

Uniform bounds for robust mean estimators

Stanislav Minsker ^{*,1}

¹*Department of Mathematics, University of Southern California*
e-mail: minske@usc.edu

Abstract: This paper is devoted to the estimators of the mean that provide strong non-asymptotic guarantees under minimal assumptions on the underlying distribution. The main ideas behind proposed techniques are based on bridging the notions of symmetry and robustness. We show that existing methods, such as median-of-means and Catoni’s estimators, can often be viewed as special cases of our construction. The main contribution of the paper is the proof of uniform bounds for the deviations of the stochastic process defined by proposed estimators. Moreover, we extend our results to the case of adversarial contamination where a constant fraction of the observations is arbitrarily corrupted. Finally, we apply our methods to the problem of robust multivariate mean estimation and show that obtained inequalities achieve optimal dependence on the proportion of corrupted samples.

Keywords and phrases: robust estimation, median-of-means estimator, concentration inequalities, adversarial contamination.

1. Introduction

Let $(S, \mathcal{S})$ be the measurable space, and $X \in S$ be a random variable with distribution P . Moreover, suppose that $X_1, \dots, X_N$ are i.i.d. copies of X . Assume that $\mathcal{F}$ is a class of measurable functions from S to $\mathbb{R}$. Many problems in mathematical statistics and statistical learning theory require simultaneously estimating $Pf := \mathbb{E}f(X)$ for all $f \in \mathcal{F}$. For example, in the maximum likelihood estimation framework, $\mathcal{F} = \{\log p_\theta(\cdot), \theta \in \Theta\}$ is a family of probability density functions with respect to a σ -finite measure μ , and $\frac{dP}{d\mu} = p_{\theta_*}$ for $\theta_* \in \Theta$. The most common way to estimate $\mathbb{E}f(X)$ is via the empirical mean $P_N f := \frac{1}{N} \sum_{j=1}^N f(X_j)$. Deviations of the resulting *empirical process* $\mathcal{F} \ni f \mapsto \sqrt{N}(P_N - P)f$ have been extensively studied, however, sharp estimates are known only under rather restrictive conditions, such as the case when functions in $\mathcal{F}$ are uniformly bounded, or when the envelope $F(x) := \sup_{f \in \mathcal{F}} |f(x)|$ of the class $\mathcal{F}$ possesses finite exponential moments [34, 20, 3, 12, 1].

Here, we consider the situation when the random variables $\{f(X), f \in \mathcal{F}\}$ indexed by $\mathcal{F}$ are allowed to be heavy-tailed, meaning that they possess finite moments of low order only (in the context of this paper, “low order” will usually mean the range between 2 and 3). In this case, the tail probabilities $P\left(\left|\frac{1}{N} \sum_{j=1}^N f(X_j) - \mathbb{E}f(X)\right| \geq t\right)$ decay polynomially, making many existing techniques unapplicable. Our approach to simultaneous mean estimation is based on replacing the sample mean by a different estimator of $\mathbb{E}f(X)$ that is “robust” to heavy tails and admits tight concentration under minimal moment assumptions. Well known examples of such estimators include the median-of-means (MOM) estimator [31, 2, 24] and Catoni’s estimator [6]. These two techniques rely on different principles for controlling the bias: while Catoni’s estimator is (informally speaking) based on delicate “truncation,” the MOM estimator exploits the fact that both the median and the mean of a symmetric distribution coincide with the center of symmetry [30]. Construction proposed in this work shows that these principles can be unified. We suggest a family of estimators that can be viewed as a “bridge” between Catoni’s estimator and the MOM technique, and prove uniform bounds for the deviations of the resulting stochastic process. We also address the more challenging framework of adversarial contamination and show that estimators of the mean of a random vector obtained using our methods admit optimal performance bounds in this case.

*Supported in part by the National Science Foundation grant DMS-1712956.

1.1. Notation and organization of the paper.

Absolute positive constants will be denoted c, C, c_1 , etc., and may take different values in different parts of the paper. For a function $h : \mathbb{R} \mapsto \mathbb{R}$, define

$$\operatorname{argmin}_{z \in \mathbb{R}} h(z) := \{z \in \mathbb{R} : h(z) \leq h(x) \text{ for all } x \in \mathbb{R}\},$$

and $\|h\|_\infty := \operatorname{ess\,sup}\{|h(x)| : x \in \mathbb{R}\}$. For a Lipschitz continuous function h , $L(h)$ will denote its Lipschitz constant. For $f \in \mathcal{F}$, denote $\sigma^2(f) = \operatorname{Var}(f(X))$ and $\sigma^2(\mathcal{F}) = \sup_{f \in \mathcal{F}} \sigma^2(f)$. 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 $N(0, \sigma^2(f))$. Additional notation and auxiliary results are introduced as necessary.

Material of the paper is organized as follows: section 2 explains the main ideas behind construction of the estimators studied in the paper. The main results are stated in section 2.1, followed by discussion and comparison to the literature on the topic in section 2.2. Section 3 discuss extensions of the results in the framework of adversarial contamination. Finally, section 3.1 explores implications of the bounds for the problem of multivariate mean estimation. Finally, the proofs are presented in section 4.

2. Construction of robust estimators of the mean.

Proposed estimators are based on the following (informally stated) principles:

- (a) If the distribution Q is symmetric, then its center of symmetry $\theta(Q)$ can be approximated by a robust estimator, such as Huber's robust M-estimator of location [18] defined via

$$\hat{\theta} := \operatorname{argmin}_{z \in \mathbb{R}} \sum_{j=1}^N \rho(z - Y_j),$$

where $Y_1, \dots, Y_N$ is an i.i.d. sample from Q and ρ is a convex, even function with bounded derivative.

- (b) In order to construct a robust estimator of a parameter $\theta^\circ(Q)$ of (not necessarily symmetric) distribution Q based on an i.i.d. sample $Y_1, \dots, Y_N$, create an auxiliary sample $\xi_1, \dots, \xi_M$ such that

- (i) it is governed by an approximately symmetric distribution;
- (ii) the center of symmetry of this distribution is close to $\tilde{\theta}(P)$.

According to (a), we then define an estimator of $\theta^\circ(Q)$ via

$$\hat{\theta}^\circ := \operatorname{argmin}_{z \in \mathbb{R}} \sum_{j=1}^M \rho(z - \xi_j).$$

The main focus of this work is the case when Q is the distribution of $f(X)$ and $\theta^\circ(Q)$ corresponds to the mean $\mathbb{E}f(X)$. To construct a “new sample” $\xi_1, \dots, \xi_M$ governed by an approximately symmetric distribution centered at $\mathbb{E}f(X)$, we rely on the fact that, under mild assumptions, the sample mean is asymptotically normal, hence asymptotically symmetric. Let k be an integer, and assume that $G_1, \dots, G_k$ are subsets of the index set $\{1, \dots, N\}$ of cardinality $|G_j| = n := \lfloor N/k \rfloor$ each, where the partition method is independent of the data $X_1, \dots, X_N$. In general, we do not require the subsets to be disjoint, and different possibilities will be discussed in the following sections. Let

$$\bar{\theta}_j(f) = \frac{1}{n} \sum_{i \in G_j} f(X_i)$$

be the empirical mean evaluated over the subsample indexed by G_j . Conditions on the suitable “loss function” ρ are summarized in the following assumption.

Assumption 1. Suppose that $\rho : \mathbb{R} \mapsto \mathbb{R}$ is a convex, even, continuously differentiable function such that

- (i) $\rho'(z) = z$ for $|z| \leq 1$ and $\rho'(z) = \operatorname{const}$ for $z \geq 2$.
- (ii) $z - \rho'(z)$ is nondecreasing;

For instance, Huber's loss [18]

$$\rho(z) = \frac{z^2}{2} I\{|z| \leq 1\} + (|z| - 1/2) I\{|z| > 1\}$$

is an example of a function satisfying Assumption 1.

Remark 2.1. Assumption 1 implies that $\rho'(2) - 2 \leq \rho'(1) - 1 = 0$, hence $\|\rho'\|_\infty \leq 2$. Moreover, for any $x > y$, $\rho'(x) - \rho'(y) = y - \rho'(y) - (x - \rho'(x)) + x - y \leq x - y$, hence ρ' is Lipschitz continuous with Lipschitz constant $L(\rho') = 1$.

Note that the loss $\rho(x) = |x|$ that leads to a “classical” median-of-means estimator does not satisfy assumption 1, in particular, ρ' is not Lipschitz continuous. Lipschitz continuity of ρ' turns out to be crucial for the derivation of our uniform bounds, however, for finite classes $\mathcal{F}$, the proofs are valid for broader class of losses that includes $\rho(x) = |x|$; see [30] for more details. The principles and assumptions stated above lead to the following definition of robust mean estimators: given $\Delta > 0$, set

$$\hat{\theta}^{(k)}(f) := \operatorname{argmin}_{z \in \mathbb{R}} \frac{1}{\sqrt{N}} \sum_{j=1}^k \rho \left(\sqrt{n} \frac{\bar{\theta}_j(f) - z}{\Delta} \right). \quad (2.1)$$

Parameter Δ , when expressed on a “natural scale” of the problem defined by $\sigma(\mathcal{F})$, can be interpreted as the truncation level. It will be shown that different combinations of the subgroup size n and “truncation level” Δ may lead to equally good bounds: in particular, when $n = 1$ and $\Delta \propto \sigma(\mathcal{F})\sqrt{N}$, we will recover Catoni's estimator, while the case of large n (e.g., $n \simeq \sqrt{N}$) and $\Delta \propto \sigma(\mathcal{F})$ leads to the MOM-type estimator.

2.1. Main results.

The collection of random variables $\{\hat{\theta}^{(k)}(f) - Pf, f \in \mathcal{F}\}$ defines a stochastic process that is a natural analogue of the empirical process in the framework of heavy-tailed data. Our main goal is to characterize the size of the supremum of the process, namely ¹

$$\sup_{f \in \mathcal{F}} |\hat{\theta}^{(k)}(f) - Pf|.$$

In particular, we will be interested in estimating the deviation probabilities $P \left(\sup_{f \in \mathcal{F}} |\hat{\theta}^{(k)}(f) - Pf| \geq t \right)$ under minimal assumptions on the process $\{f(X), f \in \mathcal{F}\}$. As a corollary of our general bounds, we will obtain new results for the problem of mean estimation in $\mathbb{R}^d$. Everywhere below, it will be assumed that $\sigma(\mathcal{F}) < \infty$.

As a first step, we introduce the main quantities appearing in our bounds. Given $f \in \mathcal{F}$ such that $\sigma(f) > 0$, $n \in \mathbb{N}$ and $t > 0$, define

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

It follows from the results of L. Chen and Q.-M. Shao [Theorem 2.2 in 8] that $g_f(t, n)$ controls the rate of convergence in the central limit theorem, namely

$$\left| \Pr \left(\frac{\sum_{j=1}^n (f(X_j) - Pf)}{\sigma(f)\sqrt{n}} \leq t \right) - \Phi(t) \right| \leq g_f(t, n), \quad (2.2)$$

¹We assume everywhere below that $\sup_{f \in \mathcal{F}} |\hat{\theta}^{(k)}(f) - Pf|$ is properly measurable. See [34, 11] for in-depth discussion of measurability issues.

given that $\sigma^2(f) < \infty$ and that an absolute constant C in the definition of $g_f(t, n)$ is large enough. The function $g_f(t, n)$ enters our bounds through the quantity that we define next. Given $\Delta > 0$, set

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

The following statement provides simple upper bounds for $g_f(t, n)$ and $G_f(n, \Delta)$.

Lemma 2.1. *Let $X_1, \dots, X_n$ be i.i.d. copies of X , and assume that $\text{Var}(f(X)) < \infty$. Then $g_f(t, n) \rightarrow 0$ as $|t| \rightarrow \infty$ and $g_f(t, n) \rightarrow 0$ as $n \rightarrow \infty$, 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 $t > 0$*

$$\begin{aligned} g_f(t, n) &\leq C' \frac{\mathbb{E}|f(X) - \mathbb{E}f(X)|^{2+\delta}}{n^{\delta/2} (\sigma(f) + |t|)^{2+\delta}} \leq C' \frac{\mathbb{E}|f(X) - \mathbb{E}f(X)|^{2+\delta}}{n^{\delta/2} |t|^{2+\delta}}, \\ G_f(n, \Delta) &\leq C'' \frac{\mathbb{E}|f(X) - \mathbb{E}f(X)|^{2+\delta}}{\Delta^{2+\delta} n^{\delta/2}}, \end{aligned} \quad (2.3)$$

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

The proof of this lemma is outlined in section 4.4. We are now ready to state the main result. Let ρ be a loss function satisfying Assumption 1. Moreover, set

$$\tilde{\Delta} := \max(\Delta, \sigma(\mathcal{F})).$$

Theorem 2.1. *Assume that $N = nk$ and that the subgroups $G_1, \dots, G_k$ are disjoint. Then there exist absolute constants $c, C > 0$ such that for all $s > 0, n$ and k satisfying*

$$\frac{1}{\Delta} \left(\frac{1}{\sqrt{k}} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{k}} \right) + \sup_{f \in \mathcal{F}} G_f(n, \Delta) + \frac{s}{k} \leq c, \quad (2.4)$$

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

$$\begin{aligned} \sup_{f \in \mathcal{F}} \left| \hat{\theta}^{(k)}(f) - Pf \right| &\leq C \left[\frac{\tilde{\Delta}}{\Delta} \left(\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{N}} \right) \right. \\ &\quad \left. + \tilde{\Delta} \left(\sqrt{n} \frac{s}{N} + \frac{\sup_{f \in \mathcal{F}} G_f(n, \Delta)}{\sqrt{n}} \right) \right]. \end{aligned}$$

The proof of the theorem is outlined in section 4.1. We note that the requirement $N = nk$ is not restrictive, as replacing N by $k \cdot \lfloor N/k \rfloor$ will only result in the change of absolute constants. The assumption that the groups sizes are equal is also not essential and is only imposed to avoid overly technical and cumbersome expressions. When the class $\mathcal{F}$ is P-Donsker [11], $\limsup_{N \rightarrow \infty} \left| \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(X_j) - Pf) \right|$ is bounded, hence condition (3.1) holds for N large enough whenever s is not too big and Δ is not too small, namely, $s \leq c'k$ and $\Delta \geq c''\sigma(\mathcal{F})$. When discussing examples in the following section, this will be our default setup.

Estimator $\hat{\theta}^{(k)}(f)$ defined in (2.1) depends on the choice of subgroups $G_1, \dots, G_k$. It is natural to ask if there exists a version of $\hat{\theta}^{(k)}(f)$ that is permutation-invariant. We address this question below and present a construction based on U-statistics. For an integer $n \leq \frac{N}{2}$, let $k = \lfloor N/n \rfloor$, and define

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

Let h be a measurable function of n variables. Recall that a U-statistic of order n with kernel h based on the i.i.d. sample $X_1, \dots, X_N$ is defined as [16]

$$U_{N,n} = \frac{1}{\binom{N}{n}} \sum_{J \in \mathcal{A}_N^{(n)}} h(\{X_j\}_{j \in J}). \quad (2.5)$$

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

$$U_{N,n}(z; f) = \frac{1}{\binom{N}{n}} \sum_{J \in A_N^{(n)}} \rho \left(\sqrt{n} \frac{\bar{\theta}(f; J) - z}{\Delta} \right),$$

and set

$$\tilde{\theta}^{(k)}(f) := \operatorname{argmin}_{z \in \mathbb{R}} U_{N,n}(z; f).$$

Theorem 2.2. *There exist absolute constants $c, C > 0$ such that for all $s > 0$, n and k satisfying*

$$\frac{1}{\Delta} \left(\frac{1}{\sqrt{k}} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{k}} \right) + \sup_{f \in \mathcal{F}} G_f(n, \Delta) + \frac{s}{k} \leq c, \quad (2.6)$$

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

$$\begin{aligned} \sup_{f \in \mathcal{F}} \left| \tilde{\theta}^{(k)}(f) - Pf \right| \leq C \left[\frac{\tilde{\Delta}}{\Delta} \left(\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{N}} \right) \right. \\ \left. + \tilde{\Delta} \left(\sqrt{n} \frac{s}{N} + \frac{\sup_{f \in \mathcal{F}} G_f(n, \Delta)}{\sqrt{n}} \right) \right]. \end{aligned}$$

The resulting deviation bounds for $\tilde{\theta}^{(k)}(f)$ are of exactly the same form as for the estimator $\hat{\theta}^{(f)}(f)$ that is based on disjoint blocks of data. The proof of this result is given in section 4.2.

2.2. Discussion and comparison with existing bounds.

A number of recent works address the problem of robust empirical risk minimization that is closely related to the question addressed in the present paper. In [5], authors prove uniform deviation bounds for robust mean estimators defined using O. Catoni's approach [6]; these bounds are limited by their dependence on the covering numbers of the class $\mathcal{F}$ with respect to the sup-norm $\|\cdot\|_\infty$. Uniform bounds for the median-of-means estimators have been obtained in several papers, including [25, 22, 26, 21]. Result that is closest to our setting has been obtained in [22]: proof of Theorem 2 of that paper implies that the estimator $\hat{\theta}_{\text{med}}^{(k)}(f)$ corresponding to $\rho(x) = |x|$ satisfies

$$\sup_{f \in \mathcal{F}} \left| \hat{\theta}_{\text{med}}^{(k)}(f) - Pf \right| \leq C \left(\frac{1}{\sqrt{N}} \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(X_j) - Pf) \right| + \sigma(\mathcal{F}) \sqrt{\frac{k}{N}} \right) \quad (2.7)$$

with probability at least $1 - e^{-ck}$ for absolute constant $c, C > 0$; a slightly stronger version of this result has appeared in [27]. The key difference between this inequality and the bound of Theorem 2.1 is the fact that the former holds only for the fixed value of the confidence parameter $s = k$, while the latter typically provides deviation guarantees over the wide range $0 < s \leq ck$ of confidence parameter s . This difference is important, as one usually wants to choose k as large as possible to improve robustness (in particular, robustness to adversarial contamination) without degrading performance of the estimator. The price that we have to pay is the necessity to tune the parameter Δ . However, as we show below, in many cases there is a wide range of “suitable” choices of Δ , so this issue is not critical.

Let us consider two examples. First, assume that $n = 1$, $k = N$ and set $\Delta = \Delta(s) := \sigma(\mathcal{F}) \sqrt{\frac{N}{s}}$. In this case, for N large enough, condition (3.1) reduces to $s \leq c'N$, and we deduce from Theorem 2.1 that

$$\sup_{f \in \mathcal{F}} \left| \hat{\theta}^{(k)}(f) - Pf \right| \leq C \left[\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{N}} \right]$$

with probability at least $1 - 2e^{-s}$. This inequality recovers, up to constants, the result of Catoni [6] for $\mathcal{F} = \{f\}$ (indeed, in this case the expected supremum is 0), and improves upon uniform bounds obtained for Catoni-type estimators in [5]. If $\Delta = \sigma(\mathcal{F})\sqrt{N}$, we get the “sub-exponential” bound

$$\sup_{f \in \mathcal{F}} \left| \widehat{\theta}^{(k)}(f) - Pf \right| \leq C \left[\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \frac{1+s}{\sqrt{N}} \right]$$

that holds with probability at least $1 - e^{-s}$ uniformly for $s \leq c'N$. Moreover, if

$$\kappa_{2+\delta}(\mathcal{F}) := \frac{\sup_{f \in \mathcal{F}} \mathbb{E} |f(X) - \mathbb{E} f(X)|^{2+\delta}}{\sigma^{2+\delta}(\mathcal{F})} < \infty$$

for some $\delta \in (0, 1]$, then the estimator is less sensitive to the choice of the parameter Δ . Specifically,

$$\sup_{f \in \mathcal{F}} \left| \widehat{\theta}^{(k)}(f) - Pf \right| \leq C \left[\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{N}} \right] \quad (2.8)$$

with probability at least $1 - 2e^{-s}$ for any Δ satisfying $\sigma(\mathcal{F}) \left(\frac{N}{s} \right)^{\frac{1}{2(1+\delta)}} (\kappa_{2+\delta}(\mathcal{F}))^{1/(1+\delta)} \leq \Delta \leq \sigma(\mathcal{F}) \sqrt{\frac{N}{s}}$. For instance, if $\delta = 1$, then any Δ in the range $\sigma(\mathcal{F})(N/s)^{1/4} \lesssim \Delta \lesssim \sigma(\mathcal{F})\sqrt{N/s}$ is suitable. Equivalently, for a given Δ , the sub-Gaussian type bound (2.8) holds uniformly for all $s \in \left[\kappa_3^2 \frac{M_\Delta^4}{N}, M_\Delta^2 \right]$ where $M_\Delta = \frac{\sigma(\mathcal{F})\sqrt{N}}{\Delta}$.

Next, assume that $N \gg n \geq 2$. For $\Delta = \sigma(\mathcal{F})\sqrt{\frac{k}{s}}$, we again recover the bound

$$\sup_{f \in \mathcal{F}} \left| \widehat{\theta}^{(k)}(f) - Pf \right| \leq C \left[\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{N}} \right]$$

that holds with probability at least $1 - 2e^{-s}$. When $\Delta \asymp \sigma(\mathcal{F})$, $\widehat{\theta}^{(k)}(f)$ most closely resembles the median-of-means estimator. In this case, the inequality that holds with probability at least $1 - 2e^{-s}$ is

$$\sup_{f \in \mathcal{F}} \left| \widehat{\theta}^{(k)}(f) - Pf \right| \leq C \left[\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \left(\sqrt{\frac{s}{N}} + \sqrt{\frac{k}{N}} \sup_{f \in \mathcal{F}} G_f(n, \sigma(\mathcal{F})) \right) \right].$$

As $\sup_{f \in \mathcal{F}} G_f(n, \sigma(\mathcal{F}))$ is small for large n , this bound is clearly better than (2.7). Finally, let us again consider the case when stronger moment assumptions hold, namely, $\kappa_3(\mathcal{F}) < \infty$. Combining the estimate for $\sup_{f \in \mathcal{F}} G_f(n, \Delta)$ provided by Lemma 2.1 and the bound of Theorem 2.1, we obtain that

$$\begin{aligned} \sup_{f \in \mathcal{F}} \left| \widehat{\theta}^{(k)}(f) - Pf \right| \leq C \left[\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{N}} \right. \\ \left. + \frac{\sup_{f \in \mathcal{F}} \mathbb{E} |f(X) - \mathbb{E} f(X)|^3}{\Delta^2} \frac{k}{N} + \Delta \frac{s}{\sqrt{kN}} \right], \end{aligned}$$

again with probability at least $1 - 2e^{-s}$. If Δ is such that $c_1 \sigma(\mathcal{F}) \left(1 \vee \sqrt{\frac{k}{\sqrt{Ns}}} \right) \leq \Delta \leq c_2 \sigma(\mathcal{F}) \sqrt{k/s}$ then this inequality implies sub-Gaussian deviation bounds at confidence level s . If for instance $1 \ll k \ll \sqrt{N}$ and $\Delta \asymp \sigma(\mathcal{F})$, then

$$\sup_{f \in \mathcal{F}} \left| \widehat{\theta}^{(k)}(f) - Pf \right| \leq C \left[\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{N}} \right]$$

for all $s \leq c'k$ uniformly.

3. Contamination with outliers.

Assume that the initial dataset of cardinality N is merged with a set of $\mathcal{O} < N$ outliers that 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 a statistician. We would like to understand performance of proposed estimators $\hat{\theta}^{(k)}(f)$ in this more challenging framework. Let $G_1, \dots, G_k$ be the disjoint partition of the index set $\{1, \dots, N^\circ\}$ that the estimators $\{\hat{\theta}^{(k)}(f), f \in \mathcal{F}\}$ are based on; as before, n will stand for the cardinality of G_j .

In the rest of the section, we will assume that $k > 2\mathcal{O}$. Let $J \subset \{1, \dots, k\}$ of cardinality $|J| \geq k - \mathcal{O}$ be the subset containing all j such that the subsample $\{X_i, j \in G_j\}$ 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

$$N_J \geq n|J| \geq \frac{kn}{2}.$$

The following analogue of Theorem 2.1 holds.

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

$$\frac{1}{\Delta} \left(\frac{1}{\sqrt{k}} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{k}} \right) + \sup_{f \in \mathcal{F}} G_f(n, \Delta) + \frac{s + \mathcal{O}}{k} \leq c, \quad (3.1)$$

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

$$\sup_{f \in \mathcal{F}} |\hat{\theta}^{(k)}(f) - Pf| \leq C \left[\frac{\tilde{\Delta}}{\Delta} \left(\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \sqrt{\frac{s}{N}} \right) + \tilde{\Delta} \left(\sqrt{n} \frac{s + \mathcal{O}}{N} + \frac{\sup_{f \in \mathcal{F}} G_f(n, \Delta)}{\sqrt{n}} \right) \right].$$

It is convenient to interpret the inequality as follows: if $\mathcal{O} \leq ck$ for a sufficiently small absolute constant c , the error $\sup_{f \in \mathcal{F}} |\hat{\theta}^{(k)}(f) - Pf|$ behaves like the maximum of 2 terms: the first term is the error bound for the case $\mathcal{O} = 0$, and the second term is of order $\tilde{\Delta} \sqrt{n} \frac{\mathcal{O}}{N}$. In the next section, we provide examples which show that the dependence on $\mathcal{O}$ in our bounds is, in general, non-improvable.

3.1. Estimators of the mean of a random vector.

Assume that $X_1, \dots, X_N$ are i.i.d. copies of a random vector $X \in \mathbb{R}^d$ with mean $\mathbb{E}X = \mu$ and covariance matrix $\mathbb{E}(X - \mu)(X - \mu)^T = \Sigma$. Let $\|\cdot\|$ be some norm in $\mathbb{R}^d$, and let B be the unit ball with respect to this norm, $B = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$. Consider the class of linear functionals $\mathcal{F} = \{f_v(x) = \langle v, x \rangle, v \in B\}$. Our goal is to estimate the mean μ , with the error measured in the norm $\|\cdot\|$. Construction that we propose is closely related to the approach employed previously by several authors [19, 27, 7, 14] that is based on combining estimators of one-dimensional projections. Assume that we are in the “adversarial contamination” framework of Theorem 3.1. Let ρ be a function satisfying assumption 1, and let $\hat{\theta}^{(k)}(v)$ be the estimator of $\langle \mu, v \rangle$, the projection of μ in direction $v \in B$:

$$\hat{\theta}^{(k)}(v) := \operatorname{argmin}_{z \in \mathbb{R}} \frac{1}{\sqrt{N}} \sum_{j=1}^k \rho \left(\sqrt{n} \frac{\bar{\mu}_j(v) - z}{\Delta} \right),$$

where $\bar{\mu}_j(v) = \frac{1}{n} \sum_{i \in G_j} \langle v, X_i \rangle$, and $\Delta \geq \sqrt{\lambda_{\max}(\Sigma)}$. Given $v \in B$ and $\varepsilon > 0$, define the closed “slab”

$$S_v(\varepsilon) := \left\{ y \in \mathbb{R}^d : \left| \langle y, v \rangle - \hat{\theta}^{(k)}(v) \right| \leq \varepsilon \right\},$$

and $M(\varepsilon) := \bigcap_{v \in B} S_v(\varepsilon)$. Finally, let $\varepsilon_* := \inf \{\varepsilon > 0 : M(\varepsilon) \neq \emptyset\}$, and take $\hat{\mu}^{(k)}$ to be any element in $M(\varepsilon_*)$ (indeed, $M(\varepsilon_*) = \bigcap_{\varepsilon > \varepsilon_*} M(\varepsilon)$ is non-empty as an intersection of nested compact sets).

Corollary 3.1. *There exist absolute constants $\tilde{c}, \tilde{C} > 0$ with the following properties: assume that*

$$\max \left(\frac{1}{\sqrt{k}\Delta} \mathbb{E} \sup_{v \in B} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N \langle v, X_j - \mu \rangle \right| + \frac{s}{k} + \sup_{v \in B} G_f(n, \Delta), \frac{\mathcal{O}}{k} \right) \leq \tilde{c}. \quad (3.2)$$

Then with probability at least $1 - 2e^{-s}$,

$$\left\| \hat{\mu}^{(k)} - \mu \right\| \leq \tilde{C} \left(\mathbb{E} \sup_{v \in B} \frac{1}{N} \sum_{j=1}^N \langle v, X_j - \mu \rangle + \sqrt{\lambda_{\max}(\Sigma)} \sqrt{\frac{s}{N}} + \Delta \left(\sup_{v \in B} \frac{G_{f_v}(n, \Delta)}{\sqrt{n}} + \sqrt{n} \frac{s + \mathcal{O}}{N} \right) \right).$$

Proof. It follows from Theorem 3.1 that on the event $\mathcal{E}$ of probability at least $1 - 2e^{-s}$, $\mu \in M(\varepsilon)$ for all

$$\varepsilon \geq \varepsilon_0 := C \left(\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^N \langle v, X_j - \mu \rangle + \sqrt{\lambda_{\max}(\Sigma)} \sqrt{\frac{s}{N}} + \Delta \left(\sup_{v \in B} \frac{G_{f_v}(n, \Delta)}{\sqrt{n}} + \sqrt{n} \frac{s + \mathcal{O}}{N} \right) \right)$$

given that (3.2) holds, hence $\varepsilon_* \leq \varepsilon_0$ on event $\mathcal{E}$. Consequently,

$$\left\| \hat{\mu}^{(k)} - \mu \right\| = \sup_{v \in B} \left| \langle \hat{\mu}^{(k)} - \mu, v \rangle \right| \leq \sup_{v \in B} \left| \langle \hat{\mu}^{(k)}, v \rangle - \hat{\theta}^{(k)}(v) \right| + \sup_{v \in B} \left| \langle \mu, v \rangle - \hat{\theta}^{(k)}(v) \right| \leq \varepsilon_* + \varepsilon_0 \leq 2\varepsilon_0$$

with probability at least $1 - 2e^{-s}$. $\square$

In the special case when $\|\cdot\|$ is the Euclidean norm $\|\cdot\|_2$, the bound of Corollary 3.1 can be further simplified. It follows from Hölder's inequality that

$$\begin{aligned} \mathbb{E} \sup_{v \in \mathbb{R}^d: \|v\|_2=1} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N \langle v, X_j - \mu \rangle \right| &\leq \mathbb{E}^{1/2} \sup_{v \in \mathbb{R}^d: \|v\|_2=1} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N \langle v, X_j - \mu \rangle \right|^2 \\ &= \sqrt{\frac{1}{N} \mathbb{E} \left\| \sum_{j=1}^N (X_j - \mu) \right\|_2^2} = \sqrt{\mathbb{E} \|X - \mu\|_2^2} = \sqrt{\text{tr} \Sigma}, \end{aligned}$$

hence $\hat{\mu}^{(k)}$ satisfies

$$\left\| \hat{\mu}^{(k)} - \mu \right\|_2 \leq C \left(\sqrt{\frac{\text{tr} \Sigma}{N}} + \sqrt{\lambda_{\max}(\Sigma)} \sqrt{\frac{s}{N}} + \Delta \left(\sup_{v \in B} \frac{G_{f_v}(n, \Delta)}{\sqrt{n}} + \sqrt{n} \frac{s + \mathcal{O}}{N} \right) \right)$$

with probability at least $1 - 2e^{-s}$ (whenever $s \leq c'k$). According to the discussion in section 2.2, in many cases the term $\Delta \left(\sup_{v \in B} \frac{G_{f_v}(n, \Delta)}{\sqrt{n}} + \sqrt{n} \frac{s + \mathcal{O}}{N} \right)$ is of order smaller than $N^{-1/2}$, whence the estimator $\hat{\mu}^{(k)}$ behaves like the sample mean of the Gaussian random variables [27].

Let us now discuss optimality with respect to adversarial contamination. Assume that $\mathcal{O} = \varepsilon N$ for $\varepsilon \geq \frac{1}{N}$; here, we assume ε to be known in advance, and the issue of adaptivity is beyond the scope of this paper. Moreover, suppose that

$$\kappa_{2+\delta} := \sup_{v: \|v\|_2=1} \frac{\mathbb{E} |\langle v, X - \mu \rangle|^{2+\delta}}{(\text{Var} \langle v, X - \mu \rangle)^{1+\delta/2}} < \infty$$

for some $\delta \in (0, 1]$. In this case, Lemma 2.1 implies that

$$\sup_{v \in B} G_{f_v}(n, \Delta) \leq C \frac{\sup_{\|v\|_2=1} \mathbb{E} |\langle X - \mu, v \rangle|^{2+\delta}}{\Delta^{2+\delta} n^{\delta/2}}$$

for some absolute constant $C > 0$. Let $M_\Delta := \frac{\Delta}{\sigma(\mathcal{F})}$, and recall that $M_\Delta \geq 1$ by assumption. Then, choosing $k = \varepsilon^{\frac{2}{2+\delta}} N \frac{M_\Delta^2}{\kappa_{2+\delta}^{\frac{2}{2+\delta}}}$, we deduce from Corollary 3.1 that

$$\left\| \hat{\mu}^{(k)} - \mu \right\|_2 \leq C \left(\sqrt{\frac{\text{tr} \Sigma}{N}} + \sqrt{\lambda_{\max}(\Sigma)} \left(\sqrt{\frac{s}{N}} + \varepsilon^{\frac{1+\delta}{2+\delta}} \kappa_{2+\delta}^{1/(2+\delta)} \right) + \frac{\kappa_{2+\delta}^{1/(2+\delta)}}{\varepsilon^{1/(2+\delta)} N} s \right)$$

with probability at least $1 - 2e^{-s}$. Since $\varepsilon \geq \frac{1}{N}$, $\varepsilon^{1/(2+\delta)}N \geq N^{-\frac{1+\delta}{2+\delta}} = o(N^{-1/2})$ whenever $\delta > 0$, hence the last term in the bound is of smaller order. According to Lemma 5.4 (see section 5), no estimator can achieve rate faster than $\varepsilon^{\frac{1+\delta}{2+\delta}}$ with respect to ε , implying that our estimator is optimal in this sense.

4. Proofs.

We will introduce some additional notation and recall useful results that we rely upon in the proofs. Denote

$$G_k(z; f) = \frac{1}{\sqrt{k}} \sum_{j=1}^k \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right)$$

and let $\hat{\theta}_0^{(k)}(f)$ be any solution of the equation $G_k(\hat{\theta}_0^{(k)}(f); f) = 0$. Clearly, $\hat{\theta}_0^{(k)}(f) = \hat{\theta}^{(k)}(f) - Pf$ is the error of the estimator $\hat{\theta}^{(k)}(f)$.

The following concentration inequality is due to Klein and Rio (see section 12.5 in [4]).

Fact 1. *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 $|Z_j(f) - \mathbb{E}Z_j(f)| \leq M$ a.s. for all $1 \leq j \leq N$ and $f \in \mathcal{F}$. Then the following inequality holds with probability at least $1 - e^{-s}$:*

$$\sup_{f \in \mathcal{F}} \left(\sum_{j=1}^N (Z_j(f) - \mathbb{E}Z_j(f)) \right) \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sum_{j=1}^N (Z_j(f) - \mathbb{E}Z_j(f)) \right) + V(\mathcal{F})\sqrt{2s} + \frac{4Ms}{3}, \quad (4.1)$$

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

It is easy to see, applying (4.1) to processes $\{-Z_j(f), f \in \mathcal{F}\}$, that

$$\inf_{f \in \mathcal{F}} \left(\sum_{j=1}^N (Z_j(f) - \mathbb{E}Z_j(f)) \right) \geq -2\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sum_{j=1}^N (\mathbb{E}Z_j(f) - Z_j(f)) \right) - V(\mathcal{F})\sqrt{2s} - \frac{4Ms}{3} \quad (4.2)$$

with probability at least $1 - e^{-s}$. Moreover, (4.1) is a corollary of the following bound for the moment generating function:

$$\log \mathbb{E} e^{\lambda \left(\sum_{j=1}^N (Z_j(f) - \mathbb{E}Z_j(f)) \right)} \leq \frac{e^{\lambda M} - \lambda M - 1}{M^2} \left(V^2(\mathcal{F}) + 2M \mathbb{E} \sup_{f \in \mathcal{F}} \left(\sum_{j=1}^N (Z_j(f) - \mathbb{E}Z_j(f)) \right) \right) \quad (4.3)$$

that holds for all $\lambda > 0$. This fact provides a straightforward extension of the concentration bounds to the case of U-statistics. Let π_N be the collection of all permutations $i : \{1, \dots, N\} \mapsto \{1, \dots, N\}$. Given $(i_1, \dots, i_N) \in \pi_N$ and a U-statistic $U_{N,n}$ defined in (2.5), let

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

It is well known (see section 5 in [17]) that the following representation holds:

$$U_{N,n} = \frac{1}{N!} \sum_{(i_1, \dots, i_N) \in \pi_N} T_{i_1, \dots, i_N}. \quad (4.4)$$

Let $U'_{N,n}(z; f) = \frac{1}{\binom{N}{n}} \sum_{J \in \mathcal{A}_N^{(n)}} \rho' \left(\sqrt{n} \frac{(\bar{\theta}(f; J) - Pf) - z}{\Delta} \right)$. Applied to $U'_{N,n}(z; f)$, relation (4.4) 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

$$T_{i_1, \dots, i_N}(z; f) = \frac{1}{k} \left(\rho' \left(\sqrt{n} \frac{\bar{\theta}(f; \{i_1, \dots, i_n\}) - Pf - z}{\Delta} \right) + \dots + \rho' \left(\sqrt{n} \frac{\bar{\theta}(f; \{i_{(k-1)n+1}, \dots, i_{kn}\}) - Pf - z}{\Delta} \right) \right).$$

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

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

hence bound (4.3) can be applied and yields that

$$\begin{aligned} \sup_{f \in \mathcal{F}} (U'_{N,n}(z; f) - \mathbb{U}'_{N,n}(z; f)) &\leq 2 \mathbb{E} \sup_{f \in \mathcal{F}} (T_{1, \dots, N}(z; f) - \mathbb{E} T_{1, \dots, N}(z; f)) \\ &\quad + \sup_{f \in \mathcal{F}} \sqrt{\text{Var} \left(\rho' \left(\sqrt{n} \frac{\bar{\theta}(f; \{1, \dots, n\}) - Pf - z}{\Delta} \right) \right)} \sqrt{\frac{2s}{k} + \frac{8s \|\rho'\|_\infty}{3k}} \quad (4.5) \end{aligned}$$

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

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

due to Lemma 5.3.

4.1. Proof of Theorem 2.1.

Recall that

$$G_k(z; f) = \frac{1}{\sqrt{k}} \sum_{j=1}^k \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right).$$

Suppose z_1, z_2 are such that on an event of probability close to 1, $G_k(z_1; f) > 0$ and $G_k(z_2; f) < 0$ for all $f \in \mathcal{F}$ simultaneously. Since G_k is decreasing in z , it is easy to see that $\hat{\theta}_0^{(k)}(f) \in (z_1, z_2)$ for all $f \in \mathcal{F}$ on this event, implying that $\left| \sup_{f \in \mathcal{F}} \hat{\theta}_0^{(k)}(f) \right| \leq \max(|z_1|, |z_2|)$. 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. We will provide detailed bounds for z_1 , while the steps to estimate z_2 are quite similar. Observe that

$$\begin{aligned} G_k(z; f) &= \frac{1}{\sqrt{k}} \sum_{j=1}^k \left(\rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \\ &\quad + \frac{1}{\sqrt{k}} \sum_{j=1}^k \left(\mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right) + \frac{1}{\sqrt{k}} \sum_{j=1}^k \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right). \end{aligned}$$

We will proceed in 3 steps: first, we will find $\varepsilon_1 > 0$ such that for any $z \in \mathbb{R}$,

$$\inf_{f \in \mathcal{F}} \frac{1}{\sqrt{k}} \sum_{j=1}^k \left(\rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \geq -\varepsilon_1 \quad (4.6)$$

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

$$\inf_{f \in \mathcal{F}} \frac{1}{\sqrt{k}} \sum_{j=1}^k \left(\mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right) \geq -\varepsilon_2,$$

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

$$\frac{1}{\sqrt{k}} \sum_{j=1}^k \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n} z_1}{\Delta} \right) > \varepsilon_1 + \varepsilon_2. \quad (4.7)$$

It follows from Lemma 4.1 that setting

$$\varepsilon_1 = \frac{8}{\Delta \sqrt{N}} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^N (f(X_j) - Pf) + \frac{\sigma(\mathcal{F})}{\Delta} \sqrt{2s} + \frac{16}{3} \frac{s}{\sqrt{k}}$$

guarantees that (4.6) holds with probability at least $1 - e^{-s}$. Next, Lemma 4.2 implies that ε_2 can be chosen as

$$\varepsilon_2 = 2\sqrt{k} \sup_{f \in \mathcal{F}} G_f(n, \Delta).$$

Finally, we apply Lemma 4.3 with

$$\varepsilon := \frac{\varepsilon_1 + \varepsilon_2}{\sqrt{k}}$$

to deduce that

$$z_1 = -\frac{1}{0.09} \left(\frac{8L(\rho')}{N} \frac{\tilde{\Delta}}{\Delta} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^N (f(X_j) - Pf) + \sigma(\mathcal{F}) \frac{\tilde{\Delta}}{\Delta} \sqrt{\frac{2s}{N}} + \tilde{\Delta} \frac{16}{3} \frac{s\sqrt{n}}{N} + 2 \frac{\sup_{f \in \mathcal{F}} G_f(n, \Delta)}{\sqrt{n}} \right)$$

satisfies (4.7) under assumption that $\varepsilon \leq 0.045$. Proceeding in a similar way, it is easy to see that setting $z_2 = -z_1$ guarantees that $G_k(z_2; f) < 0$ for all $f \in \mathcal{F}$ with probability at least $1 - e^{-s}$, hence the claim follows.

Lemma 4.1. *For any $z \in \mathbb{R}$ and ρ satisfying Assumption 1, the inequalities*

$$\begin{aligned} & \inf_{f \in \mathcal{F}} \frac{1}{\sqrt{k}} \sum_{j=1}^k \left(\rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \\ & \geq -\frac{8}{\Delta \sqrt{N}} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^N (f(X_j) - Pf) - \frac{\sigma(\mathcal{F})}{\Delta} \sqrt{2s} - \frac{16}{3} \frac{s}{\sqrt{k}}, \\ & \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{k}} \sum_{j=1}^k \left(\rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \\ & \leq \frac{8}{\Delta \sqrt{N}} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^N (f(X_j) - Pf) + \frac{\sigma(\mathcal{F})}{\Delta} \sqrt{2s} + \frac{16}{3} \frac{s}{\sqrt{k}}, \end{aligned}$$

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

Proof. We will prove the first inequality, while the second follows similarly. First observe that in view of Lemma 5.3,

$$\text{Var} \left(\rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \leq \text{Var} \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) = \frac{\sigma^2(f)}{\Delta^2},$$

hence $\sup_{f \in \mathcal{F}} \text{Var}^{1/2} \left(\rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \leq \frac{\sigma(\mathcal{F})}{\Delta}$. Next, Fact 1 (more specifically, inequality (4.2)) implies that for any fixed $z \in \mathbb{R}$,

$$\begin{aligned} & \inf_{f \in \mathcal{F}} \sum_{j=1}^k \frac{1}{\sqrt{k}} \left(\rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \\ & \geq -2 \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{k}} \sum_{j=1}^k \left(\mathbb{E} \rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \rho' \left(\frac{\sqrt{n}(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \\ & \quad - \frac{\sigma(\mathcal{F})}{\Delta} \sqrt{2s} - 8 \|\rho'\|_{\infty} \frac{s}{3\sqrt{k}} \end{aligned}$$

with probability at least $1 - e^{-s}$. It follows from Remark 2.1 that $\|\rho'\|_\infty \leq 2$, hence it remains to estimate the expected supremum. To this end, we will apply symmetrization and Talagrand's contraction inequalities ² [20, 23, 29] (where functions $h_j(x) := -\frac{\rho'(x - \sqrt{n}z/\Delta) - \rho'(-\sqrt{n}z/\Delta)}{L(\rho')}$ are contractions satisfying $h_j(0) = 0$). Let $\varepsilon_1, \dots, \varepsilon_k$ be i.i.d. Rademacher random variables independent of $X_1, \dots, X_N$, and note that

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{k}} \sum_{j=1}^k \left(-\rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) + \mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \\ \leq 2L(\rho') \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{k}} \sum_{j=1}^k \varepsilon_j h_j \left(\sqrt{n} \frac{\bar{\theta}_j(f) - Pf}{\Delta} \right) \leq \frac{4}{\Delta} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^k \varepsilon_j \frac{\sqrt{n}}{\sqrt{k}} (\bar{\theta}_j(f) - Pf), \end{aligned}$$

where we used the fact that $L(\rho') \leq 1$. Next, desymmetrization inequality [20] implies that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^k \varepsilon_j \frac{\sqrt{n}}{\sqrt{k}} (\bar{\theta}_j(f) - Pf) \leq \frac{2}{\sqrt{N}} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^N (f(X_j) - Pf).$$

and the result follows. $\square$

Lemma 4.2. 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 $|z| \leq \frac{1}{2} \frac{\Delta}{\sqrt{n}}$,

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

Proof. Let $T(x) = x - \rho'(x)$, and note that $T(x) = 0$ for $|x| \leq 1$ by Assumption 1. Moreover, T is non-decreasing. As

$$\mathbb{E} \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} - \frac{W(f) - \sqrt{n}z}{\Delta} \right) = 0,$$

it is easy to check that

$$\begin{aligned} \left| \mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right| \\ = \left| \mathbb{E} \left(T \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - T \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right) \right|. \end{aligned}$$

For any bounded non-negative function $h : \mathbb{R} \mapsto \mathbb{R}_+$ and any signed measure Q ,

$$\left| \int_{\mathbb{R}} h(x) dQ \right| = \left| \int_0^\infty Q(x : h(x) \geq t) dt \right|.$$

Since any bounded function $h : \mathbb{R} \mapsto \mathbb{R}$ can be written as $h = h_+ - h_-$, where $h_+ = \max(h, 0)$ and $h_- = \max(-h, 0)$ are both nonnegative, we deduce that

$$\left| \int_{\mathbb{R}} h(x) dQ \right| \leq \int_0^\infty |Q(x : h_+(x) \geq t)| dt + \int_0^\infty |Q(x : h_-(x) \geq t)| dt. \quad (4.8)$$

Moreover, if h is monotone, the sets $\{x : h_+(x) \geq t\}$ and $\{x : h_-(x) \geq t\}$ are half-intervals. Take $h = T$; it follows from Assumption 1 that $T_+(x) \leq \max(x - 1, 0)$, hence $T_+^{-1}(t) \geq 1 + t$ for $t > 0$. Let $\Phi_f^{(n)}(\cdot)$ stand for the cumulative distribution function of $\sqrt{n}(\bar{\theta}_j(f) - Pf)$. Applying (4.8) to the monotone function $x \mapsto T\left(\frac{x - \sqrt{n}z}{\Delta}\right)$ and $Q(\cdot) = \Phi_f^{(n)}(\cdot) - \Phi(\cdot/\sigma(f))$, we deduce that for any $f \in \mathcal{F}$,

$$\begin{aligned} \left| \mathbb{E} \left(T \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - T \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right) \right| \\ \leq \int_0^\infty \left| Q \left(x : T_+ \left(\frac{x - \sqrt{n}z}{\Delta} \right) \geq t \right) \right| dt + \int_0^\infty \left| Q \left(x : T_- \left(\frac{x - \sqrt{n}z}{\Delta} \right) \geq t \right) \right| dt. \quad (4.9) \end{aligned}$$

²We use the versions of these inequalities without absolute values inside the supremum; see [29] for the proof.

Inequality (2.2) implies that

$$\begin{aligned} \left| Q \left(x : T_+ \left(\frac{x - \sqrt{n}z}{\Delta} \right) \geq t \right) \right| &= |Q(x : x \geq \Delta T_+^{-1}(t) + \sqrt{n}z)| \\ &\leq g_f(\Delta T_+^{-1}(t) + \sqrt{n}z, n) \leq g_f \left(\Delta \left(\frac{1}{2} + t \right), n \right) \end{aligned} \quad (4.10)$$

where we used monotonicity of g_f together with the inequalities $\sqrt{n}|z| \leq \frac{1}{2}\Delta$ and $T_+^{-1}(t) \geq 1 + t$ on the last step. Integrating inequality (4.10) from 0 to ∞ , we see that

$$\int_0^\infty \left| Q \left(x : T_+ \left(\frac{x - \sqrt{n}z}{\Delta} \right) \geq t \right) \right| dt \leq \int_0^\infty g_f \left(\Delta \left(\frac{1}{2} + t \right), n \right) dt = G_f(n, \Delta).$$

Similarly, $\int_0^\infty \left| Q \left(x : T_- \left(\frac{x - \sqrt{n}z}{\Delta} \right) \geq t \right) \right| dt \leq G_f(n, \Delta)$, hence the conclusion follows from (4.9). $\square$

Lemma 4.3. *Let $\varepsilon > 0$ be such that $\varepsilon \leq 0.045$, and set*

$$z_1 = -\frac{\varepsilon}{0.09} \frac{\max(\Delta, \sigma(\mathcal{F}))}{\sqrt{n}}.$$

Then for all $f \in \mathcal{F}$,

$$\mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} \right) > \varepsilon.$$

Proof. For any bounded function h such that $h(-x) = -h(x)$ and $h(x) \geq 0$ for $x \geq 0$, and any $z \leq 0$,

$$\int_{\mathbb{R}} h(x - z) \phi_\sigma(x) dx = \int_0^\infty h(x) (\phi_\sigma(x + z) - \phi_\sigma(-x + z)) dx \geq 0,$$

where $\phi_\sigma(x) = (2\pi\sigma)^{-1/2} e^{-x^2/2\sigma^2}$. Recall that $H'(x) \geq \frac{x}{2}$ for $0 \leq x \leq 2$, and take

$$h(x) := \rho'(x) - \frac{x}{2} I\{|x| \leq 2\} = H'(x) - \frac{x}{2} I\{|x| \leq 2\}.$$

Observe that $h(x) \geq 0$ for $x \geq 0$ by assumptions on ρ , hence for any j ,

$$\begin{aligned} \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} \right) &= \frac{1}{2} \mathbb{E} \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} I \left\{ \left| \frac{W(f) - \sqrt{n}z_1}{\Delta} \right| \leq 2 \right\} \right) + \mathbb{E} h \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} \right) \\ &\geq \max \left(\frac{1}{2} \mathbb{E} \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} I \left\{ \left| \frac{W(f) - \sqrt{n}z_1}{\Delta} \right| \leq 2 \right\} \right), \mathbb{E} h \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} \right) \right), \end{aligned} \quad (4.11)$$

where we used the fact that both terms are nonnegative. Next, we will find lower bounds for each of the terms in the maximum above, starting with the first.

(1) Consider two possibilities: (a) $\Delta < \sigma(f)$ and (b) $\Delta \geq \sigma(f)$. In the first case, we will use the trivial lower bound $\mathbb{E} \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} I \left\{ \left| \frac{W(f) - \sqrt{n}z_1}{\Delta} \right| \leq 2 \right\} \right) \geq 0$. The main focus will be on the second case. To this end, note that $Z := \frac{W(f)}{\sigma(f)} \sim N(0, 1)$, hence

$$\begin{aligned} &\frac{1}{2} \mathbb{E} \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} I \left\{ \left| \frac{W(f) - \sqrt{n}z_1}{\Delta} \right| \leq 2 \right\} \right) \\ &= \frac{\sigma(f)}{2\Delta} \mathbb{E} \left(Z I \left\{ \left| Z - \frac{\sqrt{n}z_1}{\sigma(f)} \right| \leq 2 \frac{\Delta}{\sigma(f)} \right\} \right) - \frac{\sqrt{n}z_1}{2\Delta} \Pr \left(\left| Z - \frac{\sqrt{n}z_1}{\sigma(f)} \right| \leq 2 \frac{\Delta}{\sigma(f)} \right). \end{aligned} \quad (4.12)$$

Direct computation shows that for any $a \in \mathbb{R}$, $t > 0$,

$$\left| \mathbb{E} (Z I \{|Z - a| \leq t\}) \right| = \frac{1}{\sqrt{2\pi}} e^{-\frac{a^2+t^2}{2}} |e^{at} - e^{-at}|. \quad (4.13)$$

Take $a = -\frac{z_1\sqrt{n}}{\sigma(f)}$, $t = 2\frac{\Delta}{\sigma(f)}$, and observe that assumptions of the Theorem imply the inequality $|a| \leq \frac{t}{4}$. Minimum of the function $a \mapsto a^2 + t^2 - 2|a|t$ over the set $0 \leq a \leq t/4$ is attained at $a = t/4$, implying that $a^2 + t^2 - 2|a|t \geq \frac{9}{16}t^2 > \frac{t^2}{2}$. Combining this with (4.13), we deduce that

$$\left| \mathbb{E} (Z I \{ |Z - a| \leq t \}) \right| \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/4} e^{-|at|} |e^{at} - e^{-at}| = \frac{e^{-t^2/4}}{\sqrt{2\pi}} (1 - e^{-2|at|}) \leq \frac{e^{-t^2/4}}{\sqrt{2\pi}} \cdot 2|at|,$$

hence

$$\left| \frac{\sigma(f)}{2\Delta} \mathbb{E} \left(Z I \left\{ \left| Z - \frac{\sqrt{n} z_1}{\sigma(f)} \right| \leq 2 \frac{\Delta}{\sigma(f)} \right\} \right) \right| \leq \frac{2}{\sqrt{2\pi}} \left| \frac{z_1 \sqrt{n}}{\sigma(f)} \right| e^{-\frac{\Delta^2}{\sigma^2(f)}} = \frac{2}{\sqrt{2\pi}} \left| \frac{z_1 \sqrt{n}}{\Delta} \right| \frac{\Delta}{\sigma(f)} e^{-\frac{\Delta^2}{\sigma^2(f)}}.$$

Moreover, since $|z_1| \leq \frac{1}{2} \frac{\Delta}{\sqrt{n}}$ by assumptions of the lemma, it follows that

$$\Pr \left(\left| Z - \frac{\sqrt{n} z_1}{\sigma(f)} \right| \leq 2 \frac{\Delta}{\sigma(f)} \right) \geq \Pr \left(|Z| \leq \frac{3\Delta}{2\sigma(f)} \right) \geq 1 - 2\Phi(-3/2) > 0.86.$$

Together with (4.11), (4.12), the last display yields that

$$\mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n} z_1}{\Delta} \right) > \left| \frac{0.86}{2} \frac{z_1 \sqrt{n}}{\Delta} \right| - \frac{2}{\sqrt{2\pi}} \left| \frac{z_1 \sqrt{n}}{\Delta} \right| \frac{\Delta}{\sigma(f)} e^{-\frac{\Delta^2}{\sigma^2(f)}}.$$

As $x \mapsto x e^{-x^2}$ is decreasing for $x \geq 1/\sqrt{2}$, one easily checks that $\frac{\Delta}{\sigma(f)} e^{-\frac{\Delta^2}{\sigma^2(f)}} \leq e^{-1}$ as $\Delta \geq \sigma(f)$, hence

$$\mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n} z_1}{\Delta} \right) > \left(0.43 - \frac{2}{e\sqrt{2\pi}} \right) |z_1| \frac{\sqrt{n}}{\Delta} > 0.1364 |z_1| \frac{\sqrt{n}}{\Delta}.$$

(2) For the second term, we start with a simple inequality

$$\begin{aligned} \mathbb{E} h \left(\frac{W(f) - \sqrt{n} z_1}{\Delta} \right) &\geq \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n} z_1}{\Delta} \right) I \left\{ \left| \frac{W(f) - \sqrt{n} z_1}{\Delta} \right| > 2 \right\} \\ &\geq \underbrace{\rho'(2)}_{\geq 1} \mathbb{E} \left(I \left\{ \frac{W(f) - \sqrt{n} z_1}{\Delta} > 2 \right\} - I \left\{ \frac{W(f) - \sqrt{n} z_1}{\Delta} < -2 \right\} \right) \end{aligned}$$

which follows from the definition of h and assumptions on ρ . Again, we consider two possibilities: (a) $\Delta < \sigma(f)$ and (b) $\Delta \geq \sigma(f)$. In case (b), we use the trivial bound

$$\mathbb{E} \left(I \left\{ \frac{W(f) - \sqrt{n} z_1}{\Delta} > 2 \right\} - I \left\{ \frac{W(f) - \sqrt{n} z_1}{\Delta} < -2 \right\} \right) \geq 0.$$

In the first case, we see that

$$\begin{aligned} \Pr \left(\frac{W(f) - \sqrt{n} z_1}{\Delta} \geq 2 \right) - \Pr \left(\frac{W(f) - \sqrt{n} z_1}{\Delta} \leq -2 \right) \\ = \Pr \left(Z \geq \frac{\sqrt{n} z_1}{\sigma(f)} + 2 \frac{\Delta}{\sigma(f)} \right) - \Pr \left(Z \leq \frac{\sqrt{n} z_1}{\sigma(f)} - 2 \frac{\Delta}{\sigma(f)} \right) \\ = \Pr \left(Z \in \left[\frac{\sqrt{n} z_1}{\sigma(f)} + 2 \frac{\Delta}{\sigma(f)}, -\frac{\sqrt{n} z_1}{\sigma(f)} + 2 \frac{\Delta}{\sigma(f)} \right] \right). \end{aligned}$$

Lemma 5.2 implies that

$$\begin{aligned} \Pr \left(Z \in \left[\frac{\sqrt{n} z_1}{\sigma(f)} + 2 \frac{\Delta}{\sigma(f)}, -\frac{\sqrt{n} z_1}{\sigma(f)} + 2 \frac{\Delta}{\sigma(f)} \right] \right) &\geq 2e^{-\frac{2\Delta^2}{\sigma^2(f)}} \Pr \left(Z \in \left[0, \frac{\sqrt{n}|z_1|}{\sigma(f)} \right] \right) \\ &\geq 2e^{-2} \Pr \left(Z \in \left[0, \frac{\sqrt{n}|z_1|}{\sigma(f)} \right] \right), \end{aligned}$$

where we used the fact that $\Delta < \sigma(f)$ by assumption. Finally, Lemma 5.1 implies that

$$\Pr \left(Z \in \left[0, \frac{\sqrt{n}|z_1|}{\sigma(f)} \right] \right) > \frac{1}{3} \frac{\sqrt{n}|z_1|}{\sigma(f)}$$

whenever $|z_1| \leq 0.99 \frac{\sigma(f)}{\sqrt{n}}$. In conclusion, we demonstrated that in case (a)

$$\mathbb{E}h \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} \right) > \frac{2e^{-2}}{3} |z_1| \frac{\sqrt{n}}{\sigma(f)} > 0.09 |z_1| \frac{\sqrt{n}}{\sigma(f)}.$$

Combining results (1) and (2) for both terms in the maximum (4.11), we see that for any $\Delta > 0$,

$$\mathbb{E}\rho' \left(\frac{W(f) - \sqrt{n}z_1}{\Delta} \right) > \min(0.1364, 0.09) |z_1| \frac{\sqrt{n}}{\max(\Delta, \sigma(f))} = 0.09 |z_1| \frac{\sqrt{n}}{\max(\Delta, \sigma(f))} \quad (4.14)$$

given that $|z_1| \leq \frac{1}{2} \frac{\max(\Delta, \sigma(f))}{\sqrt{n}}$. Let $\varepsilon > 0$. It is easy to check that setting

$$z_1 = -\frac{1}{0.09} \max(\Delta, \sigma(f)) \frac{\varepsilon}{\sqrt{n}}$$

yields, in view of (4.14), that

$$\mathbb{E}\rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) > \varepsilon,$$

as long as condition $|z_1| \leq \frac{1}{2} \frac{\max(\Delta, \sigma(f))}{\sqrt{n}}$ holds for all j . The latter is equivalent to requirement that $\varepsilon \leq \frac{0.09}{2}$. $\square$

4.2. Proof of Theorem 2.2.

Suppose z_1, z_2 are such that on an event of probability close to 1, $U'_{N,n}(z_1; f) > 0$ and $U'_{N,n}(z_2; f) < 0$ for all $f \in \mathcal{F}$ simultaneously. For such z_1, z_2 , it is easy to see that $\tilde{\theta}^{(k)}(f) - Pf \in (z_1, z_2)$ for all $f \in \mathcal{F}$ on the corresponding event. Observe that

$$\begin{aligned} U'_{N,n}(z; f) &= \frac{1}{N!} \sum_{(i_1, \dots, i_N) \in \pi_N} T_{i_1, \dots, i_N}(z; f) - \mathbb{E} T_{i_1, \dots, i_N}(z; f) \\ &\quad + \mathbb{E}\rho' \left(\sqrt{n} \frac{\tilde{\theta}(f; \{1, \dots, n\}) - Pf - z}{\Delta} \right) - \mathbb{E}\rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) + \mathbb{E}\rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right). \end{aligned}$$

The rest of the proof mimics the steps in the proof of Theorem 2.1. Namely, we will find positive $\varepsilon_1, \varepsilon_2$ such that

$$\inf_{f \in \mathcal{F}} \frac{1}{N!} \sum_{(i_1, \dots, i_N) \in \pi_N} T_{i_1, \dots, i_N}(z; f) - \mathbb{E} T_{i_1, \dots, i_N}(z; f) = \inf_{f \in \mathcal{F}} (U'_{N,n}(z; f) - \mathbb{E} U'_{N,n}(z; f)) \geq -\varepsilon_1 \quad (4.15)$$

with high probability and

$$\inf_{f \in \mathcal{F}} \mathbb{E}\rho' \left(\sqrt{n} \frac{\tilde{\theta}(f; \{1, \dots, n\}) - Pf - z}{\Delta} \right) - \mathbb{E}\rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \geq -\varepsilon_2$$

and will choose z such that $\mathbb{E}\rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) > \varepsilon_1 + \varepsilon_2$.

In view of (4.5) and the fact that $\|\rho'\|_\infty \leq 2$, ε_1 can be chosen as

$$\varepsilon_1 = 2\mathbb{E} \sup_{f \in \mathcal{F}} \left(T_{1, \dots, N}(z; f) - \mathbb{E} T_{1, \dots, N}(z; f) \right) + \frac{\sigma(f)}{\Delta} \sqrt{\frac{2s}{k}} + \frac{16s}{3k}$$

for which inequality (4.15) holds with probability at least $1 - e^{-s}$. It is also easy to see, following the symmetrization-contraction argument of Lemma 4.1, that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(T_{1, \dots, N}(z; f) - \mathbb{E} T_{1, \dots, N}(z; f) \right) \leq \frac{8}{\Delta \sqrt{k}} \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{nk}} \sum_{j=1}^{nk} (f(X_j) - Pf) \right|.$$

Lemma 4.2 implies that

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E} \rho' \left(\sqrt{n} \frac{\bar{\theta}(f; \{1, \dots, n\}) - Pf - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right| \leq 2 \sup_{f \in \mathcal{F}} G_f(n, \Delta) := \varepsilon_2.$$

Finally, equation (4.14) in the proof of Lemma 4.3 yields that, for $z < 0$,

$$\mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) > 0.09|z| \frac{\sqrt{n}}{\max(\Delta, \sigma(f))}$$

as long as $|z| \leq \frac{1}{2} \frac{\max(\Delta, \sigma(f))}{\sqrt{n}}$. Hence, choosing $z_1 := -\frac{1}{0.09} \frac{\max(\Delta, \sigma(\mathcal{F}))}{\sqrt{n}} (\varepsilon_1 + \varepsilon_2)$ implies that $U'_{N,n}(z_1; f) > 0$ for all $f \in \mathcal{F}$ simultaneously with probability at least $1 - e^{-s}$. Similarly, $z_2 = -z_1$ satisfies $U'_{N,n}(z_2; f) < 0$ for all $f \in \mathcal{F}$ with the same probability, and the claim follows.

4.3. Proof of Theorem 3.1.

The proof closely follows the steps of the proof of Theorem 2.1. All the probabilities below are evaluated conditionally on N_J (see section 3 for the definition). Recall that

$$G_k(z; f) = \frac{1}{\sqrt{k}} \sum_{j=1}^k \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right).$$

We are looking for $z_1, z_2 \in \mathbb{R}$ with $|z_1|, |z_2|$ as small as possible such that on an event of probability close to 1, $G_k(z_1; f) > 0$ and $G_k(z_2; f) < 0$ for all $f \in \mathcal{F}$ simultaneously. Observe that

$$G_k(z; f) = \frac{1}{\sqrt{k}} \sum_{j \in J} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) + \frac{1}{\sqrt{k}} \sum_{j \notin J} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right).$$

The second sum can be estimated as

$$\left| \frac{1}{\sqrt{k}} \sum_{j \notin J} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right| \leq \|\rho'\|_\infty \sum_{j \notin J} \frac{1}{\sqrt{k}} \leq 2 \frac{\mathcal{O}}{\sqrt{k}}.$$

where we used the fact that $\|\rho'\|_\infty \leq 2$. For the first sum, we proceed as in the proof of Theorem 2.1 and decompose it as

$$\begin{aligned} & \frac{1}{\sqrt{k}} \sum_{j \in J} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \\ &= \frac{1}{\sqrt{k}} \sum_{j \in J} \left(\rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \\ & \quad + \frac{1}{\sqrt{k}} \sum_{j \in J} \left(\mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right) \\ & \quad + \frac{1}{\sqrt{k}} \sum_{j \in J} \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right). \end{aligned}$$

It follows from Lemma 4.1 that for any $z \in \mathbb{R}$,

$$\begin{aligned} & \inf_{f \in \mathcal{F}} \frac{1}{\sqrt{k}} \sum_{j \in J} \left(\rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \\ &= \sqrt{\frac{|J|}{k}} \inf_{f \in \mathcal{F}} \frac{1}{\sqrt{|J|}} \sum_{j \in J} \left(\rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) \right) \geq -\varepsilon_1 \end{aligned}$$

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

$$\varepsilon_1 = \sqrt{\frac{|J|}{k}} \left(\frac{8}{\Delta \sqrt{N_J}} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^{N_J} (f(X_j) - Pf) + \frac{\sigma(\mathcal{F})}{\Delta} \sqrt{2s} + \frac{16}{3} \frac{s}{\sqrt{|J|}} \right).$$

Next, Lemma 4.2 implies that

$$\begin{aligned} & \inf_{f \in \mathcal{F}} \sum_{j \in J} \frac{1}{\sqrt{k}} \left(\mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right) \\ &= \sqrt{\frac{|J|}{k}} \inf_{f \in \mathcal{F}} \sum_{j \in J} \frac{1}{\sqrt{|J|}} \left(\mathbb{E} \rho' \left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E} \rho' \left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right) \leq \varepsilon_2, \end{aligned}$$

where $\varepsilon_2 = 2 \frac{|J|}{\sqrt{k}} \sup_{f \in \mathcal{F}} G_f(n, \Delta)$. Finally we will choose $z_1 < 0$ such that for all $f \in \mathcal{F}$,

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

Lemma 4.3 implies that it suffices to take

$$z_1 = -\frac{1}{0.09} \frac{\tilde{\Delta}}{\sqrt{n}} \left(\varepsilon_1 + \varepsilon_2 + 2 \frac{\mathcal{O}}{\sqrt{k}} \right) \frac{\sqrt{k}}{|J|} \geq -C \frac{\tilde{\Delta}}{\sqrt{n}} \frac{\varepsilon_1 + \varepsilon_2 + \mathcal{O}/\sqrt{k}}{\sqrt{k}}.$$

To get the final form of the bound, observe that $\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^{N_J} (f(X_j) - Pf) \leq \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^N (f(X_j) - Pf)$ due to Jensen's inequality and that $N_J \geq N/2$ by assumption. Similarly, setting $z_2 = -z_1$ guarantees that $G_k(z_2; f) < 0$ for all $f \in \mathcal{F}$ with probability at least $1 - e^{-s}$, hence the claim follows.

4.4. Proof of Lemma 2.1.

For brevity, set $Y := f(X) - Pf$, let $Y_1, \dots, Y_n$ be i.i.d. copies of Y , and let $W(f)$ have normal distribution $N(0, \sigma^2(f))$. Theorem 2.2 in [8] implies that for an absolute constant $C > 0$ and any $t \in \mathbb{R}$,

$$\begin{aligned} & \left| \Pr \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n Y_j \geq t \right) - \Pr(W(f) \geq t) \right| \leq g_f(t, n) \\ &= C \left(\frac{\mathbb{E} Y^2 I \left\{ \frac{|Y|}{\sigma \sqrt{n}} > 1 + |t/\sigma| \right\}}{\sigma^2 (1 + |t/\sigma|)^2} + \frac{1}{\sqrt{n}} \frac{\mathbb{E} |Y|^3 I \left\{ \frac{|Y|}{\sigma \sqrt{n}} \leq 1 + |t/\sigma| \right\}}{\sigma^3 (1 + |t/\sigma|)^3} \right). \quad (4.16) \end{aligned}$$

It is clear that $g_f(s, n) < g_f(t, n)$ for $t < s$, $g_f(t, m) < g_f(t, n)$ for $n < m$, and that $g_f(t, n) \rightarrow 0$ as $|t| \rightarrow \infty$. To show that $g_f(t, n)$ converges to 0 as $n \rightarrow \infty$, let $\{a_n\}_{n \geq 1}$ be any sequence such that $a_n \rightarrow \infty$, $a_n \leq \sqrt{n}$ and $a_n = o(\sqrt{n})$. Then

$$\begin{aligned} & \frac{1}{\sqrt{n}} \mathbb{E} |Y|^3 I \{ |Y| \leq \sigma \sqrt{n} (1 + |t/\sigma|) \} \leq \frac{1}{\sqrt{n}} \left(\mathbb{E} |Y|^3 I \{ |Y| \leq \sigma \cdot a_n (1 + |t/\sigma|) \} \right. \\ & \quad \left. + \mathbb{E} |Y|^3 I \{ \sigma \cdot a_n (1 + |t/\sigma|) \leq |Y| \leq \sigma \cdot \sqrt{n} (1 + |t/\sigma|) \} \right) \\ & \leq (\sigma + |t|) \left(\frac{a_n}{\sqrt{n}} \mathbb{E} Y^2 + \mathbb{E} Y^2 I \{ |Y| \geq \sigma \cdot a_n (1 + |t/\sigma|) \} \right), \end{aligned}$$

hence

$$g_f(t, n) \leq \frac{C}{(\sigma + |t|)^2} \left(\mathbb{E} Y^2 I\{|Y| > a_n(\sigma + |t|)\} + \frac{a_n}{\sqrt{n}} \mathbb{E} Y^2 \right)$$

where the latter expression converges to 0 as $n \rightarrow \infty$. Next, assume that $\mathbb{E}|Y|^{2+\delta} < \infty$ for some $\delta \in (0, 1]$. Applying Hölder's inequality followed by Markov's inequality, we deduce that

$$\begin{aligned} \mathbb{E} \left(\frac{X^2}{\sigma^2} I \left\{ \frac{|X|}{\sigma\sqrt{n}} > 1 + |t/\sigma| \right\} \right) &\leq (\mathbb{E}|X/\sigma|^{2+\delta})^{2/(2+\delta)} \left(\Pr \left(\frac{|X|}{\sigma\sqrt{n}} > 1 + |t/\sigma| \right) \right)^{\delta/(2+\delta)} \\ &\leq \frac{\mathbb{E}|X/\sigma|^{2+\delta}}{n^{\delta/2}(1 + |t/\sigma|)^{\delta}}. \end{aligned}$$

Moreover, it is clear that

$$\mathbb{E} \left(\frac{|X|^3}{\sigma^3} I \left\{ \frac{|X|}{\sigma\sqrt{n}} \leq 1 + |t/\sigma| \right\} \right) \leq (\sqrt{n}(1 + |t/\sigma|))^{1-\delta} \mathbb{E}|X/\sigma|^{2+\delta}.$$

Combining these inequalities with (4.16), we deduce the bound (2.3) for $g_f(t, n)$. Finally, an upper bound for $G_f(n)$ follows by integrating the inequality (2.3).

5. Supplementary results.

Lemma 5.1. *Assume that $0 \leq \alpha \leq 0.33$ and let $z(\alpha)$ be such that $\Phi(z(\alpha)) - 1/2 = \alpha$. Then $z(\alpha) \leq 3\alpha$.*

Proof. It is a simple numerical fact that whenever $\alpha \leq 0.33$, $z(\alpha) \leq 1$; indeed, this follows as $\Phi(1) \approx 0.8413 > 1/2 + 0.33$. Since $e^{-y^2/2} \geq 1 - \frac{y^2}{2}$, we have

$$\sqrt{2\pi}\alpha = \int_0^{z(\alpha)} e^{-y^2/2} dy \geq z(\alpha) - \frac{1}{6} (z(\alpha))^3 \geq \frac{5}{6} z(\alpha), \quad (5.1)$$

Equation (5.1) implies that $z(\alpha) \leq \frac{6}{5}\sqrt{2\pi}\alpha$. Proceeding again as in (5.1), we see that

$$\sqrt{2\pi}\alpha \geq z(\alpha) - \frac{1}{6} (z(\alpha))^3 \geq z(\alpha) - \frac{12\pi}{25} \alpha^2 z(\alpha) \geq z(\alpha) (1 - 1.51\alpha^2),$$

hence $z(\alpha) \leq \frac{\sqrt{2\pi}}{1-1.51\alpha^2} \alpha$. The claim follows since $\alpha \leq 0.33$ by assumption, and $\frac{\sqrt{2\pi}}{1-1.51\alpha^2} < 3$. $\square$

The following fact is well known; we present a short proof for reader's convenience.

Lemma 5.2. *Let $\mathcal{A} \subset \mathbb{R}$ be symmetric, meaning that $\mathcal{A} = -\mathcal{A}$, and let $Z \sim N(0, 1)$. Then for all $x \in \mathbb{R}$,*

$$\Pr(Z \in \mathcal{A} - x) \geq e^{-x^2/2} \Pr(Z \in \mathcal{A}).$$

Proof. Observe that

$$\begin{aligned} \Pr(Z \in \mathcal{A}) &= \int_{\mathbb{R}} I\{z \in \mathcal{A}\} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = e^{x^2/2} \int_{\mathbb{R}} I\{z \in \mathcal{A}\} e^{-xz/2} e^{xz/2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} e^{-x^2/2} dz \\ &\leq e^{x^2/2} \sqrt{\int_{\mathbb{R}} I\{z \in \mathcal{A}\} \frac{1}{\sqrt{2\pi}} e^{-(z-x)^2/2} dz} \sqrt{\int_{\mathbb{R}} I\{z \in \mathcal{A}\} \frac{1}{\sqrt{2\pi}} e^{-(z+x)^2/2} dz} \\ &= e^{x^2/2} \int_{\mathbb{R}} I\{z \in \mathcal{A}\} \frac{1}{\sqrt{2\pi}} e^{-(z-x)^2/2} dz = e^{x^2/2} \Pr(Z \in \mathcal{A} - x), \end{aligned}$$

and the claim follows. $\square$

Lemma 5.3. *Let ρ satisfy Assumption 1. Then for any random variable Y with $\mathbb{E}Y^2 < \infty$,*

$$\text{Var}(\rho'(Y)) \leq \text{Var}(Y).$$

Proof. The function $u(t) = t - \rho'(t) - c$ is nondecreasing for any $c \in \mathbb{R}$ by Assumption 1, hence for any $s \in \mathbb{R}$, there exists $t(s)$ such that $u(t) \geq s$ iff $t \geq t(s)$. Next, consider $\tilde{u}(t) = t - \mathbb{E}Y - (\rho'(t) - \mathbb{E}\rho'(Y))$, and let t_0 be such that $\tilde{u}(t) \geq 0$ for $t \geq t_0$ and $\tilde{u}(t) \leq 0$ when $t \leq t_0$. For any $x, y \in \mathbb{R}$, $x^2 - y^2 \geq 2y(x - y)$. Letting $x = t - \mathbb{E}Y$ and $y = \rho'(t) - \mathbb{E}\rho'(Y)$, we obtain

$$(t - \mathbb{E}Y)^2 - (\rho'(t) - \mathbb{E}\rho'(Y))^2 \geq 2(\rho'(t) - \mathbb{E}\rho'(Y))(t - \mathbb{E}Y - (\rho'(t) - \mathbb{E}\rho'(Y))).$$

Consider two cases: (a) $t \geq t_0$ and (b) $t < t_0$. In the first case, $\rho'(t) - \mathbb{E}\rho'(Y) \geq \rho'(t_0) - \mathbb{E}\rho'(Y)$ and $(t - \mathbb{E}Y - (\rho'(t) - \mathbb{E}\rho'(Y))) \geq 0$, hence

$$(t - \mathbb{E}Y)^2 - (\rho'(t) - \mathbb{E}\rho'(Y))^2 \geq 2(\rho'(t_0) - \mathbb{E}\rho'(Y))(t - \mathbb{E}Y - (\rho'(t) - \mathbb{E}\rho'(Y))).$$

In the second case, $\rho'(t) - \mathbb{E}\rho'(Y) \leq \rho'(t_0) - \mathbb{E}\rho'(Y)$ and $(t - \mathbb{E}Y - (\rho'(t) - \mathbb{E}\rho'(Y))) \leq 0$, hence again

$$(t - \mathbb{E}Y)^2 - (\rho'(t) - \mathbb{E}\rho'(Y))^2 \geq 2(\rho'(t_0) - \mathbb{E}\rho'(Y))(t - \mathbb{E}Y - (\rho'(t) - \mathbb{E}\rho'(Y))). \quad (5.2)$$

Replacing t by Y in (5.2) and taking the expectation yields the result. $\square$

Lemma 5.4. *Let $p \in [2, 3]$, and assume that $X \in \mathbb{R}^d$ has distribution P from a class $\mathcal{P}$ of all distributions with bounded p -th moments of one-dimensional projections, meaning that $\sup_{\|v\|_2=1} \mathbb{E}_P |\langle X - \mathbb{E}X, v \rangle|^p \leq 1$. Let $X_1, \dots, X_N$ be a sample from $(1 - \varepsilon)P + \varepsilon Q$ where $0 \leq \varepsilon < 1/2$, $P \in \mathcal{P}$ and Q is an arbitrary distribution. Then for any estimator $\hat{\mu}(X_1, \dots, X_N)$ of the mean $\mu(P)$,*

$$\sup_{P \in \mathcal{P}, Q} \Pr \left(\|\hat{\mu} - \mu\|_2 \geq c_1 \left(\sqrt{\frac{\text{tr}(\Sigma)}{N}} \vee \varepsilon^{1-1/p} \right) \right) \geq c_2.$$

where Σ is the covariance matrix of P and c_1, c_2 are absolute constants.

Proof. If $\varepsilon = 0$, then it is well known that the minimax rate of estimating the mean is $\sqrt{\frac{\text{tr} \Sigma}{N}}$ (e.g. see the remark following Theorem 3 in [27]). Let's assume that $\varepsilon > 0$ and suppose that P_1, P_2 are two distributions supported on $\{\varepsilon^{-1/p}, 0, -\varepsilon^{1/p}\}$, namely, $P_1(0) = P_2(0) = \frac{1-2\varepsilon}{1-\varepsilon}$, $P_1(\varepsilon^{-1/p}) = \frac{\varepsilon}{1-\varepsilon}$, $P_2(-\varepsilon^{-1/p}) = 0$, $P_2(\varepsilon^{-1/p}) = 0$, $P_2(-\varepsilon^{-1/p}) = \frac{\varepsilon}{1-\varepsilon}$. Clearly, $P_1 \in \mathcal{P}$, $P_2 \in \mathcal{P}$, the means of P_1, P_2 are $\mu(P_1) = \frac{\varepsilon^{1-1/p}}{1-\varepsilon}$, $\mu(P_2) = -\frac{\varepsilon^{1-1/p}}{1-\varepsilon}$ respectively and $|\mu(P_1) - \mu(P_2)| = \frac{2\varepsilon^{1-1/p}}{1-\varepsilon}$.

Next, let Q_1 and Q_2 be Dirac measures, namely $Q_1 = \delta_{\varepsilon^{-1/p}}$ and $Q_2 = \delta_{-\varepsilon^{-1/p}}$. Then it is easy to check that $(1 - \varepsilon)P_1 + \varepsilon Q_2 = (1 - \varepsilon)P_2 + \varepsilon Q_1 = \hat{P}$, hence, given samples from $\hat{P}$, it is impossible to distinguish between P_1 and P_2 . Conclusion now follows from Theorem 5.1 in [9]. $\square$

Acknowledgements.

Part of this work was completed during my visit to ENSAE ParisTech. I would like to thank Faculty members and students at ENSAE for their hospitality.

References

- [1] ADAMCZAK, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** no. 34, 1000–1034.
- [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] BARTLETT, P. L., BOUSQUET, O., MENDELSON, S. et al. (2005). Local Rademacher complexities. *The Annals of Statistics* **33** 1497–1537.
- [4] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.
- [5] BROWNLEES, C., JOLY, E., LUGOSI, G. et al. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics* **43** 2507–2536.

- [6] 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.
- [7] CATONI, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*.
- [8] 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.
- [9] CHEN, M., GAO, C., REN, Z. et al. (2018). Robust covariance and scatter matrix estimation under Huber’s contamination model. *The Annals of Statistics* **46** 1932–1960.
- [10] DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *The Annals of Probability* 899–929.
- [11] DUDLEY, R. M. (2014). *Uniform central limit theorems* **142**. Cambridge university press.
- [12] GEER, S. A. (2000). *Empirical Processes in M-estimation* **6**. Cambridge university press.
- [13] GINÉ, E. and NICKL, R. (2015). *Mathematical foundations of infinite-dimensional statistical models* **40**. Cambridge University Press.
- [14] GIULINI, I. et al. (2018). Robust dimension-free Gram operator estimates. *Bernoulli* **24** 3864–3923.
- [15] HAUSSLER, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A* **69** 217–232.
- [16] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* 293–325.
- [17] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* **58** 13–30.
- [18] HUBER, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35** 73–101.
- [19] JOLY, E., LUGOSI, G., OLIVEIRA, R. I. et al. (2017). On the estimation of the mean of a random vector. *Electronic Journal of Statistics* **11** 440–451.
- [20] KOLTCHINSKII, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Springer Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d’Été de Probabilités de Saint-Flour.
- [21] LECUÉ, G. and LERASLE, M. (2017). Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*.
- [22] LECUÉ, G., LERASLE, M. and MATHIEU, T. (2018). Robust classification via MOM minimization. *arXiv preprint arXiv:1808.03106*.
- [23] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: isoperimetry and processes*. Springer-Verlag, Berlin.
- [24] LERASLE, M. and OLIVEIRA, R. I. (2011). Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.
- [25] LUGOSI, G. and MENDELSON, S. (2016). Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*.
- [26] LUGOSI, G. and MENDELSON, S. (2017). Regularization, sparse recovery, and median-of-means tournaments. *arXiv preprint arXiv:1701.04112*.
- [27] LUGOSI, G. and MENDELSON, S. (2018). Near-optimal mean estimators with respect to general norms. *arXiv preprint arXiv:1806.06233*.
- [28] McDIARMID, C. (1989). On the method of bounded differences. *Surveys in combinatorics* **141** 148–188.
- [29] MEDINA, A. M. and MOHRI, M. (2014). Learning theory and algorithms for revenue optimization in second price auctions with reserve. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* 262–270.
- [30] MINSKER, S. and STRAWN, N. (2017). Distributed Statistical Estimation and Rates of Convergence in Normal Approximation. *arXiv preprint arXiv:1704.02658*.
- [31] NEMIROVSKI, A. and YUDIN, D. (1983). *Problem complexity and method efficiency in optimization*. John Wiley & Sons Inc.
- [32] POLLARD, D. (2000). Asymptopia: an exposition of statistical asymptotic theory. Available at <http://www.stat.yale.edu/~pollard/Books/Asymptopia>.
- [33] TALAGRAND, M. (2014). *Upper and lower bounds for stochastic processes: modern methods and classical problems* **60**. Springer Science & Business Media.

- [34] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak convergence and empirical processes. Springer Series in Statistics.* Springer-Verlag, New York.