Doctoral Thesis

Robustness via trimmed sample means

Candidate: Lucas Resende
Advisor: Roberto Imbuzeiro Oliveira

Instituto de Matemática Pura e Aplicada

Rio de Janeiro, February, 2024.

Abstract

This thesis investigates the applications and properties of trimmed sample means in the context of robust estimation under heavy-tailed and contaminated data. We study four problems: uniform mean estimation, regression with quadratic risk, Gaussian and bootstrap approximations and vector mean estimation under arbitrary norms.

In the problem of uniform mean estimation we manage to obtain the best known bounds for this problem, with minimax-optimal dependence on moment parameters and contamination level.

Regarding the problem of regression with quadratic risk we also obtain the best known bounds and the best dependence on moment parameters and contamination level. Moreover, we also provide heuristics for robust linear regression and experimental results showing that our method outperforms similar methods available in the literature.

We also obtain Gaussian and bootstrap approximation bounds in a high-dimensional setting under weak assumptions, showing that not only the trimmed mean satisfy high-dimensional Gaussian and bootstrap approximations, but when compared with the sample mean, such approximations hold for a wider class of distributions.

To finish, we apply our results on uniform mean estimation and Gaussian approximation to study the problem of vector mean estimation under arbitrary norms, again improving all known bounds for this problem.

Keywords: sub-Gaussian estimators, trimmed mean, robustness, regression, bootstrap.

Resumo

Esta tese investiga as aplicações e propriedades das médias podadas (*trimmed sample means*) no contexto de estimação robusta com dados provenientes de distribuição de cauda pesada ou mesmo contaminados. Estudamos quatro problemas: estimação uniforme da média, regressão com risco quadrático, aproximações gaussiana e bootstrap, e estimação da média vetorial sob normas arbitrárias.

No problema de estimação uniforme da média, conseguimos obter as melhores cotas conhecidas para este problema, com dependência ótima nos parâmetros de momento e nível de contaminação.

Em relação ao problema de regressão com risco quadrático, também obtemos as melhores cotas e a melhor dependência nos parâmetros de momento e nível de contaminação. Além disso, fornecemos heurísticas para regressão linear robusta e resultados experimentais que mostram que nosso método supera métodos semelhantes disponíveis na literatura.

Obtemos também cotas de aproximação gaussiana e bootstrap em alta dimensão e sob hipóteses fracas, mostrando que não apenas a média podada realiza aproximações gaussiana e bootstrap em alta dimensão, mas, quando comparadas com a média amostral, tais aproximações valem para uma classe mais ampla de distribuições.

Para concluir, aplicamos nossos resultados de estimação uniforme da média e de aproximação gaussiana para estudar o problema de estimação da média de vetores sob normas arbitrárias, novamente melhorando as cotas conhecidas para este problema.

Palavras-chave: estimadores sub-gaussianos, média podada, robustez, regressão, bootstrap.

Contents

Chapter 1

Introduction

The sample mean is probably the most fundamental way of aggregating information in Statistics [Stigler, 2016]. Mathematically, it can be understood as a way to approximate an expected value from a random sample. Aspects of this approximation, including its convergence for large sample sizes, are described by the Law of Large Numbers, the Central Limit Theorem and other Limit Theorems.

In practice, sample means are used in several settings. One is as a way to directly estimate population parameters that correspond to expectations: means, (co)variances and other moments are natural examples.

A second way is in M-estimation. If a population parameter can be expressed as a minimizer of

$$L_P(\theta) := \mathbb{E}_{X\sim P}\, \ell(X, \theta) \;\; (\theta \in \Theta) \tag{1.1}$$

for some suitable function ℓ, it is natural to estimate this parameter by minimizing

$$\widehat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(X_i, \theta) \;\; (\theta \in \Theta) \tag{1.2}$$

instead, where $(X_1, \ldots, X_n)$ is a random sample from P. Much of classical Asymptotic Statistics [van der Vaart et al., 1996] deals with aspects of this approximation.

A somewhat related setting is "statistical learning." Here the goal is to use the random sample to find a good near-minimizer for the function of (1.1). That is, one wants to choose a data-dependent $\widehat{\theta}_n$ that makes the loss L_P small (this corresponds to performance over test data). Minimizing the sample loss (1.2) is the well-known procedure of empirical risk minimization.

Besides its use in mean estimation, M-estimation and "statistical learning," sample means also have well known Gaussian and bootstrap approximations, those are useful in hypothesis testing, designing confidence intervals, selecting penalty parameters, and other problems [Chernozhukov et al., 2023a].

In all of the above settings, the sample mean is used because it is a convenient estimator for the corresponding expectations. However, it is often *not* the best possible estimator for this purpose. This is clearly the case when there is data contamination and a single outlier can change the value of the sample mean completely. Robust Statistics [Huber, 1965, Huber and Ronchetti, 1981] takes this issue as its starting point.

More subtly, the sample mean is also not optimal in terms of a phenomenon not captured by classical (asymptotic) Robust Statistics: its fluctuations over finite samples. Consider, for instance, the basic problem of estimating the expectation of a one-dimensional random variable with variance $\sigma^2 > 0$. Asymptotically, the sample mean is Gaussian, but Catoni's seminal work [Catoni, 2012] showed its non-asymptotic bounds are much worse. Indeed, Catoni showed that Chebyshev's bound is optimal in the following sense: given a mean μ and a variance σ^2, there is a distribution P with mean μ and variance σ^2 from which an i.i.d. sample $X_1, \ldots, X_n$ from P will satisfy, for some absolute constant $c > 0$,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq c\sqrt{\frac{\sigma^2}{n\alpha}}\right] \geq \alpha, \, \forall \alpha \in (0,1).$$

For comparison, when P is sub-Gaussian,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \sqrt{\frac{2\sigma^2}{n}\ln\frac{2}{\alpha}}\right] \leq \alpha, \, \forall \alpha \in (0,1),$$

which is much better for small α.

What is really striking, however, is that there are less obvious alternatives for the sample mean with so-called "sub-Gaussian" finite-sample error bounds [Catoni, 2012, Devroye et al., 2016, Lee and Valiant, 2022]. Such alternatives attain, for a wide range of distributions P, nearly the same fluctuations the sample mean attains for sub-Gaussian distributions. Some of these estimators can also be made robust to contamination [Diakonikolas et al., 2019a].

Catoni's discovery led to a surge of interest in finite-sample mean estimation for vectors [Lugosi and Mendelson, 2019b, Lugosi and Mendelson, 2019c, Lugosi and Mendelson, 2021, Hopkins, 2020, Depersin and Lecué, 2022, Minsker, 2015], matrices [Minsker, 2018a, Mendelson and Zhivotovskiy, 2020, Abdalla and Zhivotovskiy, 2022] and

other objects under (relatively) heavy tails and contamination [Lugosi and Mendelson, 2019a]. The basic thrust behind this work is to use improvements over the sample mean to devise improved statistical methods more generally.

Starting with a breakthrough by Diakonikolas et al. [Diakonikolas et al., 2019a], a parallel line of research in Computer Science has studied computationally efficient robust estimation in high dimensions [Dong et al., 2019, Diakonikolas et al., 2022].There has also been related work for regression and other statistical tasks [Audibert and Catoni, 2011, Brownlees et al., 2015, Mourtada et al., 2021, Lecué and Lerasle, 2020, Diakonikolas et al., 2019b]. In most recent work, contamination means *adversarial contamination*, which we discuss and contrast with Huber's contamination model [Huber, 1965, Huber and Ronchetti, 1981] in §2.1.3.

1.1 The trimmed mean

A natural way to avoid the limitations of the sample mean is to use trimmed means. Suppose one is given $x_{1:n} = (x_1, \ldots, x_n) \in \mathbf{X}^n$ and a function $f : \mathbf{X} \to \mathbb{R}$. If the x_i are distributed according to a probability distribution P over $\mathbf{X}$, the sample mean corresponds to the following approximation:

$$Pf = \mathbb{E}_{X \sim P} f(X) \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i),$$

which has the aforementioned problems. By contrast, for an integer $1 \leq k < \frac{n}{2}$, the k-trimmed mean over $x_{1:n}$ is:

$$\widehat{T}_{n,k}(f, x_{1:n}) := \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} f(x_{(i)}), \tag{1.3}$$

where $(\cdot)$ is a permutation of $[n]$ such that

$$f(x_{(1)}) \leq \cdots \leq f(x_{(n)}).$$

That is, the trimmed mean is the arithmetic mean of the terms that remain once the k largest and k smallest values of $f(x_i)$ have been removed. A large k makes this estimator more robust to outliers, but also introduces some bias.

The trimmed mean is a classical estimator in Robust Statistics in the sense of Huber [Huber and Ronchetti, 1981, Stigler, 2010, Huber, 1972, Stigler, 1973, Jaeckel, 1971, Jana Jurecková, 1994, Hall, 1981]. More recently, variants of the trimmed mean have been used to estimate high dimensional means [Lugosi and Mendelson, 2021] and covariances [Oliveira and Rico, 2022] with non-asymptotic guarantees. The PhD thesis [Rico, 2022] also

proves optimality properties of the trimmed mean for estimating the mean of a single function f [Rico, 2022, Chapter 2].

1.2 Our contribution in this thesis

The main contribution of this thesis is to show that trimmed means improve the state of the art in finite-sample performance in statistical problems. The first of these problems is what we call uniform mean estimation: it is somehow "generic," in that solving it well can help improve the methods for many other problems.

1.2.1 Uniform mean estimation. To motivate this problem, consider the classical problem of bounding the supremum of an empirical process indexed by a family of functions $\mathcal{F}$:

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-Pf\right|.$$

This is a subproblem of many other problems in Statistics. To see one example, let us back to the discussion on statistical learning following (1.1) and (1.2). If $\widehat{\theta}_n$ denote the empirical risk minimizer, it is a classical fact that:

$$L_P(\widehat{\theta}_n)-\inf_{\theta\in\Theta}L_P(\theta)\leq 2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-Pf\right| \tag{1.4}$$

where $\mathcal{F}:=\{\ell(\theta,\cdot)\,:\,\theta\in\Theta\}$.

To continue with this example, for each $\theta\in\Theta$, we replace $\widehat{L}_n(\theta)$ in equation (1.2) by another estimator of $L_P(\theta)$ that we hope is "better." The analogue of (1.4) holds with a change in the right hand side:

$$\sup_{f\in\mathcal{F}}\left|\widehat{E}_f(X_1,\ldots,X_n)-Pf\right|,$$

and it is clear that *each $\widehat{E}_f$ should be designed so as to minimize the above supremum.* In other words: given a family of functions $\mathcal{F}$, we want to estimate the expectations Pf of each $f\in\mathcal{F}$ by $\widehat{E}_f(X_1,\ldots,X_n)$ while minimizing the worst-case error.

Minsker [Minsker, 2018b] seems to have been the first author to explicitly consider this problem. However, it naturally comes up in many settings beyond that of statistical learning. One of these – the case of high dimensional vector mean estimation – will be discussed separately. For now, we mention in passing another example.

Example (Integral probability metrics)**.** Let $\mathcal{F}$ be an arbitrary family of measurable functions from $\mathbf{X}$ to $\mathbb{R}$, and let $\Delta_{\mathcal{F}}$ denote the set of all probability distributions over $\mathbf{X}$ that integrate all $f \in \mathcal{F}$. $\mathcal{F}$ defines an integral probability (semi)metric over $\Delta_{\mathcal{F}}$ via the recipe:

$$D_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} |(Q - P)f| \ \ (P, Q \in \Delta_{\mathcal{F}}).$$

Popular examples of such metrics include the L^1 Wassertein metric and kernel-based metrics discussed in [Sriperumbudur et al., 2012, Sriperumbudur, 2016]. Uniform estimates for Qf and Pf naturally lead to estimates on $D_{\mathcal{F}}(P, Q)$.

Our main result on uniform mean estimation – Theorem 3.1 in Chapter 3 – shows that trimmed means give the best known bounds for this problem in the adversarial contamination setting. In particular, we improve the main result of [Minsker, 2018b] and obtain minimax-optimal dependence on moment parameters and the contamination level.

1.2.2 Regression with quadratic risk: theory and heuristics.

Our methods for uniform mean estimation also lead to new results on regression with quadratic risk. In this problem, the goal is to find a function $f \in \mathcal{F}$ making $\mathbb{E}_{(X,Y)\sim P}(Y - f(X))^2$ as small as possible.

Under suitable technical conditions, we show that a trimmed-mean-based regression method achieves optimal dependence on the contamination level and on moment parameters. This method improves on previous work by Lugosi and Mendelson [Lugosi and Mendelson, 2019c] and Lecué and Lerasle [Lecué and Lerasle, 2020]. One of our findings is that "localization" arguments from previous work extend to trimmed means.

Unfortunately, our statistical method is not computationally efficient. However, it is relatively straightforward to design trimmed-mean-based heuristics for robust linear regression. Experiments in the main text and the appendix show that our algorithm outperforms a similar method based on the median-of-means principle put forward by [Lecué and Lerasle, 2020]. These experiments also give insights on how to optimize the performance of these heuristics.

1.2.3 Gaussian and bootstrap approximations.

Classical non-asymptotic versions of the multidimensional central limit theorem, such as Berry-Essen, allow for a polynomial dependence between the dimension d and the sample size n. Even so, high-dimensional situations where d grows exponentially in n have become common in many practical domains (*e.g.* Genomics). Since the seminal work of [Chernozhukov et al., 2013], this problem has been

studied and continuously improving results have appeared in the literature providing Gaussian and bootstrap approximation bounds for the sample mean on this high-dimensional setup.

In Chapter 5 we investigate the Gaussian and bootstrap approximation properties of the trimmed mean in this high-dimensional setup, showing that it requires less restrictive assumptions to hold and thus is possible in a wider class of distributions, even allowing for adversarial contamination. We also extend our results to the infinite-dimensional case under mild assumptions on the VC dimension of the class of functions.

Besides extending the usage of trimmed means to obtain robust versions of hypothesis testing, confidence intervals and other classical applications of Gaussian and bootstrap approximations, our results also apply to the problem of vector mean estimation under general norms, as discussed in Chapter 6.

1.2.4 Vector mean estimation under arbitrary norms.

In this problem, we assume that we have a (potentially corrupted) i.i.d sample $X_1, \ldots, X_n$ from a high-dimensional distribution P over $\mathbb{R}^d$, whose mean μ_P we want to estimate. The error in our estimate will be measured by an arbitrary norm $\|\cdot\|$.

It turns out that this problem is closely connected to uniform mean estimation, as already noted by Minsker [Minsker, 2018b]. Using this connection, we present in Chapter 6 some results on this problem. More precise, the estimator we present there improves all known bounds for mean estimation under general norms [Depersin and Lecué, 2022, Lugosi and Mendelson, 2019b]. In the special setting of Euclidean norm with finite second moments, our result matches the optimal bound of Lugosi and Mendelson [Lugosi and Mendelson, 2021]. Unfortunately, the estimator we devise is not computationally efficient.

Chapter 2

Definitions and preliminary lemmata

2.1 Main definitions

2.1.1 Basics. $\mathbb{N} = \{1, 2, 3, \dots\}$ is the set of positive integers. For $n \in \mathbb{N}$, define $[n] := \{1, 2, \dots, n\}$. The cardinality of a finite set S is denoted by $|S|$.

2.1.2 Probabilities, moments and samples. Given a probability space $(\mathbf{Z}, \mathcal{Z}, P)$, we write $Z \sim P$ to denote that Z is a random element of $(\mathbf{Z}, \mathcal{Z})$ with distribution P. For $p \geq 1$, we write $L^p(P) = L^p(\mathbf{Z}, \mathcal{Z}, P)$ for the corresponding L^p space. If $f : \mathbf{Z} \to \mathbb{R}$ is measurable, we use Pf, $Pf(Z)$ or $\mathbb{E}_{Z \sim P} f(Z)$ to denote the expectation (integral) of f according to P. Moreover, we let $\widehat{P}_n$ denote the empirical measure, meaning:

$$\widehat{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \text{ for } f \in \mathcal{F}.$$

Given $n \in \mathbb{N}$, the elements of $\mathbf{Z}^n$ are denoted by $z_{1:n} = (z_1, z_2, \dots, z_n)$. We write

$$Z_{1:n} = (Z_1, \dots, Z_n) \overset{i.i.d.}{\sim} P$$

if the Z_i are independent and identically distributed (i.i.d.) random elements of $(\mathbf{Z}, \mathcal{Z})$ with common law P.

2.1.3 Adversarial contamination. Given $n \in \mathbb{N}$ and $(\mathbf{Z}, \mathcal{Z}, P)$ as above, and also a parameter $\varepsilon \in [0, 1)$, a random element $Z^{\varepsilon}_{1:n}$ of $\mathbf{Z}^n$ is an ε-contaminated i.i.d. sample from P

if the following condition holds:

$$\text{there exist } Z_{1:n} \overset{i.i.d}{\sim} P \text{ such that } |\{i \in [n] \,:\, Z_i^\varepsilon \neq Z_i\}| \leq \varepsilon\, n.$$

This is what is called *adversarial contamination* in the CS literature [Diakonikolas et al., 2019a]. Intuitively, it corresponds to a setting where an adversary inspects the i.i.d. sample $Z_{1:n}$ and then decides how to replace a fraction ε of sample points so as to make things more difficult for the statistician. This is less favorable than Huber's more classical model [Huber, 1965, Huber and Ronchetti, 1981] where P is assumed known, but a fraction of sample points may come from a different, unknown distribution. For more work on adversarial contamination, see e.g. [Depersin and Lecué, 2021, Lecué and Lerasle, 2020, Lugosi and Mendelson, 2021, Diakonikolas et al., 2019b, Diakonikolas et al., 2022, Diakonikolas et al., 2019a].

2.1.4 Compatible measures and empirical processes.

We need a technical condition to ensure the various suprema we consider are well defined. Given $p \geq 1$, we say that a probability measure P over $(\mathbf{Z}, \mathcal{Z})$ and a family $\mathcal{F}$ of $\mathcal{Z}$-measurable functions from $\mathbf{Z}$ to $\mathbb{R}$ are p-compatible if $\mathcal{F} \subset L^p(P)$ and there exists a countable subset $\mathcal{D} \subset \mathcal{F}$ such that any $f \in \mathcal{F}$ is the limit of a sequence in $\mathcal{D}$ that converges pointwise and in $L^p(P)$ norm.

For 1-compatible $\mathcal{F}$ and P as above, and exponents $p \geq 1$, we define the following (potentially infinite) moment quantities:

$$\nu_p(\mathcal{F}, P) := \sup_{f \in \mathcal{F}} (P|f - Pf|^p)^{\frac{1}{p}} = \sup_{f \in \mathcal{F}} \|f - Pf\|_{L^p(P)}. \tag{2.1}$$

We also define the expectation of the empirical process indexed by $\mathcal{F}$,

$$\mathsf{Emp}_n(\mathcal{F}, P) := \mathbb{E}_{Z_{1:n} \overset{i.i.d.}{\sim} P} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Z_i) - Pf \right| \right] \tag{2.2}$$

and the Rademacher complexity,

$$\mathsf{Rad}_n(\mathcal{F}, P) := \mathbb{E}_{\substack{\epsilon_{1:n} \overset{i.i.d.}{\sim} U(\{-1,1\}) \\ Z_{1:n} \overset{i.i.d.}{\sim} P}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(Z_i) \right| \right], \tag{2.3}$$

where it is implicit in the expectation that the $Z_{1:n}$ and $\epsilon_{1:n}$ are independent. Definitions (2.2) and (2.3) are related by the first part of the following classical results:

Theorem 2.1. *Assume that a measure P over $(\mathbf{X}, \mathcal{X})$ is 1-compatible with a family of measurable functions $\mathcal{G}$ from $\mathbf{X}$ to $\mathbb{R}$.*

1. (Symmetrization and Desymmetrization; [Boucheron et al., 2013, Lemma 11.4]) $\mathsf{Emp}_n(\mathcal{G}) \leq 2\mathsf{Rad}_n(\mathcal{G})$. *Moreover, if* $Pg = 0$ *for all* $g \in \mathcal{G}$, *then* $\mathsf{Rad}_n(\mathcal{G}) \leq 2\mathsf{Emp}_n(\mathcal{G})$.

2. (Ledoux-Talagrand contraction; [Boucheron et al., 2013, Theorem 11.6]) *If* $\tau : \mathbb{R} \to \mathbb{R}$ *is* L*-Lipschitz and satisfies* $\tau(0) = 0$, *and we set* $\tau \circ \mathcal{G} := \{\tau \circ g \,:\, g \in \mathcal{G}\}$, *then* $\mathsf{Rad}_n(\tau \circ \mathcal{G}) \leq L\,\mathsf{Rad}_n(\mathcal{G})$.

2.2 Trimming and truncation

In this section, P is a fixed probability measure over a measurable space $(\mathbf{X}, \mathcal{X})$, and $X_{1:n}^{\varepsilon}$ is an ε-contaminated sample from P. For a given $M > 0$ define the truncation function

$$\tau_M(x) := \begin{cases} M, \text{ if } x > M \\ x, \text{ if } -M \leq x \leq M \\ -M, \text{ if } x < -M \end{cases} . \tag{2.4}$$

The core observation behind all problems studied in this thesis is that with high probability under a suitable choice of the trimming level k the trimmed mean of the contaminated sample is uniformly close (for all $g \in \mathcal{G}$) to the empirical mean on the clean sample of a truncated version of $\mathcal{G}$. In other words, for a certain choice of k and M,

$$\sup_{g \in \mathcal{G}} \left| \widehat{T}_{n,k}^{\varepsilon}(g) - \widehat{P}_n(\tau_M \circ g) \right| \text{ is small.} \tag{2.5}$$

This observation is a consequence of the two lemmata that we state next. Let $\mathcal{G}$ be a family of functions and $M > 0$, define

$$V_M(\mathcal{G}) \;:=\; \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \mathbf{1}\{|g(X_i)| > M\}, \tag{2.6}$$

in words, $V_M(\mathcal{G})$ counts the number of large values of g for the worst-case function $g \in \mathcal{G}$.

The following "Counting lemma" – an abstract version of [Lugosi and Mendelson, 2021, Lemma 1] – gives a way to bound the probability that the counting function $V_M(\mathcal{G})$ exceeds a certain value t.

Lemma 2.2 (Counting lemma)**.** *Let* $\mathcal{G}$ *be a countable family of functions,* $t \in \mathbb{N}$ *and* $n > 1$. *Assume* $M > 0$ *is such that:*

$$\sup_{g \in \mathcal{G}} P\left\{ |g(X)| > \frac{M}{2} \right\} + \frac{8\mathsf{Rad}_n(\tau_M \circ \mathcal{G}, P)}{M} \leq \frac{t}{8n}. \tag{2.7}$$

Then $V_M(\mathcal{G}) \leq t$ with probability at least $1 - e^{-t}$.

The next lemma says that if $V_M(\mathcal{G})$ is small, then (2.5) can be justified.

Lemma 2.3 (Bounding lemma)**.** *Let $\mathcal{G}$ be a family of functions from $\mathbf{X}$ to $\mathbb{R}$. Also let $t \in \mathbb{N}$ and $M \geq 0$ be such that $V_M(\mathcal{G}) \leq t$. If ϕ satisfies*

$$\frac{\lfloor \varepsilon n \rfloor + t}{n} \leq \phi < \frac{1}{2},$$

then

$$\sup_{g \in \mathcal{G}} \left| \widehat{T}^{\varepsilon}_{n,\phi n}(g) - \widehat{P}_n(\tau_M \circ g) \right| \leq 6\phi M.$$

In most of our proofs we will apply variations of the counting and bounding lemmata above. This procedure is useful in two ways:

- it instantaneously deals with the contamination by approximating the trimmed mean on the contaminated sample by an empirical average on the original clean sample;
- by approximating the trimmed mean by an empirical average of truncated terms it allows us to use all the existing machinery on concentration and approximation inequalities available for empirical means of bounded empirical processes. For instance, our concentration results for the empirical mean (Theorems 3.1 and 3.3) rely strongly on Bousquet's version of Talagrand's concentration inequality for bounded empirical processes.

Theorem 2.4 (Bousquet's version of Talagrand's concentration inequality, [Bousquet, 2002])**.** *Assume that a measure P over $(\mathbf{X}, \mathcal{X})$ is 1-compatible with a family of functions $\mathcal{G}$ and $|g - Pg| \leq C$ $\forall g \in \mathcal{G}$ for some constant $C > 0$. Define:*

$$W := \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} (g(X_i) - Pg) \right| \quad \text{where } X_{1:n} \overset{i.i.d.}{\sim} P$$

Also set $\sigma^2(\mathcal{G}) := \sup_{g \in \mathcal{G}} P(g - Pg)^2$ and $v := 2C\,\mathsf{Emp}_n(\mathcal{G}, P) + \sigma^2(\mathcal{G})$. Then

$$\forall x > 0 \,:\, \mathbb{P}\left[W \geq \mathsf{Emp}_n(\mathcal{G}, P) + \sqrt{\frac{2xv}{n}} + \frac{Cx}{3n} \right] \leq e^{-x}.$$

After evoking the counting and the bounding lemmata it remains to properly choose the trimming level k (and M satisfying (2.7)) in order to optimize the bounds. In this process the

trimming level controls a bias-variance trade-off: smaller values of k induce less bias, but might not be enough to avoid outliers from heavy-tailed distributions or contaminated sample points, thus incurring in a higher variance; higher values of k discard most of the sample points inducing more bias, but can better handle outliers by reducing the variance.

2.3 Proof of the lemmata

Proof [Proof of Lemma 2.2] As in [Lugosi and Mendelson, 2021, Lemma 1], we replace $V_M(\mathcal{G})$ by a *smoother empirical process upper bound* to which we will be able to apply Ledoux-Talagrand contraction and Bousquet's version of Talagrand's concentration inequality. Specifically, we define

$$\eta_M(r) := \left(\frac{2}{M}\left(r - \frac{M}{2}\right)_+\right) \wedge 1 \ (r \in \mathbb{R}),$$

which is $2/M$-Lipschitz and satisfies

$$\forall r \geq 0 \ : \ \mathbf{1}\{r > M\} \leq \eta_M(r) \leq \mathbf{1}\{r > M/2\}.$$

Notice that $\eta_M = \eta_M \circ \tau_M$, a fact that will be useful later.

To continue, we note that

$$\forall g \in \mathcal{G} \ : \ P\left(\eta_M \circ |g|\right)^2 \leq P\left(\eta_M \circ |g|\right) \leq P\left\{|g(X)| > \frac{M}{2}\right\}. \tag{2.8}$$

One consequence of this is that

$$V_M(\mathcal{G}) = \sup_{g\in\mathcal{G}} \sum_{i=1}^n \mathbf{1}\{|g(X_i)| > M\} \leq \sup_{g\in\mathcal{G}} \sum_{i=1}^n \eta_M \circ |g(X_i)| \leq n \sup_{g\in\mathcal{G}} P\left\{|g(X)| > \frac{M}{2}\right\} + nW$$

where W is the empirical process

$$W := \sup_{g\in\mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \left(\eta_M \circ |g(X_i)| - P\left(\eta_M \circ |g|\right)\right)\right|.$$

Therefore,

$$\mathbb{P}\left[V_M(\mathcal{G}) > t\right] \leq \mathbb{P}\left[W > \frac{t}{n} - \sup_{g\in\mathcal{G}} P\left\{|g(X)| > \frac{M}{2}\right\}\right],$$

and the present lemma will follow once we bound the RHS above by e^{-t}. To do so, we apply Bousquet's version of Talagrand's concentration inequality (Theorem 2.4) to the class $\eta_M \circ |\mathcal{G}| = \{\eta_M(|g|) : g \in \mathcal{G}\}$ with $C = 1$ (as $0 \leq \eta_M \leq 1$). Letting $v = n\sigma^2(\eta_M \circ |\mathcal{G}|) + 2\mathsf{Emp}_n(\eta_M \circ |\mathcal{G}|, P)$ be as in Theorem 2.4, we deduce

$$\mathbb{P}\left[W > \mathsf{Emp}_n(\eta_M \circ |\mathcal{G}|, P) + \sqrt{\frac{2vt}{n}} + \frac{t}{3n}\right] \leq e^{-t}.$$

Therefore, we will be done once we show that

$$t \geq n \sup_{g \in \mathcal{G}} P\left\{|g(X)| > \frac{M}{2}\right\} + n\,\mathsf{Emp}_n(\eta_M \circ |\mathcal{G}|, P) + \sqrt{2vtn} + \frac{t}{3}.$$

To prove this last inequality, we bound the empirical process via symmetrization, contraction (η_M is $2/M$-Lipschitz), the fact that $\eta_M \circ |\cdot| = \eta_M \circ |\tau_M|$, and our assumption (2.7) relating t, M and n:

$$\begin{aligned}
\mathsf{Emp}(\eta_M \circ |\mathcal{G}|, P) &\leq 2\mathsf{Rad}_n(\eta_M \circ |\mathcal{G}|, P) \\
(\eta_M \circ |\cdot| = \eta_M \circ |\tau_M|) &= 2\mathsf{Rad}_n(\eta_M \circ |\tau_M \circ \mathcal{G}|, P) \\
(\text{contraction} + \eta_M \circ |\cdot| \text{ is } 2/M\text{-Lip.}) &\leq \frac{4}{M}\,\mathsf{Rad}_n(\tau_M \circ \mathcal{G}, P)
\end{aligned}$$

To bound $\sigma^2(\eta_M \circ |\mathcal{G}|)$, we use (2.8):

$$\sigma^2(\eta_M \circ |\mathcal{G}|) = \sup_{g \in \mathcal{G}} P\left(\eta_M(|g|) - P\eta_M(|g|)\right)^2 \leq \sup_{g \in \mathcal{G}} P\left\{|g(X)| > \frac{M}{2}\right\}.$$

As a consequence, the variance parameter v in Bousquet's version of Talagrand's concentration inequality can be bounded using (2.7):

$$nv \leq n \sup_{g \in \mathcal{G}} P\left\{|g(X)| > \frac{M}{2}\right\} + \frac{8n\,\mathsf{Rad}_n(\tau_M \circ \mathcal{G}, P)}{M} \leq \frac{t}{8}.$$

Combining the above bounds, we arrive at

$$\frac{t}{3} + \sqrt{2vtn} + n\,\mathsf{Emp}_n(\eta_M \circ |\mathcal{G}|, P) + n \sup_{g \in \mathcal{G}} P\left\{|g(X)| > \frac{M}{2}\right\} \leq \frac{t}{3} + \frac{t}{2} + \frac{t}{8}$$

which is enough to conclude the proof. ■

As for Lemma 2.3, we give the proof of a more general version of it, which will be useful in the next chapters. The version stated in 2.3 follows directly from the one stated below:

Lemma 2.5 (Bounding lemma)**.** *Let $m \in \mathbb{N}$ and $\mathcal{G}_1, \ldots, \mathcal{G}_m$ be families of functions from $\mathbf{X}$ to $\mathbb{R}$. Also let $t_1, t_2, \ldots, t_m \in \mathbb{N}$ and $M_1, M_2, \ldots, M_m \geq 0$ be such that $V_{M_j}(\mathcal{G}_j) \leq t_j$ for each $j \in [m]$. If ϕ satisfies*

$$\frac{\lfloor \varepsilon n \rfloor + \sum_{j=1}^m t_j}{n} \leq \phi < \frac{1}{2},$$

then for any linear combination

$$g := \sum_{j=1}^{m} a_j g_j \text{ with } a_j \in \mathbb{R} \text{ and } g_j \in \mathcal{G}_j \text{ for each } j \in [m],$$

we have

$$\left| \widehat{T}^{\varepsilon}_{n,\phi n}(g) - \sum_{j=1}^{m} a_j \widehat{P}_n\left(\tau_{M_j} \circ g_j\right) \right| \leq 6\phi \sum_{j=1}^{m} |a_j| \, M_j. \tag{2.9}$$

Proof [Proof of Lemma 2.5] Define $\tilde{g}_j := a_j g_j$ and $\tilde{M}_j := |a_j| \, M_j$ for each $j \in [m]$, so that $g = \sum_{j=1}^{m} \tilde{g}_j$. We also have

$$\forall j \in [m] \, : \, \tau_{\tilde{M}_j} \circ \tilde{g}_j = a_j \, \tau_{M_j} \circ g_j. \tag{2.10}$$

Our assumption on $V_{M_j}(\mathcal{G}_j)$ implies that for each $j \in [m]$ there are at most t_j indices such that $|\tilde{g}_j(X_i)| > \tilde{M}_j$. Therefore, the set

$$B^{\varepsilon} := \{i \in [n] \, : \, X_i \neq X_i^{\varepsilon} \text{ or } |\tilde{g}_j(X_i)| > \tilde{M}_j \text{ for some } j \in [m]\}$$

has cardinality bounded by:

$$|B^{\varepsilon}| \leq \lfloor \varepsilon n \rfloor + \sum_{j=1}^{m} t_j \leq \phi \, n < \frac{n}{2}. \tag{2.11}$$

Now define $M := \sum_{j=1}^{m} \tilde{M}_j$. Since $g = \sum_{j=1}^{m} \tilde{g}_j$, we conclude from (2.11) that there are at most ϕn indices $i \in [n]$ with $|g(X_i^{\varepsilon})| > M$. Since the ϕn largest and smallest values of $g(X_i^{\varepsilon})$ are excluded from the trimmed mean, we obtain:

$$\widehat{T}^{\varepsilon}_{n,\phi n}(g) = \widehat{T}^{\varepsilon}_{n,\phi n}(\tau_M \circ g). \tag{2.12}$$

In what follows, we use this identity to compare the trimmed mean of a sum to a sum of truncated empirical means. More specifically, we let $\widehat{P}_n^{\varepsilon}$ denote the empirical measure of the contaminated sample, and use (2.10) and (2.12) to bound:

$$\begin{aligned} \left| \widehat{T}^{\varepsilon}_{n,\phi n}(g) - \sum_{j=1}^{m} a_j \widehat{P}_n\left(\tau_{M_j} \circ g_j\right) \right| &\leq \left| \widehat{T}^{\varepsilon}_{n,\phi n}(\tau_M \circ g) - \widehat{P}_n^{\varepsilon}(\tau_M \circ g) \right| \\ &\quad + \left| \widehat{P}_n^{\varepsilon}(\tau_M \circ g) - \sum_{j=1}^{m} \widehat{P}_n\left(\tau_{\tilde{M}_j} \circ \tilde{g}_j\right) \right|. \end{aligned} \tag{2.13}$$

To bound the first term in the RHS, notice that the empirical mean of $\tau_M(g(X_i^{\varepsilon}))$ is an average over all sample points in the contaminated sample, whereas $\widehat{T}^{\varepsilon}_{n,\phi n}(\tau_M \circ g)$ is an average over a

subset of these points of size $(1-2\phi)n$. Since all terms in both averages are bounded by M in absolute value, we conclude:

$$\widehat{P}_n^{\varepsilon}(\tau_M \circ g) = (1-2\phi)\,\widehat{T}_{n,\phi n}^{\varepsilon}(\tau_M \circ g) + 2\phi\eta$$

for some $|\eta| \leq M$, from which it follows that:

$$\left|\widehat{P}_n^{\varepsilon}(\tau_M \circ g) - \widehat{T}_{n,\phi n}^{\varepsilon}(\tau_M \circ g)\right| \leq 2\phi\,(|\eta| + |\widehat{T}_{n,\phi n}^{\varepsilon}(\tau_M \circ g)|) \leq 4\phi M. \tag{2.14}$$

We now consider the difference

$$\widehat{P}_n^{\varepsilon}(\tau_M \circ g) - \sum_{j=1}^{m} \widehat{P}_n(\tau_{\tilde{M}_j} \circ \tilde{g}_j) = \frac{1}{n}\sum_{i=1}^{n}\left[\tau_M \circ g(X_i^{\varepsilon}) - \sum_{j=1}^{m} \tau_{\tilde{M}_j} \circ \tilde{g}_j(X_i)\right] \tag{2.15}$$

The n terms inside the square brackets in the RHS are bounded in absolute value by $M + \sum_{j=1}^{m} \tilde{M}_j = 2M$. Recalling (2.11), we *claim* that if $i \in [n]\backslash B^{\varepsilon}$, the corresponding term in the RHS of (2.15) is zero. To see this, fix some i and note that:

- on the one hand, for each $j \in [m]$, $|\tilde{g}_j(X_i)| \leq \tilde{M}_j$ and so $\tau_{\tilde{M}_j} \circ \tilde{g}_j(X_i) = \tilde{g}_j(X_i)$;
- on the other hand, since $X_i = X_i^{\varepsilon}$, and using the above bounds, we have $|g(X_i^{\varepsilon})| = |g(X_i)| \leq M$ and
$$\tau_M \circ g(X_i^{\varepsilon}) = g(X_i) = \sum_{j=1}^{m} \tilde{g}_j(X_i).$$

It follows from the claim that:

$$\forall i \in [n] \;:\; \left|\tau_M \circ g(X_j^{\varepsilon}) - \sum_{j=1}^{n} \tau_{\tilde{M}_j} \circ \tilde{g}_j(X_i)\right| \leq 2M\mathbf{1}\{i \in B^{\varepsilon}\},$$

and combining this with (2.11) and (2.15) gives:

$$\left|\widehat{P}_n^{\varepsilon}(\tau_M \circ g) - \sum_{j=1}^{m} \widehat{P}_n(\tau_{\tilde{M}_j} \circ \tilde{g}_j)\right| \leq \frac{2M}{n}\,|B^{\varepsilon}| \leq 2M\phi$$

The lemma follows from plugging the preceding display together with inequality (2.14) into the RHS of (2.13). ■

Chapter 3

Uniform mean estimation

3.1 Introduction

This section discusses a problem first posed in this form in [Minsker, 2018b].

Problem 3.1 (Uniform mean estimation)**.** *One is given a measurable space* $(\mathbf{X}, \mathcal{X})$*; and family* $\mathcal{F}$ *of measurable functions from* $\mathbf{X}$ *to* $\mathbb{R}$*; and a family* $\mathcal{P}$ *of probability distributions over* $(\mathbf{X}, \mathcal{X})$ *such that* $\mathcal{F}$ *and* P *are* 1*-compatible for all* $P \in \mathcal{P}$*. For a sample size* $n \in \mathbb{N}$*; a contamination parameter* $\varepsilon \in [0, 1/2)$*; and a confidence level* $1 - \alpha \in (0, 1)$*; the goal is to find a family of measurable functions (estimators)*

$$\{E_f : \mathbf{X}^n \to \mathbb{R} \,:\, f \in \mathcal{F}\}$$

with the following property: for any $P \in \mathcal{P}$*, if* $X_{1:n}^{\varepsilon}$ *is a* ε*-contaminated sample from* P *(cf. §2.1.3), then:*

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} |E_f(X_{1:n}^{\varepsilon}) - Pf| \leq \Phi_P\right] \geq 1 - \alpha; \tag{3.1}$$

with $\Phi_P = \Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ *as small as possible.*

Importantly, the estimators E_f are not allowed to depend on the specific measure $P \in \mathcal{P}$, but may depend on $\mathcal{F}$, α, n and ε. We assume implicitly that the supremum in the above event is a random variable (i.e. it is a measurable function). Many results exists for the case where E_f is the sample mean and $\varepsilon = 0$, including Gaussian approximations [van der Vaart et al., 1996, Chernozhukov et al., 2014b] and concentration inequalities [Talagrand, 1996, Bousquet, 2002].

However, the whole point of this thesis is that we do *not* expect sample means to be optimal estimators, even when there is no contamination.

3.1.1 Relevant parameters.

We consider bounds on $\Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ in terms of moment conditions and measures of "complexity" of the class $\mathcal{F}$. For 1-compatible $\mathcal{F}$ and P as above, and exponents $p \geq 1$, we use $\nu_p(\mathcal{F}, P)$ (2.1) as a measure of moment conditions. As for our complexity measure of $\mathcal{F}$ under P, we take the expectation of the supremum of the empirical process over an uncontaminated sample (2.2). The question we address is: how small can we expect $\Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ to be in terms of the above parameters?

3.1.2 Examples and their history.

Problem 3.1 is related to several results in the literature.

Single functions. Consider first the case where $\mathcal{F} = \{f\}$ consists of a single function. In this case we omit $\mathcal{F}$ from our notation. The optimal value of $\Phi_P(n, \alpha, \varepsilon)$ is (up to constant factors) a sum of two terms: a *random fluctuations term* and a *contamination term*:

$$\inf_{1 \leq q \leq 2} \nu_p(P) \left(\frac{1}{n} \ln \frac{1}{\alpha}\right)^{1-\frac{1}{q}} \text{ and } \inf_{p \geq 1} \nu_q(P)\, \varepsilon^{1-\frac{1}{p}}, \text{ respectively.} \tag{3.2}$$

The necessity of these two terms follows from [Devroye et al., 2016, Theorem 3.1] and [Minsker, 2018b, Lemma 5.4][1]. The upper bound is achieved by the trimmed mean [Rico, 2022, Chapter 2].

The median of means (MoM) construction [Alon and Spencer, 2016, Devroye et al., 2016] is often used to obtain robust mean estimators. It consists of splitting the sample into K parts, taking the sample mean of each part, and then taking the median of the K means. For $K \approx \varepsilon n + \ln(1/\alpha)$, this estimator achieves:

$$\Phi_P(n, \alpha, \varepsilon) = C \inf_{1 \leq p \leq 2} \nu_p(P) \left(\frac{1}{n} \ln \frac{1}{\alpha} + \varepsilon\right)^{1-\frac{1}{p}},$$

with $C > 0$ universal; this follows e.g. [Bubeck et al., 2013, Lemma 2] (for $\varepsilon = 0$) combined with the fact that taking $K \gg \varepsilon n$ naturally makes the estimator robust (as observed in e.g. [Lecué and Lerasle, 2020]). In general, this bound is strictly worse than the optimal (3.2).

[1] The lower bound in [Minsker, 2018b, Lemma 5.4] is given for $p \in [2, 3]$, but the same proof works for all $p > 1$.

Vector mean estimation under general norms. Consider the problem of estimating the mean μ_P of a distribution P over $\mathbb{R}^d$, with the error given by an arbitrary norm $\|\cdot\|$. This problem consists of finding a measurable function $\widehat{\mu} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ such that, if $X^{\varepsilon}_{1:n}$ is a ε-contaminated sample from some P as above,

$$\mathbb{P}\left[\|\widehat{\mu}(X^{\varepsilon}_{1:n}) - \mu_P\| \leq \Phi_P(n, \alpha, \varepsilon)\right] \geq 1 - \alpha,$$

with $\Phi_P(n, \alpha, \varepsilon)$ as small as possible.

This problem is quite close to Problem 3.1. To see this notice that the mean μ_P satisfy $\langle v, \mu_P \rangle = P\langle v, \cdot \rangle$ for all $v \in S$, where S is the dual unit ball. Thus taking $\mathcal{F} := \{\langle v, \cdot \rangle : v \in S\}$ and given a solution $\{E_f\}_{f \in \mathcal{F}}$ to Problem 3.1, one can define

$$\widehat{\mu}(x_{1:n}) \in \arg\min_{\mu \in \mathbb{R}^d} \left(\sup_{f \in \mathcal{F}} |E_f(x_{1:n}) - f(\mu_P)| \right) \ (x_{1:n} \in (\mathbb{R}^d)^n)$$

and one can show that it yields $\Phi_P(n, \alpha, \varepsilon) \leq 2\Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$. Conversely, any $\widehat{\mu} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ can be related to a mean estimator by

$$E_f := f \circ \widehat{\mu} \ \ (f \in \mathcal{F}),$$

yielding $\Phi_P(\mathcal{F}, \alpha, n, \varepsilon) = \Phi_P(n, \alpha, \varepsilon)$. We dive deeper in this problem in Chapter 6, in special we show that our solution for Problem 3.1 improves upon the best known solutions available in the literature.

Other examples. Minsker [Minsker, 2018a] motivates Problem 3.1 via maximum likelihood estimation. That paper presents an estimator based on a combination of influence functions and median of means. His estimator is optimally robust (i.e. has an optimal contamination term) under p-th moment conditions in the range $p \in [2, 3]$.

Another related problem is that of estimating probability distributions according to *integral probability metrics*, as we discussed in Example 1.2.1.

3.2 Main result

The next theorem is our main contribution in the setting of Problem 3.1.

Theorem 3.1 (Proof in §3.3.2)**.** *In the setting of Problem 3.1, let $\mathcal{P}$ denote the family of all probability distributions P over $(\mathbf{X}, \mathcal{X})$ that are 1-compatible with $\mathcal{F}$ and such that*

$\mathsf{Emp}_n(\mathcal{F}, P) < +\infty$. *If*

$$\phi := \frac{1}{n}\left(\lfloor \varepsilon n \rfloor + \left\lceil \ln \frac{2}{\alpha} \right\rceil \vee \left\lceil \frac{\left(\frac{1}{2} - \varepsilon\right) \wedge \varepsilon}{2} n \right\rceil\right) < \frac{1}{2},$$

then the family of estimators

$$E_f(x_{1:n}) := \widehat{T}_{n,\phi n}(f; x_{1:n}) \ (x_{1:n} \in \mathbf{X}^n,\ f \in \mathcal{F}),$$

satisfies the following property: if $X^{\varepsilon}_{1:n}$ is an ε-contaminated sample from some $P \in \mathcal{P}$, then

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} \left| \widehat{T}_{n,\phi n}(f, X^{\varepsilon}_{1:n}) - Pf \right| \le \Phi_P \right] \ge 1 - \alpha,$$

where $\Phi_P = \Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ is defined as follows:

$$\Phi_P = C_\varepsilon \left(8\mathsf{Emp}_n(\mathcal{F}, P) + \inf_{q \in [1,2]} \nu_q(\mathcal{F}, P) \left(\frac{\ln \frac{3}{\alpha}}{n} \right)^{1 - \frac{1}{q}} + \inf_{p \ge 1} \nu_p(\mathcal{F}, P) \varepsilon^{1 - \frac{1}{p}} \right),$$

with $C_\varepsilon := 384 \left(1 + \frac{\varepsilon}{\varepsilon \wedge \left(\frac{1}{2} - \varepsilon\right)}\right)$.

Let us comment on this result. The last two terms in the definition of $\Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ correspond to the difficulty of estimating Pf with the "worst case" choice of $f \in \mathcal{F}$. By §3.1.2, neither term can be improved by more than a constant factor.

As we discuss in Chapter 6, Theorem 3.1 leads to improvements in the problem of vector mean estimation. It also improves on Minsker's result [Minsker, 2018b]. That paper obtains "sub-Gaussian" bounds under complicated assumptions that go beyond second moments. Additionally, [Minsker, 2018b] requires approximate knowledge of the largest variance in $\mathcal{F}$ to obtain optimal results; and only controls the contamination term under p-th moment assumptions for $p \in [2, 3]$.

Remark 3.2 (Is the complexity term optimal?)**.** The complexity term $\mathsf{Emp}_n(\mathcal{F}, P)$ appears in all theorems cited above. The only case this parameter is known to be necessary is that of mean estimation for Gaussian vectors [Depersin and Lecué, 2021]. On the other hand, our Theorem follows from a more general result that does not even require $\mathsf{Emp}_n(\mathcal{F}, P) < +\infty$ or $\nu_p(\mathcal{F}, P) < +\infty$ for any $p > 1$; see Theorem 3.3 for details.

3.3 Proofs

3.3.1 Trimming and truncation: a master theorem. The main results in this thesis follow from the "master theorem" presented in this section. It may be viewed as an abstract and extended version of the arguments in [Lugosi and Mendelson, 2021], which were specific for mean estimation in Hilbert spaces.

In this subsection, P is a fixed probability measure over a measurable space $(\mathbf{X}, \mathcal{X})$, and $X_{1:n}^{\varepsilon}$ is an ε-contaminated sample from P. For a given $M > 0$, recall the definition of the truncation function from (2.4). If $\mathcal{G}$ is a family of functions, we let

$$\mathcal{G}^o := \{g - Pg : g \in \mathcal{G}\}\,,\ \mathcal{G}_M^o := \{\tau_M \circ (g - Pg) : g \in \mathcal{G}\}$$

and also $\mathrm{rem}_M(\mathcal{G}, P) := \sup_{g\in\mathcal{G}} |P\tau_M \circ g|$.

Theorem 3.3 (Master theorem for trimmed mean)**.** *Let $m \in \mathbb{N}$ and $\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_m$ be families of functions that are 1-compatible with P. Also let $X_{1:n}^{\varepsilon}$ be an ε-contaminated i.i.d. sample from P. Assume that for every $j \in [m]$ there exist $M_j > 0$, $b_j \in \{0, 1\}$ and $t_j \in \mathbb{N} \cup \{0\}$ satisfying one of the following conditions:*

- *either $b_j = 0$, $t_j = 0$ and $\mathcal{F}_j^o$ is a.s. uniformly bounded by M_j, i.e. $|f_j - Pf_j| \le M_j$ almost surely for all $f_j \in \mathcal{F}_j$;*
- *or $b_j = 1$ and we have the bound*

$$\sup_{f_j\in\mathcal{F}_j} P\left\{|f_j(X) - Pf_j| > \frac{M_j}{2}\right\} + \frac{8\,\mathsf{Rad}_n(\tau_{M_j} \circ \mathcal{F}_j^o, P)}{M_j} \le \frac{t_j}{8n}. \tag{3.3}$$

Also let ϕ be such that

$$\frac{\lfloor \varepsilon n \rfloor + \sum_{j=1}^m t_j}{n} \le \phi < \frac{1}{2}.$$

Let $x_j \ge 0$ for each $j \in [m]$. Then, with probability at least $1 - \sum_{j=1}^m (b_j\, e^{-t_j} + e^{-x_j})$, for every linear combination $f = \sum_{j=1}^m a_j f_j$, $f_j \in \mathcal{F}_j$,

$$\left|\widehat{T}_{n,\phi n}(f; X_{1:n}^{\varepsilon}) - Pf\right| \le \sum_{j=1}^n |a_j| \left\{2\mathsf{Emp}_n\left(\tau_{M_j} \circ \mathcal{F}_j^o, P\right) + \eta_j\right\},$$

where

$$\eta_j = \left(6\phi + \frac{3x_j}{n}\right) M_j + \mathrm{rem}_{M_j}\left(\mathcal{F}_j^o, P\right) + \nu_2\left(\tau_{M_j} \circ \mathcal{F}_j^o, P\right)\sqrt{\frac{2x_j}{n}}.$$

Proof We start with some notation and conventions. For brevity, we will occasionally omit P from our notation; for instance, we write $\mathsf{Emp}_n(\mathcal{F})$ instead of $\mathsf{Emp}_n(\mathcal{F}, P)$. We also let $X_{1:n} \overset{i.i.d.}{\sim} P$ be a sample from P with

$$|\{i \in [n] \,:\, X_i \neq X_i^\varepsilon\}| \leq \lfloor \varepsilon n \rfloor.$$

$\widehat{P}_n$ denotes the empirical measure of the clean sample, meaning:

$$\widehat{P}_n\, g = \frac{1}{n}\sum_{i=1}^n g(X_i) \text{ for } g : \mathbf{X} \to \mathbb{R},$$

and we use $\widehat{T}^\varepsilon_{n,\phi n}(g) := \widehat{T}_{n,\phi n}(g; X^\varepsilon_{1:n})$ to denote the trimmed mean computed over the contaminated sample.

The core observation of our proof is that if $k \in \mathbb{N}$ is properly defined and the M_j are big enough, the "trimmed empirical process" of linear combinations on a contaminated sample is close to the linear combination of truncated empirical processes on the clean sample, i.e.

$$f = \sum_{j=1}^m a_j f_j \mapsto \widehat{T}^\varepsilon_{n,k}(f) - Pf \approx \sum_{j=1}^m a_j \widehat{P}_n\, \tau_{M_j}(f_j(X_i) - Pf_j), \tag{3.4}$$

which is well understood and satisfies Bernstein-type concentration bounds. This will require two lemmata that we discuss next. Let $\mathcal{G}$ be a family of functions and $M > 0$, recall the definition of $V_M(\mathcal{G})$ from (2.6).

The "Counting Lemma" (Lemma 2.2) gives a way to bound the probability that the counting function $V_{M_j}(\mathcal{F}_j^o)$ exceeds a certain value t_j. This will be useful whenever $\mathcal{F}_j^o$ is not uniformly bounded. On the other hand, the "Bounding Lemma" (Lemma 2.5) says that if the counting functions $\{V_{M_j}(\mathcal{F}_j^o)\}_{j\in[m]}$ are all small, then (3.4) can be justified.

We come back to the proof of Theorem 3.3. Since the families $\mathcal{F}_j$ are 1-compatible with P, we may assume they are countable. Our hypotheses ensure that Lemma 2.2 can be applied to each $V_{M_j}(\mathcal{F}_j^o)$ (with the corresponding t_j) whenever $b_j = 1$. On the other hand, when $b_j = 0$, the class $\mathcal{F}_j^o$ is bounded and $V_{M_j}(\mathcal{F}_j^o) = 0 \leq t_j$ is automatic. Combining this with Lemma 2.5 (taking $\mathcal{G}_j = \mathcal{F}_j^o$ for every $j \in [m]$) we have, with probability at least $1 - \sum_{j=1}^m b_j\, e^{-t_j}$, for all $f = \sum_{j=1}^m a_j f_j$,

$$\left|\widehat{T}^\varepsilon_{n,\phi n}(f) - Pf\right| \leq \left|\sum_{j=1}^m a_j \widehat{P}_n \tau_{M_j}(f_j - Pf_j)\right| + 6\phi \sum_{j=1}^m |a_j|\, M_j.$$

We can bound

$$\left|\sum_{j=1}^{m} a_j \widehat{P}_n \tau_{M_j}(f_j - Pf_j)\right| \leq \sum_{j=1}^{m} |a_j| \left\{ \sup_{f_j \in \mathcal{F}_j} \left|(\widehat{P}_n - P)\tau_{M_j}(f_j - Pf_j)\right| + \mathrm{rem}_{M_j}(\mathcal{F}_j^o, P) \right\}$$

We bound the suprema on the RHS using Bousquet's version of Talagrand's concentration inequality for each class $\mathcal{F}_j^o$ (Theorem 2.4) and observing that

$$\left|\tau_{M_j}(f_j - Pf_j) - P\tau_{M_j}(f_j - Pf_j)\right| \leq 2M_j.$$

We finish using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\sqrt{ab} \leq a + \frac{b}{4}$ to bound:

$$\sqrt{\frac{8x_j M_j}{n} \mathsf{Emp}_n\left(\tau_{M_j} \circ \mathcal{F}_j^o, P\right)} \leq \frac{2x_j}{n} M_j + \mathsf{Emp}_n\left(\tau_{M_j} \circ \mathcal{F}_j^o, P\right).$$

■

3.3.2 Bounds for uniform mean estimation.

We now prove our main result on uniform mean estimation, Theorem 3.1. This will require converting the error bounds in Theorem 3.3 into moment-based quantities. The next lemma does this; it is somewhat stronger than what we need.

Lemma 3.4. *Let $\mathcal{F}_j$, t_j, b_j, x_j and M_j satisfy the hypothesis of Theorem 3.3, define*

$$\bar{\varepsilon} := \frac{\left(\frac{1}{2} - \varepsilon\right) \wedge \varepsilon}{1 + \sum_{j=1}^m b_j}, \quad t_j = b_j \left(\lceil x_j \rceil \vee \lceil \bar{\varepsilon} n \rceil\right) \text{ and } \phi := \frac{\lfloor \varepsilon n \rfloor + \sum_{j=1}^m t_j}{n}.$$

Also assume $\phi < \frac{1}{2}$, $x_j \geq \frac{1}{3}$ for all j with $b_j = 1$ and let

$$C_j^{\varepsilon} := 192 \left(1 + \frac{\sum_{l=1}^m t_l}{t_j} + \frac{\varepsilon}{\bar{\varepsilon}}\right).$$

Then, for every j with $b_j = 1$, it is possible to choose M_j satisfying (3.3) such that

$$2\mathsf{Emp}_n\left(\tau_{M_j} \circ \mathcal{F}_j^o, P\right) + \eta_j \leq C_j^{\varepsilon} \left\{ 8\mathsf{Emp}_n(\mathcal{F}_j, P) + \inf_{q \in [1,2]} \nu_q(\mathcal{F}_j, P) \left(\frac{x_j}{n}\right)^{1-\frac{1}{q}} + \inf_{p \geq 1} \nu_p(\mathcal{F}_j, P) \left(\frac{2\varepsilon}{1 + \sum_{l=1}^m b_l}\right)^{1-\frac{1}{p}} \right\}. \tag{3.5}$$

Proof [Proof of Theorem 3.1] Apply Theorem 3.3 using $m = 1$, $b_1 = 1$, $x_1 := \ln \frac{3}{\alpha}$ with t_1 and M_1 as chosen in the Lemma 3.4. Inspection reveals that this leads to the bound claimed in Theorem 3.1. ■

Proof [Proof of Lemma 3.4] Our goal is to find, for the values j with $b_j = 1$, M_j such that (3.3) holds and the bound (3.5) is valid. To start, define

$$b = \sum_{j=1}^{m} b_j \text{ and } t = \sum_{j=1}^{m} t_j.$$

First step: choose M_j as a function of t_j and $\mathcal{F}_j$.

Notice that contraction and symmetrization gives

$$M_j \geq \frac{256\, n}{t_j} \mathsf{Emp}_n(\mathcal{F}_j) \Rightarrow \frac{8\, \mathsf{Rad}_n(\tau_{M_j} \circ \mathcal{F}_j^o)}{M_j} \leq \frac{16\, \mathsf{Emp}_n(\mathcal{F}_j)}{M_j} \leq \frac{t_j}{16n}.$$

So, (3.3) follows if we can define

$$M_j = m_j(t_j) \vee \left(\frac{256n}{t_j} \mathsf{Emp}_n(\mathcal{F}_j) \right),$$

with $m_j(t_j)$ such that

$$\forall f \in \mathcal{F}_j \,:\, P\left\{ |f(X) - Pf| > \frac{m_j(t_j)}{2} \right\} \leq \frac{t_j}{16n}.$$

We can explicitly find a choice of $m_j(t)$. Markov's inequality gives, for every $p \geq 1$,

$$\sup_{f \in \mathcal{F}_j} \mathbb{P}\left(|f(X) - Pf| > \frac{m_j(t_j)}{2} \right) \leq 2^p \frac{\nu_p^p(\mathcal{F}_j)}{m_j(t_j)^p}$$

and so we can take $m_j(t_j) = 2\nu_p(\mathcal{F}_j) \left(\frac{16n}{t_j}\right)^{\frac{1}{p}}$. Thus, we define

$$\begin{aligned} M_j = M_j(t) &:= \left(2\nu_p(\mathcal{F}_j) \left(\frac{16n}{t_j} \right)^{\frac{1}{p}} \right) \vee \left(\frac{256\, n}{t_j} \mathsf{Emp}_n(\mathcal{F}_j) \right) \\ &\leq 32\nu_p(\mathcal{F}_j) \left(\frac{t_j}{n} \right)^{-\frac{1}{p}} + \frac{256\, n}{t_j} \mathsf{Emp}_n(\mathcal{F}_j). \end{aligned} \tag{3.6}$$

Second step: bound $2\mathsf{Emp}_n\left(\tau_{M_j} \circ \mathcal{F}_j^o, P\right) + \eta_j$.

Given our choice of M_j we bound the terms of η_j. We can easily bound

$$\mathrm{rem}_{M_j}(\mathcal{F}_j^o) = \sup_{f\in\mathcal{F}_j} P(f-Pf)\mathbf{1}_{|f-Pf|>M_j} \leq \frac{\nu_p^p(\mathcal{F}_j)}{M_j^{p-1}} \leq \frac{\nu_p^p(\mathcal{F}_j)}{m_j(t_j)^{p-1}} \leq \nu_p(\mathcal{F}_j)\left(\frac{t_j}{n}\right)^{1-\frac{1}{p}}$$

and, using contraction and symmetrization,

$$\mathsf{Emp}_n\left(\tau_{M_j}\circ\mathcal{F}_j^o\right) \leq 2\mathsf{Rad}_n\left(\tau_{M_j}\circ\mathcal{F}_j^o\right) \leq 2\mathsf{Rad}_n\left(\mathcal{F}_j^o\right) \leq 4\mathsf{Emp}_n(\mathcal{F}_j).$$

We now bound the largest variance in $\tau_{M_j}\circ\mathcal{F}_j^o$ in terms of the moment parameters of $\mathcal{F}_j$. For $f\in\mathcal{F}_j$,

$$\mathbb{V}\left(\tau_{M_j}\circ(f-Pf)\right) \leq P\left(\tau_{M_j}\circ(f-Pf)\right)^2 \leq P\left(M_j\wedge|f-Pf|\right)^2,$$

where the first inequality follows from bounding the variance by the second moment, and the second is a consequence of $|\tau_{M_j}\circ(f-Pf)| = M_j\wedge|f-Pf|$. Now, for any $1\leq q\leq 2$,

$$\left(M_j\wedge|f-Pf|\right)^2 \leq M_j^{2-q}\,|f-Pf|^q,$$

so that

$$\mathbb{V}\left(\tau_{M_j}\circ(f-Pf)\right) \leq M_j^{2-q}\|f-Pf\|_{L^q(P)}^q \leq M_j^{2-q}\,\nu_q^q(\mathcal{F}_j).$$

It gives

$$\nu_2\left(\tau_{M_j}\circ\mathcal{F}_j^o\right)\sqrt{\frac{2x_j}{n}} \leq \sqrt{2}\left(\frac{M_jx_j}{n}\right)^{1-\frac{q}{2}}\left(\nu_q(\mathcal{F}_j)\left(\frac{x_j}{n}\right)^{1-\frac{1}{q}}\right)^{\frac{q}{2}}$$

and we can bound, using Young's inequality and $\sqrt{2}<2$,

$$\begin{aligned}\nu_2\left(\tau_{M_j}\circ\mathcal{F}_j^o\right)\sqrt{\frac{2x_j}{n}} &\leq \sqrt{2}\left(1-\frac{q}{2}\right)\frac{M_jx_j}{n} + \sqrt{2}\frac{q}{2}\nu_q(\mathcal{F}_j)\left(\frac{x_j}{n}\right)^{1-\frac{1}{q}}\\ &\leq 2\left(\frac{M_jx_j}{2n} + \nu_q(\mathcal{F}_j)\left(\frac{x_j}{n}\right)^{1-\frac{1}{q}}\right).\end{aligned}$$

Notice that

$$6\phi + \frac{4x_j}{n} = \left(\frac{6t}{t_j} + \frac{6\lfloor\varepsilon n\rfloor}{t_j} + \frac{4x_j}{t_j}\right)\frac{t_j}{n} \leq \left(4 + 6\frac{t}{t_j} + 6\frac{\varepsilon}{\bar{\varepsilon}}\right)\frac{t_j}{n},$$

taking $C_j' = 32\left(4+6\frac{t}{t_j}+6\frac{\varepsilon}{\bar{\varepsilon}}\right)$ and using (3.6), we have

$$M_j\left(6\phi+\frac{4x_j}{n}\right) \leq C_j'\left(\nu_p(\mathcal{F}_j)\left(\frac{t_j}{n}\right)^{1-\frac{1}{p}} + 8\mathsf{Emp}_n(\mathcal{F}_j)\right)$$

Combining the bounds gives

$$\begin{aligned}2\mathsf{Emp}_n\left(\tau_{M_j}\circ\mathcal{F}_j^o,P\right)+\eta_j &\le 8\mathsf{Emp}_n(\mathcal{F}_j)+2\nu_q(\mathcal{F}_j)\left(\frac{x_j}{n}\right)^{1-\frac{1}{q}}\\ &\quad+M_j\left(6\phi+\frac{4x_j}{n}\right)+\nu_p(\mathcal{F}_j)\left(\frac{t_j}{n}\right)^{1-\frac{1}{p}}\\ &\le 8(1+C_j')\mathsf{Emp}_n(\mathcal{F}_j)+2\nu_q(\mathcal{F}_j)\left(\frac{x_j}{n}\right)^{1-\frac{1}{q}}\\ &\quad+(1+C_j')\nu_p(\mathcal{F}_j)\left(\frac{t_j}{n}\right)^{1-\frac{1}{p}}\end{aligned}$$

Final step: finish proof by case analysis on t_j.

Recall

$$C_j^\varepsilon := 192\left(1+\frac{t}{t_j}+\frac{\varepsilon}{\bar\varepsilon}\right).$$

Since $x_j \ge \frac{1}{3}$ and $q\in[1,2]$ we have

$$\lceil x_j\rceil^{1-\frac{1}{q}} \le (x_j+1)^{1-\frac{1}{q}} \le 2x_j^{1-\frac{1}{q}},$$

and we bound

$$\left(\frac{\lceil x_j\rceil}{n}\right)^{1-\frac{1}{q}} \le 2\left(\frac{x_j}{n}\right)^{1-\frac{1}{q}}.$$

We also have $\lceil a\rceil\le 2a$ when $a\ge 1$. Thus, if $\varepsilon n\ge 1$ (i.e., when there is a contaminated sample point) we have

$$\frac{\lceil\bar\varepsilon n\rceil}{n} \le \frac{1}{n}\left\lceil\frac{\varepsilon n}{1+b}\right\rceil \le \frac{2\varepsilon}{1+b}.$$

The case $\varepsilon\, n<1$ means that there is no contamination and we might replace ε with 0, obtaining the same bound.

We now take the infimum over $q\in[1,2]$ and $p\ge 1$. If $t_j=\lceil x_j\rceil$,

$$2\mathsf{Emp}_n\left(\tau_{M_j}\circ\mathcal{F}_j^o,P\right)+\eta_j \le 8C_j^\varepsilon\mathsf{Emp}_n(\mathcal{F}_j)+C_j^\varepsilon\inf_{q\in[1,2]}\nu_q(\mathcal{F}_j)\left(\frac{x_j}{n}\right)^{1-\frac{1}{q}}.$$

The case $t_j=\lceil\bar\varepsilon n\rceil$ gives

$$\begin{aligned}2\mathsf{Emp}_n\left(\tau_{M_j}\circ\mathcal{F}_j^o,P\right)+\eta &\le 8C_j^\varepsilon\mathsf{Emp}_n(\mathcal{F}_j)\\ &\quad+4\inf_{q\in[1,2]}\nu_q(\mathcal{F}_j)\left(\frac{x_j}{n}\right)^{1-\frac{1}{q}}+C_j^\varepsilon\inf_{p\ge1}\nu_p(\mathcal{F}_j)\left(\frac{2\varepsilon}{1+b}\right)^{1-\frac{1}{p}}.\end{aligned}$$

The final bound follows from considering the two possible values of t_j and performing some overestimates. ■

Chapter 4

Regression with quadratic risk: theory and heuristics

4.1 Introduction

This chapter discusses the problem of regression with quadratic loss as studied by Lugosi and Mendelson [Lugosi and Mendelson, 2019c] and Lecué and Lerasle [Lecué and Lerasle, 2020]. In this setting, we consider probability measures P over product spaces $\mathbf{X} \times \mathbb{R}$; and we use $P_{\mathbf{X}}$ and $P_{\mathbb{R}}$ to denote the respective marginals.

Problem 4.1 (Regression with quadratic loss)**.** *One is given a measurable space $(\mathbf{X}, \mathcal{X})$; a convex family $\mathcal{F}$ of measurable functions from $\mathbf{X}$ to $\mathbb{R}$; and a family of probability measures $\mathcal{P}$ over $(\mathbf{X} \times \mathbb{R}, \mathcal{X} \times \mathcal{B})$ with the following property: for all $P \in \mathcal{P}$, $\mathcal{F}$ and $P_{\mathbf{X}}$ are 2-compatible; $\mathcal{F}$ is closed in $L^2(P_{\mathbf{X}})$; and $P_{\mathbb{R}}$ has a finite second moment.*

In this setup, define (for each $P \in \mathcal{P}$):

$$f_P^\star := \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)\sim P} \left(f(X) - Y\right)^2. \tag{4.1}$$

For a sample size $n \in \mathbb{N}$, a contamination parameter $\varepsilon \in [0, 1/2)$; and a confidence level $1 - \alpha \in (0, 1)$; find a mapping

$$F_n : (\mathbf{X} \times \mathbb{R})^n \to \mathcal{F}$$

with the following property: for any $P \in \mathcal{P}$, if $Z_{1:n}^\varepsilon := \{(X_i^\varepsilon, Y_i^\varepsilon)\}_{i\in[n]}$ is an ε-contaminated

sample from P (cf. §2.1.3), then the function $\widehat{f}_n^{\varepsilon} := F_n(Z_{1:n}^{\varepsilon})$ achieves

$$\mathbb{P}\left[\left\|\widehat{f}_n^{\varepsilon} - f_P^{\star}\right\|_{L^2(P_{\mathbf{X}})} \leq \Phi_P\right] \geq 1 - \alpha, \tag{4.2}$$

with $\Phi_P := \Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ as small as possible.

Let us make some technical comments about this problem. Firstly, to avoid delicate measurability issues regarding F_n, the probability in (4.2) should be interpreted as a inner probability $\mathbb{P}_*$. Secondly, it follows from the fact that $\mathcal{F}$ is convex and closed in $L^2(P_{\mathbf{X}})$ that $f_P^{\star}$ is uniquely defined up to a $P_{\mathbf{X}}$-null set. With these observations, it is clear that Problem 4.1 is well-posed. Our third comment is that we will also consider a variant of Problem 4.1 where the goal is to minimize *excess risk*. Letting:

$$R_P(f) := P\,(f(X) - Y)^2 \ (f \in \mathcal{F}),$$

one can check that

$$\forall f \in \mathcal{F} \,:\, \|f - f_P^{\star}\|^2_{L^2(P_{\mathbf{X}})} \leq R_P(f) - R_P(f_P^{\star}).$$

Therefore, results on the excess risk of $f = \widehat{f}_n^{\varepsilon}$ also bound the distance between $\widehat{f}_n^{\varepsilon}$ and f_P^*. Our main result on the regression problem, Theorem 4.4 below, will give both types of bounds. In what follows we present the definitions and conditions we need to present our solution to Problem 4.1. The following concrete example will serve to illustrate our discussion.

Example (Linear regression with independent errors)**.** Let $\mathbf{X} = \mathbb{R}^d$ with the Borel σ-field $\mathcal{X}$. We consider the family of linear functions

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

$\mathcal{P}$ consists of the family of distributions P such that $\mathbb{E}_{(X,Y)\sim P}[\|X\|_2^2 + Y^2] < +\infty$ and, given $(X, Y) \sim P$, there exists $\beta_P^{\star} \in \Theta$ with:

$$Y = \langle \beta_P^{\star}, X \rangle + \xi_P, \text{ where } \xi_P \text{ is mean zero and independent from } X.$$

In this case, $f_P^{\star}(\cdot) = \langle \cdot, \beta_P^{\star} \rangle$ and the excess risk is:

$$R_P(f_\beta) - R_P(f_P^{\star}) = \|\langle \cdot, \beta - \beta_P^{\star} \rangle\|^2_{L^2(P_{\mathbf{X}})} = \langle \beta - \beta_P^{\star}, \Sigma_P\,(\beta - \beta_P^{\star}) \rangle,$$

where $\Sigma_P = \mathbb{E}_{(X,Y)\sim P} X X^T$ is the population design matrix. We set $\sigma_P^2 = \mathbb{E}\,\xi_P^2$.

4.1.1 Relevant parameters.

As with uniform mean estimation, the analysis of Problem 4.1 will depend on moment bounds and complexity parameters of the function class $\mathcal{F}$. As in previous work, the right parameters to consider are localized [Massart, 2000, Mendelson, 2015, Lugosi and Mendelson, 2019b, Lecué and Lerasle, 2020].

Some notation. For $f \in \mathcal{F}$, let

$$\ell_f(x,y) := (f(x)-y)^2((x,y) \in \mathbf{X} \times \mathbb{R}). \tag{4.3}$$

The excess risk of f is

$$R_P(f) - R_P(f_P^\star) = P\left(\ell_f - \ell_{f_P^\star}\right).$$

Now let $\xi_P(x,y) := y - f_P^\star(x)$ denote the "regression residual" at the optimal $f_P^\star$ (this coincides with the ξ_P in Example 4.1), and set:

$$\mathsf{m}_f(x,y) = \mathsf{m}_{f,P}(x,y) := \xi_P(x,y)\left(f(x) - f_P^\star(x)\right) \ (f \in \mathcal{F},\ (x,y) \in \mathbf{X} \times \mathbb{R}). \tag{4.4}$$

Then the difference

$$\ell_f(x,y) - \ell_{f_P^\star}(x,y) = (f(x) - f_P^\star(x))^2 - 2\mathsf{m}_{f,P}(x,y)$$

consists of a "quadratic term" and a "multiplicative term." As in [Lecué and Mendelson, 2013], we note that analyzing the standard empirical risk minimizer – or other risk minimization procedures – requires *lower bounds* on the quadratic part and *upper bounds* on the multiplier part.

Localization, complexities and critical radii. Since Massart [Massart, 2000] it has been known that local analyses of regression problems lead to the best results. Following Mendelson [Mendelson, 2015], we consider the local parameters that are relevant to the analysis of quadratic and multiplier parts of our process. What makes these parameters "local" is that they are parameterized by the distance to $f_P^\star$. Specifically, define, for $r > 0$:

$$\begin{aligned}
\mathcal{F}_{\mathsf{q}}(r,P) &:= \left\{f - f_P^\star \,:\, f \in \mathcal{F},\ \|f - f_P^\star\|_{L^2(P_{\mathbf{X}})} = r\right\}, \\
\mathcal{F}_{\mathsf{m}}(r,P) &:= \left\{\mathsf{m}_{f,P} - P\mathsf{m}_{f,P} : f \in \mathcal{F},\ \|f - f_P^\star\|_{L^2(P_{\mathbf{X}})} \le r\right\}.
\end{aligned}$$

As in [Lecué and Lerasle, 2020], we use Rademacher complexities (2.3) to measure the size of these function families. Given constants $\delta_{\mathsf{q}}, \delta_{\mathsf{m}} > 0$, we define the *critical radii*

$$\begin{aligned}
r_{\mathsf{q}}(\delta_{\mathsf{q}}, \mathcal{F}, P) &:= \inf\{r > 0 \,:\, \mathcal{F}_{\mathsf{q}}(r,P) \neq \emptyset \text{ and } \mathsf{Rad}_n(\mathcal{F}_{\mathsf{q}}(r,P),P) \le \delta_{\mathsf{q}}\, r\}, \\
r_{\mathsf{m}}(\delta_{\mathsf{m}}, \mathcal{F}, P) &:= \inf\{r > 0 \,:\, \mathsf{Rad}_n(\mathcal{F}_{\mathsf{m}}(r,P),P) \le \delta_{\mathsf{m}}\, r^2\},
\end{aligned}$$

where (by convention) the infimum of an empty set is $+\infty$. Typically, we will take $\delta_{\mathsf{q}}, \delta_{\mathsf{m}} < 1$. In this case, the intuition is that r_{q} is the smallest radius r at which the quadratic process is significantly larger than r^2. The other critical radius r_{m} is the smallest radius r at which the multiplier empirical processes becomes small when compared to r^2.

Remark 4.1 (Critical radii in linear regression with independent errors)**.** In Example 4.1, one can check that:

$$\mathsf{Rad}_n(\mathcal{F}_{\mathsf{q}}(r,P),P) \leq r\sqrt{\frac{\mathrm{tr}(\Sigma_P)}{n}} \text{ and } \mathsf{Rad}_n(\mathcal{F}_{\mathsf{m}}(r,P),P) \leq r\,\sigma_P\sqrt{\frac{\mathrm{tr}(\Sigma_P)}{n}}.$$

Therefore, $r_{\mathsf{q}}(\delta_{\mathsf{q}}, \mathcal{F}, P) = 0$ when $n \geq \mathrm{tr}(\Sigma_P)/\delta_{\mathsf{q}}^2$. Moreover,

$$r_{\mathsf{m}}(\delta_{\mathsf{m}}, \mathcal{F}, P) \leq \frac{\sigma_P}{\delta_{\mathsf{m}}}\sqrt{\frac{\mathrm{tr}(\Sigma_P)}{n}}.$$

If there is no condition on n, it is possible that $r_{\mathsf{q}}(\delta_{\mathsf{q}}, \mathcal{F}, P) = +\infty$; for instance, this will be the case if $P_{\mathbf{X}}$ is Gaussian and $n \leq c\,\mathrm{tr}(\Sigma_P)/\delta_{\mathsf{q}}^2$.

Moment parameters. Our results also require the introduction of two moment-related quantities. The first of these is:

$$\theta_0(\mathcal{F}, P) := \sup\left\{\frac{\|f - f_P^\star\|_{L^2(P_{\mathbf{X}})}}{\|f - f_P^\star\|_{L^1(P_{\mathbf{X}})}} \;:\; f \in \mathcal{F},\, \|f - f_P^\star\|_{L^1(P_{\mathbf{X}})} > 0\right\} \tag{4.5}$$

As shown in Proposition 2 of [Lecué and Lerasle, 2019], a bound on $\theta_0(\mathcal{F}, P) < +\infty$ is essentially equivalent to a "small ball condition" on the functions $f - f_P^\star$:

$$\mathbb{P}_{X\sim P_{\mathbf{X}}}\{|f(X) - f_P^\star(X)| \geq c_0\|f - f_P^\star\|_{L^2(P_{\mathbf{X}})}\} \geq \alpha_0 \tag{4.6}$$

for $c_0, \alpha_0 > 0$. This will give a convenient way to control the quadratic part of the excess risk.

The second moment parameter applies to the multiplier part. Given $p \geq 1$, let

$$\kappa_p(\mathcal{F}, P) := \sup\left\{\frac{\|\mathsf{m}_{f,P} - P\,\mathsf{m}_{f,P}\|_{L^p(P)}}{\|f - f_P^\star\|_{L^2(P_{\mathbf{X}})}} \;:\; f \in \mathcal{F},\, \|f - f_P^\star\|_{L^2(P_{\mathbf{X}})} > 0\right\}.$$

Remark 4.2 (Moment conditions and linear regression)**.** In the setting of Example 4.1,

$$\begin{aligned}\theta_0(\mathcal{F}, P)^{-1} &= \inf_{\beta\in\mathbb{R}^d\,:\,\langle\beta,\Sigma_P\,\beta\rangle=1} \|\langle\cdot,\beta\rangle\|_{L^1(P_{\mathbf{X}})} \text{ and} \\ \kappa_p(\mathcal{F}, P) &= \left(\frac{\|\xi_P\|_{L^p(P)}}{\sigma_P}\right) \sup_{\beta\in\mathbb{R}^d\,:\,\langle\beta,\Sigma_P\beta\rangle=1} \|\langle\cdot,\beta\rangle\|_{L^p(P_{\mathbf{X}})}.\end{aligned}$$

In particular, for $p > 2$ the parameter $\kappa_p(\mathcal{F}, P)$ depends on hypercontractivity properties of the random variable ξ_P and of the one-dimensional marginals of $X \sim P_{\mathbf{X}}$.

4.1.2 Examples and history.

The study of regression with quadratic loss is of course quite classical. In the linear- and ridge- regression setting, a seminal contribution by Audibert and Catoni [Audibert and Catoni, 2011] gives an estimator with strong finite-sample performance under weak moment assumptions. However, [Audibert and Catoni, 2011] does not consider contamination, requires restricting their regression method to a bounded set, and achieves dimension-dependent bounds.

Lugosi and Mendelson [Lugosi and Mendelson, 2019b] gives regressors with "sub-Gaussian guarantees" in Problem 4.1 under weak moment assumptions, and for more general classes of functions than [Audibert and Catoni, 2011]. In particular, the function class may be unbounded and infinite-dimensional. Following up on [Lugosi and Mendelson, 2019b], Lecué and Lerasle [Lecué and Lerasle, 2020] obtained the best known results on this problem. In our notation, their bound for $\Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ in Problem 4.1 takes the following form (up to constant factors):

$$\theta_0(\mathcal{F}, P)^2 \left(r_{\mathsf{q}}(\delta_{\mathsf{q}}, \mathcal{F}, P) \vee r_{\mathsf{m}}(\delta_{\mathsf{m}}, \mathcal{F}, P) + \kappa_2(\mathcal{F}, P) \sqrt{\varepsilon + \frac{1}{n} \ln\left(\frac{1}{\alpha}\right)} \right), \tag{4.7}$$

with $\delta_{\mathsf{m}} = \frac{1}{6144\theta_0(\mathcal{F},P)^2}$ and $\delta_{\mathsf{q}} = \frac{1}{384\theta_0(\mathcal{F},P)}$. They also (implicitly) require a condition on the contamination level of the form $\varepsilon \leq \frac{1}{768\theta_0(\mathcal{F},P)^2}$. Both papers [Lugosi and Mendelson, 2019b, Lecué and Lerasle, 2020] rely on median of means type constructions. Specifically, [Lugosi and Mendelson, 2019b] requires knowledge of Φ_P and [Lecué and Lerasle, 2020, Theorem 7] requires a number of blocks bounded below by a quantity depending on the critical radii.

Remark 4.3. Lecué and Lerasle [Lecué and Lerasle, 2020] also consider regularized versions of Problem 4.1, and allow for non-identically distributed data, albeit in a way that requires equalities between some expectations relating different data points (X_i, Y_i). We do not consider either generalization in this thesis.

4.2 Main result

We present a trimmed-mean-based estimator for Problem 4.1 that satisfies improved bounds. Like [Lecué and Lerasle, 2020], we use the observation that

$$f_P^\star = \operatorname*{arg\,min}_{f \in \mathcal{F}} P\,\ell_f = \operatorname*{arg\,min}_{f \in \mathcal{F}} \left(\sup_{g \in \mathcal{F}} P\left(\ell_f - \ell_g\right) \right),$$

and define F_n via

$$\forall z_{1:n} \in (\mathbf{X} \times \mathbb{R})^n \,:\, F_n(z_{1:n}) \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \left(\sup_{g \in \mathcal{F}} \widehat{T}_{n,\phi n} \left(\ell_f - \ell_g, z_{1:n} \right) \right) \tag{4.8}$$

with the convention $\widehat{f}_n^{\varepsilon} := F_n\left(Z_{1:n}^{\varepsilon}\right)$.

Theorem 4.4 (Proof in §4.5.1)**.** *In the setting of Problem 4.1, let $\mathcal{P}$ denote the family of all probability distributions P over $(\mathbf{X}, \mathcal{X})$ that are 2-compatible with $\mathcal{F}$. Given $\alpha \in (0,1)$ and $\varepsilon > 0$ define*

$$\phi := \frac{\lfloor \varepsilon n \rfloor + \left\lceil \ln \frac{3}{\alpha} \right\rceil \vee \left\lceil \frac{\varepsilon n}{2} \right\rceil}{n} \text{ and assume } \phi + \frac{1}{2n} \ln \frac{3}{\alpha} \leq \frac{1}{96 \theta_0^2}. \tag{4.9}$$

Then, there is an event E with probability at least $1 - \alpha$ where

$$\left\| \widehat{f}_n^{\varepsilon} - f_P^{\star} \right\|_{L^2(P_{\mathbf{X}})} \leq \Phi_P \tag{4.10}$$

with $\Phi_P = \Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ given by

$$\begin{aligned} \Phi_P := 49152 & \left[r_{\mathsf{q}} \left(\frac{1}{32 \theta_0(\mathcal{F}, P)}, \mathcal{F}, P \right) \vee 16 r_{\mathsf{m}} \left(\frac{1}{448 \theta_0^2(\mathcal{F}, P)}, \mathcal{F}, P \right) \right] \\ & + 49152\, \theta_0(\mathcal{F}, P)^2 \left[\inf_{1 \leq q \leq 2} \kappa_q(\mathcal{F}, P) \left(\frac{\ln \frac{3}{\alpha}}{n} \right)^{1 - \frac{1}{q}} + \inf_{p \geq 1} \kappa_p(\mathcal{F}, P)\, \varepsilon^{1 - \frac{1}{p}} \right]. \end{aligned} \tag{4.11}$$

Moreover, in the same event E the following inequality holds:

$$R_P\left(\widehat{f}_n^{\varepsilon} \right) - R_P\left(f_P^{\star} \right) \leq \left(1 + \frac{1}{16 \theta_0(\mathcal{F}, P)^2} \right) \Phi_P^2. \tag{4.12}$$

When compared to the bound on $\Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ from (4.7), we see that our result matches the dependence on the critical radii from [Lecué and Lerasle, 2020]. In fact, Theorem 4.4 obtains improved values of δ_{m} and δ_{q}. The dependence on the contamination ε and on the moment parameters $\kappa_p(\mathcal{F}, P)$ is improved in our result and is optimal. The optimal dependence follows from the optimality of 3.1, since, when second moments exists, mean estimation can be viewed as a regression with quadratic loss over the class of constant functions.

Remark 4.5. As in [Lecué and Lerasle, 2020], Theorem 4.4 requires a restriction that $\varepsilon \leq c\, \theta_0(\mathcal{F}, P)^{-2}$ for some positive $c > 0$. In §4.5.2 we explain why some such restriction is necessary for any method, and show this relates to §4.4.2.

4.3 Algorithms for robust linear regression

We now present some heuristics for linear regression that are related to the theoretical results in Theorem 4.4. Our main finding, presented in §4.4, is that trimmed-mean-based methods tend to outperform ordinary least squares and robust alternatives based on median-of-means.

Before describing our heuristics, we note that Theorem 4.4 does not translate directly into a practical method. First of all, the definition of $\hat{f}_n^\varepsilon$ requires a choice of confidence level $1-\alpha$, which is (theoretically) unavoidable for any estimator [Devroye et al., 2016, Theorem 3.2]. Theorem 4.4 also requires some knowledge of the contamination level ε, which is also unavoidable for trimmed-mean-based methods. Finally, the min-max problem (4.8) defining $\hat{f}_n^\varepsilon$ poses computational challenges. Recall that the estimator $\hat{f}_n^\varepsilon = F_n(Z_{1:n}^\varepsilon)$ is defined by the min-max problem

$$F_n(z_{1:n}) \in \arg\min_{f\in\mathcal{F}} \left(\max_{g\in\mathcal{F}} \widehat{T}_{n,k} \left(\ell_f - \ell_g, z_{1:n} \right) \right). \tag{4.13}$$

The difference $\ell_f - \ell_g$ is convex in f and concave in g, which suggests that the min-max problem above can be efficiently solved. Unfortunately, the trimmed mean operation $\widehat{T}_{n,k}$ introduces complications; e.g. it is not obvious how to compute sub- and super-gradients. Besides that, the TM estimator requires knowledge of the (usually unknown) contamination level ε to choose the trimming parameter $k = \phi n$.

Nevertheless, we will argue that there are practical ways to choose the trimming parameter k and to compute regressors that seem to circumvent these challenges. This requires several algorithmic choices that we describe below. We also discuss an adaptation of the median-of-means regression procedure of [Lecué and Lerasle, 2020] that will be compared to our own method in the next section.

4.3.1 Preliminaries. We work in the setting of linear regression, corresponding to Example 4.1 above. For $\beta \in \mathbb{R}^d$, we write $\ell_\beta(x,y) := (\langle \beta, x\rangle - y)^2$ to denote the loss associated with the linear regression function $\langle \beta, \cdot \rangle$ on a point $(x,y) \in \mathbb{R}^d \times \mathbb{R}$. Given a trimming parameter $k \in [n]$; points $z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i \in [n]$; and vectors $\beta^m, \beta^M \in \mathbb{R}^d$, we may write:

$$\widehat{T}_{n,k} \left(\ell_{\beta^m} - \ell_{\beta^M}; z_{1:n} \right) = \frac{1}{n-2k} \sum_{i \in I_k(\beta^m, \beta^M, z_{1:n})} \left((\langle x_i, \beta^m \rangle - y_i)^2 - (\langle x_i, \beta^M \rangle - y_i)^2 \right),$$

where $I_k(\beta^m, \beta^M, z_{1:n})$ – called the *active set* for $(\beta^m, \beta^M, z_{1:n})$ – is the set of indices $i \in [n]$ that appear in the trimmed mean, i.e., the set obtained once the k largest and k smallest values of

$(\langle x_i, \beta^m\rangle - y_i)^2 - (\langle x_i, \beta^M\rangle - y_i)^2$ are removed (with ties broken arbitrarily). Our heuristic will be described in two steps. First, we describe how to optimize the choice of β^m and β^M for a specific k. Second, we present a cross-validation procedure to choose the trimming parameter.

4.3.2 Optimization for a fixed trimming level.

Assume that the trimming parameter is fixed at some value k. We consider two algorithms to evaluate (4.13): the Plug-in method (Algorithm 1) and the Alternating Direction Method of Multipliers (ADMM, Algorithm 2), which is an adaptation of the best performing method in [Lecué and Lerasle, 2020].

input : $(x_1, y_1), \cdots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$: the data
$\beta_0^m, \beta_0^M \in \mathbb{R}^d$: initial guesses
ϕ: trimming level
$T_{\max}$: number of iterations

output : $\beta_\phi^\star$: an approximate solution for the min-max problem

$t \leftarrow 0$
$k \leftarrow \lfloor \phi n \rfloor$
while $t < T_{\max}$ **do**
  $I_t^m \leftarrow I_k\left(\beta_t^m, \beta_t^M, z_{1:n}\right)$ (get active indices)
  $\beta_{t+1}^m \leftarrow \texttt{Fit}\left(\{(x_i, y_i)\}_{i \in I_t^m}\right)$
  $I_t^M \leftarrow I_k\left(\beta_{t+1}^m, \beta_t^M, z_{1:n}\right)$ (update active indices)
  $\beta_{t+1}^M \leftarrow \texttt{Fit}\left(\{(x_i, y_i)\}_{i \in I_t^M}\right)$
  $t \leftarrow t + 1$
end
$\beta_\phi^\star \leftarrow \arg\min\left\{\widehat{T}_{n,k}(\ell_\beta, z_{1:n}) \,:\, \beta \in \bigcup_{t=1}^{T_{\max}} \{\beta_t^m, \beta_t^M\}\right\}$.

Algorithm 1: Plug-in algorithm.

The Plug-in method relies on the existence of a black-box function `Fit` that on input $((x_i, y_i))_{i \in I}$ (with $I \subset [n]$ nonempty) will return a vector

$$\beta((x_i, y_i)_{i \in I}) \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \frac{1}{|I|} \sum_{i \in I} \left(\langle \beta, x_i\rangle - y_i\right)^2.$$

The idea is that the set I will correspond to the set of active indices, of size $(1-2\phi)n = n - 2k$. Te algorithm alternates between optimizing β^m and β^M for a fixed set I, and updating I to be the current active set. The output of the algorithm is vector produced in the iterations that minimizes the trimmed mean estimate of the loss.

input : $(x_1, y_1), \cdots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$: the data
$\beta_0^m, \beta_0^M \in \mathbb{R}^d$: a initial guess
ϕ: trimming level
$T_{\max}$: number of iterations
ρ: multiplier parameter

output : $\beta_\phi^\star$: an approximate solution for the min-max problem

$t \leftarrow 0$

$k \leftarrow \lfloor \phi n \rfloor$

while $t \leq T_{\max}$ **do**

$I_t^m \leftarrow I_k\left(\beta_t^m, \beta_t^M, z_{1:n}\right)$ (get active indices)

$\beta_{t+1}^m \leftarrow \left(x_{I_t^m}^T x_{I_t^m} + \rho I_d\right)^{-1}\left(x_{I_t^m}^T y_{I_t^m} + \rho \beta_t^m\right)$

$I_t^M \leftarrow I_k\left(\beta_{t+1}^m, \beta_t^M, z_{1:n}\right)$ (get updated active indices)

$\beta_{t+1}^M \leftarrow \left(x_{I_t^M}^T x_{I_t^M} + \rho I_d\right)^{-1}\left(x_{I_t^M}^T y_{I_t^M} + \rho \beta_t^M\right)$

$t \leftarrow t + 1$

end

$\beta_\phi^\star \leftarrow \arg\min\left\{\widehat{T}_{n,k}(\ell_\beta, z_{1:n}) \,:\, \beta \in \bigcup_{t=1}^{T_{\max}}\{\beta_t^m, \beta_t^M\}\right\}$.

Algorithm 2: Alternating Direction Method of Multipliers.

Algorithm 2 differs from its analogue in [Lecué and Lerasle, 2020] in three ways. First, their estimator is based on the MoM principle. Second, that paper considers a sparse regression setting with an ℓ_1 penalty. Third, the final output $\beta_\phi^\star$ in [Lecué and Lerasle, 2020] is simply the final iterate $\beta_{T_{\max}}^m$. Preliminary experiments show that choosing $\beta_\phi^\star$ as a minimizer of the trimmed empirical risk improves the performance of both algorithms more than 90% of the time in all settings considered in what follows.

For the Plug-in method, we set $T_{\max} = 20$ in all experiments, as solutions do not improve beyond this number iterations. For the same reasons, we set $T_{\max} = 50$ for ADMM. The parameter ρ in Algorithm 2 is set to 5 as in [Lecué and Lerasle, 2020]. The initial choices for β_0^m, β_0^M are random perturbations of OLS solutions on the full data.

4.3.3 Cross-validation. We now consider the problem of choosing the trimming level ϕ for best-possible performance. Following [Lecué and Lerasle, 2020] we choose k via a cross-validation procedure.

Let $(z_i)_{i=1}^n = (x_i, y_i)_{i=1}^n$ be the data points. Assume that $\phi_1 \leq \phi_2 \leq \cdots \leq \phi_m$ is an ordered grid

of possible choices for ϕ. Let $v \leq n$ be the number of folds: that is, $[n]$ is partitioned into sets $\{B_l\}_{l=1}^v$ with respective sizes $n_l := |B_l| \in \{\lfloor n/v \rfloor, \lfloor n/v \rfloor + 1\}$. The procedure works as follows.

1. For each choice of $(j, l) \in [m] \times [v]$, let $\beta^\star_{\phi_j}([n] - B_l)$ be the output of Algorithm 1 (or Algorithm 2) on the $n - n_l$ data points $(x_i, y_i)_{i \notin B_l}$.

2. For each choice of $(j, l) \in [m] \times [v]$, estimate the loss of $\beta^\star_{\phi_j}([n] - B_l)$ via a trimmed mean with level ϕ_j on the fold B_l:

$$L(j, l) := \widehat{T}_{n_l, \phi_j n_l} \left(\ell_{f_{j,l}}, (z_i)_{i \in B_l} \right).$$

3. For each $j \in [m]$, associate a loss with trimming level ϕ_j via

$$L(j) := \text{median}(L(j, l) \, : \, l \in [v]).$$

4. Choose $\phi^\star = \phi_{j^\star}$ where $j^\star = \arg\max_{j=2,\ldots,m} \frac{L(j-1)}{L(j)}$.

5. Compute the final estimator $\beta^\star_{\phi^\star}$ by running Algorithm 1 (or Algorithm 2) with trimming level $\phi^\star$ on the full dataset $(z_i)_{i \in [n]}$.

Intuitively, the choice of trimming level $\phi^\star = \phi_{j^\star}$ corresponds to the point at which increasing j induces the largest drop in the loss. This is inspired by the Slope Heuristic [Birgé and Massart, 2001]. We here call the choice of $\phi^\star$ made using step (4) above as choice by slope maximization (abbr. **max slope**). We also test a variant of step (4), that is used in [Lecué and Lerasle, 2020]:

4′. Choose $\phi^\star = \phi_{j^\star}$ where $j^\star = \arg\min_{j=1,\ldots,m} L(j)$.

We call this last variant choice by loss minimization (abbr. **min loss**). Both approaches will be compared in the next section. We note in passing that we have fixed $v = 5$ folds in all our experiments.

4.3.4 Median of Means (MoM). We compare the performance of the trimmed mean estimator against a variant of the Median of Means procedure from [Lecué and Lerasle, 2020]. MoM requires splitting the n data points into K buckets of approximately equal size. These splits are performed randomly at each iteration, as recommended in [Lecué and Lerasle, 2020]. The parameter K is a close analogue of the trimming parameter k in our procedure. This allows us to adapt the optimization and cross validation procedures described above. In brief:

- To adapt Algorithm 1, we use a different concept of active set. Consider the blocks $A_1, \ldots, A_K \subset [n]$ used by the median-of-means construction. Then the active set $I_k(\beta^m, \beta^M)$ is the block of indices A_r for which

$$\begin{aligned}&\frac{1}{\#A_r}\sum_{i\in A_r}(\ell_{\beta^m}(x_i,y_i)-\ell_{\beta^M}(x_i,y_i))\\&=\text{median}\left\{\frac{1}{\#A_s}\sum_{i\in A_s}(\ell_{\beta^m}(x_i,y_i)-\ell_{\beta^M}(x_i,y_i))\ :\ s\in[K]\right\},\end{aligned}$$

  with ties between blocks broken arbitrarily.

- When performing the optimization iterations, the blocks of median-of-means are resampled uniformly at random at each step.

- The cross validation procedure is now over choices of K_j. However, the loss estimate $L(j,l)$ are performed via a MoM estimator using the data in fold B_l with K_j/v blocks.

One important difference should be noted. The original algorithm in [Lecué and Lerasle, 2020] was designed for sparse linear regression in a $d \gg n$ setting. By contrast, our own analysis is restricted to $d \ll n$ and our algorithmic choices for MoM were optimized for this case. This explains why our version of the MoM method is somewhat different from that in [Lecué and Lerasle, 2020].

4.4 Experiments with linear regression

We perform experiments in two different data generation setups. Setup A (§4.4.1) favors robust methods and is analogous to [Lecué and Lerasle, 2020]. Setup B (§4.4.2) is closely related to Remark 4.5 and favors the OLS. In our experiments the cross-validation procedure (§4.3.3) uses $v = 5$ folds and selects the TM parameter ϕ among the values in

$$\Lambda := \left\{\varepsilon' + \frac{1}{30} : \varepsilon' \in \left\{0, \frac{2}{100}, \frac{4}{100}, \frac{6}{100}, \frac{8}{100}, \frac{10}{100}, \frac{15}{100}, \frac{20}{100}, \frac{30}{100}, \frac{40}{100}\right\}\right\},$$

and the number of buckets for the MoM from the values in $\{2\phi n + 1 : \phi \in \Lambda\}$. Each parameter combination underwent 96 runs of the experiment. A GitHub repository with code to reproduce all figures and experiments here can be found at

`github.com/lucasresenderc/trimmedmean`.

4.4.1 Setup A. We consider a linear model. Let $d \geq 1$ be an integer and $X_1, X_2, \cdots, X_n$ be i.i.d. standard Gaussian's. For $i \in [n]$, define

$$Y_i = \langle X_i, \beta^\star \rangle + \xi_i \text{ where } \beta^\star = \frac{1}{\sqrt{d}} [1, 1, \cdots, 1] \in \mathbb{R}^d$$

and the ξ_i are errors. To obtain the error random variables, we first sample

$$\eta_{1:n} \overset{i.i.d.}{\sim} \texttt{Normal}(0,1) \text{ (light tails) or } \eta_{1:n} \overset{i.i.d.}{\sim} \texttt{Student}(\nu) \text{ with } \nu \in \{1, 2, 4\} \text{ (heavy tails)}$$

independently from $X_{1:n}$. We then define $\xi_{1:n}$ in one of four ways.

1. Homoscedastic/non-skewed: simply set $\xi_i = \eta_i$ for each $i \in [n]$.
2. Homoscedastic/skewed: set $\xi_i = T(\eta_i)$ for each $i \in [n]$, where
$$T(x) = \log(x)\mathbf{1}_{x>1} + (x-1)\mathbf{1}_{x \leq 1} (x \in \mathbb{R}).$$
3. Heteroscedastic/non-skewed: in this case we take $\xi_i = \exp(\|X_i\|^2)/2)\, \eta_i$ for each $i \in [n]$.
4. Heteroscedastic/skewed: set $\xi_i = \exp(\|X_i\|^2)/2)\, T(\eta_i)$ for each $i \in [n]$, with T as above.

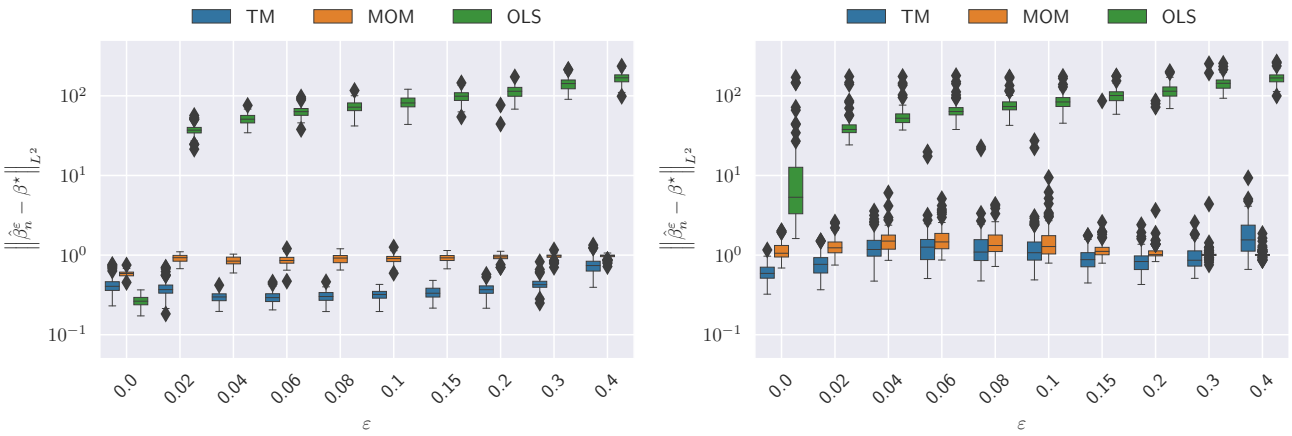
In total, the four different choices for the law of the $\eta_{1:n}$ and the four items above correspond to 16 cases. As is implicit above, the different choices for $\eta_{1:n}$ lead to different tail behaviors: Student with $\nu = 1$ is symmetric but does not have first moment; $\nu = 2$ gives a finite mean but not a finite variance; $\nu = 4$ induces finite mean, variance and third moment; and the normal law has finite moments of all orders.

The contamination model for Setup A is defined as follows: a set of indices $\mathcal{O} \subset [n]$ of size $\lfloor \varepsilon n \rfloor$ is chosen uniformly at random, and then one sets

$$(X_i^\varepsilon, Y_i^\varepsilon) = (X_i, Y_i) \; \forall i \notin \mathcal{O} \text{ and } (X_i^\varepsilon, Y_i^\varepsilon) = (\beta^\star, 10000) \; \forall i \in \mathcal{O}.$$

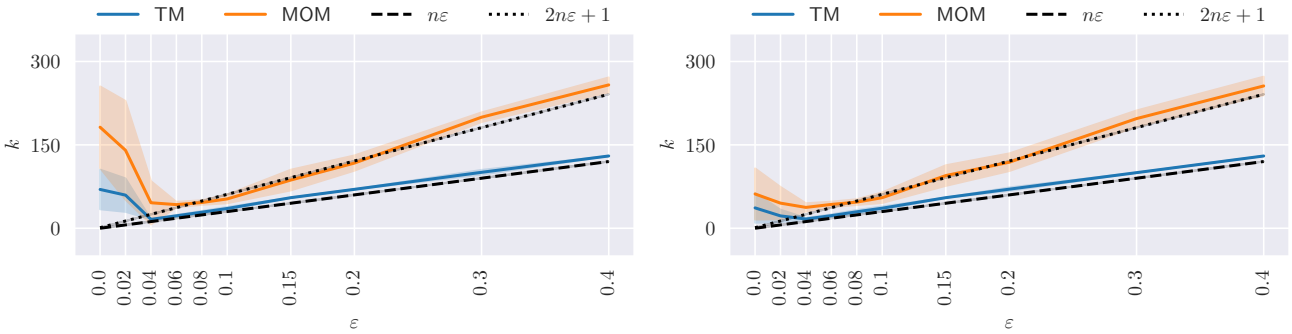
Remark 4.6. Before proceeding, notice that our setup is such that the L^2 error $\|\langle \beta, \cdot \rangle - \langle \beta^\star, \rangle\|_{L^2(P_{\mathbf{X}})}$ equals the Euclidean distance between β and $\beta^\star$.

Figure 4.1 presents two extreme cases of Setup A, both using $d = 20$ and $n = 300$ and on the homoscedastic/non-skewed configuration. As expected, the OLS performs poorly when $\varepsilon > 0$ and also when the error is heavy tailed. The TM performs well even when the contamination is high, with a slight increase in error compared to the no-contamination case. The MoM regression error is larger in all cases, with the exception of $\varepsilon \geq 0.2$ under heavy-tailed noise.



(a) Normal error ($\xi \sim \texttt{Normal}(0, 1)$). (b) Heavy-tailed error ($\xi \sim \texttt{Student}(\nu = 1)$).

Figure 4.1: L_2 error behavior varying the contamination proportion under Setup A, on the homoscedastic/non-skewed configuration.



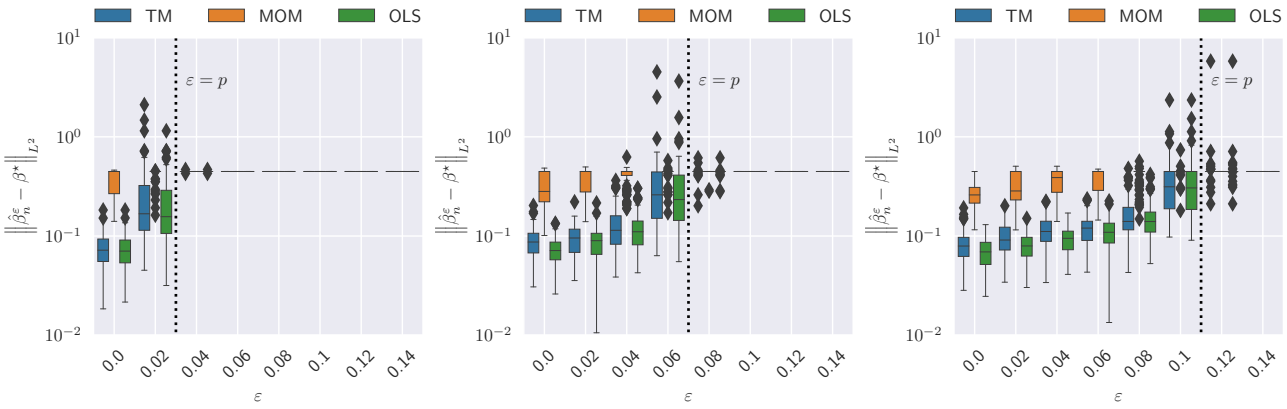
(a) Parameter choices for 4.1a. (b) Parameter choices for 4.1b.

Figure 4.2: Trimming/bucket parameters selected by cross-validation varying ε under Setup A.

Figure 4.2 offers insights on how the two methods select their parameters (the trimming level for TM and the number of buckets for MoM). Intuitively, in order to mitigate the effect of contamination, one expects cross-validation to choose $k = \varepsilon n$ for the TM and (by the pigeonhole principle) $K = 2\varepsilon n + 1$ buckets for the MoM. This is observed in Figure 4.2 for $\varepsilon > 0.02$, coinciding with the cases where the OLS performs poorly in Figure 4.1. For smaller ε, both TM and MoM make conservative parameter choices.

4.4.2 Setup B. This distribution and contamination model are a caricature of missing data and favors OLS. Let $p \in (0, 1]$. Take independent

$$X'_{1:n} \overset{i.i.d.}{\sim} \texttt{Normal}(0_d, I_{d\times d}),\ \xi_{1:n} \overset{i.i.d.}{\sim} \texttt{Normal}(0, 1) \text{ and } B'_{1:n} \overset{i.i.d.}{\sim} \texttt{Ber}(p).$$



(a) $p = 0.03$ (b) $p = 0.07$ (c) $p = 0.11$

Figure 4.3: Experiments varying the contamination proportion using Setup B.

The uncontaminated data is given by

$$X_i := B_i \frac{X_i'}{\sqrt{p}} \text{ and } Y_i = \langle X_i, \beta^\star \rangle + \xi_i \ \forall i \in [n],$$

where $\beta^\star = \frac{1}{d}[1, 1, \ldots, 1]$ as in the previous case. The contamination model is defined as follows. Take a subset $\mathcal{O} \subset \{i \in [n] \ : \ B_i = 1\}$ that satisfies $|\mathcal{O}| \leq \varepsilon n$ and is as large as possible. Set $B_i^\varepsilon = 0$ when $B_i = 0$ or $i \in \mathcal{O}$, and $B_i^\varepsilon = 1$ otherwise. The contaminated sample is

$$X_i^\varepsilon = B_i^\varepsilon \frac{X_i'}{\sqrt{p}} \text{ and } Y_i^\varepsilon = \langle X_i^\varepsilon, \beta^\star \rangle + \xi_i \ \forall i \in [n].$$

Figure 4.3 displays the performance of OLS, TM and MoM in Setup B with $d = 5$ and $n = 1000$. It contrasts with Figure 4.1 since the OLS is no longer losing to the robust estimators in most cases. Intuitively, OLS ignores points with $B_i^\varepsilon = 0$. The performance of TM is slightly worse than that of OLS, whereas MoM can often be significantly worse.

4.4.3 A more comprehensive comparison between methods. We design an experiment to simultaneously compare: the performance of the Plug-in method and the ADMM; the two cross-validation parameter selection variations (max slope and min loss) both in terms of its ability to estimate the contamination level and in terms of the error obtained; the overall performance of the trimmed-mean-based-regression against the median-of-means-based-regression.

For that experiment we use Setup A (§4.4.1). The different choices for the $\xi_{1:n}$ lead to 16 different possibilities for the uncontaminated data distribution. We vary the contamination

level ε in

$$\left\{0, \frac{2}{100}, \frac{4}{100}, \frac{6}{100}, \frac{8}{100}, \frac{10}{100}, \frac{15}{100}, \frac{20}{100}, \frac{30}{100}, \frac{40}{100}\right\}.$$

Thus giving 160 combinations of data distributions and contamination levels. In all cases, we set $d = 20$, $n = 300$, and we evaluate performance by performing 96 independent trials. Moreover, cross-validation with always be performed with $v = 5$ folds.

For each trial, and each regression method (TM vs MoM), we evaluated the performance of the four different combinations of optimization and cross-validation methods: ADMM with max slope, Plug-in with max slope, ADMM with min loss and Plug-in with min loss. The grid of values for ϕ used during the cross-validation for the trimmed-mean-based-regression was

$$\left\{\varepsilon' + \frac{1}{30} : \varepsilon' \in \left\{0, \frac{2}{100}, \frac{4}{100}, \frac{6}{100}, \frac{8}{100}, \frac{10}{100}, \frac{15}{100}, \frac{20}{100}, \frac{30}{100}, \frac{40}{100}\right\}\right\}$$

and for the selection of the number of buckets K for the median-of-means-based-regression was

$$\left\{2\left(\varepsilon' + \frac{1}{30}\right) + 1 : \varepsilon' \in \left\{0, \frac{2}{100}, \frac{4}{100}, \frac{6}{100}, \frac{8}{100}, \frac{10}{100}, \frac{15}{100}, \frac{20}{100}, \frac{30}{100}, \frac{40}{100}\right\}\right\}.$$

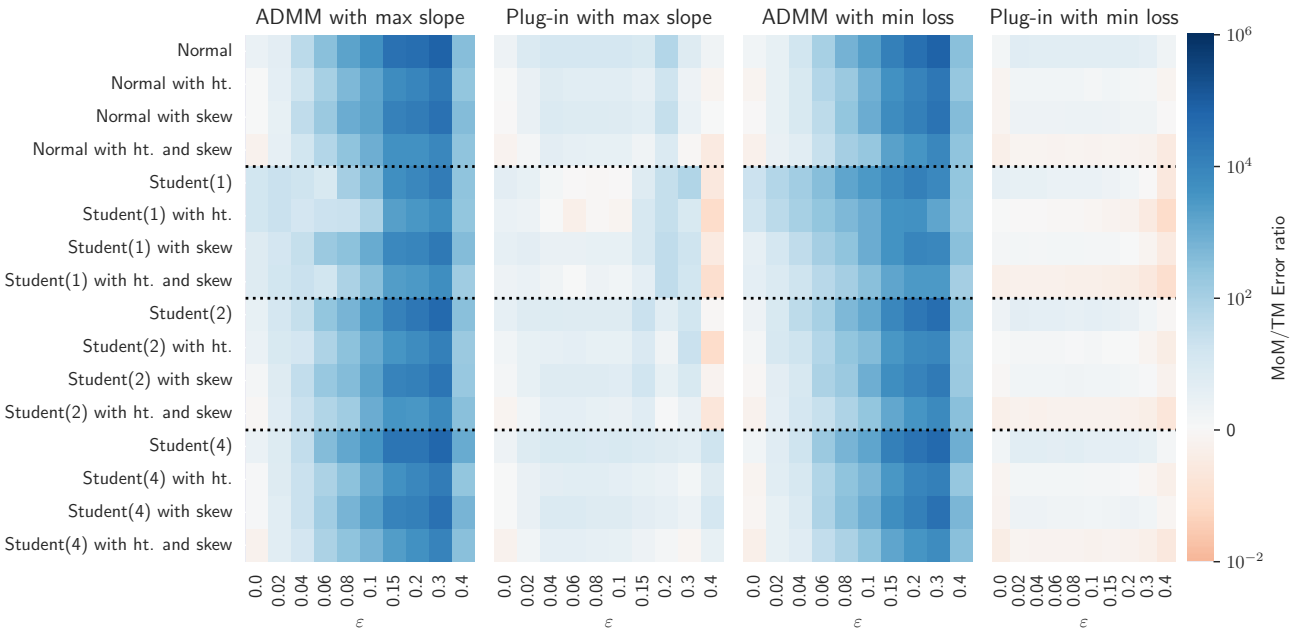


Figure 4.4: Ratio between the L^2 error of the median-of-means-based-regression and the error of the trimmed-mean-based-regression on the four combinations of algorithm and cross-validation strategy. Blue: TM outperforms MoM; Orange: MoM outperforms TM.

Figure 4.4 compares how median-of-means-based-regression and trimmed-mean-based-regression perform in each combination of algorithm and cross-validation alternative. The

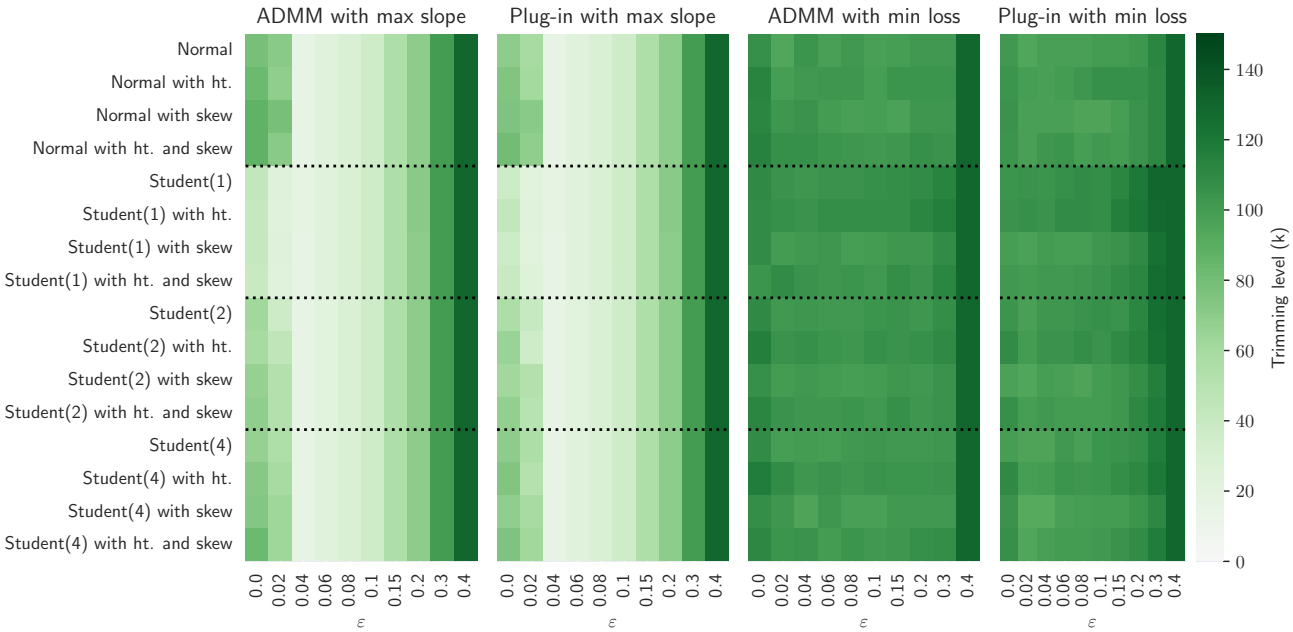


Figure 4.5: Cross-validation procedure for trimmed-mean-based-regression: average trimming level ($k = \phi n$) selected by cross-validation on the four combinations of algorithm and cross-validation strategy.

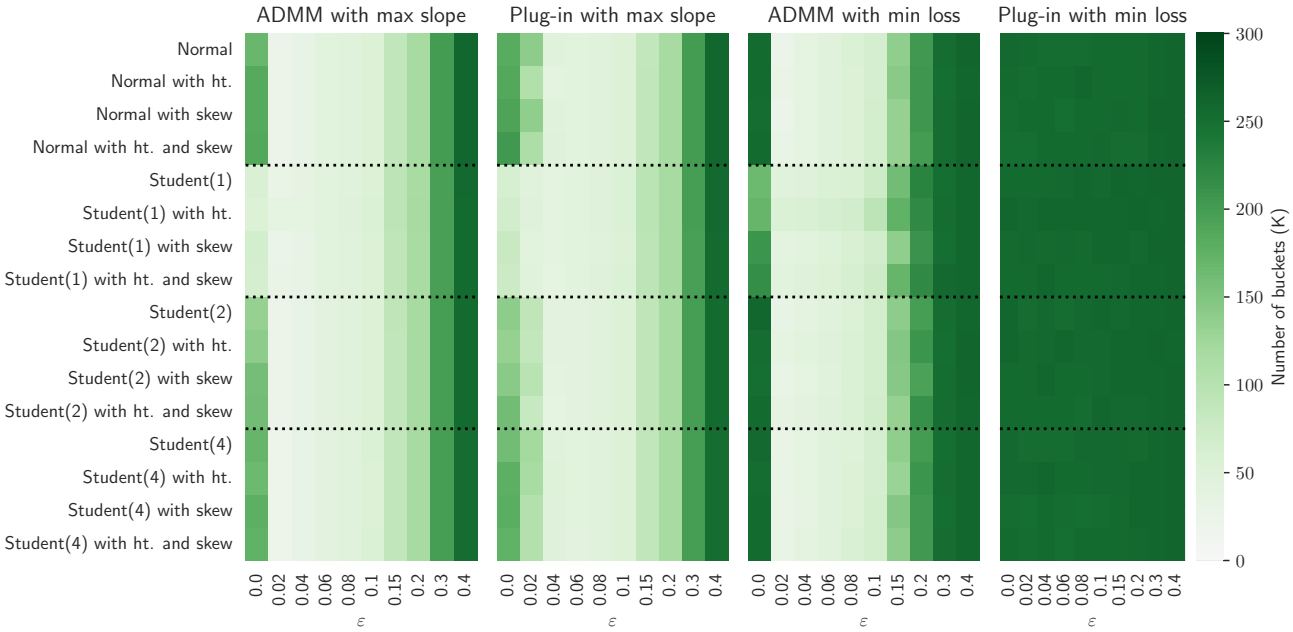


Figure 4.6: Cross-validation procedure for median-of-means-based-regression: average number of buckets (K) selected by cross-validation on the four combinations of algorithm and cross-validation strategy.

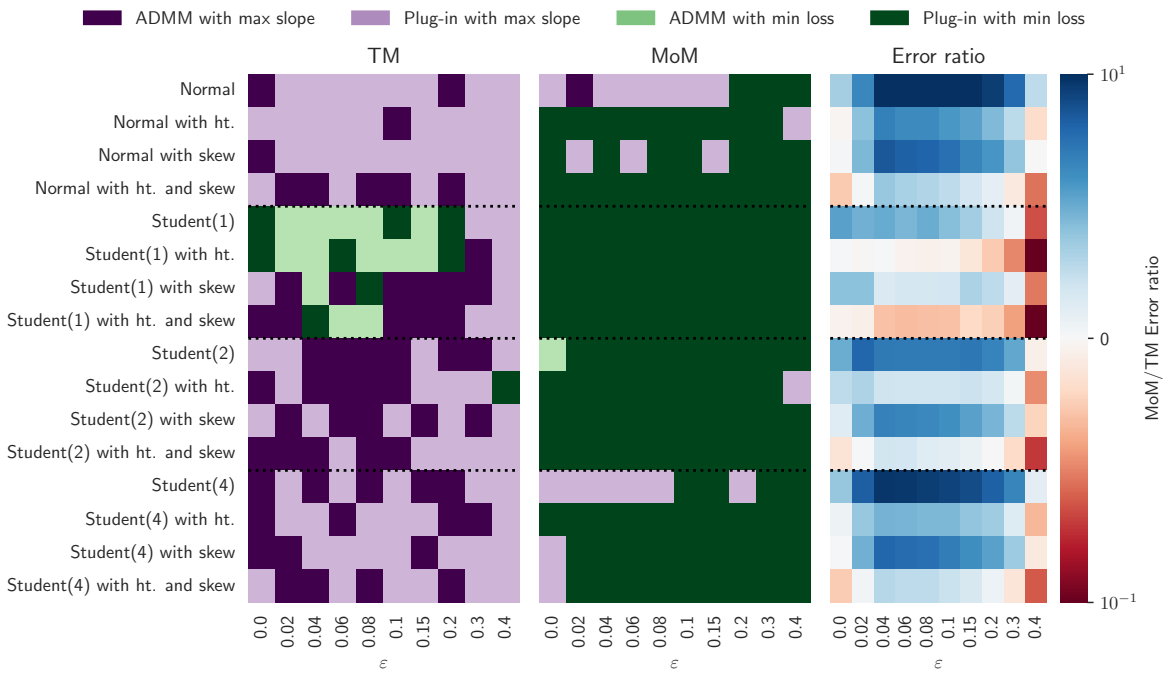


Figure 4.7: First panel: best choice of algorithm and cross-validation strategy to Trimmed Mean. Second panel: best choice of algorithm and cross-validation strategy to Median of Means. Last panel: ratio of the errors using the best choices for both Median of Means and Trimmed Mean.

colors in the plot represent the ratio of the L^2 errors of MoM and TM estimators in each setup: the bluer the color, the bigger the advantage of TM over MoM. When ADMM is used, the trimmed mean significantly outperforms MoM. By contrast, the difference between the MoM and the TM with the Plug-in method is less noticeable.

Figures 4.5 and 4.6 shows the trimming level $k = \phi n$ the average number of buckets K and selected by the cross-validation procedures for the median-of-means-based-regression and the trimmed-mean-based-regression procedures, respectively. In both cases, the min slope heuristic helps the method choose a parameter that is more closely related to the contamination level. This effect is more pronounced for TM, but also visible for MoM. We do note that both cross-validation methods are conservative for small contamination levels.

The leftmost panels of Figure 4.7 shows the best combination of optimization algorithm and cross-validation strategy for each choice of distribution and contamination level, both for TM and MoM. In the rightmost panel, we plot the ratio of the L^2 errors of the best combinations for MoM and TM. We highlight the following conclusions:

- For the trimmed mean, the combination of Plug-in optimization and max slope cross-validation is the best-performing method.
- For median-of-means, by contrast, plug-in with min-loss cross-validation is the best performing method.
- The best TM outperforms the best MoM in most cases. Exceptions occur when the error has a very heavy tail and heteroscedasticity is present, or when the contamination level is very high ($\varepsilon \geq 0.3$).

4.5 Proofs

4.5.1 Bounds for regression. We prove Theorem 4.4 below. The notation introduced in §4.1.1 and in the statement of Theorem 4.4 will be used throughout this subsection. The probability measure P will often be implicit throughout the section. We will sometimes use the following property,

$$\forall f \in \mathcal{F} \,:\, P\,\mathsf{m}_f = P\,\xi_P(X,Y)(f(X) - f_P^\star(X)) \leq 0, \tag{4.14}$$

where $\mathsf{m}_f = \mathsf{m}_{f,P}$ is the "multiplier term" from (4.4). Indeed, (4.14) is the first-order optimality condition for $f_P^\star = \arg\min_{f\in\mathcal{F}} P\,(Y - f(X))^2$.

In what follows, $Z^\varepsilon_{1:n}$ is an ε-contaminated sample from P. Similarly to the proof of Theorem 3.3, we write $\widehat{T}^\varepsilon_{n,k}(\cdot) := \widehat{T}_{n,k}(\cdot, Z^\varepsilon_{1:n})$. The estimator $\widehat{f}^\varepsilon_n \in \mathcal{F}$ of $f_P^\star \in \mathcal{F}$ is obtained solving the minimization problem

$$\widehat{f}^\varepsilon_n \in \arg\min_{f\in\mathcal{F}} \left(\sup_{g\in\mathcal{G}} \widehat{T}^\varepsilon_{n,\phi n}\left(\ell_f - \ell_g\right)\right).$$

The next Lemma reduces Theorem 4.4 to proving localized upper and lower bounds on certain trimmed means. Introduce the notation:

$$r_f := \|f - f_P^\star\|_{L_2(P_{\mathbf{X}})}.$$

Lemma 4.7. *If $r > 0$ and $\gamma > 0$ are such that*

$$\inf_{r_f = r} \widehat{T}^\varepsilon_{n,\phi n}\left(\ell_f - \ell_{f_P^\star}\right) \geq \gamma \geq 2 \sup_{r_f \leq r} \widehat{T}^\varepsilon_{n,\phi n}\left(\mathsf{m}_f - P\mathsf{m}_f\right), \tag{4.15}$$

then $r_{\widehat{f}^\varepsilon_n} = \left\|\widehat{f}^\varepsilon_n - f_P^\star\right\|_{L^2(P_{\mathbf{X}})} \leq r$ and $R(\widehat{f}^\varepsilon_n) - R(f_P^\star) \leq r^2 + 2\gamma$.

Proof The proof proceeds in three stages that follow the "localization + fixed point" outline from previous work [Lecué and Lerasle, 2020, Mendelson, 2015]. In the first stage, we use a localization argument and show:

$$\forall f \in \mathcal{F} : \widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_f - \ell_{f_P^\star}\right) \begin{cases} > \gamma, & r_f > r; \\ \geq -\gamma, & r_f \leq r. \end{cases} \tag{4.16}$$

In the second stage, we bound $r_{\hat{f}^{\varepsilon}_n}$ via (4.16) and a "fixed point" argument. In the final stage, we notice that an excess risk bound is implicit in the first two steps.

Localization: our goal here is to prove (4.16). Since $\mathcal{F}$ is convex, we can scale down elements $f \in \mathcal{F}$ with $r_f > r$ to elements $\bar{f} \in \mathcal{F}$ with $r_{\bar{f}} = r$ and bound the corresponding trimmed mean via our assumption. Explicitly, assume $q_f = \frac{r_f}{r} > 1$ and define $\bar{f} = f_P^\star + \frac{f - f_P^\star}{q_f}$, so that $\bar{f} \in \mathcal{F}$ by convexity and $r_{\bar{f}} = r$. We obtain

$$\begin{aligned} \widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_f - \ell_{f_P^\star}\right) &= \widehat{T}^{\varepsilon}_{n,\phi n}\left(q_f^2\,(\bar{f} - f_P^\star)^2 - 2\mathsf{m}_{\bar{f}}\right) \\ &\geq q_f\,\widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_{\bar{f}} - \ell_{f_P^\star}\right) > \gamma \text{ using (4.15) and } q_f > 1. \end{aligned}$$

This proves (4.16) for $r_f > r$. For $r_f \leq r$, we use (4.14) and $(f - f_P^\star)^2 \geq 0$ to obtain

$$\begin{aligned} \widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_f - \ell_{f_P^\star}\right) &= \widehat{T}^{\varepsilon}_{n,\phi n}\left((f - f_P^\star)^2 - 2\mathsf{m}_f\right) \\ &\geq \widehat{T}^{\varepsilon}_{n,\phi n}\left((f - f_P^\star)^2 - 2\left(\mathsf{m}_f - P\mathsf{m}_f\right)\right) \\ &\geq -2\widehat{T}^{\varepsilon}_{n,\phi n}\left(\mathsf{m}_f - P\mathsf{m}_f\right) \geq -\gamma \text{ by (4.15)}. \end{aligned}$$

Fixed point argument: for any $f \in \mathcal{F}$, let

$$\Delta(f) = \sup_{g \in \mathcal{F}} \widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_f - \ell_g\right),$$

so that $\hat{f}^{\varepsilon}_n$ minimizes Δ over $\mathcal{F}$. Therefore,

$$\widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_{\hat{f}^{\varepsilon}_n} - \ell_{f_P^\star}\right) \leq \Delta(\widehat{f}^{\varepsilon}_n) \leq \Delta(f_P^\star) \tag{4.17}$$

The bounds in (4.16) show that for any $g \in \mathcal{F}$:

$$\widehat{T}^{\varepsilon}_{n,\phi n}(\ell_{f_P^\star} - \ell_g) \leq \begin{cases} \gamma, & r_g \leq r; \\ -\gamma, & r_g > r. \end{cases}$$

Since $\gamma > 0$, we obtain $\Delta(f_P^\star) \leq \gamma$. This implies via (4.17) that $\widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_{\widehat{f}^{\varepsilon}_n} - \ell_{f_P^\star}\right) \leq \gamma$. We deduce that $r_{\widehat{f}^{\varepsilon}_n} \leq r$ via assumption (4.15).

Excess risk: we have:

$$R_P(\hat{f}_n^\varepsilon) - R_P(f_P^\star) = r^2_{\hat{f}_n^\varepsilon} - 2\, P\, \mathsf{m}_{\hat{f}_n^\varepsilon}.$$

The first term in the RHS is $\leq r^2$ by the above. The second can be bounded by:

$$\begin{aligned} -2P\mathsf{m}_{\widehat{f}_n^\varepsilon} &= 2\widehat{T}^\varepsilon_{n,\phi n}\left(\mathsf{m}_{\widehat{f}_n^\varepsilon} - P\mathsf{m}_{\widehat{f}_n^\varepsilon}\right) - 2\widehat{T}^\varepsilon_{n,\phi n}\left(\mathsf{m}_{\widehat{f}_n^\varepsilon}\right) \\ &\leq 2\widehat{T}^\varepsilon_{n,\phi n}\left(\mathsf{m}_{\widehat{f}_n^\varepsilon} - P\mathsf{m}_{\widehat{f}_n^\varepsilon}\right) + \widehat{T}^\varepsilon_{n,\phi n}\left(\ell_{\hat{f}_n^\varepsilon} - \ell_{f_P^\star}\right) \\ &\leq \gamma + \Delta(\hat{f}_n^\varepsilon) \leq 2\gamma, \end{aligned}$$

where the last line follows from (4.15) combined with $r_{\widehat{f}_n^\varepsilon} \leq r$ and the calculations in the previous step. ■

We now apply Lemma 4.7 to prove our main result on regression.

Proof [Proof of Theorem 4.4] We continue to use the notational conventions introduced above. Our goal is to find an event E and a constant $\gamma > 0$ such that assumption (4.15) of Lemma 4.7 holds in E, with the value of $r := \Phi_P(\mathcal{F}, n, \alpha, \varepsilon)$ defined in (4.11), and a suitable $\gamma > 0$. Let $x := \ln \frac{3}{\alpha} \geq \frac{1}{3}$. Following (4.11), we set

$$\theta_0 := \theta_0(\mathcal{F}, P),\ \delta_{\mathsf{q}} := \frac{1}{32\theta_0} \text{ and } \delta_{\mathsf{m}} = \frac{1}{448\theta_0^2}. \tag{4.18}$$

First step: lower bound quadratic part by Lipschitz term.

Lemma 4.7 requires control from below of $\ell_f - \ell_{f_P^\star}$, which includes a quadratic term. However, our assumptions are on the process $f - f_P^\star$ without the square. Therefore, our first step will be to find a bounded Lipschitz minorant for the quadratic term, to which we can apply concentration and contraction. Specifically, consider $a, c > 0$ with $2c > a$ (specific values to be chosen later). Define

$$\psi(y) = r^2 \left(2a\left(\frac{|y|}{r} \wedge c\right) - a^2\right)_+ \quad (y \in \mathbb{R}).$$

Then:

1. $0 \leq \psi(y) \leq y^2$ for all $y \in \mathbb{R}$: this follows from the fact that the graph of $y \mapsto y^2$ is lower bounded by the tangent line at $y = ar$;
2. ψ is $2ar$-Lipschitz with $\psi(0) = 0$; and
3. ψ is bounded above by the constant $M_q := r^2(2ac - a^2)$.

Thus, for all $f \in \mathcal{F}$ with $r_f = r$:

$$\widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_f - \ell_{f^\star_P}\right) \geq \widehat{T}^{\varepsilon}_{n,\phi n}\left(\psi(f - f^\star_P) - 2(\mathsf{m}_f - P\mathsf{m}_f)\right), \tag{4.19}$$

where we also used (4.14).

Second step: define E and lower bound $\mathbb{P}\left[(\right] E)$ via Theorem 3.3.

Recall that our goal is to prove the existence of an event E with $\mathbb{P}\left[(\right] E) \geq 1 - \alpha$ so that, when E holds, both (4.10) and (4.12) are satisfied. To do this, we recall the definition of $\mathcal{F}_{\mathsf{q}}(r) = \mathcal{F}_{\mathsf{q}}(r, P)$ and $\mathcal{F}_{\mathsf{m}}(r) = \mathcal{F}_{\mathsf{m}}(r, P)$ from §4.1.1, and set:

$$\mathcal{F}_1 \;:=\; \psi \circ \mathcal{F}_{\mathsf{q}}(r) = \left\{\psi(f - f^\star_P) \,:\, f \in \mathcal{F},\, \|f - f^\star_P\|_{L^2(P_{\mathbf{X}})} = r\right\}; \tag{4.20}$$

$$\mathcal{F}_2 \;:=\; \mathcal{F}_{\mathsf{m}}(r) = \left\{\mathsf{m}_f - P\mathsf{m}_f : f \in \mathcal{F},\; \|f - f^\star_P\|_{L^2(P_{\mathbf{X}})} \leq r\right\}. \tag{4.21}$$

Now define

$$\eta_{\mathsf{q}} := \left(6\phi + \frac{3x}{n}\right) M_{\mathsf{q}} + \nu_2\left(\mathcal{F}_1\right)\sqrt{\frac{2x}{n}}, \tag{4.22}$$

with M_q as above, and

$$\eta_{\mathsf{m}} := \left(6\phi + \frac{3x}{n}\right) M_{\mathsf{m}} + \mathrm{rem}_{M_{\mathsf{m}}}\left(\mathcal{F}_2\right) + \nu_2\left(\tau_{M_{\mathsf{m}}} \circ \mathcal{F}_2\right)\sqrt{\frac{2x}{n}}, \tag{4.23}$$

with M_{m} soon to be defined. The event E is defined as follows:

$$E = \left\{ \begin{array}{c} \forall a_{\mathsf{q}}, a_{\mathsf{m}} \in \mathbb{R}, f_{\mathsf{q}} \in \mathcal{F}_{\mathsf{q}}(r),\; f_{\mathsf{m}} \in \mathcal{F}_{\mathsf{m}}(r) \\ \left|\widehat{T}^{\varepsilon}_{n,\phi n}(a_{\mathsf{q}}\psi(f_{\mathsf{q}}) + a_{\mathsf{m}} f_{\mathsf{m}}) - a_{\mathsf{q}} P\psi(f_{\mathsf{q}})\right| \leq \\ |a_{\mathsf{q}}|\left\{2\mathsf{Emp}_n\left(\mathcal{F}_1^o\right) + \eta_{\mathsf{q}}\right\} + |a_{\mathsf{m}}|\left\{2\mathsf{Emp}_n\left(\tau_{M_{\mathsf{m}}} \circ \mathcal{F}_2\right) + \eta_{\mathsf{m}}\right\} \end{array} \right\}$$

The event E is measurable because $\mathcal{F}$ has a countable dense subset. We now argue that E is precisely the kind of event whose probability is bounded in Theorem 3.3. To see this, we follow the notation in that theorem, set $m := 2$, $x_1 = x_2 = x$ and $\mathcal{F}_1$ and $\mathcal{F}_2$ as above. We make the following choices of M_i, b_i and t_i:

- The functions in $\mathcal{F}_1 = \psi \circ \mathcal{F}_{\mathsf{q}}(r)$ take values in $[0, M_{\mathsf{q}}]$, so their expectations are also in this range. It follows that all functions in the centered class $\mathcal{F}_1^o$ are bounded by $M_1 := M_{\mathsf{q}}$ in absolute value. This means we can take $t_1 = b_1 = 0$.
- Now consider $\mathcal{F}_2 = \mathcal{F}_{\mathsf{m}}(r)$. Note that $Pf_{\mathsf{m}} = 0$ for all $f_{\mathsf{m}} \in \mathcal{F}_{\mathsf{m}}(r)$, so $\mathcal{F}_2 = \mathcal{F}_2^o$. Since $\mathcal{F}_2$ may be unbounded, we will take $b_2 = 1$, and use Lemma 3.4 to obtain $M_{\mathsf{m}} := M_2$ and

$t_2 \geq x_2$ satisfying the assumptions of Theorem 3.3 and also the bound

$$2\mathsf{Emp}_n\left(\tau_{M_\mathsf{m}} \circ \mathcal{F}_2\right) + \eta_\mathsf{m} \leq C_\varepsilon \left\{ 8\mathsf{Emp}_n(\mathcal{F}_2) + \inf_{q\in[1,2]} \nu_q(\mathcal{F}_2) \left(\frac{\ln\frac{3}{\alpha}}{n}\right)^{1-\frac{1}{q}} + \inf_{p\geq 1} \nu_p(\mathcal{F}_2)\varepsilon^{1-\frac{1}{p}} \right\},$$

where $C_\varepsilon := 384\left(1 + \frac{\varepsilon}{\varepsilon \wedge \left(\frac{1}{2}-\varepsilon\right)}\right)$ can be bounded by 768 noticing that (4.9) implies $\varepsilon \leq \frac{1}{96\theta_0^2} < \frac{1}{4}$. Using further that $\nu_p(\mathcal{F}_2) = \nu_p(\mathcal{F}_\mathsf{m}(r)) \leq r\,\kappa_p(\mathcal{F})$, we obtain the bound:

$$2\mathsf{Emp}_n\left(\tau_{M_\mathsf{m}} \circ \mathcal{F}_2\right) + \eta_\mathsf{m} \leq 768 \left\{ 8\mathsf{Emp}_n(\mathcal{F}_2) + r \inf_{q\in[1,2]} \kappa_q(\mathcal{F}) \left(\frac{\ln\frac{3}{\alpha}}{n}\right)^{1-\frac{1}{q}} + r \inf_{p\geq 1} \kappa_p(\mathcal{F})\varepsilon^{1-\frac{1}{p}} \right\}, \tag{4.24}$$

The upshot of this discussion is that Theorem 3.3 can indeed be used to bound the probability of E, and we obtain:

$$\mathbb{P}\left[(\right] E) \geq 1 - 3e^{-x} \geq 1 - \alpha.$$

From now on, we perform all calculations deterministically while assuming that E holds.

Third step: bounds assuming E holds.

We combine the lower bound from the first step with the one defining the event E. Taking $a_\mathsf{q} = 0$, $a_\mathsf{m} = 2$ in E gives, for $r_f \leq r$,

$$2\widehat{T}^{\varepsilon}_{n,\phi n}\left(\mathsf{m}_f - P\mathsf{m}_f\right) \leq \underbrace{2\left\{2\mathsf{Emp}_n\left(\tau_{M_\mathsf{m}} \circ \mathcal{F}_2\right) + \eta_\mathsf{m}\right\}}_{\text{(i)}}. \tag{4.25}$$

Similarly, to consider $r_f = r$ we take $a_\mathsf{q} = 1$, $a_\mathsf{m} = -2$ in E and obtain

$$\begin{aligned} \widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_f - \ell_{f^\star_P}\right) &\geq \widehat{T}^{\varepsilon}_{n,\phi n}\left(\psi(f - f^\star_P) - 2\left(\mathsf{m}_f - P\mathsf{m}_f\right)\right) \\ &\geq \underbrace{\inf_{f_\mathsf{q}\in\mathcal{F}_\mathsf{q}(r)} P\psi(f_\mathsf{q})}_{\text{(ii)}} - \left\{ \underbrace{2\mathsf{Emp}_n\left(\mathcal{F}_1^o\right)}_{\text{(iii)}} + \underbrace{\eta_\mathsf{q}}_{\text{(iv)}} \right\} \\ &\quad - \underbrace{2\left\{2\mathsf{Emp}_n\left(\tau_{M_\mathsf{m}} \circ \mathcal{F}_2\right) + \eta_\mathsf{m}\right\}}_{\text{(i)}}, \end{aligned} \tag{4.26}$$

where the last inequality is where we need E to hold.

The bounds in (4.26) and (4.25) are still unwieldy. To obtain more useful bounds for (i - iv), we need a few calculations that are quite messy and not too enlightening.

Bound (i). Note that, since $r \geq r_{\mathsf{m}}(\delta_{\mathsf{m}})$ one has $\mathcal{F}_{\mathsf{m}}(r)/r \subset \mathcal{F}_{\mathsf{m}}(r_{\mathsf{m}}(\delta_{\mathsf{m}}))/r_{\mathsf{m}}(\delta_{\mathsf{m}})$. Combining this with symmetrization and the definition of $r_{\mathsf{m}}(\delta_{\mathsf{m}})$, we obtain:

$$\mathsf{Emp}(\mathcal{F}_{\mathsf{m}}(r)) \leq 2\mathsf{Rad}(\mathcal{F}_{\mathsf{m}}(r)) \leq \frac{2r}{r_{\mathsf{m}}(\delta_{\mathsf{m}})}\mathsf{Rad}(\mathcal{F}_{\mathsf{m}}(r_{\mathsf{m}}(\delta_{\mathsf{m}}))) \leq 2\delta_{\mathsf{m}}\, r\, r_{\mathsf{m}}(\delta_{\mathsf{m}}).$$

By (4.24), we obtain:

$$\text{(i)} \leq 1536\, r \left\{ 16\delta_{\mathsf{m}} r_{\mathsf{m}}(\delta_{\mathsf{m}}) + \inf_{q\in[1,2]} \kappa_q(\mathcal{F}) \left(\frac{x}{n}\right)^{1-\frac{1}{q}} + \inf_{p\geq 1} \kappa_p(\mathcal{F}) \varepsilon^{1-\frac{1}{p}} \right\} \leq 14\delta_{\mathsf{m}} r^2, \tag{4.27}$$

where the upper bound is provided by the choice of $r = \Phi_P(\mathcal{F}, \alpha, n, \varepsilon)$ in (4.11) and the choice of δ_{m} in (4.18).

Bound (ii). Using $y \wedge c \geq y - \frac{y^2}{c}$ and the definition of θ_0 we can bound, for $r_f = r$,

$$P\psi(f - f_P^\star) \geq P\left\{ r^2 \left(2a \left(\frac{|f - f_P^\star|}{r} - \frac{|f - f_{\mathsf{q}}|^2}{r^2 c} \right) - a^2 \right) \right\} \geq r^2 \left\{ 2a \left(\frac{1}{\theta_0} - \frac{1}{c} \right) - a^2 \right\}.$$

Bound (iii). Using contraction and symmetrization (Theorem 2.1) together with the fact that $r \geq r_{\mathsf{q}}(\delta_{\mathsf{q}})$ give:

$$\mathsf{Emp}_n\left(\mathcal{F}_1^o\right) = \mathsf{Emp}_n\left(\psi \circ \mathcal{F}_{\mathsf{q}}(r)\right) \leq 2\mathsf{Rad}_n\left(\psi \circ \mathcal{F}_{\mathsf{q}}(r)\right) \leq 4ar\mathsf{Rad}_n\left(\mathcal{F}_{\mathsf{q}}(r)\right) \leq 4a\delta_{\mathsf{q}} r^2.$$

Bound (iv). Since $\mathcal{F}_1 = \psi \circ \mathcal{F}_{\mathsf{q}}(r)$, $\psi(0) = 0$ and ψ is $2ar$-Lipschitz,

$$\nu_2\left(\mathcal{F}_1^o\right) \leq 2ar \sup_{f_{\mathsf{q}} \in \mathcal{F}_{\mathsf{q}}(r)} \|f_{\mathsf{q}}\|_{L^2(P_{\mathbf{X}})} = 2ar^2.$$

Observe that $\mathrm{rem}_{M_{\mathsf{q}}}\left(\mathcal{F}_1^o\right) = 0$ because $\mathcal{F}_1^o$ is uniformly bounded by M_{q}, thus

$$\eta_{\mathsf{q}} \leq \left(6\phi + \frac{3x}{n}\right) M_{\mathsf{q}} + 2ar^2\sqrt{\frac{2x}{n}}.$$

End of third step. From (4.25) and (4.27) we have

$$2 \sup_{r_f \leq r} \widehat{T}^{\varepsilon}_{n,\phi n}\left(\mathsf{m}_f - P\mathsf{m}_f\right) \leq 14\delta_{\mathsf{m}} r^2. \tag{4.28}$$

Observe that (4.9) and our choice $x = \ln\frac{3}{\alpha}$ imply

$$\sqrt{\frac{2x}{n}} \leq \sqrt{\frac{1}{4}\left(6\phi + \frac{3x}{n}\right)} \leq \frac{1}{8\theta_0},$$

so, (4.26) and the bounds on (i-iv) give

$$\widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_f - \ell_{f^\star_P}\right) \geq r^2\left\{2a\left(\frac{1}{\theta_0}-\frac{1}{c}\right)-a^2\right\} - 8a\delta_{\mathsf{q}}r^2 \tag{4.29}$$
$$- \left(6\phi + \frac{3x}{n}\right)M_{\mathsf{q}} - 2ar^2\sqrt{\frac{2x}{n}} - 14\delta_{\mathsf{m}}r^2$$
$$((4.9) + M_q \leq 2acr^2) \geq r^2\left\{2a\left(\frac{1}{\theta_0}-\frac{1}{c}-\frac{c}{16\theta_0^2}-\frac{1}{8\theta_0}-4\delta_{\mathsf{q}}\right)-a^2\right\} - 14\delta_{\mathsf{m}}r^2.$$

Final step: apply Lemma 4.7 via choices of constants.

We finish the proof via an application of Lemma 4.7, assuming as before that E holds. Recall $r := \Phi_P$ and take $\gamma := (32\theta_0^2)^{-1}r^2$. We defined $\delta_{\mathsf{m}} = (448\theta_0^2)^{-1}$ in (4.18); therefore, (4.28) gives condition (4.15) for $r_f \leq r$. Now consider the lower bound for $\widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_f - \ell_{f^\star_P}\right)$ when $r_f = r$. Recall from (4.18) that $\delta_{\mathsf{q}} = (32\theta_0)^{-1}$. Insert this into the RHS of (4.29) and optimize over $2c > a > 0$. This leads to the choices $c = 4\theta_0$ and $a = c^{-1}$, giving

$$r^2 \sup_{a,c>0:\ 2c>a} 2a\left(\frac{1}{\theta_0}-\frac{1}{c}-\frac{c}{16\theta_0^2}-\frac{1}{8\theta_0}-4\delta_{\mathsf{q}}\right)-a^2 = \frac{r^2}{16\theta_0^2} = 2\gamma.$$

Finally, combine (4.29) and (4.25) to obtain:

$$\text{when } r_f = r:\ \widehat{T}^{\varepsilon}_{n,\phi n}\left(\ell_f - \ell_{f^\star_P}\right) \geq \gamma.$$

This gives the missing half of (4.15). Therefore, Lemma 4.7 may be applied, and this finishes the proof. ■

4.5.2 The relation between contamination level and the small-ball assumption.

This section corresponds to Remark 4.5, where we note that a restriction of the form $\varepsilon \leq c\,\theta_0(\mathcal{F}, P)^{-2}$ is necessary in the setting of robust regression with quadratic loss (as in Theorem 4.1).

To prove this, we use a family of distributions and a contamination model discussed in §4.4.2. Given a dimension $d \in \mathbb{N}$ and a parameter $p \in (0,1)$, let $X' \sim \texttt{Normal}(0_{\mathbb{R}^d}, I_{d\times d})$, $\xi \sim \texttt{Normal}(0,1)$ be independent. Given $\beta \in \mathbb{R}^d$, we let P_β denote the distribution of the random pair (X,Y) given by

$$X = B_i\frac{X'}{\sqrt{p}} \text{ and } Y = \langle X, \beta\rangle + \xi.$$

Because X is isotropic, robust linear regression in this setting consists of estimating β in the

Euclidean norm from an ε-contaminated i.i.d. sample from P_β. Now, clearly,

$$\forall \beta \in \mathbb{R}^d \,:\, \theta_0(\mathcal{F}, P_\beta) = \sqrt{\frac{\pi}{2p}}.$$

The next Lemma roughly says that a contaminated sample from P_β with $\varepsilon > c\,\theta_0(\mathcal{F}, P)^{-2}$ essentially contains no information about β. More precisely, the Lemma implies via standard arguments that for any $R > 0$, one can find at least one $\beta \in \mathbb{R}^d$ for which the error of any estimator for β will be larger than R, with probability $\geq 1 - \alpha$.

Lemma 4.8. *For any $\beta \in \mathbb{R}^d$, and parameters $n \in \mathbb{N}$, $\varepsilon, p, \alpha \in (0,1)$ satisfying*

$$\varepsilon \geq 2p = \frac{\theta_0(\mathcal{F}, P_\beta)^2}{\pi} \text{ and } n \geq \left(\frac{2(1-p)+2\varepsilon}{p}\right) \ln \frac{1}{\alpha}$$

one can define ε-contaminated samples $Z_{1:n}^{\varepsilon,\beta}$ from P_β and an i.i.d. (uncontaminated) sample $Z_{1:n}^{0,0_{\mathbb{R}^d}}$ from $P_{0_{\mathbb{R}^d}}$ such that

$$\mathbb{P}\left[Z_{1:n}^{\varepsilon,\beta} = Z_{1:n}^{0,0_{\mathbb{R}^d}}\right] \geq 1 - \alpha.$$

Proof Let $Z_{1:n}^{\beta} \overset{i.i.d.}{\sim} P_\beta$, with each $Z_i^\beta = (X_i, Y_i)$. We may assume that $X_i = B_i\, X_i'/\sqrt{p}$ and $Y_i = \langle X_i, \beta\rangle + \xi_i$ with $(B_i, X_i', \xi) \sim (B, X', \xi)$ as above. As a result, a $\texttt{Ber}(p)$ proportion of X_i in the sample are non-zero. To define an ε-contaminated sample from P_β, we choose a subset $\mathcal{O} \subset \{i \in [n] \,:\, B_i \neq 0\}$ of size $\#\mathcal{O} \leq \varepsilon\, n$ that is as large as possible, and then set:

$$Z_i^{\varepsilon,\beta} := \begin{cases} (X_i, Y_i), & i \in [n]\backslash\mathcal{O} \\ (0_{\mathbb{R}^d}, \xi_i), & i \in \mathcal{O}. \end{cases}$$

Now, the random sample consisting of the points $Z_i^{0,0_{\mathbb{R}^d}} := (0_{\mathbb{R}^d}, \xi_i)$ $(i \in [n])$ is i.i.d. from P_0. Moreover, by our definition of the contamination,

$$\left\{Z_{1:n}^{\varepsilon,\beta} = Z_{1:n}^{\varepsilon,0_{\mathbb{R}^d}}\right\} \supset \{\mathcal{O} = \{i \in [n] \,:\, B_i \neq 0\}\} = \left\{\sum_{i=1}^n B_i \leq \varepsilon n\right\}.$$

The sum $\sum_{i=1}^n B_i$ is a binomial random variable with mean pn. Chernoff bounds give

$$\mathbb{P}\left[\sum_{i=1}^n B_i \leq \varepsilon n\right] \geq 1 - e^{-\frac{(\varepsilon-p)^2 n}{2\,p(1-p)+2\varepsilon}} \geq 1 - \alpha$$

where the our assumptions on ε and n were used in the last two inequalities. ■

Chapter 5

Gaussian and bootstrap approximations

5.1 Introduction

We consider the problems of Gaussian and bootstrap approximations for the trimmed mean when the sample size n is much smaller than the number of features (or dimension) d, even considering $d = \infty$. These problems are fundamental for many statistical tasks (e.g. finding confidence intervals, hypothesis testing, penalty selection) and data sets where $n \ll d$ have become common in many practical domains. Moreover, Gaussian and bootstrap approximations for the empirical average are very sensitive to contamination and heavy-tailed distributions.

Let $\mathcal{X}$ be a set, P be a probability over $\mathcal{X}$ and $\mathcal{F}$ be a class of square-integrable functions from $\mathcal{X}$ to $\mathbb{R}$ satisfying $Pf = 0$. Also let $\Sigma_{\mathcal{F},P}$ be the covariance matrix (or kernel) of $\mathcal{F}$, i.e.,

$$\Sigma_{\mathcal{F},P}(f,g) = P(fg)\ \forall f,g \in \mathcal{F}.$$

Given an i.i.d. sample $X_{1:n}$ from P, Classical weak convergence results, such as the Central Limit Theorem (for $|\mathcal{F}| < \infty$) or Donsker's theorem (for $|\mathcal{F}| = \infty$) state that, under certain assumptions,

$$\mathbb{G}_n(f) := \sqrt{n}\widehat{P}_n(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f(X_i),\ f \in \mathcal{F}$$

weakly converges, as $n \to \infty$, to a centered multivariate Gaussian (resp. Gaussian process) $\{G_P(f) : f \in \mathcal{F}\}$ with the same covariance $\Sigma_{\mathcal{F},P}$. Assuming $d = |\mathcal{F}| < \infty$, high-dimensional quantitative versions of the CLT typically focus on bounding

$$\varrho^E := \sup_{\lambda \in \mathbb{R}} \left|\mathbb{P}\left[Z_n(\mathcal{F}) \le \lambda\right] - \mathbb{P}\left[Z(\mathcal{F}) \le \lambda\right]\right|,$$

where

$$Z_n(\mathcal{F}) = \max_{f\in\mathcal{F}} \mathbb{G}_n(f) \text{ and } Z(\mathcal{F}) = \max_{f\in\mathcal{F}} G_P(f).$$

Under strong moment assumptions, the pioneering work of [Chernozhukov et al., 2013] first bounded $\varrho^E \leq C\left(\frac{\ln(nd)}{n}\right)^{\frac{1}{6}}$. This bound was previously improved by [Koike, 2021] and [Chernozhuokov et al., 2022] to $\varrho^E \leq C\left(\frac{\ln(nd)}{n}\right)^{\frac{1}{5}}$. Under additional assumptions, such as $\Sigma_{\mathcal{F},P}$ being positive definite, ϱ^E can be further bounded, as discussed in §5.2.4.

Bounds are also available for infinite function families $\mathcal{F}$, which correspond to $d = \infty$. Assuming the existence of an envelope function and some moment conditions, a line of work started by [Chernozhukov et al., 2014b, Chernozhukov et al., 2016] derived bounds depending on the metric entropy of the class $\mathcal{F}$ and on the weak-variance assumption

$$\underline{\sigma}_{\mathcal{F},P} := \inf_{f\in\mathcal{F}} \Sigma_{\mathcal{F},P}(f,f) > 0.$$

More recently, entropy-free and weak-variance-free bounds were obtained by [Giessing, 2023], although also requiring an envelope function with finite third moment.

Bootstrap approximation bounds are also presented in all previously discussed works. Two versions of the bootstrap are typically studied: the empirical and the multiplier versions. Both versions are defined given random variables $\tilde{X}_{1:n}$ and $\xi_{1:n}$ as follows:

- for the empirical bootstrap $\tilde{X}_{1:n}$ is an i.i.d. sample from the empirical measure $\widehat{P}_n$ of the sample and $\xi_i = 1$ for every $i \in [n]$;
- for the multiplier bootstrap $\tilde{X}_{1:n} = X_i$ and $\xi_{1:n}$ are i.i.d. standard Gaussian's.

Conditionally on a sample $X_{1:n}$ one can define

$$\tilde{\mathbb{G}}_n(f) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \left(f(\tilde{X}_i) - \widehat{P}_n(f) \right), \ f \in \mathcal{F}.$$

Bounds similar to the ones discussed for ϱ^E are also available for

$$\tilde{\varrho}^E := \sup_{\lambda\in\mathbb{R}} \left| \mathbb{P}\left[\tilde{Z}_n(\mathcal{F}) \leq \lambda \,\middle|\, X_{1:n} \right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda \right] \right|,$$

where $\tilde{Z}_n(\mathcal{F}) = \max_{f\in\mathcal{F}} \tilde{\mathbb{G}}_n(f)$. In [Chernozhukov et al., 2014a, Koike, 2021, Chernozhuokov et al., 2022] one can found such bounds for the case $d < \infty$ and in [Chernozhukov et al., 2014b, Chernozhukov et al., 2016, Giessing, 2023] for the case $d = \infty$.

5.1.1 Contributions.

In this chapter we study the Gaussian and bootstrap approximations of the trimmed mean. For the Gaussian approximation we replace the empirical process $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ by its trimmed-mean based counterpart

$$\mathbb{T}^{\varepsilon}_{n,k}(f) = \sqrt{n}\,\widehat{T}^{\varepsilon}_{n,k}(f, X^{\varepsilon}_{1:n}),\ (f \in \mathcal{F})$$

and for the bootstrap approximation we replace $\{\tilde{\mathbb{G}}_n(f) : f \in \mathcal{F}\}$ with

$$\tilde{\mathbb{T}}^{\varepsilon}_{n,k}(f) := \frac{\sqrt{n}}{n-2k} \sum_{i=k+1}^{n-k} \xi_{(i)} \left(f(\tilde{X}^{\varepsilon}_{(i)}) - \widehat{T}_{n,k}(f, X^{\varepsilon}_{1:n}) \right),\ (f \in \mathcal{F}),$$

where $(\cdot)$ is a permutation satisfying

$$\xi_{(1)} \left(f(X_{(1)}) - \widehat{T}_{n,k}(f, X^{\varepsilon}_{1:n}) \right) \leq \cdots \leq \xi_{(n)} \left(f(X_{(n)}) - \widehat{T}_{n,k}(f, X^{\varepsilon}_{1:n}) \right).$$

We are interested in bounding quantities such as

$$\varrho := \sup_{\lambda \in \mathbb{R}} \left| \mathbb{P}\left[Z^{\varepsilon}_{n,k}(\mathcal{F}) \leq \lambda \right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda \right] \right|,$$

where $Z^{\varepsilon}_{n,k}(\mathcal{F}) = \max_{f \in \mathcal{F}} \mathbb{T}_n(f)$, and its analogous for the bootstrap approximation. Similar bounds to the ones available in the literature for the empirical mean will be derived in the following sections for both problems, but under less restrictive moment assumptions and also considering adversarial contamination. Our main contributions are:

- First, this work is the first to explore both Gaussian and bootstrap approximations for high-dimensional trimmed means. It is also the first to explore the effects of sample contamination;
- In the case $d < \infty$ our results allow for exponential dependence of d on n assuming only that $\nu_p(\mathcal{F}) < \infty$ for some $p > 2$. In [Kock and Preinerstorfer, 2023], an example is constructed satisfying $\nu_p(\mathcal{F}) < \infty$ for some $p > 2$, but for which the empirical mean does not satisfy Gaussian approximation when $d \gg n^{\frac{p}{2}-1}$ (see §5.2.2 for a discussion). Thus, our results show that Gaussian approximations for the trimmed mean are still valid even when they are not valid for the empirical average.
- In the case $d = \infty$ our results provide better rates then the ones in [Chernozhukov et al., 2016]. We also provide results for $p > 2$ and do not require the existence of an envelope function $F \in L^4(P)$.

We also point out that our proof technique, which relies on relating trimming and truncation (see §2.2), can be easily coupled with new improved bounds for ϱ^E and $\tilde{\varrho}^E$ and might as well provide better bounds than the ones here obtained.

5.1.2 Notation. Besides the definitions previously discussed in Chapter 2, in this chapter we introduce a few new definitions. Let $\mathcal{F}$ and P be 2-compatible, define the weak variance as

$$\underline{\sigma}^2_{\mathcal{F},P} := \inf_{f\in\mathcal{F}} \Sigma_{\mathcal{F},P}(f,f). \tag{5.1}$$

Moreover, if $\mathcal{F}'$ and P' are also 2-compatible and a map $\pi : \mathcal{F} \to \mathcal{F}'$ is given, let

$$\Delta_\pi(\Sigma_{\mathcal{F},P}, \Sigma_{\mathcal{F}',P'}) := \sup_{f,g\in\mathcal{F}} \left|\Sigma_{\mathcal{F},P}(f,g) - \Sigma_{\mathcal{F}',P'}(\pi(f),\pi(g))\right|. \tag{5.2}$$

Notice that when $|\mathcal{F}| = |\mathcal{F}'| < \infty$ and π is a permutation, then $\Delta_\pi(\Sigma_{\mathcal{F},P}, \Sigma_{\mathcal{F}',P'})$ is simply the entry-wise $\|\cdot\|_\infty$ norm of the difference of the two covariance matrices. To see this, let $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$, $X \sim P$ and $X' \sim P'$, then $\Delta_\pi(\Sigma_{\mathcal{F},P}, \Sigma_{\mathcal{F}',P'})$ is the difference between the covariance matrix of $(f_j(X))_{j=1}^d$ and the covariance matrix of $(\pi(f_j)(X'))_{j=1}^d$. We sometimes omit the map π when there is a natural map, mainly when we are mapping a class $\mathcal{F}$ with a modification of that class, such as $\mathcal{F}^o_M = \{\tau_M \circ f - P\tau_M \circ f : f \in \mathcal{F}\}$ and so π can be naturally taken to be the map $f \mapsto \tau_M \circ f - P\tau_M \circ f$.

For convenience we also define

$$\delta_{n,d,B} = \left(\frac{B^2 \ln^5(nd)}{n}\right)^{\frac{1}{4}} \text{ and } \delta_{n,q,d,B} = \sqrt{\frac{B^2 \ln^{3-\frac{2}{p}}(nd)}{n^{1-\frac{2}{p}}}}.$$

5.2 High-dimensional results

In this section we let $\mathcal{F}$ be finite with size d and a 2-compatible distribution P. We also require $\mathcal{F}$ to be centered, i.e., $Pf = 0$ for every $f \in \mathcal{F}$.

5.2.1 Gaussian approximation. Our main result on the finite-dimensional context is the following Gaussian approximation result:

Theorem 5.1 (High-dimensional Gaussian approximation for trimmed means)**.** *Assume $n \geq 3$, $d \geq 2$ and that $\nu_p := \nu_P(\mathcal{F}, P) < \infty$ for some $p \in (2,\infty)$. If*

$$k := \lfloor \varepsilon n \rfloor + \left\lceil 3\ln(1+d) + 7n^{\frac{p-2}{4p-2}} \ln(nd) \right\rceil < \frac{n}{2},$$

and

$$\nu_p^2\, n^{-\frac{3p-6}{4p-2}} (\ln(nd))^{1-\frac{2}{p}} \leq \frac{3}{8} \underline{\sigma}_{\mathcal{F},P}. \tag{5.3}$$

Then

$$\varrho := \sup_{\lambda \in \mathbb{R}} \left| \mathbb{P}\left[Z^{\varepsilon}_{n,k}(\mathcal{F}) \leq \lambda \right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda \right] \right|$$

satisfies

$$\varrho \leq C \left(\nu_p \vee \nu_p^{\frac{1}{2}} \right) \left(\frac{\ln^{6-\frac{4}{p}}(nd)}{n^{\frac{2p-4}{2p-1}}} \right)^{\frac{1}{4}} + 15 \frac{\nu_p}{\underline{\sigma}_{\mathcal{F},P}} \varepsilon n^{\frac{1}{2}+\frac{3}{4p-2}} \ln^{\frac{1}{2}-\frac{1}{p}}(nd) + 2\exp\left\{ -n^{\frac{p-2}{4p-2}} \ln(nd) \right\}$$

for a constant C that depends only on $\nu_2(\mathcal{F}, P)$ and $\underline{\sigma}_{\mathcal{F},P}$.

This result contrasts with the following Gaussian approximation result for the empirical average:

Theorem 5.2 (Adapted from Theorem 2.4 and Lemma 4.3 of [Chernozhuokov et al., 2022])**.** *Suppose that $Pf^4 \leq B^2 \nu_2^2(\mathcal{F}, P)$ for every $f \in \mathcal{F}$ for some $B > 0$ and that $\underline{\sigma}_{\mathcal{F},P} > 0$. There exists a constant C depending only on $\nu_2(\mathcal{F}, P)$ and $\underline{\sigma}_{\mathcal{F},P}$ such that*

(i) if $f(X_1)$ is sub-exponential with Orlicz norm bounded by B for all $f \in \mathcal{F}$, then

$$\sup_{\lambda \in \mathbb{R}} \left| \mathbb{P}\left[Z_n(\mathcal{F}) \leq \lambda \right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda \right] \right| \leq C \delta_{n,d,B};$$

(ii) if $\mathbb{E}\left[\max_{f \in \mathcal{F}} |f(X_1)|^p \right] \leq B^p$ for some $p \in (2, \infty)$, then

$$\sup_{\lambda \in \mathbb{R}} \left| \mathbb{P}\left[Z_n(\mathcal{F}) \leq \lambda \right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda \right] \right| \leq C \left(\delta_{n,d,B} \vee \delta_{n,q,d,B} \right).$$

Remark 5.3 (Dependence on $\nu_2(\mathcal{F}, P)$ and $\underline{\sigma}_{\mathcal{F},P}$ in Theorem 5.2)**.** Although we stated the theorem saying that the constant C depends on $\nu_2(\mathcal{F}, P)$ and $\underline{\sigma}_{\mathcal{F},P}$, it was originally stated saying that C depends on constants $a, b > 0$ such that $\nu_2(\mathcal{F}, P) \geq a$ and $\underline{\sigma}_{\mathcal{F},P} \geq b$.

Remark 5.4 (Comparison with the empirical average)**.** As expected, if one has an uncontaminated sample and a light-tailed envelope, Theorem 5.2 provides a better Gaussian approximation for the empirical mean than the obtained in 5.1 for the trimmed mean: Theorem 5.2 allows for $d \ll \exp\left\{ n^{\frac{1}{5}} \right\}$ and Theorem 5.1 allows for $d \ll \exp\left\{ n^{\frac{1}{6}+O(p^{-1})} \right\}$. This minor difference is expected in this scenario as a consequence of the unnecessary discarding of sample points by the trimmed mean.

Now if one only has $\nu_p(\mathcal{F}, P) < \infty$ for some $p \in (2, \infty)$ the picture changes. Since

$$\mathbb{E}\left[\max_{f \in \mathcal{F}} |f(X_1)|^p \right] \leq \mathbb{E}\left[\sum_{f \in \mathcal{F}} |f(X_1)|^p \right] \leq \nu_p^p(\mathcal{F}, P) d$$

the value of B in case (ii) of Theorem 5.2 can be of order $B = \nu_p(\mathcal{F}, P)d^{\frac{1}{p}}$, which allows only for a polynomial dependence of $d \ll n^{\frac{p}{2}-1}$. Meanwhile, the Gaussian approximation for the trimmed mean still allows for $d \ll \exp\left\{n^{\frac{1}{6}+O(p^{-1})}\right\}$. Indeed, as discussed in §5.2.2, this polynomial dependence is not an artifact of Theorem 5.2, but a property of the empirical average itself.

Remark 5.5 (Choice of trimming level k)**.** We also observe that the choice of k depends only on the number d of features, the sample size and the contamination level ε. Thus, if an upper bound for ε is given, the trimming level k can be easily computed.

Remark 5.6 (Dependence of ε on n)**.** Theorem 5.1 suggests a relation between ε and n for obtaining Gaussian approximation:

$$\frac{\nu_p(\mathcal{F}, P)}{\underline{\sigma}_{\mathcal{F},P}}\varepsilon \ll n^{-\frac{1}{2}-\frac{3}{4p-2}} \ln^{-\frac{1}{2}+\frac{1}{p}}(nd).$$

It is easy to see that under the adversarial contamination setup if $\sqrt{n}\varepsilon \gg 1$, then it is impossible to obtain Gaussian approximation bounds for any given estimator. Let $X_{1:n}$ have an i.i.d. distribution with $X_i \sim \mathcal{N}(0, I_d)$, $W \sim \mathcal{N}(0, 1)$, and let $X_i' = \mathbf{1}_d B_i W + (1 - B_i)X_i$, with B_i a independent Bernoulli with rate $q = n^{-\frac{1}{2}}$ for all $i \in [n]$. The covariance of X_1 is the identity and of X_1' is $(1 - q)I_d + q\mathbf{1}_d\mathbf{1}_d^t$, which are the covariance matrices of a standard multivariate normal Y and of $Y' = \sqrt{1-q}\,Y + \sqrt{q}\mathbf{1}_d W$, respectively. Since the maximum of Y concentrates around $\sqrt{2\ln(2d)}$ and the maximum of $\sqrt{1-q}\,Y$ around $\sqrt{2(1-q)\ln(2d)}$ we can take $\lambda = \frac{\sqrt{2\ln(2d)}+\sqrt{2(1-q)\ln(2d)}}{2}$ in order to have, as $d \to \infty$,

$$\mathbb{P}\left[\max_{j\in[d]} Y_j \le \lambda\right] \to 0 \text{ and } \mathbb{P}\left[\max_{j\in[d]} Y_j' \le \lambda\right] \to 1.$$

Now notice that if $\sum_{i=1}^n B_i \le \varepsilon n$ it is possible to contaminate $X_{1:n}$ in order to obtain $X_{1:n}^{\varepsilon} = X_{1:n}'$. Taking n large enough and $\varepsilon n^{-\frac{1}{2}} \gg 1$

$$\mathbb{P}\left[\sum_{i=1}^n B_i \le \varepsilon n\right] \approx \mathbb{P}\left[W \le \sqrt{n}\frac{\varepsilon - q}{\sqrt{q(1-q)}}\right].$$

Showing that no Gaussian approximation bounds are possible for any given estimator when $\varepsilon n^{-\frac{1}{2}} \gg 1$.

5.2.2 A threshold phenomenon by Kock and Preinerstorfer.

As discussed in Remark 5.4, the polynomial dependence of d on n for the Gaussian approximation of the

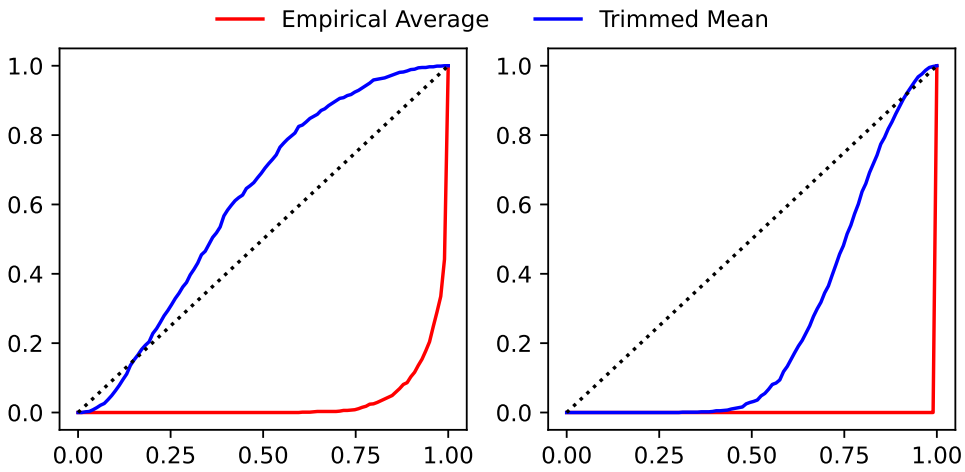


Figure 5.1: P-P plots comparing $Z_n(\mathcal{F})$ (red) and $Z^{\varepsilon}_{n,k}(\mathcal{F})$ (blue) vs. $Z(\mathcal{F})$ (dashed black) when $\varepsilon = 0$. Both figures assume $n = 100$, the left one has $d = 1000$ and the right one $d = 10000$. The trimming level k was $k = 15$ for $d = 1000$ and $k = 20$ for $d = 10000$. Samples were drawn assuming i.i.d. marginals with Student's distribution having 3 degrees of freedom, thus ensuring only moments $p \in [1, 3)$. Curves were obtained empirically from 1000 repetitions of the experiment.

empirical average is no coincidence. [Kock and Preinerstorfer, 2023] analysed this phenomenon by constructing an example. Let $p \in (2, \infty)$ and Ψ_p be a c.d.f function given by

$$\Psi_p(x) = \begin{cases} \frac{1}{2}\frac{1}{|x|^p(\ln|x|\vee 1)^2}, \text{ if } |x| \geq 1 \\ \frac{1}{2}, \text{ if } |x| < 1 \end{cases}.$$

As shown by [Kock and Preinerstorfer, 2023], this distribution is centered, has p-th moment but no higher moments. Moreover, the following holds:

Theorem 5.7 (Threshold phenomenon for the Gaussian approximation for the empirical average; adapted from Theorems 2.1 and 2.2 of [Kock and Preinerstorfer, 2023])**.** *Let P be a distribution over $\mathbb{R}^d$, take $\mathcal{F}$ to be the family of coordinate projections, and let $p \in (2, \infty)$.*

(i) If P has i.i.d. marginals with distribution Ψ_p and $\delta > 0$ satisfies $\limsup_{n\to\infty} dn^{1-\frac{p}{2}-\delta} > 0$, then

$$\limsup_{n\to\infty} \sup_{\lambda\in\mathbb{R}} |\mathbb{P}[Z_n(\mathcal{F}, P) \leq \lambda] - \mathbb{P}[Z(\mathcal{F}, P) \leq \lambda]| = 1.$$

(ii) Moreover, given $0 < c \leq C^{\frac{2}{p}} < \infty$, let $\mathrm{P}(c, C, p)$ be the class of all centered distributions

P over $\mathbb{R}^d$ with $\underline{\sigma}_{\mathcal{F},P} \geq c$ and $\nu_p^p(\mathcal{F},P) \leq C$. If $\delta > 0$ satisfies $\lim_{n\to\infty} dn^{1-\frac{p}{2}+\delta} = 0$, then

$$\lim_{n\to\infty} \sup_{\lambda\in\mathbb{R}} \sup_{P\in \mathrm{P}(c,C,p)} \left|\mathbb{P}\left[Z_n(\mathcal{F},P) \leq \lambda\right] - \mathbb{P}\left[Z(\mathcal{F},P) \leq \lambda\right]\right| = 0$$

Thus, the Gaussian approximation for the empirical average has a threshold at $d = n^{\frac{p}{2}-1}$, being feasible for a large class of distributions when $d \ll n^{\frac{p}{2}-1-\delta}$, but not when $d \gg n^{\frac{p}{2}-1+\delta}$. Figure 5.1 is a P-P plot of $Z_n(\mathcal{F})$ and $Z_{n,k}^{\varepsilon}(\mathcal{F})$ vs. $Z(\mathcal{F})$. It was constructed taking i.i.d. marginals with Student's distribution having 3 degrees of freedom. It illustrates a scenario with weak moment guarantees were the Gaussian approximation for the empirical average performs poorly when compared with the trimmed mean. This empirical result adds to the previous discussion on threshold phenomenon and motivates the usage and study of trimmed means in the context of Gaussian approximation.

5.2.3 Bootstrap approximations.

Our bootstrap approximation bounds are quite analogous to the Gaussian approximation bounds.

Theorem 5.8 (High-dimensional bootstrap approximations for trimmed means)**.** *Assume $n \geq 3$, $d \geq 2$ and that $\nu_p := \nu_P(\mathcal{F},P) < \infty$ for some $p \in (2,\infty)$. Let $\tilde{Z}_{n,k}^{\varepsilon}(\mathcal{F})$ be obtained via the empirical or the multiplier bootstrap. Suppose that*

$$\nu_p^2\, n^{-\frac{3p-6}{4p-2}} (\ln(nd))^{1-\frac{2}{p}} \leq \frac{3}{8}\underline{\sigma}_{\mathcal{F},P} \tag{5.4}$$

and let

$$\tilde{\varrho} := \sup_{\lambda\in\mathbb{R}} \left|\mathbb{P}\left[\tilde{Z}_{n,k}^{\varepsilon}(\mathcal{F}) \leq \lambda \,\middle|\, X_{1:n}^{\varepsilon}\right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda\right]\right|.$$

The following holds:

(i) If $\tilde{Z}_{n,k}^{\varepsilon}(\mathcal{F})$ is obtained via the empirical bootstrap, then with probability at least $1 - C\nu_p^{\frac{1}{2}} n^{-\frac{p-2}{4p-2}} \ln^{\frac{5p-2}{4p}}(nd) - \exp\left\{-n^{\frac{p-2}{4p-2}}\ln(nd)\right\}$,

$$\tilde{\varrho} \leq C\left(\nu_p \vee \nu_p^{\frac{1}{2}}\right)\left(\frac{\ln^{6-\frac{4}{p}}(nd)}{n^{\frac{2p-4}{2p-1}}}\right)^{\frac{1}{4}} + 30\frac{\nu_p}{\underline{\sigma}_{\mathcal{F},P}}\varepsilon n^{\frac{1}{2}+\frac{3}{4p-2}} \ln^{\frac{1}{2}-\frac{1}{p}}(nd) + 3\exp\left\{-6n^{\frac{p-2}{4p-2}}\ln(nd)\right\}$$

for a constant C that depends only on $\nu_2(\mathcal{F},P)$ and $\underline{\sigma}_{\mathcal{F},P}$.

(ii) If $\tilde{Z}_{n,k}^{\varepsilon}(\mathcal{F})$ is obtained via the Gaussian multiplier bootstrap, then with probability at least $1 - \exp\left\{-n^{\frac{p-2}{4p-2}}\ln(nd)\right\}$,

$$\tilde{\varrho} \leq C\left(\nu_p \vee \nu_p^{\frac{1}{2}}\right)\left(\frac{\ln^{8-\frac{4}{p}}(nd)}{n^{\frac{2p-4}{2p-1}}}\right)^{\frac{1}{4}} + 110\frac{\nu_p}{\underline{\sigma}_{\mathcal{F},P}}\varepsilon n^{\frac{1}{2}+\frac{3}{4p-2}} \ln^{1-\frac{1}{p}}(nd) + \frac{2}{n}$$

for a constant C that depends only on $\nu_2(\mathcal{F}, P)$ and $\underline{\sigma}_{\mathcal{F},P}$.

One may compare our bounds with the ones by [Chernozhuokov et al., 2022], which have more restrictive assumptions:

Theorem 5.9 (Adapted from Lemma 4.5 and Lemma 4.6 of [Chernozhuokov et al., 2022])**.** *Make the same assumptions as in Theorem 5.2. Let $\tilde{Z}_n$ be obtained from the Gaussian multiplier bootstrap or the empirical bootstrap. There exists a constant C depending only on $\nu_2(\mathcal{F})$ and $\underline{\sigma}_{\mathcal{F},P}$ such that*

(i) if $f(X_1)$ is sub-exponential with Orlicz norm bounded by B for all $f \in \mathcal{F}$, then with probability at least $1 - C\delta_{n,d,B}$,

$$\sup_{\lambda \in \mathbb{R}} \left| \mathbb{P}\left[\tilde{Z}_n(\mathcal{F}) \leq \lambda \,\middle|\, X_{1:n} \right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda \right] \right| \leq C\delta_{n,d,B};$$

(ii) if $\mathbb{E}\left[\max_{f \in \mathcal{F}} |f(X_1)|^p\right] \leq B^p$ for some $p \in (2, \infty)$, then with probability at least $1 - C\left(\delta_{n,d,B} \vee \delta_{n,q,d,B}\right)$,

$$\sup_{\lambda \in \mathbb{R}} \left| \mathbb{P}\left[\tilde{Z}_n(\mathcal{F}) \leq \lambda \,\middle|\, X_{1:n} \right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda \right] \right| \leq C\left(\delta_{n,d,B} \vee \delta_{n,q,d,B}\right).$$

Remark 5.10 (Comparison with the usual empirical and multiplier bootstrap)**.** Assume $\varepsilon = 0$. In order to obtain convergence to zero as $n \to \infty$, Theorem 5.8 allows for $d \ll \exp\left\{n^{\frac{1}{6}+O(p^{-1})}\right\}$ for the empirical bootstrap and $d \ll \exp\left\{n^{\frac{1}{8}+O(p^{-1})}\right\}$ for the Gaussian multiplier bootstrap. Meanwhile, Theorem 5.9 allows for $d \ll \exp\left\{n^{\frac{1}{5}}\right\}$, but under more restrictive assumptions.

5.2.4 Further background.

The only assumption made in Theorem 5.1 (or Theorem 5.2) about $\Sigma_{\mathcal{F},P}$ is the positiveness of the weak covariance. In special, we do not require $\Sigma_{\mathcal{F},P}$ to be positive definite. Both [Kuchibhotla and Rinaldo, 2020] and [Chernozhukov et al., 2023b] have shown that better bounds are possible for the Gaussian approximation of the empirical average when $\Sigma_{\mathcal{F},P}$ is positive definite. In fact, [Chernozhukov et al., 2023b] obtained a bound of order $(\ln n)\sqrt{\frac{\ln^3 d}{n}}$ assuming that the coordinates of X_i are all bounded. They also proved that this order is optimal up to the $\ln n$ factor. Similar results are also available in the literature considering symmetric distributions [Chernozhuokov et al., 2022] or even variance decay [Lopes et al., 2020]. Although this is out of the scope of this thesis, all such results can be adapted for the trimmed mean using our proof techniques (see §5.4).

5.3 Gaussian approximation for empirical processes.

Before discussing our results for empirical processes we need a few definitions. Given a measure Q and a class $\mathcal{F}$ of functions 2-compatible with Q we let $\mathcal{N}(\mathcal{F}, d_Q, \delta)$ be the δ-covering number of $\mathcal{F}$ with respect to the semi-metric $d_Q(f,g) = Q(f-g)^2$. We also say that F is an envelope of $\mathcal{F}$ if $|f(x)| < F(x)$ for all $x \in \mathbf{X}, f \in \mathcal{F}$. In this section we assume the existence of a centered Gaussian process $\{G_P f : f \in \mathcal{F}\}$ with covariance $\Sigma_{\mathcal{F},P}$.

Definition 5.11 (VC-subgraph class)**.** We say that a class $\mathcal{F}$ of functions $f : \mathbf{X} \to \mathbb{R}$ is a VC subgraph class with dimension $v = \mathrm{vc}(\mathcal{F})$ if $v < \infty$ is the VC dimension of its subgraphs, i.e.,

$$v := \mathrm{vc}\left(\{\{(x,t) \in \mathbf{X} \times \mathbb{R} : t < f(x)\} : f \in \mathcal{F}\}\right) < \infty.$$

Definition 5.12 (VC-type class)**.** We say that a class $\mathcal{F}$ of functions $f : \mathbf{X} \to \mathbb{R}$ is a VC-type class with envelope F if there are constants $A, v > 0$ such that

$$\sup_Q \mathcal{N}\left(\mathcal{F}, d_Q, \delta \|F\|_{L^2(Q)}\right) \leq \left(\frac{A}{\delta}\right)^v \quad \forall \delta \in (0,1].$$

Where the supremum is taken over all probability measures over $(\mathbf{X}, \mathcal{X})$ with finite support.

The following lemma relates both definitions:

Lemma 5.13. *Assume $\mathcal{F}$ is a VC-subgraph class and let $q \geq 1$. Then $\mathcal{F}_M^q = \{(\tau_M \circ f)^q : f \in \mathcal{F}\}$ is VC-type with envelope M^q and constants $(A, v) = (8e, 2vc(\mathcal{F}))$.*

Proof First notice that $\mathrm{vc}(\mathcal{F}_M^q) \leq \mathrm{vc}(\mathcal{F})$. Then use Theorem 5.11 of [Zhang, 2023]. ■

Our result on the Gaussian approximation for empirical processes can now be stated:

Theorem 5.14. *Assume $\mathcal{F}$ is a VC-subgraph class and let $p \in (2, \infty)$ be such that $\nu_p(\mathcal{F}) < \infty$. Define*

$$K_n := K_n(\mathcal{F}) = 2vc(\mathcal{F})\left(\ln n \vee \ln 8e\right) \text{ and } \Xi(\delta) = \mathbb{E}\left[\sup_{\substack{f,g \in \mathcal{F} \\ d_P(f,g) < \delta}} G_P(f-g)\right].$$

Suppose that

$$16K_n \leq n^{\frac{p-2}{4p-2}} \tag{5.5}$$

and

$$\nu_p^2(\mathcal{F})\, n^{-\frac{3p-6}{4p-2}} K_n^{1-\frac{2}{p}} \leq \frac{3}{8}\underline{\sigma}_{\mathcal{F},P}. \tag{5.6}$$

Then, there is an absolute constant C such that taking

$$k := \lfloor \varepsilon n \rfloor + \left\lceil C n^{\frac{p-2}{4p-2}} K_n \right\rceil < \frac{n}{2},$$

implies that

$$\varrho := \sup_{\lambda \in \mathbb{R}} \left| \mathbb{P}\left[Z_{n,k}^{\varepsilon}(\mathcal{F}) \leq \lambda \right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda \right] \right|$$

satisfies

$$\varrho \leq C' \left(\nu_p \vee \nu_p^{\frac{1}{2}} \right) \left(\frac{K_n^{6-\frac{4}{p}}}{n^{\frac{2p-4}{2p-1}}} \right)^{\frac{1}{4}} + 18 \frac{\nu_p}{\underline{\sigma}_{\mathcal{F},P}} \varepsilon n^{\frac{1}{2}+\frac{3}{4p-2}} K_n^{\frac{1}{2}-\frac{1}{p}} + \frac{3\sqrt{K_n}}{\underline{\sigma}_{\mathcal{F},P}} \Xi\left(5\nu_p K_n^{\frac{1}{2}-\frac{1}{p}} n^{\frac{9}{16p-8}-\frac{3}{8}} \right)$$

for a constant C' depending only on $\nu_2(\mathcal{F})$ and on $\underline{\sigma}_{\mathcal{F},P}$.

Remark 5.15. The quantity $\Xi(\delta)$ is typically of second order in this bound. First notice that if G_P has uniformly d_P-continuous sample paths then $\Xi(\delta) \to 0$ as $\delta \to 0$. Moreover, non-asymptotic bounds can be derived using entropy integral bounds. For instance, if $\mathcal{F}$ is a VC-type class (as in Definition 5.12) one can easily see that

$$\Xi(\delta) \leq C \int_0^{\delta} \sqrt{v \ln \frac{A\|F\|_{L^2(P)}}{s}}\, ds$$

for some absolute constant $C > 0$, thus $\Xi(\delta)$ is linear on δ.

Theorem 5.14 can be compared with the following one by [Chernozhukov et al., 2016]:

Theorem 5.16 (Theorem 2.1 of [Chernozhukov et al., 2016])**.** *Suppose that $\mathcal{F}$ satisfies*

- *$\mathcal{F}$ and P are 2-compatible;*
- *$\mathcal{F}$ is VC-type with envelope F and constants $A \geq 2$, $v \geq 1$;*
- *There exist constants $b \geq \sigma > 0$ and $p \in [4, \infty)$ such that $\nu_k^k \leq \sigma^2 b^{k-2}$ for $k = 2, 3, 4$ and $\|F\|_{L^p(P)} \leq b$.*

and let

$$K_n' = v \left(\ln n \vee \ln \frac{Ab}{\sigma} \right) \text{ be such that } K_n' \leq n^{\frac{1}{3}}.$$

Then, for every $\gamma \in (0,1)$,

$$\varrho \leq \Psi\left(C_1\left\{\frac{bK_n'}{\gamma^{\frac{1}{p}} n^{\frac{1}{2}-\frac{1}{p}}} + \frac{(b\sigma^2(K_n')^2)^{\frac{1}{3}}}{\gamma^{\frac{1}{3}} n^{\frac{1}{6}}}\right\}\right) + C_2\left(\gamma + \frac{1}{n}\right)$$

where C_1, C_2 *are positive constants that depend only on* p *and*

$$\Psi(\eta) = \inf_{\delta, r>0}\left\{\frac{2}{\underline{\sigma}_{\mathcal{F},P}}(\eta + \Xi(\delta) + r\delta)\left(\sqrt{2v\ln\frac{Ab}{\delta}} + 2\right) + e^{-\frac{r^2}{2}}\right\}.$$

Remark 5.17. Theorem 5.16 is originally stated in terms of a Gaussian coupling, but the authors also provide an anti-concentration lemma to obtain a representation in Kolmogorov distance. Thus, to obtain Theorem 5.16 we combined its original form with the anti-concentration lemma the authors provided (Lemma 2.2 of [Chernozhukov et al., 2016]).

Remark 5.18 (Comparison between Theorem 5.14 and Theorem 5.16)**.** We now compare Theorem 5.14 and Theorem 5.16. Taking $\delta = \frac{\sigma}{bn^\alpha}$ for some $\alpha > 0$ yields a bound (in Theorem 5.16) of order at least

$$\frac{b(K_n')^{\frac{3}{2}}}{\gamma^{\frac{1}{p}} n^{\frac{1}{2}-\frac{1}{p}}} + \frac{(b\sigma^2)^{\frac{1}{3}}(K_n')^{\frac{7}{6}}}{\gamma^{\frac{1}{3}} n^{\frac{1}{6}}} + \gamma$$

and optimizing in γ yields $\left(\frac{(K_n')^{\frac{12p}{p+1}}}{n}\right)^{\frac{1}{8}}$. Thus, the rate of convergence of Theorem 5.14 is better. Moreover, Theorem 5.14 does not require the existence of an envelope and holds for $p \in (2,\infty)$. Meanwhile, Theorem 5.16 requires an envelope in $L^q(P)$ for $p \in [4,\infty)$.

Remark 5.19. One might also want to compare Theorem 5.14 with the more recent results by [Giessing, 2023] for sample means. Although [Giessing, 2023] present bounds independent of the metric entropy and does not require $\mathcal{F}$ to be a VC-subgraph class, they require the existence of an envelope in $L^3(P)$ and attain a worse dependence in n.

5.4 Proof ideas

Our proofs are based on the trimming and truncation strategy discussed in §2.2, where the goal is to adjust the trimming level k (and the corresponding truncation level M) in order to control the approximation of $\widehat{T}^{\varepsilon}_{n,k}(f)$ by $\widehat{P}_n(\tau_M \circ f)$ over the class $\mathcal{F}$. In this section we illustrate the main proof ideas discussing the case $d := |\mathcal{F}| < \infty$.

5.4.1 Gaussian approximation. We begin discussing the proof of Theorem 5.1. The proof start using the trimming and truncation relation discussed in §2.2.

First step: counting and bounding.

Recall that

$$V_M(\mathcal{F}) = \max_{f\in\mathcal{F}} \sum_{i=1}^{n} \mathbf{1}_{\{|f(X_i)|>M\}}.$$

We start using the following finite-dimensional version of Lemma 2.2:

Lemma 5.20 (Finite-dimensional counting lemma)**.** *Let $d := |\mathcal{F}| < \infty$, given $M > 0$ we have*

$$\mathbb{P}\left[V_M(\mathcal{F}) \geq 3\ln(1+d) + 7n\frac{\nu_p^p}{M^p}\right] \leq 2\exp\left\{-n\frac{\nu_p^p}{M^p}\right\}.$$

Proof It follows directly from Lemma A.3 noticing that

$$\max_{j\in[d]} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{1}_{\{|f(X_i)|>M\}}\right] = \max_{j\in[d]} \mathbb{P}\left[|f(X_1)| > M\right] \leq \frac{\nu_p^p}{M^p},$$

where the last inequality holds by Markov's inequality. ■

Thus, taking $t \geq 3\ln(1+d) + 7n\frac{\nu_p^p}{M^p}$ one can ensure that, with high probability, $V_M(\mathcal{F}) \leq t$. Set

$$k := \phi n \text{ where } \phi = \frac{\lfloor \varepsilon n \rfloor + t}{n},$$

we proceed using the Bounding Lemma (Lemma 2.3) to obtain, for all $\lambda \in \mathbb{R}$,

$$\mathbb{P}\left[Z_{n,k}^{\varepsilon}(\mathcal{F}) \leq \lambda\right] \leq \mathbb{P}\left[Z_n(\tau_M \circ \mathcal{F}) \leq \lambda + 6\phi M\sqrt{n}\right] + 2\exp\left\{-n\frac{\nu_p^p}{M^p}\right\}.$$

Second step: use the Gaussian approximation result for the empirical average.

We are almost ready to approximate $Z_n(\tau_M \circ \mathcal{F})$ by $Z(\tau_M \circ \mathcal{F})$ via Theorem 5.2, but first we need to center the class $\tau_M \circ \mathcal{F}$. Let $\mathcal{F}_M^o := \{\tau_M \circ f - P(\tau_M \circ f) : f \in \mathcal{F}\}$, thus

$$\mathbb{P}\left[Z_{n,k}^{\varepsilon}(\mathcal{F}) \leq \lambda\right] \leq \mathbb{P}\left[Z_n(\mathcal{F}_M^o) \leq \lambda + \left(6\phi M + \sup_{f\in\mathcal{F}} P(\tau_M \circ f)\right)\sqrt{n}\right] + 2\exp\left\{-n\frac{\nu_p^p}{M^p}\right\}.$$

Now, Theorem 5.2 yields, for all $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}\left[Z_{n,k}^{\varepsilon}(\mathcal{F}) \leq \lambda\right] \leq & \mathbb{P}\left[Z(\mathcal{F}_M^o) \leq \lambda + \left(6\phi M + \sup_{f\in\mathcal{F}} P(\tau_M \circ f)\right)\sqrt{n}\right] \\ & + C\left(\frac{M^2\ln^5(nd)}{n}\right)^{\frac{1}{4}} + 2\exp\left\{-n\frac{\nu_p^p}{M^p}\right\}\end{aligned}$$

where the constant C depends on $\nu_2(\mathcal{F})$ and on $\underline{\sigma}_{\mathcal{F}_M^o,P}$. As will become clear in the end of the proof, (5.3) implies $\underline{\sigma}_{\mathcal{F}_M^o,P} \geq \frac{1}{2}\underline{\sigma}_{\mathcal{F},P}$ we can say that C depends only on $\nu_2(\mathcal{F})$ and $\underline{\sigma}_{\mathcal{F},P}$.

Third step: use a Gaussian to Gaussian comparison inequality.

The next step is to use the Gaussian to Gaussian comparison inequality (Lemma A.4), to obtain

$$\sup_{\lambda\in\mathbb{R}} |\mathbb{P}\left[Z(\mathcal{F},P)\leq\lambda\right] - \mathbb{P}\left[Z(\mathcal{F}_M^o,P)\leq\lambda\right]| \leq C'(\ln d)\sqrt{\Delta_\pi(\Sigma_{\mathcal{F},P} - \Sigma_{\mathcal{F}_M^o,P})},$$

where C' depends only on $\underline{\sigma}_{\mathcal{F},P}$ and $\pi(f) = \tau_M\circ f - P(\tau_M\circ f)\in\mathcal{F}_o^M$ for all $f\in\mathcal{F}$. The RHS of the inequality above is bounded using the following Lemma:

Lemma 5.21 (Covariance bounds)**.** *Let $\mathcal{F}$ and P be 2-compatible and $M>0$,*

$$\Delta_\pi(\Sigma_{\mathcal{F},P},\Sigma_{\mathcal{F}_M^o,P}) \leq 4\nu_p^p M^{2-p}.$$

This bound also holds when $|\mathcal{F}|=\infty$.

Proof Given $f,g\in\mathcal{F}$ we need to bound,

$$P\left(fg - \pi(f)\pi(g)\right) = P\left(fg - (\tau_M\circ f)(\tau_M\circ g)\right) + P(\tau_M\circ f)P(\tau_M\circ g).$$

We start using Hölder to get $P(\tau_M\circ f) = P\left(\tau_M\circ f - f\right) \leq P|f|\mathbf{1}_{|f|>M} \leq \nu_p^p M^{1-p}$. Using Holder two more times yields:

$$\begin{aligned}
P\left|fg - (\tau_M\circ f)(\tau_M\circ g)\right| &\leq MP\left|f-\tau_M\circ f\right|\mathbf{1}_{|f|>M} + MP\left|g-\tau_M\circ g\right|\mathbf{1}_{|g|>M} \\
&\quad + P\left(|f|\mathbf{1}_{|f|>M}|g|\mathbf{1}_{|g|>M}\right) \\
&\leq 2\nu_p^p M^{2-p} + P\left(|f|\mathbf{1}_{|f|>M}|g|\mathbf{1}_{|g|>M}\right) \\
&\leq 2\nu_p^p M^{2-p} + \sup_{j\in[d]} P|f|^2\mathbf{1}_{|f|>M} \\
&\leq 2\nu_p^p M^{2-p} + \left(\nu_p^p\right)^{\frac{2}{p}}\left(\frac{\nu_p^p}{M^p}\right)^{1-\frac{2}{p}} \leq 3\nu_p^p M^{2-p}.
\end{aligned}$$

■

We can now bound

$$\begin{aligned}
&\sup_{\lambda\in\mathbb{R}}\left|\mathbb{P}\left[Z_{n,k}^\varepsilon(\mathcal{F})\leq\lambda\right] - \mathbb{P}\left[Z(\mathcal{F})\leq\lambda + \left(6\phi M + \sup_{f\in\mathcal{F}} P(\tau_M\circ f)\right)\sqrt{n}\right]\right| \\
&\leq C'(\ln d)\sqrt{4\nu_p^p M^{2-p}} + C\left(\frac{M^2\ln^5(nd)}{n}\right)^{\frac{1}{4}} + 2\exp\left\{-n\frac{\nu_p^p}{M^p}\right\}
\end{aligned}$$

Fourth step: use a Gaussian anti-concentration inequality.

We are almost done. To get rid of the term $\left(6\phi M + \sup_{f\in\mathcal{F}} P(\tau_M \circ f)\right)\sqrt{n}$ we bound

$$\sup_{f\in\mathcal{F}} P(\tau_M \circ f) \leq \nu_p^p M^{1-p} \tag{5.7}$$

and make use of a Gaussian anti-concentration inequality, in this case we use Lemma A.5,

$$\begin{aligned}\sup_{\lambda\in\mathbb{R}} \left|\mathbb{P}\left[Z_{n,k}^{\varepsilon}(\mathcal{F}) \leq \lambda\right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda\right]\right| \leq & \left(6\phi M + \nu_p^p M^{1-p}\right) \frac{2+\sqrt{2\ln d}}{\underline{\sigma}_{\mathcal{F},P}}\sqrt{n} \\ & + C'(\ln d)\sqrt{4\nu_p^p M^{2-p}} \\ & + C\left(\frac{M^2 \ln^5(nd)}{n}\right)^{\frac{1}{4}} + 2\exp\left\{-n\frac{\nu_p^p}{M^p}\right\}\end{aligned} \tag{5.8}$$

Final step: optimize M.

Once we arrive at a bound such as Equation (5.8), we finish by optimizing M, as done in §5.5.1.

5.4.2 Bootstrap approximations.

Here we discuss the proof of Theorem 5.8.

Part (i). We start with the empirical bootstrap approximation, in this case we have

$$\tilde{\mathbb{T}}_{n,k}^{\varepsilon}(f) := \frac{\sqrt{n}}{n-2k}\sum_{i=k+1}^{n-k} f(\tilde{X}_{(i)}^{\varepsilon}) - \widehat{T}_{n,k}(f, X_{1:n}^{\varepsilon}),\ (f\in\mathcal{F}),$$

where the points $\tilde{X}_i^{\varepsilon}$ are independently and uniformly drawn from $\{X_i^{\varepsilon} : i \in [n]\}$.

The strategy to prove the empirical bootstrap approximation is to approximate $\tilde{Z}_{n,k}^{\varepsilon}(\mathcal{F})$ by $\tilde{Z}_n(\tau_M \circ \mathcal{F})$ and then use Lemma 5.9. This approximation requires a conditional version of the counting lemma.

First step: conditional counting.

Recall that given $X_{1:n}$, we sample $\tilde{X}_{1:n}$ from the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, define

$$\tilde{V}_M(\mathcal{F}) = \max_{f\in\mathcal{F}} \sum_{i=1}^n \mathbf{1}_{\{|f(\tilde{X}_i)|>M\}}$$

to be the analogous of $V_M(\mathcal{F})$ for the empirical bootstrap.

Lemma 5.22 (Conditional finite-dimensional counting lemma)**.** *Assume $d = |\mathcal{F}| < \infty$. Given $M > 0$ and $t \geq 0$ we have*

$$\mathbb{P}\left[\tilde{V}_M(\mathcal{F}) \geq 3\ln(1+d) + 7t \,\middle|\, V_M(\mathcal{F}) \leq t\right] \leq 2e^{-t}.$$

Proof Follows using Lemma A.3 conditionally on $X_{1:n}$. ■

By Lemma 5.20 we know that the event

$$E = \left\{ V_M(\mathcal{F}) < 3\ln(1+d) + 7n\frac{\nu_p^p}{M^p} \right\} \tag{5.9}$$

satisfies $\mathbb{P}[E] \geq 1 - e^{-n\frac{\nu_p^p}{M^p}}$. From now on we will be conditioning on E.

Set $t = \left\lceil 24\ln(1+d) + 49n\frac{\nu_p^p}{M^p} \right\rceil$ use Lemma 5.22 to get

$$\mathbb{P}\left[\tilde{V}_M(\mathcal{F}) < t \mid E \right] \geq 1 - 2\exp\left\{ -\left(3\ln(1+d) + 7n\frac{\nu_p^p}{M^p} \right) \right\} \geq 1 - 2e^{-\frac{t}{8}}.$$

Second step: control the contamination level on the bootstrap sample.

Let $\tilde{\varepsilon}$ be the ratio of copies of contaminated samples on $\tilde{X}^{\varepsilon}_{1:n}$, i.e.,

$$\tilde{\varepsilon} = \frac{|\{i : \tilde{X}_i^{\varepsilon} \neq \tilde{X}_i\}|}{n}.$$

Notice that $\tilde{\varepsilon}$ is independent of $X_{1:n}$. By Chernoff's bound for the binomial distribution:

$$\mathbb{P}[\tilde{\varepsilon}n - \varepsilon n \geq \varepsilon n \vee t] \leq \exp\left\{ -\frac{\varepsilon n}{3}\left(1 \vee \frac{t}{\varepsilon n} \right)^2 \right\} \leq e^{-\frac{t}{3}} \leq e^{-\frac{t}{8}}$$

Third step: bounding as usual.

Now we can use Lemma 2.3 to approximate $\tilde{Z}^{\varepsilon}_{n,k}(\mathcal{F})$ by $Z_n(\tau_M \circ \mathcal{F})$ with high probability conditioned on E, this is done using the lemma twice:

$$\begin{aligned}\frac{1}{\sqrt{n}}\left| \tilde{\mathbb{T}}^{\varepsilon}_{n,k}(f) - \tilde{\mathbb{G}}_n(f) \right| &\leq \left| \widehat{T}_{n,k}(f, \tilde{X}^{\varepsilon}_{1:n}) - \widehat{P}_n(\tau_M \circ f, \tilde{X}_{1:n}) \right| + \left| \widehat{T}_{n,k}(f, X^{\varepsilon}_{1:n}) - \widehat{P}_n(\tau_M \circ f, X_{1:n}) \right| \\ &\leq 12\phi M.\end{aligned}$$

Up to now we have shown that, with probability at least $1 - e^{-n\frac{\nu_p^p}{M^p}}$ taking

$$k = \phi n \text{ where } \phi = \frac{\lfloor \varepsilon n + (\varepsilon n \vee t) \rfloor + t}{n}$$

yields

$$\mathbb{P}\left[\left| \tilde{Z}^{\varepsilon}_{n,k}(\mathcal{F}) - \tilde{Z}_n(\tau_M \circ \mathcal{F}) \right| \leq 6\phi M\sqrt{n} \,\middle|\, E \right] \geq 1 - 3e^{-\frac{t}{8}}.$$

Next steps: from now on we follow as on the proof of Gaussian approximation.

Centering the functions $\tau_M \circ \mathcal{F}$ and using (5.7) gives:

$$\mathbb{P}\left[\left|\tilde{Z}^{\varepsilon}_{n,k}(\mathcal{F}) - \tilde{Z}_n(\mathcal{F}^o_M)\right| \leq \left(6\phi M + \nu_p^p M^{1-p}\right)\sqrt{n}\,\middle|\, E\right] \geq 1 - 3e^{-\frac{t}{8}}. \tag{5.10}$$

The proof now follows in a similar fashion as the proof of the Gaussian approximation: we use Lemma 5.9 to obtain a approximation of $\tilde{Z}_n(\mathcal{F}^o_M)$ by $Z(\mathcal{F}^o_M)$ and then we make use of Gaussian to Gaussian comparison and Gaussian anti-concentration inequalities.

Part (ii). Finally, we discuss the proof of the Gaussian multiplier bootstrap. This proof is simpler than the one we have just discussed for the empirical bootstrap. Its main ingredient is the following version of the bounding lemma capable of dealing with Gaussian weights.

Lemma 5.23 (Gaussian bounding lemma)**.** *Let t and $M \geq 0$ be such that $V_M(\mathcal{F}) \leq t$. Assume that $\tilde{X}^{\varepsilon}_{1:n}$ come from a Gaussian bootstrap.*

$$\frac{\lfloor \varepsilon n \rfloor + t}{n} \leq \phi < \frac{1}{2}.$$

Then, with probability at least $1 - \frac{2}{n}$,

$$\sup_{f\in\mathcal{F}} \left|\tilde{\mathbb{T}}^{\varepsilon}_{n,k}(f) - \tilde{\mathbb{G}}_n(\tau_M \circ f)\right| \leq 28\phi M\sqrt{2n\ln 2n} + 6\phi M\sqrt{2\ln n}. \tag{5.11}$$

Proof The proof is similar to the proof of the original bounding lemma (Lemma 2.3), but we now have to bound some random elements. Let $k = \phi n$ and $f \in \mathcal{F}$ be fixed. Let $S \subset [n]$ be the set of active indexes in $\tilde{\mathbb{T}}^{\varepsilon}_{n,k}(f)$, by the triangular inequality:

$$\left|\tilde{\mathbb{T}}^{\varepsilon}_{n,k}(f) - \tilde{\mathbb{G}}_n(\tau_M \circ f)\right| \leq \overbrace{\left|\frac{1}{\sqrt{n}(1-2\phi)}\sum_{i\in S}\xi_i f(X^{\varepsilon}_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \tau_M(f(X_i))\right|}^{(b)}$$
$$+ \underbrace{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\left(\widehat{P}_n(\tau_M\circ f) - \widehat{T}^{\varepsilon}_{n,k}(f)\right)\right|}_{(a)}$$

First step: control probabilities.

Recall that the Gaussian weights ξ_i are standard i.i.d. Gaussians, so

$$E = \left\{\max_{i\in[n]}\xi_i \leq 2\sqrt{2\ln 2n} \text{ and } \sum_{i=1}^n \xi_i \leq \sqrt{2n\ln n}\right\} \text{ satisfies } \mathbb{P}[E] \geq 1 - \frac{2}{n}.$$

In the remainder of the proof we assume that E happens.

Second step: bound (a).

Since the term $\widehat{P}_n(\tau_M \circ f) - \widehat{T}^{\varepsilon}_{n,k}(f)$ is constant in the sum one can bound

$$(a) \leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \right| \left| \widehat{P}_n(\tau_M \circ f) - \widehat{T}^{\varepsilon}_{n,k}(f) \right|.$$

The sum of the weights ξ is bounded in the event E and the term $\left| \widehat{P}_n(\tau_M \circ f) - \widehat{T}^{\varepsilon}_{n,k}(f) \right|$ is bounded by the original bounding lemma (Lemma 2.3), thus

$$(a) \leq 6\phi M \sqrt{2 \ln n}.$$

Third step: bound (b).

We decompose (b) the same way we do in the proof of the original bounding lemma, so

$$\begin{aligned}(b) \leq & \left| \frac{1}{\sqrt{n}(1-2\phi)} \sum_{i \in S} \xi_i f(X_i^{\varepsilon}) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \tau_M(f(X_i^{\varepsilon})) \right| \\ & + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \tau_M(f(X_i^{\varepsilon})) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \tau_M(f(X_i)) \right|\end{aligned}$$

The last term in the RHS can be bounded observing that at most $\lfloor \varepsilon n \rfloor$ terms differ and all terms are bounded by $2M\sqrt{2 \ln 2n}$, thus

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \tau_M(f(X_i^{\varepsilon})) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \tau_M(f(X_i)) \right| \leq 4 \lfloor \varepsilon n \rfloor M \sqrt{\frac{2 \ln 2n}{n}}.$$

To bound the first term in the RHS we note that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \tau_M(f(X_i^{\varepsilon})) = (1-2\phi) \frac{1}{\sqrt{n}(1-2\phi)} \sum_{i \in S} \xi_i \tau_M(f(X_i^{\varepsilon})) + 2\phi\eta \tag{5.12}$$

for some η satisfying $|\eta| \leq 2M\sqrt{2n \ln 2n}$. In addition,

$$\left| \frac{1}{\sqrt{n}(1-2\phi)} \sum_{i \in S} \xi_i \left(\tau_M(f(X_i^{\varepsilon})) - f(X_i^{\varepsilon}) \right) \right| \leq \frac{4t}{\sqrt{n}(1-2\phi)} 2M\sqrt{2 \ln 2n}. \tag{5.13}$$

Where the last inequality happens since $V_M(\mathcal{F}) \leq t$, $\left| \widehat{T}^{\varepsilon}_{n,k}(f) \right| \leq M$ and for every $i \in S$

$$\left| \xi_i \left(f(X_i^{\varepsilon}) - \widehat{T}^{\varepsilon}_{n,k}(f) \right) \right| \leq 4M\sqrt{2 \ln 2n},$$

implying that

$$|\xi_i(\tau_M(f(X_i^\varepsilon)) - f(X_i^\varepsilon))| \leq \left|\xi_i\left(f(X_i^\varepsilon) - \widehat{T}_{n,k}^\varepsilon(f)\right)\right| + \left|\xi_i\left(\widehat{T}_{n,k}^\varepsilon(f) - \tau_M(f(X_i^\varepsilon))\right)\right| \leq 8M\sqrt{2\ln 2n}.$$

It also follows that, for every $i \in S$,

$$|\xi_i f(X_i^\varepsilon)| \leq \left|\xi_i\left(f(X_i^\varepsilon) - \widehat{T}_{n,k}^\varepsilon(f)\right)\right| + \left|\xi_i\widehat{T}_{n,k}^\varepsilon(f)\right| \leq 8M\sqrt{2\ln 2n}.$$

From Equations (5.12) and (5.13) we obtain

$$\left|\frac{1}{\sqrt{n}(1-2\phi)}\sum_{i\in S}\xi_i f(X_i^\varepsilon) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\tau_M(f(X_i^\varepsilon))\right| \leq \left(10\phi + \frac{4t}{n}\right)2M\sqrt{2n\ln 2n}.$$

Thus, $(b) \leq 28\phi M\sqrt{2n\ln 2n}$. ■

The proof follows applying Lemma 5.20 followed by 5.23 to approximate $\tilde{Z}_{n,k}^\varepsilon(\mathcal{F})$ by $\tilde{Z}_n(\mathcal{F}_M^o)$. To finish, we use Lemma 5.9 to approximate $\tilde{Z}_n(\mathcal{F}_M^o)$ by $Z(\mathcal{F}_M^o)$ and conclude using Gaussian to Gaussian comparison and Gaussian anti-concentration inequalities to approximate $Z(\mathcal{F}_M^o)$ by $Z(\mathcal{F})$. We conclude properly selecting M to minimize errors.

5.5 Proofs

5.5.1 High-dimensional results.

Gaussian approximation. Here we finish the proof of Theorem 5.1. Recall Equation (5.8), which is repeated above

$$\varrho \leq \left(6\phi M + \nu_p^p M^{1-p}\right)\frac{2+\sqrt{2\ln d}}{\underline{\sigma}_{\mathcal{F},P}}\sqrt{n} + C'(\ln d)\sqrt{4\nu_p^p M^{2-p}} + C\left(\frac{M^2\ln^5(nd)}{n}\right)^{\frac{1}{4}} + 2e^{-n\frac{\nu_p^p}{M^p}}$$

and holds taking

$$t = \left\lceil 3\ln(1+d) + 7n\frac{\nu_p^p}{M^p}\right\rceil \text{ and } k := \phi n \text{ where } \phi = \frac{\lfloor \varepsilon n\rfloor + t}{n}.$$

Since $d \geq 2$ and $n \geq 3$, we can use $3\ln(1+d)+1 \leq 3\ln(nd)$ and $2+\sqrt{2\ln d} \leq \frac{5}{2}\sqrt{\ln(nd)}$ to bound

$$k \leq \varepsilon n + 3\ln(nd) + 7n\frac{\nu_p^p}{M^p}$$

and

$$\left(6\phi M + \nu_p^p M^{1-p}\right) \frac{2+\sqrt{2\ln d}}{\underline{\sigma}_{\mathcal{F},P}}\sqrt{n} \leq \frac{5}{2\underline{\sigma}_{\mathcal{F},P}} \left(6\varepsilon M + 18\frac{\ln(nd)}{n}M + 43\nu_p^p M^{1-p}\right)\sqrt{n\ln(nd)}.$$

The previous bounds combined yield

$$\begin{aligned}\varrho \leq \frac{15}{\underline{\sigma}_{\mathcal{F},P}}\varepsilon M\sqrt{n\ln(nd)} + \frac{108}{\underline{\sigma}_{\mathcal{F},P}}\left(\frac{\ln(nd)}{n}M + \nu_p^p M^{1-p}\right)\sqrt{n\ln(nd)} \\ + C'\nu_p^{\frac{p}{2}} M^{1-\frac{p}{2}}\ln d + C\left(\frac{M^2\ln^5(nd)}{n}\right)^{\frac{1}{4}} + 2\exp\left\{-n\frac{\nu_p^p}{M^p}\right\}.\end{aligned}$$

Let $C'' = \frac{108}{\underline{\sigma}_{\mathcal{F},P}} \vee C \vee C'$. To finish we let $M = n^\alpha \nu_p^\beta \ln^\gamma(nd)$ and explore the choices of α, β, γ to minimize our bound for ϱ. Notice that the term in the exponential can be neglected during the optimization as it will be of smaller order. We can write

$$\begin{aligned}\varrho \leq \frac{15}{\underline{\sigma}_{\mathcal{F},P}}\varepsilon n^{\frac{1}{2}+\alpha}\nu_p^\beta \ln^{\frac{1}{2}+\gamma}(nd) + C''\Big[n^{\alpha\left(1-\frac{p}{2}\right)}\nu_p^{\beta\left(1-\frac{p}{2}\right)+\frac{p}{2}}(\ln(nd))^{\gamma\left(1-\frac{p}{2}\right)+1} \\ + n^{\alpha-\frac{1}{2}}\nu_p^\beta(\ln(nd))^{\gamma+\frac{3}{2}} + n^{\alpha(1-p)+\frac{1}{2}}\nu_p^{\beta(1-p)+p}(\ln(nd))^{\gamma(1-p)+\frac{1}{2}} + n^{\frac{2\alpha-1}{4}}\nu_p^{\frac{\beta}{2}}(\ln(nd))^{\frac{2\gamma+5}{4}}\Big].\end{aligned}$$

One can check that

$$\alpha = \frac{3}{4p-2},\ \beta = 1,\ \text{and}\ \gamma = -\frac{1}{p}$$

yields the desired bound on ϱ and also that (5.3) holds with the previous choice of α, β and γ.

Bootstrap approximations. Here we complete the proof of Theorem 5.8.

Part (i). We start finishing the proof of the empirical bootstrap approximation. Recalling the discussion in §5.4, (5.10) holds with probability at least $1 - e^{-n\frac{\nu_p^p}{M^p}}$ given that

$$k = \phi n \text{ where } \phi = \frac{\lfloor \varepsilon n + (\varepsilon n) \vee t \rfloor + t}{n} \text{ and } t = \left\lceil 24\ln(1+d) + 49n\frac{\nu_p^p}{M^p}\right\rceil.$$

Lemma 5.9 (case (i)) gives, with probability at least $1 - C\left(\frac{M^2\ln^5(nd)}{n}\right)^{\frac{1}{4}}$,

$$\sup_{\lambda\in\mathbb{R}}\left|\mathbb{P}\left[\tilde{Z}_n(\mathcal{F}_M^o) \leq \lambda \,\middle|\, X_{1:n}\right] - \mathbb{P}\left[Z(\mathcal{F}_M^o) \leq \lambda\right]\right| \leq C\left(\frac{M^2\ln^5(nd)}{n}\right)^{\frac{1}{4}}$$

for some absolute constant C depending only on $\nu_2(\mathcal{F}_M^o) \leq \nu_2(\mathcal{F})$ and on $\underline{\sigma}_{\mathcal{F}_M^o,P}$. The weak variance $\underline{\sigma}_{\mathcal{F}_M^o,P}$ can be chosen such that $\underline{\sigma}_{\mathcal{F}_M^o,P} \geq \frac{1}{2}\underline{\sigma}_{\mathcal{F},P}$, it will be a consequence of our final

choice of M, of assumption (5.4), and of Lemma 5.21. Using the Gaussian to Gaussian comparison inequality (Lemma A.4) and the bound from Lemma 5.21 we get, again with probability at least $1 - C\left(\frac{M^2 \ln^5(nd)}{n}\right)^{\frac{1}{4}}$,

$$\sup_{\lambda \in \mathbb{R}} \left| \mathbb{P}\left[\tilde{Z}_n(\mathcal{F}_M^o) \leq \lambda \,\middle|\, X_{1:n} \right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda \right] \right| \leq C\left(\frac{M^2 \ln^5(nd)}{n}\right)^{\frac{1}{4}} + C'(\ln d)\sqrt{4\nu_p^p M^{2-p}}$$

Using (5.10) and Nazarov's Gaussian anti-concentration inequality (Lemma A.5) yields, with probability at least $1 - C\left(\frac{M^2 \ln^5(nd)}{n}\right)^{\frac{1}{4}} - e^{-n\frac{\nu_p^p}{M^p}}$,

$$\tilde{\varrho} \leq \left(6\phi M + \nu_p^p M^{1-p}\right) \frac{2 + \sqrt{2 \ln d}}{\underline{\sigma}_{\mathcal{F},P}} \sqrt{n} + C'(\ln d)\sqrt{4\nu_p^p M^{2-p}} + C\left(\frac{M^2 \ln^5(nd)}{n}\right)^{\frac{1}{4}} + 3e^{-\frac{t}{8}}.$$

The remainder of the proof is analogous to the Gaussian approximation. Since $d \geq 2$ and $n \geq 3$ we can use $24 \ln(1+d) + 1 \leq 25 \ln(nd)$ and $2 + \sqrt{2 \ln d} \leq \frac{5}{2}\sqrt{\ln(nd)}$ to bound

$$k \leq 2\varepsilon n + 50 \ln(nd) + 98 n \frac{\nu_p^p}{M^p}$$

and

$$\left(6\phi M + \nu_p^p M^{1-p}\right) \frac{2 + \sqrt{2 \ln d}}{\underline{\sigma}_{\mathcal{F},P}} \sqrt{n} \leq \frac{5}{2\underline{\sigma}_{\mathcal{F},P}} \left(12\varepsilon M + 300 \frac{\ln(nd)}{n} M + 589 \nu_p^p M^{1-p}\right) \sqrt{n \ln(nd)}.$$

Taking $C'' = \frac{1473}{\underline{\sigma}_{\mathcal{F},P}} \vee C \vee C'$ and $M = n^\alpha \nu_p^\beta \ln^\gamma(nd)$ we proceed as for the Gaussian approximation minimizing our bound for $\tilde{\varrho}$. Notice that the terms in the exponential can be neglected during the optimization as it will be of smaller order. We have

$$\begin{aligned}\tilde{\varrho} \leq \frac{30}{\underline{\sigma}_{\mathcal{F},P}} \varepsilon n^{\frac{1}{2}+\alpha} \nu_p^\beta \ln^{\frac{1}{2}+\gamma}(nd) + C'' \Bigg[& n^{\alpha\left(1-\frac{p}{2}\right)} \nu_p^{\beta\left(1-\frac{p}{2}\right)+\frac{p}{2}} (\ln(nd))^{\gamma\left(1-\frac{p}{2}\right)+1} \\ & + n^{\alpha-\frac{1}{2}} \nu_p^\beta (\ln(nd))^{\gamma+\frac{3}{2}} + n^{\alpha(1-p)+\frac{1}{2}} \nu_p^{\beta(1-p)+p} (\ln(nd))^{\gamma(1-p)+\frac{1}{2}} + n^{\frac{2\alpha-1}{4}} \nu_p^{\frac{\beta}{2}} (\ln(nd))^{\frac{2\gamma+5}{4}} \Bigg].\end{aligned}$$

One can check that

$$\alpha = \frac{3}{4p-2},\ \beta = 1,\ \text{and}\ \gamma = -\frac{1}{p}$$

yields the desired bound. Moreover, it gives

$$t = \left\lceil 24 \ln(1+d) + 49 n^{\frac{p-2}{4p-2}} \ln(nd) \right\rceil$$

which implies

$$3e^{-\frac{t}{8}} \leq 3 \exp\left\{ -6 n^{\frac{p-2}{4p-2}} \ln(nd) \right\}.$$

Part (ii). We now proceed to the Gaussian bootstrap approximation. Thus, assume that $\tilde{X}^{\varepsilon}_{1:n}$ was obtained via the Gaussian bootstrap and let

$$k = \phi n \text{ where } \phi = \frac{\lfloor \varepsilon n \rfloor + t}{n} \text{ and } t = \left\lceil 3\ln(1+d) + 7n\frac{\nu_p^p}{M^p} \right\rceil.$$

Again, we let E be the event defined in (5.9), which has probability at least $1 - e^{-n\frac{\nu_p^p}{M^p}}$. Lemma 5.23 yields

$$\mathbb{P}\left[\left| \tilde{Z}^{\varepsilon}_{n,k}(\mathcal{F}) - \tilde{Z}_n(\mathcal{F}^o_M) \right| \leq 44\phi M\sqrt{n\ln 2n} \;\middle|\; E \right] \geq 1 - \frac{2}{n},$$

where we used

$$28\sqrt{2n\ln 2n} + 6\sqrt{2\ln n} \leq 44\sqrt{n\ln 2n}$$

for all $n \geq 3$.

Lemma 5.9 (case (i)) gives, with probability at least $1 - C\left(\frac{M^2\ln^5(nd)}{n}\right)^{\frac{1}{4}}$,

$$\sup_{\lambda\in\mathbb{R}} \left| \mathbb{P}\left[\tilde{Z}_n(\mathcal{F}^o_M) \leq \lambda \;\middle|\; X_{1:n} \right] - \mathbb{P}\left[Z(\mathcal{F}^o_M) \leq \lambda \right] \right| \leq C\left(\frac{M^2\ln^5(nd)}{n}\right)^{\frac{1}{4}}$$

for some absolute constant C depending only on $\nu_2(\mathcal{F}) \leq \nu_2(\mathcal{F})$ and on $\underline{\sigma}_{\mathcal{F}^o_M,P}$. As before, the weak variance $\underline{\sigma}_{\mathcal{F}^o_M,P}$ will be chosen such that $\underline{\sigma}_{\mathcal{F}^o_M,P} \geq \frac{1}{2}\underline{\sigma}_{\mathcal{F},P}$, as consequence of our final choice of M, of assumption (5.4), and of Lemma 5.21.

Using the Gaussian to Gaussian comparison inequality (Lemma A.4) and Nazarov's Gaussian anti-concentration inequality (Lemma A.5) yields

$$\tilde{\varrho} \leq 44\phi M\frac{2+\sqrt{2\ln d}}{\underline{\sigma}_{\mathcal{F},P}}\sqrt{n\ln 2n} + C'(\ln d)\sqrt{4\nu_p^p M^{2-p}} + C\left(\frac{M^2\ln^5(nd)}{n}\right)^{\frac{1}{4}} + \frac{2}{n}.$$

Again we bound

$$k \leq \varepsilon n + 3\ln(nd) + 7n\frac{\nu_p^p}{M^p}$$

and

$$44\phi M\frac{2+\sqrt{2\ln d}}{\underline{\sigma}_{\mathcal{F},P}}\sqrt{n\ln 2n} \leq \frac{110}{\underline{\sigma}_{\mathcal{F},P}}\left(\varepsilon M + 4\frac{\ln(nd)}{n}M + 7\nu_p^p M^{1-p}\right)\sqrt{n}\ln(nd).$$

Taking $C'' = \frac{770}{\underline{\sigma}_{\mathcal{F},P}} \vee C' \vee C$ and $M = n^{\alpha}\nu_p^{\beta}\ln^{\gamma}(nd)$ we proceed minimizing our bound for $\tilde{\varrho}$. We have

$$\begin{aligned}\tilde{\varrho} \leq \frac{110}{\underline{\sigma}_{\mathcal{F},P}}\varepsilon n^{\frac{1}{2}+\alpha}\nu_p^{\beta}\ln^{1+\gamma}(nd) + C''\Bigg[& n^{\alpha\left(1-\frac{p}{2}\right)}\nu_p^{\beta\left(1-\frac{p}{2}\right)+\frac{p}{2}}(\ln(nd))^{\gamma\left(1-\frac{p}{2}\right)+1} \\ & + n^{\alpha-\frac{1}{2}}\nu_p^{\beta}(\ln(nd))^{2+\gamma} + n^{\alpha(1-p)+\frac{1}{2}}\nu_p^{\beta(1-p)+p}(\ln(nd))^{\gamma(1-p)+1} + n^{\frac{2\alpha-1}{4}}\nu_p^{\frac{\beta}{2}}(\ln(nd))^{\frac{2\gamma+5}{4}}\Bigg] + \frac{2}{n}.\end{aligned}$$

And our result follows taking

$$\alpha = \frac{3}{4p-2},\ \beta = 1,\ \text{and}\ \gamma = -\frac{1}{p}.$$

5.5.2 Gaussian approximation for empirical processes.

Here we proof Theorem 5.14. The overall argument is to use a δ-net transforming the infinite-dimensional problem into a finite-dimensional problem and then controlling the errors.

First step: counting and bounding.

In order to use the original counting lemma (Lemma 2.2) for a class $\mathcal{G}$ we need to find t, M satisfying

$$\underbrace{\sup_{g\in\mathcal{G}} P\left\{|g(X)| > \frac{M}{2}\right\}}_{\text{(i)}} + \underbrace{\frac{8\mathsf{Rad}_n(\tau_M \circ \mathcal{G}, P)}{M}}_{\text{(ii)}} \leq \frac{t}{8n}.$$

It will imply that $V_M(\mathcal{G}) \leq t$ with probability at least $1 - e^{-t}$. We will apply it to the class $\mathcal{G} = \{f^{\frac{p}{2}} : f \in \mathcal{F}\}$, for that end we must bound (i) and (ii) in order to choose t.

Bound (i). Using Markov's bound:

$$\sup_{g\in\mathcal{G}} P\left\{|g(X)| > \frac{M^{\frac{p}{2}}}{2}\right\} = \sup_{f\in\mathcal{F}} P\left\{|f(X)|^p > \frac{M^p}{4}\right\} \leq 4\frac{\nu_p^p(\mathcal{F})}{M^p}$$

Bound (ii). Our strategy will be to apply symmetrization (Lemma 2.1) and then Lemma A.1. To apply symmetrization we need first to center the class $\tau_{M^{\frac{p}{2}}} \circ \mathcal{G}$. Let

$$\mathcal{G}^o_{M^{\frac{p}{2}}} = \left\{\tau_{M^{\frac{p}{2}}} \circ g - P\left(\tau_{M^{\frac{p}{2}}} \circ g\right) : g \in \mathcal{G}\right\}.$$

We have

$$\mathsf{Rad}_n\left(\tau_{M^{\frac{p}{2}}} \circ \mathcal{G}, P\right) \leq \mathsf{Rad}_n\left(\mathcal{G}^o_{M^{\frac{p}{2}}}, P\right) + \frac{1}{\sqrt{n}} \sup_{g\in\mathcal{G}} P\left(\tau_{M^{\frac{p}{2}}} \circ g\right)$$

and $\sup_{g\in\mathcal{G}} P\left(\tau_{M^{\frac{p}{2}}} \circ g\right) \leq \nu_{p/2}^{p/2}(\mathcal{F}) \leq \nu_p^{\frac{p}{2}}(\mathcal{F})$. We can now apply symmetrization to obtain

$$\frac{8\mathsf{Rad}_n\left(\tau_{M^{\frac{p}{2}}} \circ \mathcal{G}, P\right)}{M^{\frac{p}{2}}} \leq \frac{16\mathsf{Emp}_n\left(\mathcal{G}^o_{M^{\frac{p}{2}}}, P\right)}{M^{\frac{p}{2}}} + 8\sqrt{\frac{\nu_p^p(\mathcal{F})}{nM^p}}.$$

To use Lemma A.1 we must first bound the entropy of $\tau_{M^{\frac{p}{2}}} \circ \mathcal{G}$, this is done by Lemma 5.13:

$$\sup_Q \mathcal{N}\left(\tau_{M^{\frac{p}{2}}} \circ \mathcal{G}, d_Q, \delta M^{\frac{p}{2}}\right) \leq \left(\frac{8e}{\delta}\right)^v \ \forall \delta \in (0, 1],$$

where $v = 2vc(\mathcal{F})$. By Lemma A.1 we have

$$\mathsf{Emp}_n\left(\mathcal{G}^o_{M^{\frac{p}{2}}}, P\right) \leq C_1 \left\{ \sqrt{\frac{v\nu_p^p(\mathcal{F})}{n} \ln \frac{8eM^{\frac{p}{2}}}{\nu_p^{\frac{p}{2}}}} + \frac{vM^{\frac{p}{2}}}{n} \ln \frac{8eM^{\frac{p}{2}}}{\nu_p^{\frac{p}{2}}} \right\}$$

for some absolute constant C_1.

Choose t. Using the AM-GM inequality we can now bound

$$\begin{aligned}(i)+(ii) &\leq 16C_1 \left\{ \sqrt{\frac{v\nu_p^p(\mathcal{F})}{nM^p} \ln \frac{8eM^{\frac{p}{2}}}{\nu_p^{\frac{p}{2}}}} + \frac{v}{n} \ln \frac{8eM^{\frac{p}{2}}}{\nu_p^{\frac{p}{2}}} \right\} + 8\sqrt{\frac{\nu_p^p(\mathcal{F})}{nM^p}} + 4\frac{\nu_p^p(\mathcal{F})}{M^p} \\ &\leq 24C_1 \frac{v}{n} \ln \frac{8eM^{\frac{p}{2}}}{\nu_p^{\frac{p}{2}}} + (4+8C_1)\frac{\nu_p^p(\mathcal{F})}{M^p} + 8\sqrt{\frac{\nu_p^p(\mathcal{F})}{nM^p}}\end{aligned}$$

Assuming that (we will verify it latter)

$$n\frac{\nu_p^p(\mathcal{F})}{M^p} \geq 1 \tag{5.14}$$

one can take $C = 8\left(24C_1 \vee (12+8C_1)\right)$ and

$$t \geq C\left(v \ln \frac{8eM^{\frac{p}{2}}}{\nu_p^{\frac{p}{2}}} + n\frac{\nu_p^p(\mathcal{F})}{M^p} \right).$$

It gives $V_M(\mathcal{F}) \leq t$ with probability at least $1 - e^{-t}$. Moreover, the bounding lemma (Lemma 2.3) can be used taking

$$k = \phi n \text{ where } \phi = \frac{\lfloor \varepsilon n \rfloor + t}{n},$$

to obtain

$$\sup_{f\in\mathcal{F}} \left| \mathbb{T}^{\varepsilon}_{n,k}(f) - \mathbb{G}_n(\tau_M \circ f) \right| \leq 6\phi M\sqrt{n}.$$

In addition, centering the class $\tau_M \circ \mathcal{F}$ gives

$$\left| Z^{\varepsilon}_{n,k}(\mathcal{F}) - Z_n(\mathcal{F}^o_M) \right| \leq \left(6\phi M + \nu_p^p M^{1-p}\right)\sqrt{n}.$$

Second step: approximating by a δ-net.

Let $\delta = n^{-\frac{3p}{8p-4}}$ and let $\mathcal{H}_M$ be a δM-net of $\tau_M \circ \mathcal{F}$, also let $\mathcal{H}$ be the corresponding set in $\mathcal{F}$ (notice that it may not be a δM-net for $\mathcal{F}$), i.e.,

$$\mathcal{H}_M = \{\tau_M \circ f : f \in \mathcal{H}\}.$$

By Lemma 5.13 we have

$$|\mathcal{H}| = |\mathcal{H}_M| \leq \left(\frac{8e}{\delta}\right)^v \ \forall \delta \in (0,1]$$

and so $\ln|\mathcal{H}| \leq 2K_n(\mathcal{F})$.

We can bound the error of approximating by a δ-net by

$$Z_n(\mathcal{F}_M^o) - Z_n(\mathcal{H}_M^o) \leq \sup_{f,g\in\mathcal{F}: d_P(\tau_M\circ f, \tau_M\circ g)\leq \delta M} \mathbb{G}_n(\tau_M\circ f - P\tau_M\circ f) - \mathbb{G}_n(\tau_M\circ g - P\tau_M\circ g).$$

We now use Lemma A.2 (notice that its assumptions hold because of (5.5)) to bound the RHS by

$$Z_n(\mathcal{F}_M^o) - Z_n(\mathcal{H}_M^o) \leq C_2\delta M\sqrt{K_n(\mathcal{F})}$$

with probability at least $1-\frac{1}{n}$ for some universal constant C_2.

Third step: approximate $Z_n(\mathcal{H}_M^o)$ by $Z(\mathcal{H})$.

Apply Theorem 5.2, Lemma A.4 and Lemma 5.21 to get:

$$\sup_{\lambda\in\mathbb{R}} |\mathbb{P}[Z_n(\mathcal{H}_M^o)\leq\lambda] - \mathbb{P}[Z(\mathcal{H})\leq\lambda]| \leq C_3\left(\frac{M^2K_n(\mathcal{F})^5}{n}\right)^{\frac{1}{4}} + C_4K_n(\mathcal{F})\sqrt{\nu_p^pM^{2-p}}$$

for constants C_3 and C_4 that will depend only on $\underline{\sigma}_{\mathcal{F},P}$ and $\nu_2(\mathcal{F},P)$ by (5.6).

Forth step: approximate $Z_{n,k}(\mathcal{F})$ by $Z(\mathcal{F})$.

By the second and third steps together with Nazarov's inequality (Lemma A.5) we also have, for every $\lambda\geq 0$,

$$\begin{aligned}\sup_{\lambda>0}\left|\mathbb{P}\left[Z_{n,k}^{\varepsilon}(\mathcal{F})\leq\lambda\right] - \mathbb{P}[Z(\mathcal{H})\leq\lambda]\right| \leq &\left((6\phi M + \nu_p^pM^{1-p})\sqrt{n} + C_2Mn^{-\frac{3p}{8p-4}}\sqrt{K_n}\right)\frac{2+\sqrt{2K_n}}{\underline{\sigma}_{\mathcal{F},P}}\\ &+ C_3\left(\frac{M^2K_n^5}{n}\right)^{\frac{1}{4}} + C_4K_n\sqrt{\nu_p^pM^{2-p}} + e^{-t} + \frac{1}{n}\end{aligned}$$

On the other hand, if $d_P(\tau_M\circ f, \tau_M\circ g)\leq\delta M$, then

$$P(f-g)^2 = Pf^2 - 2Pfg + Pg^2 \leq d_P(\tau_M\circ f,\tau_M\circ g)^2 + 16\nu_p^pM^{2-p} \leq \delta^2M^2 + 16\nu_p^pM^{2-p}.$$

Thus the d_P distance between a point $f\in\mathcal{F}$ and its closest point $g\in\mathcal{H}$ is at most $\sqrt{\delta^2M^2+16\nu_p^pM^{2-p}}$. Borell-TIS inequality gives

$$\mathbb{P}\left[Z(\mathcal{F}) - Z(\mathcal{H}) \leq \Xi\left(\sqrt{\frac{M^2}{n}+16\nu_p^pM^{2-p}}\right) + \sqrt{2\left(\frac{M^2}{n}+16\nu_p^pM^{2-p}\right)\ln n}\right] \geq 1-\frac{1}{n}.$$

Up to redefining the constant C_4,

$$\sqrt{32\nu_p^p M^{2-p}\ln n}\frac{2+\sqrt{2K_n}}{\underline{\sigma}_{\mathcal{F},P}} \leq C_4 K_n \sqrt{\nu_p^p M^{2-p}}.$$

Using Nazarov's inequality again and $2+\sqrt{2K_n} \leq 3\sqrt{K_n}$ yields (up to redefining C_2)

$$\begin{aligned}\sup_{\lambda>0}\left|\mathbb{P}\left[Z_{n,k}^{\varepsilon}(\mathcal{F}) \leq \lambda\right] - \mathbb{P}\left[Z(\mathcal{F}) \leq \lambda\right]\right| \leq & \frac{3\sqrt{nK_n}}{\underline{\sigma}_{\mathcal{F},P}}\left(6\phi M + \nu_p^p M^{1-p}\right) + C_2 K_n M n^{-\frac{3p}{8p-4}} \\ & \frac{3\sqrt{K_n}}{\underline{\sigma}_{\mathcal{F},P}}\Xi\left(\sqrt{\frac{M^2}{n}+16\nu_p^p M^{2-p}}\right) \\ & + 3C_3\left(\frac{M^2K_n^5}{n}\right)^{\frac{1}{4}} + 2C_4K_n\sqrt{\nu_p^p M^{2-p}} + e^{-t} + \frac{1}{n}\end{aligned}$$

Taking $M = n^{\alpha}\nu_p^{\beta}K_n^{\gamma}$ and optimizing gives

$$\alpha = \frac{3}{4p-2},\ \beta = 1,\ \text{and}\ \gamma = -\frac{1}{p}.$$

We now verify (5.14). Since $K_n \geq \ln 8e \geq 1$ and $n \geq 1$

$$n\frac{\nu_p^p(\mathcal{F})}{M^p} = n^{\frac{p-2}{4p-2}}K_n \geq 1.$$

Moreover $v\ln\frac{8eM^{\frac{p}{2}}}{\nu_p^{\frac{p}{2}}} \leq 2K_n$ and so we can take $t = \left\lceil Cn^{\frac{p-2}{4p-2}}K_n\right\rceil \geq n^{\frac{p-2}{4p-2}}K_n$ up to redefining C.

Chapter 6

Vector mean estimation under arbitrary norms

6.1 Introduction

In this chapter we dive deeper in the problem of vector mean estimation under general norms, which was discussed in §3.1.2. This will be an application of our uniform mean estimation and Gaussian approximation results. We here let $\mathbf{X} = \mathbb{R}^d$ and P be a distribution over $\mathbf{X}$. Our goal is to estimate the mean μ_P of P given samples $X_{1:n}^{\varepsilon}$. Let $\|\cdot\|$ be a norm in $\mathbb{R}^d$, and $S \subset \mathbb{R}^d$ be a symmetric set that spans $\mathbb{R}^d$ such that

$$\forall x \in \mathbb{R}^d \,:\, \|x\| = \sup_{v \in S} \langle x, v \rangle.$$

Such a set S always exists: for instance, one can take it to be the unit ball of the dual norm. Recall that the mean of distribution P in $\mathbb{R}^d$ is characterized by

$$\forall v \in S \,:\, \langle v, \mu_P \rangle = P\langle v, \cdot \rangle.$$

In this problem we are interested in finding a measurable function $\hat{\mu}(x_{1:n})$ satisfying

$$\mathbb{P}\left[\|\hat{\mu}(X_{1:n}^{\varepsilon}) - \mu_P\| \leq \Phi_P(n, \alpha, \varepsilon)\right] \geq 1 - \alpha \tag{6.1}$$

for the smallest possible $\Phi_P(n, \alpha, \varepsilon)$. Also notice that there is a natural family $\mathcal{F}$ of functions associated with the norm $\|\cdot\|$ given by

$$\mathcal{F} := \{\langle \cdot, v \rangle : v \in S\}.$$

As discussed in §3.1.2, the problem of estimating μ_P is closely related to the problem of uniform mean estimation over the class $\mathcal{F}$. To see this let $\hat{E}_f$ be any given estimator for $f \in \mathcal{F}$. Assume there is a measurable map $\hat{\mu} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ such that

$$\forall x_{1:n} \in (\mathbb{R}^d)^n \,:\, \hat{\mu}(x_{1:n}) \in \arg\min_{\mu \in \mathbb{R}^d} \left(\sup_{f \in \mathcal{F}} \left| \hat{E}_f(x_{1:n}) - f(\mu) \right| \right). \tag{6.2}$$

Then

$$\begin{aligned}
\|\hat{\mu}(X_{1:n}) - \mu_P\| &= \sup_{f \in \mathcal{F}} |f(\hat{\mu}(X_{1:n}) - \mu_P)| \\
&= \sup_{f \in \mathcal{F}} |f(\hat{\mu}(X_{1:n})) - f(\mu_P)| \\
&= \sup_{f \in \mathcal{F}} \left| f(\hat{\mu}(X_{1:n})) - \hat{E}_f(X_{1:n}) \right| + \sup_{f \in \mathcal{F}} \left| \hat{E}_f(X_{1:n}) - f(\mu_P) \right| \\
&\leq \sup_{f \in \mathcal{F}} \left| f(\mu_P) - \hat{E}_f(X_{1:n}) \right| + \sup_{f \in \mathcal{F}} \left| \hat{E}_f(X_{1:n}) - f(\mu_P) \right| \\
&= 2 \sup_{f \in \mathcal{F}} \left| \hat{E}_f(X_{1:n}) - f(\mu_P) \right|
\end{aligned}$$

where in the inequality we used the definition of $\hat{\mu}$ as a minimizer.

On the other hand, given a mean estimator $\hat{\mu}$ for μ_P, we can easily construct estimators $\hat{E}_f(x_{1:n}) = f(\hat{\mu}(x_{1:n}))$ for each $f \in \mathcal{F}$, and obtain

$$\sup_{f \in \mathcal{F}} \left| \hat{E}_f(X_{1:n}) - f(\mu_P) \right| = \|\hat{\mu}(X_{1:n}) - \mu_P\|.$$

6.1.1 Relevant parameters.

The parameters considered on Problem 3.1 are still relevant for this problem. Namely,

$$\nu_P(\mathcal{F}, P) = \sup_{f \in \mathcal{F}} (P\,|f(X - \mu_P)|^p)^{\frac{1}{p}} = \sup_{v \in S} (P\,|\langle X - \mu_P, v \rangle|^p)^{\frac{1}{p}}, \tag{6.3}$$

$$\mathsf{Emp}_n(\mathcal{F}, P) = \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n X_i - \mu_P \right\| = \mathbb{E} \left[\sup_{v \in S} \left| \frac{1}{n} \sum_{i=1}^n \langle X_i - \mu_P, v \rangle \right| \right]. \tag{6.4}$$

In the special case were $\|\cdot\|$ is the euclidean norm we have, for P with covariance Σ_P,

$$\nu_P(\mathcal{F}, P) = \|\Sigma_P\|_{op} \text{ and } \mathsf{Emp}_n(\mathcal{F}, P) \leq \sqrt{\frac{\mathrm{tr}(\Sigma_P)}{n}}.$$

We also define the Gaussian width of a set $A \subset \mathbb{R}^d$ as

$$w(A) := \mathbb{E} \left[\sup_{v \in A} \langle v, W \rangle \right]$$

where $W \sim \mathcal{N}(0, I_d)$.

6.1.2 Historical background.

Early work on vector mean estimation in the sense of (6.1) includes [Minsker, 2015, Joly et al., 2017]. In the Hilbert space setting, a breakthrough result by Lugosi and Mendelson [Lugosi and Mendelson, 2019d] presented a higher-dimensional analogue of median of means. A more recent estimator by the same authors [Lugosi and Mendelson, 2021], based on a "high-dimensional trimmed mean" [1], gives better results[2]:

$$\Phi_P(n,\alpha,\varepsilon) = C\left(\sqrt{\frac{\operatorname{tr}(\Sigma_P) + \|\Sigma_P\|_{\mathrm{op}} \ln(1/\alpha)}{n}} + \inf_{p>1} \nu_p(\mathcal{F},P)\varepsilon^{1-\frac{1}{p}}\right). \tag{6.5}$$

Notice that the contamination term is optimal by §3.1.2, the other term is also known to be optimal if we only assume finite second moments (see Theorem 6.1 and the discussion following it).

The problem of general norms was studied by Lugosi and Mendelson [Lugosi and Mendelson, 2019b] and Depersin and Lecué [Depersin and Lecué, 2021]. Both papers present median of means based estimators; [Depersin and Lecué, 2021] gives the following upper bound:

$$\Phi_P(n,\alpha,\varepsilon) = C\left(\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mu_P\right\| + \nu_2(\mathcal{F},P)\sqrt{\varepsilon + \frac{\ln(1/\alpha)}{n}}\right). \tag{6.6}$$

They also present a matching lower bound for Gaussian distributions with $\varepsilon = 0$, as we discuss below.

6.1.3 A characterization in terms of the Gaussian width.

The following result characterizes $\Phi_P(n,\alpha,\varepsilon)$ for P Gaussian and $\varepsilon = 0$ in terms of the Gaussian width of $\Sigma^{\frac{1}{2}}S$ and a fluctuation term.

Theorem 6.1 (Theorem 3 of [Depersin and Lecué, 2021]). *Let Σ be a positive-definite matrix. Assume that S, $\mathcal{F}$ and $\|\cdot\|$ are as above. If $\hat{\mu} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ satisfies for all $\alpha \in \left(0, \frac{1}{4}\right]$ and for all $\mu \in \mathbb{R}^d$,*

$$\mathbb{P}_{X_{1:n}\sim\otimes_{i=1}^n \mathcal{N}(\mu,\Sigma)}\left[\|\hat{\mu}(X_{1:n}) - \mu\| \le \Phi_\Sigma(n,\alpha,0)\right] \ge 1-\alpha,$$

then

$$\Phi_\Sigma(n,\alpha,0) \ge c\left(\frac{w(\Sigma^{\frac{1}{2}}S)}{\sqrt{n}} + \nu_2(\mathcal{F})\sqrt{\frac{1}{n}\ln\frac{1}{\alpha}}\right)$$

[1] Their estimator involves a truncation at sample quantiles, and is more properly called a "Winsorized mean."

[2] They only derive explicit bounds for $p = 2$, but the more general bounds follow from the same methods.

for some absolute constant $c > 0$. Moreover, taking $\hat{\mu}$ as the empirical mean yields

$$\Phi_\Sigma(n, \alpha, 0) \leq C \left(\frac{w(\Sigma^{\frac{1}{2}} S)}{\sqrt{n}} + \nu_2(\mathcal{F}) \sqrt{\frac{1}{n} \ln \frac{1}{\alpha}} \right)$$

for some absolute constant $C > 0$.

An interpretation of Theorem 6.1 is that the correct complexity term for the problem of vector mean estimation is the Gaussian width of $\Sigma_P^{\frac{1}{2}} S$. When $\|\cdot\|$ is the euclidean norm we have (see Proposition 2.5.1 of [Talagrand, 2014])

$$\mathsf{Emp}_n(\mathcal{F}, P) \approx \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} \approx \frac{w(\Sigma^{\frac{1}{2}} \mathbb{S}^{d-1})}{\sqrt{n}},$$

and so the complexity term $\mathsf{Emp}_n(\mathcal{F}, P)$ matches the Gaussian width. However, it is unclear if it also happens under general norms. In addition, Theorem 6.1 also points to the optimality of the fluctuation term

$$\nu_2(\mathcal{F}) \sqrt{\frac{1}{n} \ln \frac{1}{\alpha}}.$$

6.2 Our estimator

To define our estimator we follow the strategy outlined in (6.2). Ideally we would like to define

$$\hat{\mu}_{n,k}(x_{1:n}) \in \operatorname*{arg\,min}_{\mu \in \mathbb{R}^d} \left(\sup_{f \in \mathcal{F}} \left| \widehat{T}_{n,k}(f, x_{1:n}) - f(\mu) \right| \right),$$

but for that end we must check if it is possible to define $\hat{\mu}_{n,k}(x_{1:n})$ in a way that the estimator is measurable. This is done in the next lemma.

Lemma 6.2. *Given $1 \leq k < n/2$, there exists a measurable function $\hat{\mu}_{n,k} : (\mathbb{R}^d)^n \to \mathbb{R}$ such that*

$$\forall x_{1:n} \in (\mathbb{R}^d)^n \, : \, \hat{\mu}_{n,k}(x_{1:n}) \in \operatorname*{arg\,min}_{\mu \in \mathbb{R}^d} \left(\sup_{f \in \mathcal{F}} \left| \widehat{T}_{n,k}(f, x_{1:n}) - f(\mu) \right| \right). \tag{6.7}$$

Proof Recall that S is a symmetric set and so $\mathcal{F}$ is a symmetric class of functions, thus

$$\sup_{f \in \mathcal{F}} \left| \widehat{T}_{n,k}(f, x_{1:n}) - f(\mu) \right| = \sup_{f \in \mathcal{F}} \widehat{T}_{n,k}(f, x_{1:n}) - f(\mu).$$

Define

$$F(\mu, x_{1:n}) = \sup_{f \in \mathcal{F}} \widehat{T}_{n,k}(f, x_{1:n}) - f(\mu),$$

which is convex in μ for fixed $x_{1:n}$ because it is a supremum of affine functions. F is also a measurable function of $(\mu, x_{1:n})$ because the supremum can be taken over a countable dense subset $\mathcal{D} \subset \mathcal{F}$. For fixed $x_{1:n}$, the values $\widehat{T}_{n,k}(f, x_{1:n})$ are uniformly bounded, and one can deduce from this that the set

$$K(x_{1:n}) := \arg\min_{\mu \in \mathbb{R}^d} F(\mu, x_{1:n})$$

is convex, compact and nonempty.

It remains to show that we can take a measurable function $\hat{\mu}_{n,k}$ with $\hat{\mu}_{n,k}(x_{1:n}) \in K(x_{1:n})$. In order to use Kuratowski and Ryll-Nardzewski measurable selection theorem [Kuratowski and Ryll-Nardzewski, 1965] we need to show that for every open set $U \subset \mathbb{R}^d$, the set

$$A_U := \{x_{1:n} : K(x_{1:n}) \cap U \neq \emptyset\}$$

is measurable. If $U = \emptyset$ we are done. Otherwise, we can write $U = \bigcup_{i=1}^n K_i$ where K_i is compact and has non-empty interior for every $i \in \mathbb{N}$. Thus, it suffices to show that

$$A_K := \{x_{1:n} : K(x_{1:n}) \cap K \neq \emptyset\}$$

is measurable for every compact set K with non-empty interior. Let D be dense and countable in $\mathbb{R}^d$, and also assume that $K \cap D$ is dense in K. Notice that $K(x_{1:n}) \cap K \neq \emptyset$ if there is some minimizer of $F(\cdot, x_{1:n})$ in K. For a given $x_{1:n} \in (\mathbb{R}^d)^n$,

$$\begin{gathered}
K(x_{1:n}) \cap K \neq \emptyset \\
\Leftrightarrow \\
\exists \mu \in K \text{ such that } F(\mu, x_{1:n}) \leq F(\mu', x_{1:n}) \, \forall \mu' \in \mathbb{R}^d \\
\Leftrightarrow \\
\exists \mu \in K \text{ such that } F(\mu, x_{1:n}) \leq F(\mu', x_{1:n}) \; \forall \mu' \in D \\
\Leftrightarrow \\
\forall m \in \mathbb{N}, \exists \mu \in K \cap D \text{ such that } F(\mu, x_{1:n}) \leq F(\mu', x_{1:n}) + \tfrac{1}{m} \; \forall \mu' \in D.
\end{gathered}$$

The first equivalence follows from our previous observation. The second equivalence follows noticing that F is continuous on μ. The last equivalence is less obvious and uses that K is compact:

- ($\Rightarrow$): assume $\mu \in K$ is such that $F(\mu, x_{1:n}) \leq F(\mu', x_{1:n}) \; \forall \mu' \in D$; since $K \cap D$ is dense in K there exists $(\mu_k)_{k=1}^\infty \subset K \cap D$ satisfying $\mu_k \to \mu$ as $k \to \infty$. Since F is continuous on μ we have $F(\mu_k, x_{1:n}) \to F(\mu, x_{1:n})$ as $k \to \infty$, given $m \in \mathbb{N}$ exists k_m such that $F(\mu_{k_m}, x_{1:n}) \leq F(\mu, x_{1:n}) + \frac{1}{m}$ and so $F(\mu_{k_m}, x_{1:n}) \leq F(\mu', x_{1:n}) + \frac{1}{m} \; \forall \mu' \in D$.

- ($\Leftarrow$): Take a sequence $(\mu_m)_{m=1}^{\infty} \subset K \cap D$ satisfying $F(\mu_m, x_{1:n}) \leq F(\mu', x_{1:n}) + \frac{1}{m}\ \forall \mu' \in D$. Since K is compact there is a sub-sequence $(\mu_{k_m})_{m=1}^{\infty} \subset (\mu_m)_{m=1}^{\infty}$ such that $\mu_{k_m} \to \mu$ as $m \to \infty$ for some $\mu \in K$. Moreover, $F(\mu, x_{1:n}) = \lim_{m\to\infty} F(\mu_{k_m}, x_{1:n})$ and so $F(\mu, x_{1:n}) \leq F(\mu', x_{1:n})\ \forall \mu' \in D$.

Writing the previous observations using the appropriate set operation for each logical quantifier yields

$$A_K = \bigcap_{m=1}^{\infty} \bigcup_{\mu \in K \cap D} \bigcap_{\mu' \in D} \left\{ x_{1:n} : F(\mu, x_{1:n}) \leq F(\mu', x_{1:n}) + \frac{1}{m} \right\}$$

and so A_K is measurable. ■

6.3 Main results

We now apply our results for uniform mean estimation (Theorem 3.1) and for Gaussian approximation (Theorem 5.14) to this problem.

6.3.1 Consequences of our uniform mean estimation results.

Our result on uniform mean estimation (Theorem 3.1) gives, for $k = \phi n$ with

$$\phi := \frac{1}{n}\left(\lfloor \varepsilon n \rfloor + \left\lceil \ln \frac{2}{\alpha} \right\rceil \vee \left\lceil \frac{\left(\frac{1}{2} - \varepsilon\right) \wedge \varepsilon}{2} n \right\rceil \right) < \frac{1}{2},$$

the following bound

$$\Phi_P(n, \alpha, \varepsilon) = C_\varepsilon \left(8\mathsf{Emp}_n(\mathcal{F}, P) + \inf_{q \in [1,2]} \nu_q(\mathcal{F}, P) \left(\frac{\ln \frac{3}{\alpha}}{n} \right)^{1-\frac{1}{q}} + \inf_{p \geq 1} \nu_p(\mathcal{F}, P) \varepsilon^{1-\frac{1}{p}} \right),$$

where C_ε is a constant depending only on ε.

In the general case, it is also an improvement over (6.6), as it allows for a better dependence on the contamination level. In the case of the Euclidean norm, this bound is optimal (by Theorem 6.1) and provides a slight improvement over the main result of [Lugosi and Mendelson, 2021] showed in (6.5) as it allows for a better fluctuation term.

6.3.2 Consequences of our Gaussian approximation results.

Assume that $\nu_p(\mathcal{F}, P) < \infty$ for some $p > 2$. To apply Theorem 5.14 to the mean vector estimation problem

we first notice that the VC-subgraph dimension of $\mathcal{F}$ is at most $d+2$, this is because for a given $f = \langle \cdot, v \rangle$ its subgraph

$$sg(f) := \{(x,t) : \langle \cdot, v \rangle \leq t\} \in \mathbb{R}^{d+1}$$

is a half-space and the VC dimension of all half-spaces is $d+2$, thus $vc(\mathcal{F}) \leq d+2$. Let

$$K_n = (2d+4)(\ln n \vee \ln 8e)$$

and assume n is large enough so it satisfies $16K_n \leq n^{\frac{p-2}{4p-2}}$ and $\nu_p^2(\mathcal{F})\, n^{-\frac{3p-6}{4p-2}} K_n^{1-\frac{2}{p}} \leq \frac{3}{8}\underline{\sigma}_{\mathcal{F},P}$. Then, Theorem 5.14 says that exists an absolute constant C such that taking

$$k := \lfloor \varepsilon n \rfloor + \left\lceil C n^{\frac{p-2}{4p-2}} K_n \right\rceil < \frac{n}{2},$$

yields

$$\mathbb{P}\left[\|\hat{\mu}_{n,k}(X_{1:n}^{\varepsilon}) - \mu_P\| \leq C\left(\frac{w(\Sigma^{\frac{1}{2}} S)}{\sqrt{n}} + \nu_2(\mathcal{F}) \sqrt{\frac{1}{n} \ln \frac{1}{\alpha}} \right) \right] \geq 1 - \alpha - \varrho$$

where

$$\varrho \leq C'\left(\nu_p \vee \nu_p^{\frac{1}{2}}\right)\left(\frac{K_n^{6-\frac{4}{p}}}{n^{\frac{2p-4}{2p-1}}}\right)^{\frac{1}{4}} + 18\frac{\nu_p}{\underline{\sigma}_{\mathcal{F},P}} \varepsilon n^{\frac{1}{2}+\frac{3}{4p-2}} K_n^{\frac{1}{2}-\frac{1}{p}} + \frac{3\sqrt{K_n}}{\underline{\sigma}_{\mathcal{F},P}} \Xi\left(5\nu_p K_n^{\frac{1}{2}-\frac{1}{p}} n^{\frac{9}{16p-8}-\frac{3}{8}}\right)$$

for a constant C' depending only on $\nu_2(\mathcal{F})$ and on $\underline{\sigma}_{\mathcal{F},P}$.

Notice that for a given δ we can use an entropy integral bound and the fact that S is contained in a euclidean ball to obtain, for some absolute constant $C'' > 0$,

$$\Xi(\delta) \leq C'' \delta \sqrt{d \ln \frac{\operatorname{diam}(S)}{\delta}}.$$

Thus, if n is high enough to satisfy $\varrho \leq \alpha$ we match the optimal bound given by Theorem 6.1 up to constant factors.

Chapter 7

Conclusions

We now discuss a few research questions that this work inspired. Most of this research questions focus on the extension of our proof techniques to other problems.

> ***Research question 1.*** *When the data is contaminated, the percentage ε of contaminated sample points is typically unknown. Under which assumptions we can estimate the mean even when ε is unknown? What about non-adversarial contamination models?*

In terms of informational-theoretical bounds, this question is related to the discussion on the dependence on the confidence level made in [Devroye et al., 2016]. In [Devroye et al., 2016] is proved that previous knowledge of the confidence level α is necessary to construct a mean estimator E that, as asked in Question 1, behaves as if data were Gaussian irrespective of its distribution. This is reflected in the choices of $k \approx \ln \frac{1}{\alpha}$ necessary for the trimmed mean in the problem of uniform mean estimation and regression (Theorems 3.1 and 4.4). An analogous question can be made for the dependence on the contamination level ε and the answer might come from similar techniques.

The Huber contamination model might also be considered. Experiments using cross-validation in §4.4.1, §4.4.2 and also in [Lecué and Lerasle, 2020] to select the contamination level for robust regression in artificial data contaminated via Huber's model have shown to perform well. An enlightening result may come from studying the cross-validation parameter selection in terms of the distributional distance between the clean sample distribution and the distribution used for contamination.

Research question 2. *Can we apply the trimming and truncation proof techniques described in §2.2 to other problems and estimators?*

The proofs in Chapters 3, 4 and 5 are similar, as the same basic arguments are used to translate the study of the trimmed mean to the study of a truncated empirical process related to it. I'm currently working with Zoraida Rico to apply the same techniques to the problem of robust null space estimation using an estimator similar to the least trimmed squares (a version of the trimmed mean where only the k largest entries are removed before taking the average).

Research question 3. *Can we derive bootstrap approximation bounds in the infinite-dimensional setting? What about entropy-free and weak-variance-free bounds?*

The arguments used to proof the high-dimensional bootstrap approximation results in §5.2, together with the δ-net approximation argument used to proof the Gaussian approximation result for the infinite dimensional case presented in §5.3, might be useful to obtain infinite-dimensional bootstrap approximation results.

Moreover, in a recent paper [Giessing, 2023], entropy-free and weak-variance-free bounds are obtained for the Gaussian and bootstrap approximations of the empirical mean, assuming the existence of an envelope $F \in L^3(P)$. Our proof techniques might also be capable of translating these results to the trimmed mean while removing the requirement of an envelope function.

Appendix A

Auxiliary results

A.1 Inequalities for empirical processes

We state here some classical results used on our proofs. Recall the definitions from §2.1.4 and from §5.3.

Lemma A.1 (Expectation bound for VC-type classes, Corollary 5.1 of [Chernozhukov et al., 2014b]). *Let $\mathcal{G}$ be a VC-type class with a square-integrable envelope G and constants $A \geq e$, $v \geq 1$. Then,*

$$\mathsf{Emp}_n(\mathcal{G}, P) \leq C \left\{ \sqrt{\frac{v\nu_2^2(\mathcal{G})}{n} \ln \frac{A\|G\|_{L^2(P)}}{\nu_2(\mathcal{G})}} + \frac{v\|M\|_{L^2(P)}}{n} \ln \frac{A\|G\|_{L^2(P)}}{\nu_2(\mathcal{G})} \right\}$$

for some absolute constant C and where

$$M = \max_{i \in [n]} G(X_i).$$

Lemma A.2 (Talagrand's concentration inequality for VC-type classes, as in Theorem B.1 of [Chernozhukov et al., 2014a]). *Let $\mathcal{G}$ be a VC-type class with a constant envelope function b and constants $A \geq e$, $v \geq 1$. Consider also that σ^2 satisfies $\nu_2^2(\mathcal{G}) \leq \sigma^2 \leq b^2$. If $b^2 v \ln \frac{Ab}{\sigma} \leq n\sigma^2$, then for all $t \leq \frac{n\sigma^2}{b^2}$,*

$$\mathbb{P}\left[\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - Pg \right| \geq C \sqrt{\frac{\sigma^2}{n} \left(t \vee \left(v \ln \frac{Ab}{\sigma} \right) \right)} \right] \leq e^{-t},$$

where $C > 0$ is an absolute constant.

A.2 Tail bounds for the maxima of random variables

The following theorem allow us to control the tails of the maxima of random variables when the set $\mathcal{G}$ is finite, this result will be used to proof our finite-dimensional counting lemmata.

Lemma A.3 (Counting lemma for boolean random variables)**.** *Let $X_1, \dots, X_n$ be independent random vectors taking values in $\{0,1\}^d$. Then,*

$$\mathbb{P}\left[\max_{j\in[d]} \frac{1}{n}\sum_{i=1}^{n} X_{i,j} \geq 3\frac{\ln(1+d)}{n} + 7\rho\right] \leq 2e^{-n\rho},$$

where $\rho = \max_{j\in[d]} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_{i,j}\right]$.

Proof Let $m \geq 2$ and $p_{i,j} = \mathbb{E}\left[X_{i,j}\right]$, then

$$\mathbb{E}\left[|X_{i,j} - p_{i,j}|^m\right] = p_{i,j}(1-p_{i,j})\left[(1-p_{i,j})^{m-1} + p_{i,j}^{m-1}\right] \leq p_{i,j} \leq \frac{m!}{2}\left(\frac{2}{3}\right)^{m-2} p_{i,j}.$$

It follows directly from Lemma 4 of [van de Geer and Lederer, 2013] taking $\tau = \sqrt{\frac{6\sigma^2}{n}}$, $L = \sqrt{\frac{6K^2}{\sigma^2 n}}$, $K = \frac{2}{3}$, $\sigma^2 = \rho$ that for every $\theta > 0$

$$\mathbb{P}\left[\max_{j\in[d]} \left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \mathbb{E}\left[X_{i,j}\right]\right| \geq \sqrt{6\rho\frac{\ln(1+d)}{n}} + 2\frac{\ln(1+d)}{n} + \sqrt{6\rho\frac{\theta}{n}} + \frac{2\theta}{n}\right] \leq 2e^{-\theta}.$$

Using the AM-GM inequality and taking $\theta = n\rho$ we can simplify it to

$$\mathbb{P}\left[\max_{j\in[d]} \left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \mathbb{E}\left[X_{i,j}\right]\right| \geq 3\frac{\ln(1+d)}{n} + 6\rho\right] \leq 2e^{-n\rho}.$$

The desired result follows removing the centering at a cost of ρ. ■

A.3 Gaussian comparison and anti-concentration inequalities

In our proofs it is sometimes required to compare two Gaussian's with different covariances. For that end we make use of the Gaussian comparison inequality.

Lemma A.4 (Proposition 2.1 of [Chernozhuokov et al., 2022]). *Let P, P' be distributions over $\mathbf{X}$ and $\mathcal{G}, \mathcal{G}'$ be classes of P and P' centered functions, respectively. Let $d := |\mathcal{G}| < \infty$ and $\pi : \mathcal{G} \to \mathcal{G}'$. Assume $\underline{\sigma}_{\mathcal{G},P} > 0$, then*

$$\sup_{\lambda \in \mathbb{R}} \left| \mathbb{P}\left[Z(\mathcal{G}, P) \leq \lambda \right], \mathbb{P}\left[Z(\mathcal{G}', P') \leq \lambda \right] \right| \leq C(\ln d) \sqrt{\Delta_\pi(\Sigma_{\mathcal{G},P} - \Sigma_{\mathcal{G}',P'})}$$

for some constant C depending only on $\underline{\sigma}_{\mathcal{G},P}$.

To prove Gaussian approximation results, in both the high-dimensional and the infinite-dimensional cases anti-concentration inequalities are required to control errors. The following lemma, which is due to [Nazarov, 2003] and has a self-contained proof in [Chernozhukov et al., 2017].

Lemma A.5 (Nazarov's inequality). *Let $\mathcal{G}$ be a family of P-centered functions with $|\mathcal{G}| = d$. If $\underline{\sigma}_{\mathcal{G},P} > 0$, then for all $\delta > 0$,*

$$\sup_{\lambda \in \mathbb{R}} \mathbb{P}\left[\lambda \leq Z(\mathcal{G}) \leq \lambda + \delta \right] \leq \frac{\delta}{\underline{\sigma}_{\mathcal{G},P}} \left(2 + \sqrt{2 \ln d} \right).$$

Bibliography

[Abdalla and Zhivotovskiy, 2022] Abdalla, P. and Zhivotovskiy, N. (2022). Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails.

[Alon and Spencer, 2016] Alon, N. and Spencer, J. (2016). *The Probabilistic Method*. Wiley Series in Discrete Mathematics and Optimization. Wiley.

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

[Birgé and Massart, 2001] Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3:203–268.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Cambridge University Press.

[Bousquet, 2002] Bousquet, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500.

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

[Bubeck et al., 2013] Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Trans. Inf. Theory*, 59(11):7711–7717.

[Catoni, 2012] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4).

[Chernozhukov et al., 2013] Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786.

[Chernozhukov et al., 2014a] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818.

[Chernozhukov et al., 2014b] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014b). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564 – 1597.

[Chernozhukov et al., 2016] Chernozhukov, V., Chetverikov, D., and Kato, K. (2016). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings. *Stochastic Processes and their Applications*, 126(12):3632–3651.

[Chernozhukov et al., 2017] Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Detailed proof of nazarov's inequality. *arXiv preprint arXiv:1711.10696*.

[Chernozhukov et al., 2023a] Chernozhukov, V., Chetverikov, D., Kato, K., and Koike, Y. (2023a). High-dimensional data bootstrap. *Annual Review of Statistics and Its Application*, 10:427–449.

[Chernozhukov et al., 2023b] Chernozhukov, V., Chetverikov, D., and Koike, Y. (2023b). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *The Annals of Applied Probability*, 33(3):2374–2425.

[Chernozhuokov et al., 2022] Chernozhuokov, V., Chetverikov, D., Kato, K., and Koike, Y. (2022). Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5):2562–2586.

[Depersin and Lecué, 2022] Depersin, J. and Lecué, G. (2022). Robust sub-gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1).

[Depersin and Lecué, 2021] Depersin, J. and Lecué, G. (2021). Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms.

[Devroye et al., 2016] Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6).

[Diakonikolas et al., 2019a] Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019a). Robust estimators in high dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864.

[Diakonikolas et al., 2019b] Diakonikolas, I., Kamath, G., Kane, D., Li, J., Steinhardt, J., and

Stewart, A. (2019b). Sever: A robust meta-algorithm for stochastic optimization. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1596–1606. PMLR.

[Diakonikolas et al., 2022] Diakonikolas, I., Kane, D. M., Pensia, A., and Pittas, T. (2022). Streaming algorithms for high-dimensional robust statistics. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S., editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5061–5117. PMLR.

[Dong et al., 2019] Dong, Y., Hopkins, S., and Li, J. (2019). Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Giessing, 2023] Giessing, A. (2023). Gaussian and bootstrap approximations for suprema of empirical processes. *arXiv preprint arXiv:2309.01307*.

[Hall, 1981] Hall, P. (1981). Large sample property of Jaeckel's adaptive trimmed mean. *Annals of the Institute of Statistical Mathematics*, 33(A):449–462.

[Hopkins, 2020] Hopkins, S. B. (2020). Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193 – 1213.

[Huber, 1965] Huber, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758.

[Huber, 1972] Huber, P. J. (1972). The 1972 Wald Lecture Robust Statistics: A Review. *The Annals of Mathematical Statistics*, 43(4):1041 – 1067.

[Huber and Ronchetti, 1981] Huber, P. J. and Ronchetti, E. M. (1981). Robust statistics john wiley & sons. *New York*, 1(1).

[Jaeckel, 1971] Jaeckel, L. A. (1971). Some Flexible Estimates of Location. *The Annals of Mathematical Statistics*, 42(5):1540 – 1552.

[Jana Jurecková, 1994] Jana Jurecková, Roger Koenker, A. H. W. (1994). Adaptive choice of trimming proportions. *Annals of the Institute of Statistical Mathematics*, 46(4):737–755.

[Joly et al., 2017] Joly, E., Lugosi, G., and Oliveira, R. I. (2017). On the estimation of the

mean of a random vector. *Electronic Journal of Statistics*, 11(1).

[Kock and Preinerstorfer, 2023] Kock, A. B. and Preinerstorfer, D. (2023). Moment-dependent phase transitions in high-dimensional gaussian approximations. *arXiv preprint arXiv:2310.12863*.

[Koike, 2021] Koike, Y. (2021). Notes on the dimension dependence in high-dimensional central limit theorems for hyperrectangles. *Japanese Journal of Statistics and Data Science*, 4:257–297.

[Kuchibhotla and Rinaldo, 2020] Kuchibhotla, A. K. and Rinaldo, A. (2020). High-dimensional clt for sums of non-degenerate random vectors: $n^{-\frac{1}{2}}$-rate. *arXiv preprint arXiv:2009.13673*.

[Kuratowski and Ryll-Nardzewski, 1965] Kuratowski, K. and Ryll-Nardzewski, C. (1965). A general theorem on selectors. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys*, 13(6):397–403.

[Lecué and Lerasle, 2019] Lecué, G. and Lerasle, M. (2019). Learning from mom's principles: Le cam's approach. *Stochastic Processes and their applications*, 129(11):4385–4410.

[Lecué and Lerasle, 2020] Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2).

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

[Lee and Valiant, 2022] Lee, J. C. and Valiant, P. (2022). Optimal sub-gaussian mean estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683. IEEE.

[Lopes et al., 2020] Lopes, M. E., Lin, Z., and Müller, H.-G. (2020). Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional and multinomial data. *The Annals of Statistics*, 48(2):1214.

[Lugosi and Mendelson, 2019a] Lugosi, G. and Mendelson, S. (2019a). Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.*, 19(5):1145–1190.

[Lugosi and Mendelson, 2019b] Lugosi, G. and Mendelson, S. (2019b). Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields*, 175(3-4):957–973.

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

[Lugosi and Mendelson, 2019d] Lugosi, G. and Mendelson, S. (2019d). Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2).

[Lugosi and Mendelson, 2021] Lugosi, G. and Mendelson, S. (2021). Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1).

[Massart, 2000] Massart, P. (2000). Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303.

[Mendelson, 2015] Mendelson, S. (2015). Learning without concentration. *Journal of the ACM (JACM)*, 62(3):1–25.

[Mendelson and Zhivotovskiy, 2020] Mendelson, S. and Zhivotovskiy, N. (2020). Robust covariance estimation under $L_4 - L_2$ norm equivalence. *The Annals of Statistics*, 48(3):1648 – 1664.

[Minsker, 2015] Minsker, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4).

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

[Minsker, 2018b] Minsker, S. (2018b). Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*.

[Mourtada et al., 2021] Mourtada, J., Vaskevicius, T., and Zhivotovskiy, N. (2021). Distribution-free robust linear regression. *CoRR*, abs/2102.12919.

[Nazarov, 2003] Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a gaussian measure. In *Geometric Aspects of Functional Analysis: Israel Seminar 2001-2002*, pages 169–187. Springer.

[Oliveira and Rico, 2022] Oliveira, R. I. and Rico, Z. F. (2022). Improved covariance estimation: optimal robustness and sub-gaussian guarantees under heavy tails.

[Rico, 2022] Rico, Z. F. (2022). *Optimal statistical estimation: sub-Gaussian properties, heavy-tailed data, and robustness*. PhD thesis, Instituto de Matemática Pura e Aplicada (IMPA).

[Sriperumbudur, 2016] Sriperumbudur, B. (2016). On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3).

[Sriperumbudur et al., 2012] Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(none).

[Stigler, 2016] Stigler, S. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard University Press.

[Stigler, 1973] Stigler, S. M. (1973). The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, 1(3):472–477.

[Stigler, 2010] Stigler, S. M. (2010). The changing history of robustness. *The American Statistician*, 64(4):277–281.

[Talagrand, 1996] Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563.

[Talagrand, 2014] Talagrand, M. (2014). *Upper and lower bounds for stochastic processes*. Springer.

[van de Geer and Lederer, 2013] van de Geer, S. and Lederer, J. (2013). The bernstein–orlicz norm and deviation inequalities. *Probability theory and related fields*, 157(1-2):225–250.

[van der Vaart et al., 1996] van der Vaart, A., van der Vaart, A., and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

[Zhang, 2023] Zhang, T. (2023). *Mathematical analysis of machine learning algorithms*. Cambridge University Press.