 2$. Moreover, $\mathrm{Var}\left(\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right) \leqslant \frac{\sigma^2(\ell,f)}{\Delta^2}$ by Lemma 5.1, hence with the same probability

$$|\widehat{G}_k(0;f)| \leqslant \sqrt{k}\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right| + 2\left(\frac{\sigma(\ell,f)}{\Delta}\sqrt{s} + \frac{2s}{\sqrt{k}} + \frac{\mathcal{O}}{\sqrt{k}}\right).$$

Lemma 6.2 in [35] implies that

$$\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right| \leqslant \underbrace{\mathbb{E}\rho'\left(\frac{W(\ell(f))}{\Delta}\right)}_{=0} + 2G_f(n,\Delta),$$

hence the claim follows.

A.3. Proof of Lemma 5.8.

Since F is smooth, for any $x, y \in \mathbb{R}$, $F(y)-F(x) = \int_0^1 F'(x+\alpha(y-x))d\alpha\cdot(y-x)$. Let $S_n^\xi = \sum_{j=1}^n \xi_j$, $S_n^\eta = \sum_{j=1}^n \eta_j$. Then

$$F\left(S_n^\xi\right) - F\left(S_n^\eta\right) = \left(S_n^\xi - S_n^\eta\right)\int_0^1 F'\left(S_n^\eta + \alpha\left(S_n^\xi - S_n^\eta\right)\right)d\alpha,$$

hence

$$\mathbb{E}\left(F\left(S_n^\xi\right)-F\left(S_n^\eta\right)\right)=\int_0^1\mathbb{E}\left[\left(S_n^\xi-S_n^\eta\right)F'\left(S_n^\eta+\alpha\left(S_n^\xi-S_n^\eta\right)\right)\right]d\alpha.$$

Hölder's inequality yields that

$$\begin{aligned}\left|\mathbb{E}\left(S_n^\xi-S_n^\eta\right)F'\left(S_n^\eta+\alpha\left(S_n^\xi-S_n^\eta\right)\right)\right|&\leqslant\left(\mathbb{E}\left|S_n^\xi-S_n^\eta\right|^2\right)^{1/2}\left(\mathbb{E}\left|F'\left(S_n^\eta+\alpha\left(S_n^\xi-S_n^\eta\right)\right)\right|^2\right)^{1/2}\\&\leqslant\sqrt{n}\operatorname{Var}^{1/2}\left(\xi_1-\eta_1\right)\left(\mathbb{E}\left|F'\left(S_n^\eta+\alpha\left(S_n^\xi-S_n^\eta\right)\right)\right|^2\right)^{1/2},\end{aligned}$$

implying the first inequality. The rest of the proof is devoted to the second inequality of the lemma. Let (W,Z) be a centered Gaussian vector with the same covariance as (ξ_1,η_1), and let $(W_1,Z_1),\dots,(W_n,Z_n)$ be i.i.d. copies of (W,Z). We also set $S_n^W=\sum_{j=1}^n W_j$, $S_n^Z=\sum_{j=1}^n Z_j$. As $\mathbb{E}F\left(S_n^W\right)=\mathbb{E}F\left(S_n^Z\right)=0$ for bounded odd F, it is easy to see that

$$\begin{aligned}\left|\mathbb{E}\left(F\left(S_n^\xi\right)-F\left(S_n^\eta\right)\right)\right|&=\left|\mathbb{E}\left(F\left(S_n^\xi\right)-F\left(S_n^\eta\right)\right)-\mathbb{E}\left(F\left(S_n^W\right)-F\left(S_n^Z\right)\right)\right|\\&=\left|\int_0^1\left(\mathbb{E}\left(S_n^\xi-S_n^\eta\right)F'\left(S_n^\eta+\alpha\left(S_n^\xi-S_n^\eta\right)\right)-\mathbb{E}\left(S_n^W-S_n^Z\right)F'\left(S_n^Z+\alpha\left(S_n^W-S_n^Z\right)\right)\right)d\alpha\right|\\&\leqslant\int_0^1\left|\mathbb{E}\left(S_n^\xi-S_n^\eta\right)F'\left(S_n^\eta+\alpha\left(S_n^\xi-S_n^\eta\right)\right)-\mathbb{E}\left(S_n^W-S_n^Z\right)F'\left(S_n^Z+\alpha\left(S_n^W-S_n^Z\right)\right)\right|d\alpha.\end{aligned}$$

Next we will estimate, for each $\alpha\in[0,1]$, the expression

$$\left|\mathbb{E}\left(S_n^\xi-S_n^\eta\right)F'\left(S_n^\eta+\alpha\left(S_n^\xi-S_n^\eta\right)\right)-\mathbb{E}\left(S_n^W-S_n^Z\right)F'\left(S_n^Z+\alpha\left(S_n^W-S_n^Z\right)\right)\right|. \tag{A.1}$$

To this end, we will use Lindeberg's replacement method. For $i=0,\dots,n$, denote

$$T_i=(\xi_1-\eta_1,\dots,\xi_i-\eta_i,W_{i+1}-Z_{i+1},\dots,W_n-Z_n,\eta_1,\dots,\eta_i,Z_{i+1},\dots,Z_n).$$

Then the expression in (A.1) is equal to $|\mathbb{E}G(T_n)-\mathbb{E}G(T_0)|$, where

$$G(T)=\left(\sum_{i=1}^n T^{(i)}\right)F'\left(\sum_{j=1}^n\left(T^{(j+n)}+\alpha T^{(j)}\right)\right)$$

and $T^{(j)}$ stands for the j-th coordinate of T. Clearly,

$$\left|\mathbb{E}G(T_n)-\mathbb{E}G(T_0)\right|\leqslant\sum_{i=1}^n\left|\mathbb{E}G(T_i)-\mathbb{E}G(T_{i-1})\right|. \tag{A.2}$$

Fix i, and consider the Taylor expansions of $G(T_i)$ and $G(T_{i-1})$ at the point

$$T_i^0=(\xi_1-\eta_1,\dots,\xi_{i-1}-\eta_{i-1},0,W_{i+1}-Z_{i+1},\dots,W_n-Z_n,\eta_1,\dots,\eta_{i-1},0,Z_{i+1},\dots,Z_n)$$

(note that T_i^0 does not depend on ξ_i, η_i, W_i and Z_i). For $G(T_i)$ we get, setting $\delta_i=\xi_i-\eta_i$,

$$\begin{aligned}G(T_i)&=G(T_i^0)+\partial_iG(T_i^0)\cdot\delta_i+\partial_{n+i}G(T_i^0)\cdot\eta_i\\&\quad+\frac{1}{2}\left(\partial_{i,i}^2G(T_i^0)\cdot\delta_i^2+2\partial_{i,n+i}^2G(T_i^0)\cdot\delta_i\eta_i+\partial_{n+i,n+i}^2G(T_i^0)\cdot\eta_i^2\right)\\&\quad+\frac{1}{6}\left(\partial_{i,i,i}^3G(\tilde T_i^0)\cdot\delta_i^3+\partial_{n+i,n+i,n+i}^3G(\tilde T_i^0)\cdot\eta_i^3+\partial_{n+i,n+i,i}^3G(\tilde T_i^0)\cdot\eta_i^2\delta_i+\partial_{n+i,i,i}^3G(\tilde T_i^0)\cdot\eta_i\delta_i^2\right),\end{aligned}$$

where $\tilde T_i^0$ is a point on a line segment between T_i^0 and T_i. Similarly, setting $\Delta_i=W_i-Z_i$,

$$\begin{aligned}G(T_{i-1})&=G(T_i^0)+G(T_i^0)+\partial_iG(T_i^0)\cdot\Delta_i+\partial_{n+i}G(T_i^0)\cdot Z_i\\&\quad+\frac{1}{2}\left(\partial_{i,i}^2G(T_i^0)\cdot\Delta_i^2+\partial_{i,n+i}^2G(T_i^0)\cdot\Delta_iZ_i+\partial_{n+i,n+i}^2G(T_i^0)\cdot Z_i^2\right)\\&\quad+\frac{1}{6}\left(\partial_{i,i,i}^3G(\tilde T_i^0)\cdot\Delta_i^3+\partial_{n+i,n+i,n+i}^3G(\tilde T_i^0)\cdot Z_i^3+3\partial_{n+i,n+i,i}^3G(\tilde T_i^0)\cdot Z_i^2\Delta_i+3\partial_{n+i,i,i}^3G(\tilde T_i^0)\cdot Z_i\Delta_i^2\right),\end{aligned} \tag{A.3}$$

where $\bar{T}_i^0$ is a point on a line segment between T_i^0 and T_{i-1}. Using independence of T_i^0 and $(\xi_i, \eta_i, W_i, Z_i)$ and the fact that covariance structures of (ξ_i, η_i) and (W, Z) are the same, we deduce that

$$|\mathbb{E}G(T_i) - \mathbb{E}G(T_{i-1})| \leqslant \frac{1}{6}\mathbb{E}\Big|\partial^3_{i,i,i}G(\tilde{T}_i^0)\cdot\delta_i^3 + \partial^3_{n+i,n+i,n+i}G(\tilde{T}_i^0)\cdot\eta_i^3 + 3\partial^3_{n+i,n+i,i}G(\tilde{T}_i^0)\cdot\eta_i^2\delta_i + 3\partial^3_{n+i,i,i}G(\tilde{T}_i^0)\cdot\eta_i\delta_i^2\Big|$$
$$+ \frac{1}{6}\mathbb{E}\Big|\partial^3_{i,i,i}G(\tilde{T}_i^0)\cdot\Delta_i^3 + \partial^3_{n+i,n+i,n+i}G(\tilde{T}_i^0)\cdot Z_i^3 + 3\partial^3_{n+i,n+i,i}G(\tilde{T}_i^0)\cdot Z_i^2\Delta_i + 3\partial^3_{n+i,i,i}G(\tilde{T}_i^0)\cdot Z_i\Delta_i^2\Big|.$$

It remains estimate each of the terms above. Assume that $\tau \in [0,1]$ is such that

$$\tilde{T}_i^0 = (\xi_1 - \eta_1, \ldots, \xi_{i-1} - \eta_{i-1}, \tau(\xi_i - \eta_i), W_{i+1} - Z_{i+1}, \ldots, W_n - Z_n, \eta_1, \ldots, \eta_{i-1}, \tau\eta_i, Z_{i+1}, \ldots, Z_n).$$

1. Direct computation implies that

$$\partial^3_{i,i,i}G(\tilde{T}_i^0) = 3\alpha^2 F'''\left(\sum_{j\neq i}\Big(\eta_j + \alpha\delta_j\Big) + \tau(\eta_i + \alpha\delta_i)\right) + \alpha^3 F''''\left(\sum_{j\neq i}\Big(\eta_j + \alpha\delta_j\Big) + \tau(\eta_i + \alpha\delta_i)\right)\Big(\sum_{j\neq i}\delta_j + \tau\delta_i\Big),$$

hence

$$\mathbb{E}\Big|\partial^3_{i,i,i}G(\tilde{T}_i^0)\cdot\delta_i^3\Big| \leqslant 3\alpha^2\|F'''\|_\infty\mathbb{E}|\delta_i^3| + \alpha^3\|F''''\|_\infty\left(\mathbb{E}\Big|\sum_{j\neq i}\delta_j\Big|\,\mathbb{E}|\delta_i|^3 + \mathbb{E}|\delta_i|^4\right)$$
$$\leqslant 3\alpha^2\|F'''\|_\infty\left(\mathbb{E}\delta_i^2\right)^{1/2}\left(\mathbb{E}\delta_i^4\right)^{1/2} + \alpha^3\|F''''\|_\infty\left(\sqrt{\sum_{j\neq i}\mathbb{E}\delta_j^2}\,\left(\mathbb{E}\delta_i^2\right)^{1/2}\left(\mathbb{E}\delta_i^4\right)^{1/2} + \mathbb{E}|\delta_i|^4\right), \quad \text{(A.4)}$$

where we used Hölder's inequality in the last step.

2. Next,

$$\partial^3 G_{\eta_i,\eta_i,\eta_i}(\tilde{T}_i^0) = F''''\left(\sum_{j\neq i}\Big(\eta_j + \alpha\delta_j\Big) + \tau(\eta_i + \alpha\delta_i)\right)\Big(\sum_{j\neq i}\delta_j + \tau\delta_i\Big),$$

hence Hölder's inequality, together with the identity $\|F''''\|_\infty = M^{-3}\|H''''\|_\infty$, imply that

$$\mathbb{E}\Big|\partial^3_{n+i,n+i,n+i}G(\tilde{T}_i^0)\cdot\eta_i^3\Big| \leqslant \|F''''\|_\infty\left(\mathbb{E}|\eta_i|^3\mathbb{E}\left|\sum_{j\neq i}\delta_j\right| + \mathbb{E}|\delta_i\eta_i^3|\right)$$
$$\leqslant \|F''''\|_\infty\left(\mathbb{E}|\eta_i|^3\sqrt{\sum_{j\neq i}\mathbb{E}\delta_j^2} + \left(\mathbb{E}\delta_i^4\right)^{1/4}\left(\mathbb{E}\eta_i^4\right)^{3/4}\right). \quad \text{(A.5)}$$

3. Proceeding in a similar fashion, we deduce that

$$\partial^3 G_{n+i,n+i,i}(\tilde{T}_i^0) = F'''\left(\sum_{j\neq i}\Big(\eta_j + \alpha\delta_j\Big) + \tau(\eta_i + \alpha\delta_i)\right) + \alpha F''''\left(\sum_{j\neq i}\Big(\eta_j + \alpha\delta_j\Big) + \tau(\eta_i + \alpha\delta_i)\right)\Big(\sum_{j\neq i}\delta_j + \tau\delta_i\Big),$$

so that, applying Hölder's inequality, we obtain

$$\mathbb{E}\Big|\partial^3_{n+i,n+i,i}G(\tilde{T}_i^0)\cdot\eta_i^2\delta_i\Big| \leqslant \|F'''\|_\infty\left(\mathbb{E}\eta_i^4\right)^{1/2}\left(\mathbb{E}\delta_i^2\right)^{1/2} + \alpha\|F''''\|_\infty\mathbb{E}\Big|\eta_i^2\delta_i\Big(\sum_{j\neq i}\delta_j + \tau\delta_i\Big)\Big|$$
$$\leqslant \|F'''\|_\infty\left(\mathbb{E}\eta_i^4\right)^{1/2}\left(\mathbb{E}\delta_i^2\right)^{1/2} + \alpha\|F''''\|_\infty\left(\sqrt{\sum_{j\neq i}\mathbb{E}\delta_j^2}\,\left(\mathbb{E}\eta_i^4\right)^{1/2}\left(\mathbb{E}\delta_i^2\right)^{1/2} + \sqrt{\mathbb{E}\delta_i^4\,\mathbb{E}\eta_i^4}\right). \quad \text{(A.6)}$$

4. Finally,

$$\partial^3 G_{n+i,i,i}(\tilde{T}_i^0) = 2\alpha F''' \left(\sum_{j\neq i} \left(\eta_j + \alpha\delta_j\right) + \tau(\eta_i + \alpha\delta_i) \right) + \alpha^2 F'''' \left(\sum_{j\neq i} \left(\eta_j + \alpha\delta_j\right) + \tau(\eta_i + \alpha\delta_i) \right) \left(\sum_{j\neq i} \delta_j + \tau\delta_i \right).$$

Hölder's inequality implies that $\mathbb{E}|\eta_i\delta_i^2| = \mathbb{E}|\eta_i\delta_i\delta_i| \leqslant \left(\mathbb{E}\delta_i^2\right)^{1/2}\left(\mathbb{E}\delta_i^4\right)^{1/4}\left(\mathbb{E}\eta_i^4\right)^{1/4}$, hence

$$\begin{aligned}\left|\mathbb{E}\partial^3_{n+i,i,i} G(\tilde{T}_i^0)\cdot\eta_i\delta_i^2\right| &\leqslant 2\alpha\|F'''\|_\infty \left(\mathbb{E}\delta_i^2\right)^{1/2}\left(\mathbb{E}\delta_i^4\right)^{1/4}\left(\mathbb{E}\eta_i^4\right)^{1/4} + \alpha^2\|F''''\|_\infty \mathbb{E}\left|\eta_i\delta_i^2\Big(\sum_{j\neq i}\delta_j + \tau\delta_i\Big)\right| \\ &\leqslant 2\alpha\|F'''\|_\infty \left(\mathbb{E}\delta_i^2\right)^{1/2}\left(\mathbb{E}\delta_i^4\right)^{1/4}\left(\mathbb{E}\eta_i^4\right)^{1/4} \\ &\quad + \alpha^2\|F''''\|_\infty \left(\sqrt{\sum_{j\neq i}\mathbb{E}\delta_j^2}\left(\mathbb{E}\delta_i^2\right)^{1/2}\left(\mathbb{E}\delta_i^4\right)^{1/4}\left(\mathbb{E}\eta_i^4\right)^{1/4} + \left(\mathbb{E}\delta_i^4\right)^{3/4}\left(\mathbb{E}\eta_i^4\right)^{1/4}\right). \quad \text{(A.7)}\end{aligned}$$

Similar calculations yield an analogous bound for the terms in the expansion (A.3) of $G(T_{i-1})$. The equivalence of the moments of Gaussian random variables together with the fact that the covariance structure of (W, Z) matches that of (ξ_1, η_1) imply that the upper bounds (A.4),(A.5),(A.6),(A.7) remain valid for the terms in (A.3), up to an additional absolute multiplicative constant. Hence, combination of (A.2), (A.4),(A.5),(A.6), (A.7) and straightforward application of Hölder's inequality yields the result.

A.4. Proof of Lemma 5.9.

Define

$$D(\delta) := \sup_{\ell(f)\in\mathcal{F}(\delta)} \mathbb{E}^{1/2}\left(\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right) - \rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)^2.$$

Recall that ρ' is Lipschitz continuous and $L(\rho') = 1$, hence

$$\left(\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) - \rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)^2 \leqslant \left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\bar{\mathcal{L}}_1(f_*)-(\mathcal{L}(f)-\mathcal{L}(f_*))}{\Delta}\right)^2, \quad \text{(A.8)}$$

which implies that

$$D(\delta) \leqslant \frac{\nu(\delta)}{\Delta}. \quad \text{(A.9)}$$

Next, observe that $\widehat{G}_k(0;f) = \frac{1}{\sqrt{k}}\sum_{j\in J}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right) + \frac{1}{\sqrt{k}}\sum_{j\notin J}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)$, hence application of the triangle inequality yields that

$$\begin{aligned}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_k(0;f) - \widehat{G}_k(0;f_*)\right| &\leqslant \sup_{f\in\mathcal{F}(\delta)}|G_k(0;f) - G_k(0;f_*)| \\ &\quad + \sqrt{\frac{|J|}{k}}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(0;f) - \widehat{G}_{|J|}(0;f_*) - \mathbb{E}\left(\widehat{G}_{|J|}(0;f) - \widehat{G}_{|J|}(0;f_*)\right)\right| + 4\frac{\mathcal{O}}{\sqrt{k}}, \quad \text{(A.10)}\end{aligned}$$

where $\widehat{G}_{|J|}(0;f) := \frac{1}{\sqrt{|J|}}\sum_{j\in J}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)$. Talagrand's concentration inequality (specifically, the bound of Lemma 5.5) implies, together with the inequalities $\|\rho'\|_\infty \leqslant 2$ and $|J| > k/2$, that for any

$s > 0$

$$\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|\leqslant$$
$$2\left[\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|+D(\delta)\sqrt{\frac{s}{2}}+\frac{32\sqrt{2}s}{3\sqrt{k}}\right] \quad \text{(A.11)}$$

with probability at least $1-2e^{-s}$. According to (A.9), $D(\delta)\leqslant\frac{L(\rho')}{\Delta}\nu(\delta)$. Hence, it remains to estimate the expected supremum. Sequential application of symmetrization, contraction and desymmetrization inequalities implies that

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|\\
&\qquad\leqslant 2\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{|J|}}\sum_{j\in J}\varepsilon_j\left(\rho'\left(\sqrt{n}\,\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)-\rho'\left(\sqrt{n}\,\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right|\\
&\qquad\leqslant\frac{4L(\rho')}{\Delta}\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{\sqrt{n}}{\sqrt{|J|}}\sum_{j\in|J|}\varepsilon_j\left((\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))(X_j)-(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))(X_j)\right)\right|\\
&\qquad\leqslant\frac{8\sqrt{2}L(\rho')}{\Delta}\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N_J}\left((\ell(f)-\ell(f_*))(X_j)-P(\ell(f)-\ell(f_*))\right)\right|\leqslant\frac{8\sqrt{2}}{\Delta}\omega(\delta) \quad \text{(A.12)}
\end{aligned}$$

since $L(\rho')=1$. To estimate $\sup_{f\in\mathcal{F}(\delta)}|G_k(0;f)-G_k(0;f_*)|$, we consider 2 cases: the first case when only 2 finite moments of $\ell(f(X))$, $f\in\mathcal{F}$ exist, and the second case when 4 moments are finite. To obtain the bound in the first case, we observe that, since $\mathbb{E}\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)=0$ for any $f\in\mathcal{F}$,

$$\begin{aligned}
&\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right|\\
&\qquad=\left|\mathbb{E}T\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}T\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right|
\end{aligned}$$

where $T(x)=x-\rho'(x)$. Next, we apply Lemma 5.8 with $F=T$, $\xi_j=\frac{\ell(f(X_j)-\mathbb{E}\ell(f(X_j))}{\Delta\sqrt{n}}$ and $\eta_j=\frac{\ell(f_*(X_j)-\mathbb{E}\ell(f_*(X_j))}{\Delta\sqrt{n}}$. The first inequality of the lemma implies that

$$\begin{aligned}
&\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right|\leqslant\sqrt{\operatorname{Var}\left(\frac{\ell(f(X))-\ell(f_*(X))}{\Delta}\right)}\\
&\qquad\times\max_{\alpha\in[0,1]}\sqrt{\mathbb{E}\left(T'\left(\alpha\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}+(1-\alpha)\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)^2}.
\end{aligned}$$

Observe that $T'(x)=1-\rho''(x)\leqslant I\{|x|\geqslant 1\}$ by Assumption 1. It implies that for any $\alpha\in[0,1]$,

$$\begin{aligned}
&\mathbb{E}\left(T'\left(\alpha\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}+(1-\alpha)\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)^2\\
&\qquad\leqslant\Pr\left(\left|\alpha\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}+(1-\alpha)\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right|\geqslant 1\right)\\
&\qquad\leqslant\sup_{f\in\mathcal{F}}\operatorname{Var}\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)=\sup_{f\in\mathcal{F}}\frac{\sigma^2(\ell,f)}{\Delta^2}.
\end{aligned}$$

by Chebyshev's inequality. Hence

$$\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right|\leqslant\operatorname{Var}^{1/2}\left(\ell(f(X))-\ell(f_*(X))\right)\frac{\sigma(\ell,\mathcal{F})}{\Delta^2}.$$

and, taking supremum over $f \in \mathcal{F}(\delta)$ and recalling that $\Delta = M_\Delta \cdot \sigma(\ell, \mathcal{F})$ for $M_\Delta \geqslant 1$, we obtain the inequality

$$\sup_{f\in\mathcal{F}(\delta)} |G_k(0;f) - G_k(0;f_*)| \leqslant \sqrt{k}\frac{\nu(\delta)}{\Delta}\frac{1}{M_\Delta} \leqslant \sqrt{k}\,\widetilde{B}(\delta). \tag{A.13}$$

On the other hand, under the assumption of existence of 4 moments, we get that

$$\begin{aligned}\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) - \mathbb{E}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right| \\ \leqslant \frac{C(\rho)}{\sqrt{n}\Delta}\left(\mathrm{Var}^{1/2}(\ell(f(X))-\ell(f_*(X)))\left(\frac{R_4^2(\ell,\mathcal{F})}{\Delta^2}+\frac{R_4^3(\ell,\mathcal{F})}{\Delta^3}\right)\right. \\ \left. + \frac{\mathbb{E}^{1/4}(\ell(f(X))-\ell(f_*(X)))^4}{\sqrt{n}}\frac{R_4^3(\ell,\mathcal{F})}{\Delta^3}\right),\end{aligned}$$

Again, taking supremum over $f \in \mathcal{F}(\delta)$ and recalling that $\Delta = M_\Delta \cdot \sigma(\ell, \mathcal{F})$ for $M_\Delta \geqslant 1$, we deduce that

$$\begin{aligned}\sup_{f\in\mathcal{F}(\delta)} |G_k(0;f) - G_k(0;f_*)| &\leqslant C(\rho)\sqrt{\frac{k}{n}}\left(\frac{\nu(\delta)}{\Delta}\left(\frac{\mathfrak{B}^3(\ell,\mathcal{F})}{M_\Delta^3}\vee\frac{\mathfrak{B}^2(\ell,\mathcal{F})}{M_\Delta^2}\right)+\frac{\nu_4(\delta)}{\Delta}\frac{\mathfrak{B}^3(\ell,\mathcal{F})}{M_\Delta^3\sqrt{n}}\right) \\ &\leqslant C(\rho)\sqrt{\frac{k}{n}}\mathfrak{B}^3(\ell,\mathcal{F})\left(\frac{\nu(\delta)}{\Delta}\frac{1}{M_\Delta^2}+\frac{\nu_4(\delta)}{\Delta}\frac{1}{M_\Delta^3\sqrt{n}}\right) \leqslant C(\rho)\sqrt{k}\,\widetilde{B}(\delta),\end{aligned} \tag{A.14}$$

implying the result.

A.5. Proof of Lemma 5.10.

Recall that $\widehat{G}_{|J|}(0;f) := \frac{1}{\sqrt{|J|}}\sum_{j\in J}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)$. Given $\delta \geqslant \delta_{\min}$, define

$$\begin{aligned}\widehat{Q}_{|J|}(\delta) &:= \sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right|, \\ \widehat{T}_{|J|}(\delta_{\min}) &:= \sup_{\delta\geqslant\delta_{\min}}\widehat{Q}_{|J|}(\delta).\end{aligned}$$

Observe that for any $\delta \geqslant \delta_{\min}$,

$$\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right| \leqslant \frac{\delta}{\delta_{\min}}\widehat{T}_{|J|}(\delta_{\min}). \tag{A.15}$$

Hence, our goal will be to find an upper bound for $\widehat{T}_{|J|}(\delta_{\min})$. To this end, note that

$$\begin{aligned}\widehat{T}_{|J|}(\delta_{\min}) \leqslant \sup_{\delta\geqslant\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right| \\ + \sup_{\delta\geqslant\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\left|G_k(0;f)-G_k(0;f_*)\right|.\end{aligned} \tag{A.16}$$

It remains to estimate both terms in the inequality above. Inequality (A.8) implies the bound

$$\begin{aligned}\sup_{\delta\geqslant\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\mathrm{Var}^{1/2}\left(\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right) \\ \leqslant \frac{L(\rho')}{\Delta}\sup_{\delta\geqslant\delta_{\min}}\frac{\delta_{\min}}{\delta}\nu(\delta) \leqslant \frac{L(\rho')}{\Delta}\sup_{\delta\geqslant\delta_{\min}}\frac{\delta_{\min}}{\delta}\widetilde{\nu}(\delta) \leqslant \frac{1}{\Delta}\widetilde{\nu}(\delta_{\min})\end{aligned}$$

since $\widetilde{\nu}$ is a function of concave type. Moreover, it is clear that for any $\delta \geqslant \delta_{\min}$,

$$\left|\frac{\delta_{\min}}{\delta}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\frac{\delta_{\min}}{\delta}\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right| \leqslant 2\left\|\rho'\right\|_\infty \leqslant 4$$

almost surely. Now, Talagrand's concentration inequality implies that for any $s > 0$,

$$\sup_{\delta\geqslant\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|$$
$$\leqslant 2\Big[\mathbb{E}\sup_{\delta\geqslant\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|$$
$$+\frac{L(\rho')}{\Delta}\widetilde{\nu}(\delta_{\min})\sqrt{\frac{s}{2}}+\frac{32\sqrt{2}s}{3\sqrt{k}}\Big] \quad (A.17)$$

with probability at least $1-e^{-s}$. To estimate the expectation, we proceed as follows: for $j\in\mathbb{Z}$, set $\delta_j := 2^{-j}$, and observe that

$$\mathbb{E}\sup_{\delta\geqslant\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|$$
$$\leqslant \mathbb{E}\sup_{j:\delta_j\geqslant\delta_{\min}}\sup_{\delta\in(\delta_{j+1},\delta_j]}\frac{\delta_{\min}}{\delta}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|$$
$$\leqslant \sum_{j:\delta_j\geqslant\delta_{\min}}\frac{\delta_{\min}}{\delta_{j+1}}\mathbb{E}\sup_{\delta\in(\delta_{j+1},\delta_j]}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|$$
$$\leqslant 2\sum_{j:\delta_j\geqslant\delta_{\min}}\frac{\delta_{\min}}{\delta_j}\mathbb{E}\sup_{f\in\mathcal{F}(\delta_j)}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|,$$

where the last inequality relied on the fact that $\mathcal{F}(\delta)\subseteq\mathcal{F}(\delta')$ for $\delta\leqslant\delta'$. It follows from (A.12) that

$$\mathbb{E}\sup_{f\in\mathcal{F}(\delta_j)}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|\leqslant\frac{8\sqrt{2}L(\rho')}{\Delta}\omega(\delta_j)\leqslant\frac{8\sqrt{2}}{\Delta}\widetilde{\omega}(\delta_j),$$

where $\widetilde{\omega}(\cdot)$ is an upper bound on $\omega(\cdot)$ of strictly concave type (with exponent γ for some $\gamma\in(0,1)$). Hence, applying Proposition 4.2 in [24], we deduce that

$$\mathbb{E}\sup_{\delta\geqslant\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|$$
$$\leqslant\frac{16}{\Delta}\delta_{\min}\sum_{j:\delta_j\geqslant\delta_{\min}}\frac{\widetilde{\omega}(\delta_j)}{\delta_j}\leqslant\frac{c(\gamma)}{\Delta}\delta_{\min}\frac{\widetilde{\omega}(\delta_{\min})}{\delta_{\min}}=\frac{c(\gamma)}{\Delta}\widetilde{\omega}(\delta_{\min}),$$

and (A.17) yields the inequality

$$\sup_{\delta\geqslant\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\left|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\right)\right|\leqslant\widetilde{U}(\delta_{\min},s), \quad (A.18)$$

where $\widetilde{U}(\delta,s)$ was defined in (5.17). For the second term in (A.16), inequality (A.14) implies that

$$\sup_{\delta\geqslant\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\left|G_k(0;f)-G_k(0;f_*)\right|$$
$$\leqslant C(\rho)\delta_{\min}\sqrt{\frac{k}{n}}\bar{R}^3(\ell,\mathcal{F},\Delta)\sup_{\delta\geqslant\delta_{\min}}\left(\frac{\nu(\delta)}{\delta\Delta}\frac{1}{M_\Delta^2}+\frac{\nu_4(\delta)}{\delta\,\Delta}\frac{1}{M_\Delta^3\sqrt{n}}\right)$$
$$\leqslant C(\rho)\sqrt{k}\,\mathfrak{B}^3(\ell,\mathcal{F})\left(\frac{\widetilde{\nu}(\delta_{\min})}{\Delta}\frac{1}{M_\Delta^2}+\frac{\widetilde{\nu}_4(\delta_{\min})}{\Delta}\frac{1}{M_\Delta^3\sqrt{n}}\right)$$

since $\nu(\delta)\leqslant\widetilde{\nu}(\delta)$, $\nu_4(\delta)\leqslant\widetilde{\nu}_4(\delta)$ and $\widetilde{\nu}(\delta)$, $\widetilde{\nu}_4(\delta)$ are functions of concave type. Combining the bound above with (A.18), we deduce that

$$\widehat{T}_{|J|}(\delta_{\min})\leqslant\widetilde{U}(\delta_{\min},s)+C(\rho)\sqrt{k}\widetilde{B}(\delta_{\min}),$$

hence (A.10) and (A.15) imply that for all $\delta \geqslant \delta_{\min}$ simultaneously,

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

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

A.6. Proof of Lemma 5.11.

The following identity is immediate:

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

Assumptions on ρ imply that for any $f\in\mathcal{F}$ and $j=1,\dots,k$, there exists $\tau_j\in[0,1]$ such that

$$\rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)-\widehat{e}^{(k)}(f)}{\Delta}\right) = \rho'\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right) - \frac{\sqrt{n}}{\Delta}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\cdot\widehat{e}^{(k)}(f) + \frac{n}{\Delta^2}\rho'''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)-\tau_j\widehat{e}^{(k)}(f)}{\Delta}\right)\cdot\left(\widehat{e}^{(k)}(f)\right)^2, \quad (A.19)$$

hence

$$\widehat{G}_k\left(\widehat{e}^{(k)}(f);f\right) - \widehat{G}_k\left(0;f\right) = -\frac{\sqrt{n}}{\Delta}\frac{\widehat{e}^{(k)}(f)}{\sqrt{k}}\sum_{j=1}^k\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right) + \frac{n}{\Delta^2}\frac{\left(\widehat{e}^{(k)}(f)\right)^2}{\sqrt{k}}\sum_{j=1}^k\rho'''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)-\tau_j\widehat{e}^{(k)}(f)}{\Delta}\right),$$

and

$$R_N(f) = \frac{\sqrt{n}}{\Delta}\frac{\widehat{e}^{(k)}(f)}{\sqrt{k}}\sum_{j=1}^k\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right) - \mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right) - \frac{n}{\Delta^2}\frac{\left(\widehat{e}^{(k)}(f)\right)^2}{\sqrt{k}}\sum_{j=1}^k\rho'''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)-\tau_j\widehat{e}^{(k)}(f)}{\Delta}\right). \quad (A.20)$$

We will need the following modification of Theorem 3.1 that is stated below and proved in Section A.7.

Lemma A.1. *Then there exist positive constants $c(\rho)$, $C(\rho)$ with the following properties. Fix $\delta_{\min}>0$. Then for all $s>0$, $\delta\geqslant\delta_{\min}$, positive integers n and k such that*

$$\delta\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}\sqrt{k}} + \sup_{f\in\mathcal{F}}G_f(n,\Delta) + \frac{s+\mathcal{O}}{k} \leqslant c(\rho), \quad (A.21)$$

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

$$\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{e}^{(k)}(f)\right| \leqslant C(\rho)\widetilde{\Delta}\left[\frac{\delta}{\sqrt{N}}\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}} + \frac{\sigma(\ell,f_*)}{\Delta}\sqrt{\frac{s}{N}} + \frac{\sup_{f\in\mathcal{F}}G_f(n,\Delta)}{\sqrt{n}} + \frac{(s+\mathcal{O})\sqrt{n}}{N}\right]. \quad (A.22)$$

In the rest of the proof, we will assume that conditions of Lemma A.1 and Theorem 3.1 hold, and let Θ' be an event of probability at least $1-4e^{-s}$ on which inequalities (A.22) and (3.4) are valid. On event

Θ', the last term in (A.20) can thus be estimated as

$$\sup_{f\in\mathcal{F}(\delta)}\left|\frac{n}{\Delta^2}\frac{\big(\widehat{e}^{(k)}(f)\big)^2}{\sqrt{k}}\sum_{j=1}^k\rho'''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)-\tau_j\widehat{e}^{(k)}(f)}{\Delta}\right)\right|\leqslant C_1(\rho)\frac{\sqrt{nN}}{\Delta^2}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{e}^{(k)}(f)\right|^2$$
$$\leqslant C_2(\rho)\sqrt{N}\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{n^{1/2}\delta^2}{N}\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}\right)^2\bigvee\frac{\sigma^2(\ell,f_*)}{\Delta^2}\frac{n^{1/2}\,s}{N}\right.$$
$$\left.\bigvee n^{1/2}\left(\sup_{f\in\mathcal{F}}\frac{G_f(n,\Delta)}{\sqrt{n}}\right)^2\bigvee n^{3/2}\frac{s^2+\mathcal{O}^2}{N^2}\right). \quad \text{(A.23)}$$

where we used the fact that $\|\rho'''\|_\infty<\infty$. It remains to estimate the first term in (A.20). The required bound will follow from the combination of Theorem A.1 and the following lemma that is proved in Section A.8.

Lemma A.2. *Fix $\delta_{\min}>0$. With probability at least $1-3e^{-s}$, for all $\delta\geqslant\delta_{\min}$ simultaneously,*

$$\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{k}}\sum_{j=1}^k\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)\right|$$
$$\leqslant C(\rho)\left(\delta\,\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}+\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{s}+\frac{s+\mathcal{O}}{\sqrt{k}}\right).$$

Let Θ'' be the event of probability at least $1-3e^{-2s}$ on which the inequality of Lemma A.2 holds. Then simple algebra yields that on event $\Theta'\cap\Theta''$ of probability at least $1-7e^{-s}$,

$$\sup_{f\in\mathcal{F}(\delta)}\left|\frac{\sqrt{n}}{\Delta}\frac{\widehat{e}^{(k)}(f)}{\sqrt{k}}\sum_{j=1}^k\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)\right|$$
$$\leqslant C_3(\rho)\sqrt{N}\frac{\widetilde{\Delta}}{\Delta}\left(\frac{n^{1/2}\delta^2}{N}\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}\right)^2\bigvee\frac{\sigma^2(\ell,f_*)}{\Delta^2}\frac{n^{1/2}\,s}{N}\right.$$
$$\left.\bigvee n^{1/2}\left(\sup_{f\in\mathcal{F}}\frac{G_f(n,\Delta)}{\sqrt{n}}\right)^2\bigvee n^{3/2}\frac{s^2+\mathcal{O}^2}{N^2}\right). \quad \text{(A.24)}$$

Combination of inequalities (A.23) and (A.24) that hold with probability at leat $1-7e^{-s}$ yields the result.

A.7. Proof of Lemma A.1.

In the situation when δ is fixed, the argument mimics the proof of Theorem 4.1 in [35], with minor modifications outlined below. Recall that

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

Let z_1,z_2 be such that on an event of probability close to 1, $\widehat{G}_k(z_1;f)>0$ and $\widehat{G}_k(z_2;f)<0$ for all $f\in\mathcal{F}(\delta)$ simultaneously. Since $\widehat{G}_k$ is decreasing in z, it is easy to see that $\widehat{e}^{(k)}(f)\in(z_1,z_2)$ for all $f\in\mathcal{F}(\delta)$ on this event. Hence, our goal is to find z_1,z_2 satisfying conditions above and such that $|z_1|$, $|z_2|$ are as small as possible. Observe that

$$\widehat{G}_k(z;f)=\frac{1}{\sqrt{k}}\sum_{j\in J}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)+\frac{1}{\sqrt{k}}\sum_{j\notin J}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)$$

and $\left|\frac{1}{\sqrt{k}}\sum_{j\notin J}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)\right| \leqslant 2\frac{\mathcal{O}}{\sqrt{k}}$. Moreover,

$$\begin{aligned}
&\frac{1}{\sqrt{k}}\sum_{j\in J}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)\\
&\qquad=\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)-\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)\right.\\
&\qquad\quad\left.-\mathbb{E}\left[\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)-\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)\right]\right)\\
&\qquad+\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)\right)\\
&\qquad\quad+\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)-\mathbb{E}\rho'\left(\frac{W(\ell(f))-\sqrt{n}z}{\Delta}\right)\right)\\
&\qquad\qquad+\frac{1}{\sqrt{k}}\sum_{j\in J}\mathbb{E}\rho'\left(\frac{W(\ell(f))-\sqrt{n}z}{\Delta}\right).
\end{aligned}$$

We will proceed in 4 steps: first, we will find $\varepsilon_1>0$ such that for any $z\in\mathbb{R}$ and all $f\in\mathcal{F}(\delta)$,

$$\begin{aligned}
&\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)-\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)\right.\\
&\qquad\left.-\mathbb{E}\left[\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)-\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)\right]\right)\leqslant\varepsilon_1
\end{aligned}$$

with high probability, then $\varepsilon_2>0$ such that

$$\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)\right)\leqslant\varepsilon_2,$$

ε_3 satisfying

$$\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)-\mathbb{E}\rho'\left(\frac{W(\ell(f))-\sqrt{n}z}{\Delta}\right)\right)\right|\leqslant\varepsilon_3,$$

and finally we will choose $z_1<0$ such that for all $f\in\mathcal{F}(\delta)$,

$$\frac{1}{\sqrt{k}}\sum_{j\in J}\mathbb{E}\rho'\left(\frac{W(\ell(f))-\sqrt{n}z}{\Delta}\right)>\varepsilon_1+\varepsilon_2+\varepsilon_3+2\frac{\mathcal{O}}{\sqrt{k}}. \tag{A.25}$$

Talagrand's concentration inequality [44, Corollary 16.1], together with the bound $\|\rho'\|_\infty\leqslant 2$, implies that for any $s>0$,

$$\begin{aligned}
&\sqrt{\frac{|J|}{k}}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)\right)\right|\leqslant\\
&\qquad 2\left[\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)\right)\right|+D(\delta)\sqrt{\frac{s}{2}}+\frac{32}{3}\frac{s}{\sqrt{k}}\right]
\end{aligned}$$

with probability at least $1-2e^{-s}$. It has been observed in (A.9) that $D(\delta)\leqslant\frac{\nu(\delta)}{\Delta}$. It remains to estimate the expected supremum. Sequential application of symmetrization, contraction and desymmetrization

inequalities, together with the fact that $L(\rho') = 1$, implies that

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)\right)\right| \\
&\quad\leqslant 2\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{|J|}}\sum_{j\in J}\varepsilon_j\left(\rho'\left(\sqrt{n}\,\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)-z}{\Delta}\right)\right)-\rho'\left(\sqrt{n}\,\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)-z}{\Delta}\right)\right| \\
&\quad\leqslant \frac{4}{\Delta}\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{\sqrt{n}}{\sqrt{|J|}}\sum_{j\in J}\varepsilon_j\left((\bar{\mathcal{L}}_j(f)-\mathcal{L}(f))-(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))\right)\right| \\
&\quad\leqslant \frac{8\sqrt{2}}{\Delta}\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N_J}\Big((\ell(f)-\ell(f_*))(X_j)-P(\ell(f)-\ell(f_*))\Big)\right| \leqslant \frac{8\sqrt{2}}{\Delta}\,\omega(\delta).
\end{aligned}$$

Hence, it suffices to choose

$$\varepsilon_1 = \frac{8\sqrt{2}}{\Delta}\,\omega(\delta)+\frac{\nu(\delta)}{\Delta}\sqrt{s}+\frac{32}{3}\frac{s}{\sqrt{k}}.$$

Next, Bernstein's inequality and Lemma 5.1 together yield that with probability at least $1-2e^{-s}$,

$$\begin{aligned}
&\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)\right) \\
&\qquad\qquad\leqslant 2\left(\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{s}+\frac{3s}{\sqrt{k}}\right),
\end{aligned}$$

thus we can set $\varepsilon_2 = 2\left(\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{s}+3\frac{s}{\sqrt{k}}\right)$. Lemma 5.3 impliies that ε_3 can be chosen as

$$\varepsilon_3 = \sqrt{k}\sup_{f\in\mathcal{F}(\delta)}G_f(n,\Delta).$$

Finally, we apply Lemma 6.3 of [35] with

$$\varepsilon := \varepsilon_1+\varepsilon_2+\varepsilon_3+2\frac{\mathcal{O}}{\sqrt{k}}$$

to deduce that

$$z_1 = -C\frac{\widetilde{\Delta}}{\sqrt{N}}\cdot\left(\varepsilon_1+\varepsilon_2+\varepsilon_3+2\frac{\mathcal{O}}{\sqrt{k}}\right),$$

satisfies (A.25) under assumption that $\frac{\varepsilon_1+\varepsilon_2+\varepsilon_3}{\sqrt{k}}+\frac{\mathcal{O}}{k}\leqslant c$ for some absolute constants $c, C>0$. Proceeding in a similar way, it is easy to see that setting $z_2 = -z_1$ guarantees that $\widehat{G}_k(z_2;f)<0$ for all $f\in\mathcal{F}(\delta)$ with probability at least $1-e^{-s}$, hence the claim follows.

It remains to make the bound uniform in $\delta\geqslant\delta_{\min}$. To this end, we need to repeat the "slicing argument" of Lemma 5.10 below (specifically, see equation (A.18)) to deduce that with probability at least $1-2e^{-s}$,

$$\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)-\mathbb{E}\left(\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)\right)\right|\leqslant\delta\,\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}$$

uniformly for all $\delta\geqslant\delta_{\min}$, hence the value of ε_1 should be replaced by $\varepsilon_1 = \delta\,\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}$.

A.8. Proof of Lemma A.2.

Observe that

$$\begin{aligned}
&\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)\\
&\qquad=\frac{1}{\sqrt{k}}\sum_{j\notin J}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)\\
&\qquad+\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right.\\
&\qquad\left.-\mathbb{E}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)\right)\\
&\qquad+\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)-\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right).
\end{aligned}$$

Clearly, as $\|\rho''\|_\infty\leqslant 1$, $\left|\frac{1}{\sqrt{k}}\sum_{j\notin J}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)\right|\leqslant 2\frac{\mathcal{O}}{\sqrt{k}}$. Next, repeating the "slicing argument" of Lemma 5.10, it is not difficult to deduce that with probability at least $1-2e^{-2s}$,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right.\right.\\
&\qquad\left.\left.-\mathbb{E}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)\right)\right|\leqslant C(\rho)\,\delta\,\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}
\end{aligned}$$

uniformly for all $\delta\geqslant\delta_{\min}$. Next, we will apply Bernstein's inequality to estimate the remaining term. Since ρ is convex, ρ'' is nonnegative, moreover, it follows from Assumption 1 that $\rho''(x)\neq 0$ for $|x|\leqslant 2$, $\rho''(x)=1$ for $|x|\leqslant 1$, and $\|\rho''\|_\infty=1$, hence $\left(\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)^2\geqslant\left(\Pr\left(\left|\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\leqslant 1\right)\right)^2$,

$$\mathbb{E}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)^2\leqslant\Pr\left(\left|\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\leqslant 1\right)+\Pr\left(\left|\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\in[1,2]\right),$$

and

$$\begin{aligned}
\mathrm{Var}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)&\leqslant\Pr\left(\left|\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\leqslant 1\right)-\left(\Pr\left(\left|\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\leqslant 1\right)\right)^2\\
&\quad+\Pr\left(\left|\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\geqslant 1\right)\\
&\leqslant 2\Pr\left(\left|\sqrt{n}\frac{\bar{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\geqslant 1\right)\leqslant 2\frac{\mathrm{Var}\left(\ell(f(X))\right)}{\Delta^2}.
\end{aligned}$$

Bernstein's inequality implies that with probability at least $1-e^{-s}$,

$$\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)-\mathbb{E}\rho''\left(\sqrt{n}\frac{\bar{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)\leqslant 2\left(\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{s}+\frac{s}{\sqrt{k}}\right),$$

hence the desired conclusion follows.

A.9. Proof of Lemma 4.1.

In the context of regression with quadratic loss, $\omega(\delta)$ takes the form

$$\omega(\delta)=\mathbb{E}\sup_{\ell(f)\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left((Y_j-f(Z_j))^2-(Y_j-f_*(Z_j))^2-\mathbb{E}\left((Y_j-f(Z_j))^2-(Y_j-f_*(Z_j))^2\right)\right)\right|.$$

In view of Bernstein's assumption verified above, $\omega(\delta)$ is bounded by

$$\mathbb{E}\sup_{\|f-f_*\|^2_{L_2(\Pi)}\leqslant 2\delta}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N\Big((Y_j-f(Z_j))^2-(Y_j-f_*(Z_j))^2-\mathbb{E}\left((Y_j-f(Z_j))^2-(Y_j-f_*(Z_j))^2\right)\Big)\right|.$$

To estimate the latter quantity, we will use the approach based on the $L_\infty(\Pi_n)$-covering numbers of the class $\mathcal{F}$ (e.g., see [9]). We will also set

$$B(\mathcal{F};\tau):=\{f\in\mathcal{F}:\|f-f_*\|^2_{L_2(\Pi)}\leqslant\tau\}.$$

It is easy to see that

$$(Y-f(X))^2-(Y-f_*(X))^2=(f(X)-f_*(X))^2+2(f(X)-f_*(X))(f_*(X)-Y),$$

hence

$$w(\delta)\leqslant\mathbb{E}\sup_{B(\mathcal{F};2\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N(f(Z_j)-f_*(Z_j))^2-\mathbb{E}(f(Z_j)-f_*(Z_j))^2\right|$$
$$+2\,\mathbb{E}\sup_{B(\mathcal{F};2\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N(f(Z_j)-f_*(Z_j))(Y_j-f_*(Z_j)\right|. \quad \text{(A.26)}$$

We will estimate the two terms separately. By assumption, the covering numbers of the class $\mathcal{F}$ satisfy the bound

$$N\left(\mathcal{F},L_2(\Pi_N),\varepsilon\right)\leqslant\left(\frac{A\|F\|_{L_2(\Pi_N)}}{\varepsilon}\right)^V\vee 1 \quad \text{(A.27)}$$

for some constants $A\geqslant 1$, $V\geqslant 1$ and all $\varepsilon>0$. We apply bound of Lemma 5.4 to the first term in (A.26) to get that

$$\mathbb{E}\sup_{B(\mathcal{F};2\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N(f(Z_j)-f_*(Z_j))^2-\mathbb{E}(f(Z_j)-f_*(Z_j))^2\right|$$
$$\leqslant C\left(\sqrt{2\delta}\sqrt{\Gamma_{N,\infty}(B(\mathcal{F};2\delta))}\bigvee\frac{\Gamma_{N,\infty}(B(\mathcal{F};2\delta))}{\sqrt{N}}\right).$$

To estimate $\Gamma_{n,\infty}(B(\mathcal{F};2\delta)):=\mathbb{E}\gamma_2^2(B(\mathcal{F};2\delta);L_\infty(\Pi_N))$, we will use Dudley's entropy integral bound. Observe that

$$\operatorname{diam}\left(B(\mathcal{F};2\delta);L_\infty(\Pi_N)\right)\leqslant 2\|F\|_{L_\infty(\Pi_N)}.$$

Moreover, for any $f,g\in\mathcal{F}$,

$$\frac{1}{N}\sum_{j=1}^N(f(Z_j)-g(Z_j))^2\geqslant\frac{1}{N}\max_{1\leqslant j\leqslant N}(f(Z_j)-g(Z_j))^2,$$

hence $N(B(\mathcal{F};2\delta),L_\infty(\Pi_N),\varepsilon)\leqslant N\left(B(\mathcal{F};2\delta),L_2(\Pi_N),\frac{\varepsilon}{\sqrt{N}}\right)$ and, whenever (A.27) holds,

$$\log N(B(\mathcal{F};2\delta),L_\infty(\Pi_N),\varepsilon)\leqslant V\log_+\left(\frac{A\sqrt{N}\|F\|_{L_2(\Pi_N)}}{\varepsilon}\right),$$

where $\log_+(x):=\max(\log x,0)$. It yields that

$$\Gamma_{N,\infty}(B(\mathcal{F};2\delta))\leqslant\mathbb{E}\left(\sqrt{V}\int_0^{2\|F\|_{L_\infty(\Pi_N)}}\log_+^{1/2}\left(\frac{A\|F\|_{L_2(\Pi_N)}\sqrt{N}}{\varepsilon}\right)d\varepsilon\right)^2$$
$$\leqslant C\,V\mathbb{E}\left(\|F\|^2_{L_\infty(\Pi_N)}\log\left(\frac{A\sqrt{N}\|F\|_{L_2(\Pi_N)}}{\|F\|_{L_\infty(\Pi_N)}}\vee e\right)\right)\leqslant C\,V\log(A\sqrt{N})\mathbb{E}\,\|F\|^2_{L_\infty(\Pi_N)}$$

for an absolute constant $C > 0$. Finally, since $\|F\|_{\psi_2} < \infty$,

$$\mathbb{E}\left\|F^2\right\|_{L_\infty(\Pi_N)} \leqslant C_1 \log(N)\|F^2\|_{\psi_1} = C_1 \log(N)\|F\|_{\psi_2}^2,$$

hence

$$\mathbb{E}\sup_{B(\mathcal{F};2\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N (f(Z_j)-f_*(Z_j))^2 - \mathbb{E}(f(Z_j)-f_*(Z_j))^2\right| \\ \leqslant C_2\left(\sqrt{\delta}\sqrt{V}\log(A^2N)\|F\|_{\psi_2} \bigvee \frac{V\|F\|_{\psi_2}^2\log^2(A^2N)}{\sqrt{N}}\right). \quad \text{(A.28)}$$

Next, the multiplier inequality [46] implies that

$$\mathbb{E}\sup_{B(\mathcal{F};2\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N (f(Z_j)-f_*(Z_j))(Y_j-f_*(Z_j)\right| \\ \leqslant C\|\eta\|_{2,1}\max_{k=1,\dots,N}\mathbb{E}\sup_{B(\mathcal{F};2\delta)}\left|\frac{1}{\sqrt{k}}\sum_{j=1}^k (f(Z_j)-f_*(Z_j))\right|.$$

Using symmetrization inequality and applying Dudley's entropy integral bound, we deduce that for any k

$$\mathbb{E}\sup_{B(\mathcal{F};2\delta)}\left|\frac{1}{\sqrt{k}}\sum_{j=1}^k (f(Z_j)-f_*(Z_j))\right| \leqslant C\sqrt{V}\mathbb{E}\int_0^{\sigma_k}\log^{1/2}\left(\frac{A\|F_{2\delta}\|_{L_2(\Pi_k)}}{\varepsilon}\right)d\varepsilon \\ \leqslant C_1\sqrt{V}\mathbb{E}\left(\sigma_k\log^{1/2}\left(\frac{eA\|F_{2\delta}\|_{L_2(\Pi_k)}}{\sigma_k}\right)\right),$$

where $F_{2\delta}$ is the envelope of the class $B(\mathcal{F};2\delta)$ and $\sigma_k^2 := \sup_{f\in B(\mathcal{F};2\delta)}\|f-f_*\|_{L_2(\Pi_k)}^2$. Cauchy-Schwarz inequality, together with an elementary observation that $k\sigma_k^2 \geqslant \|F_{2\delta}\|_{L_2(\Pi_k)}^2$, gives

$$\mathbb{E}\left(\sigma_k\log^{1/2}\left(\frac{eA\|F_{2\delta}\|_{L_2(\Pi_k)}}{\sigma_k}\right)\right) \leqslant \sqrt{\mathbb{E}\sigma_k^2}\,\log^{1/2}(eA\sqrt{k}).$$

According to (A.28),

$$\mathbb{E}\sigma_k^2 \leqslant 2\delta + C_2\left(\sqrt{\delta}\sqrt{\frac{V}{N}}\log(A^2N)\|F\|_{\psi_2} \bigvee \frac{V\|F\|_{\psi_2}^2\log^2(A^2N)}{N}\right).$$

Simple algebra now yields that

$$\mathbb{E}\sup_{B(\mathcal{F};2\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N (f(Z_j)-f_*(Z_j))(Y_j-f_*(Z_j)\right| \\ \leqslant C\|\eta\|_{2,1}\sqrt{V\log(e^2A^2N)}\left(\sqrt{\delta}+\sqrt{\frac{V}{N}}\log(A^2N)\|F\|_{\psi_2}\right). \quad \text{(A.29)}$$

Finally, combination of inequalities (A.28) and (A.29) implies that

$$w(\delta) \leqslant \widetilde{\omega}(\delta) := C\left(\sqrt{\delta}\sqrt{V}\log(A^2N)(\|F\|_{\psi_2}+\|\eta\|_{2,1}) \bigvee \frac{V(\|F\|_{\psi_2}^2+\|\eta\|_{2,1}^2)\log^2(A^2N)}{\sqrt{N}}\right),$$

where $\widetilde{\omega}(\delta)$ is of strictly concave type, hence

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

thus proving the claim.

Appendix B: Further numerical study.

We present additional sresults of numerical experiments omitted in the main text.

B.1. Application to the "Communities and Crime" data.

We compare performance of our methods with the ordinary least squares regression applied to a real dataset. The dataset we chose is called "Communities and Crime Unnormalized Data Set" and is available through the UCI Machine Learning Repository. These data contain 2215 observations from a census and law enforcement records. The task we devised was to predict the crime activity (represented as the count of incidents) using the following features: the population of the area, the per capita income, the median family income, the number of vacant houses, and the land area. The choice of this specific dataset was motivated by the fact that it likely contains a non-negligible number of outliers due to the nature of the features and the fact that the data have not been preprocessed, hence the advantages of proposed approach could be highlighted. Figure 9 presents a pairplot of the dataset; specifically, a pairplot shows all the different scatter plots of one feature versus another (hence, the diagonal consists of the histograms of an individual feature). Such a pairplot offers a visual confirmation of the fact that the data likely contains outliers.

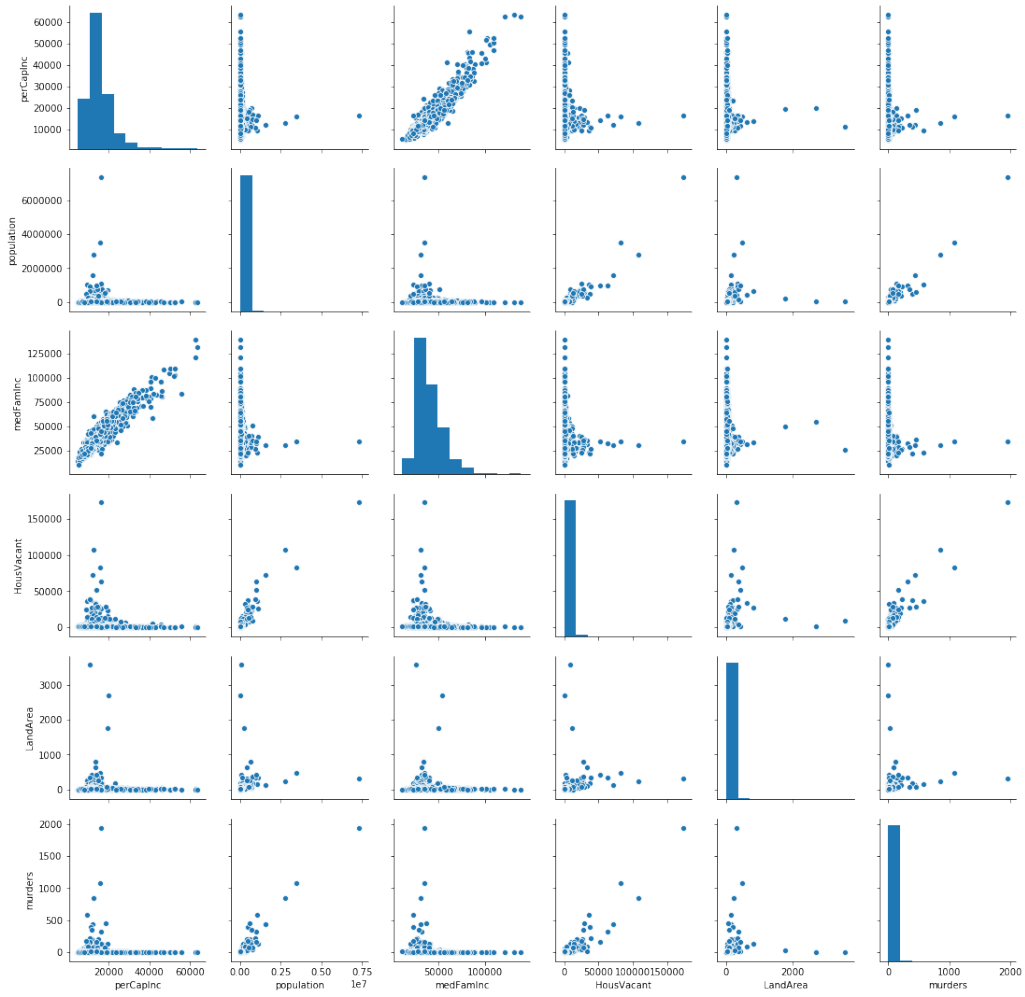


Fig 9: Pairplot detailling the 2D marginals of the dataset.

We studied the dependency of the MSE with k. Similarly to Figure 6a, we plotted the MSE as a function of k (figure 10).

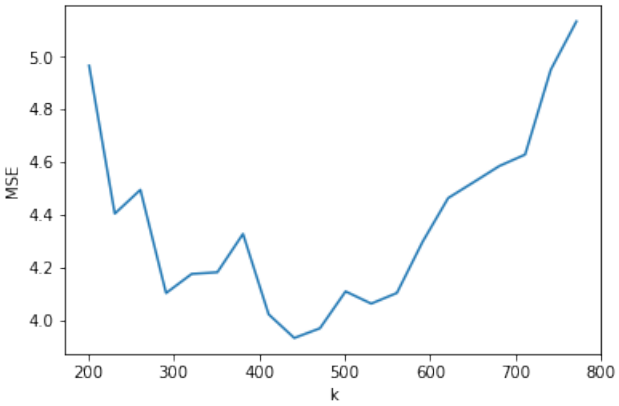


Fig 10: Plot of the number of blocks k (x-axis) vs the test mean squared error (y-axis) obtained with Algorithm 3 on 500 folds.

B.1.1. A remark on cross-validation in a corrupted setting.

Cross-validation is a common way to assess the performance of a machine learning algorithm. However, cross-validation is not robust when the method itself is not robust (as it is the case here with regression with quadratic loss). For our purposes, we slightly changed the way we approach cross validation. Namely, we still partition the data into m parts used separately for training and testing, however, once we obtain the m scores associated with the m folds, we evaluate the median of these scores instead of the mean. The rationale behind this approach is that if at least half of the folds do not contain outliers, the results of cross-validation will be robust. To use this approach, we choose m, the number of folds, to be large (in the example above, $m = 500$).

Fig 11: Robust cross-validation with the median.

Input: the dataset $(X_i, Y_i)_{1 \leqslant i \leqslant N}$.
Construct the blocks $G_1, \ldots, G_m$, partition of $\{1, \ldots, N\}$.
for all $j = 1, \ldots, m$ **do**
  Train $\hat{f}$ on the dataset $(X_l, Y_l), l \in \bigcup_{i \neq j} G_i$.
  Compute the test MSE $\text{Score}_j = \frac{1}{|G_j|} \sum_{l \in G_j} (\hat{f}(X_l) - Y_l)^2$
end for
Output: Median $(\text{Score}_1, \ldots, \text{Score}_m)$.

We compared the three algorithms using robust cross-validation with median described above Our method (based on Algorithm 3) yields MSE of $\simeq e^{4.2}$ while the MSE for the ordinary least squares regression is of order $e^{22.1}$, while the Huber Regression leads to MSE $\simeq e^{8.9}$. The empirical density of the logarithm of the MSE over 500 folds is shown in Figure 12.

B.2. Comparison of Algorithm 1 and Algorithm 3.

We present a numerical evidence that the permutation-invariant estimator $\widehat{f}_N^U$ is superior to the the estimator $\widehat{f}_N$ based on fixed partition of the dataset. Evaluation was performed for the regression task where the data contained outliers of type (a), as described in Section 2.3. Average MSE was evaluated over 500 repetitions of the experiment, and the standard deviation of the MSE was also recored. Results are presented in Figure 13 and confirm the significant improvements achieved by Algorithm 3 over Algorithm 1. We set $k = 71$ and $\Delta = 1$ for both algorithms.

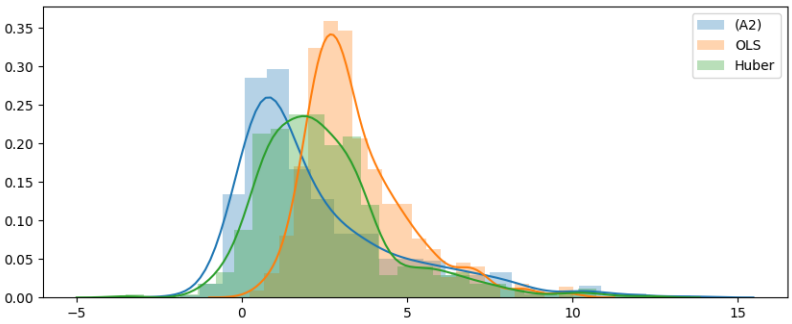


Fig 12: Histogram of densities of the logarithm of the MSE for the different methods (light blue corresponds to the approach of this paper (Algorithm 3), orange - to the standard least squares regression, and green - to Huber's regression).

	Algorithm 1	Algorithm 3
average MSE	97.8	2
standard deviation of MSE	577.3	13

Fig 13: Comparison of Algorithms 1 and 3.