Distribution-Free Robust Linear Regression

Jaouad Mourtada* Tomas Vaškevičius† Nikita Zhivotovskiy‡

October 25, 2021

Abstract

We study random design linear regression with no assumptions on the distribution of the covariates and with a heavy-tailed response variable. In this distribution-free regression setting, we show that boundedness of the conditional second moment of the response given the covariates is a necessary and sufficient condition for achieving nontrivial guarantees. As a starting point, we prove an optimal version of the classical in-expectation bound for the truncated least squares estimator due to Györfi, Kohler, Krzyżak, and Walk. However, we show that this procedure fails with constant probability for some distributions despite its optimal in-expectation performance. Then, combining the ideas of truncated least squares, median-of-means procedures, and aggregation theory, we construct a non-linear estimator achieving excess risk of order d/n with an optimal sub-exponential tail. While existing approaches to linear regression for heavy-tailed distributions focus on proper estimators that return linear functions, we highlight that the improperness of our procedure is necessary for attaining nontrivial guarantees in the distribution-free setting.

MSC2020 Subject Classifications: 62J05; 62G35; 68Q32.

Keywords: Least squares, random design linear regression, robust estimation, improper learning, median-of-means tournaments.

1 Introduction

In the random design regression problem, one has access to n input-output pairs $(X_i, Y_i) \in \mathbf{R}^d \times \mathbf{R}$ sampled i.i.d. from some unknown distribution P. We call any function $g : \mathbf{R}^d \to \mathbf{R}$ a *predictor* and measure its quality via the expected squared error $R(g) = \mathbf{E}(g(X) - Y)^2$, also called *risk*. Based on the sample $S_n = (X_i, Y_i)_{i=1}^n$, we aim to construct a *predictor* $\widehat{g}$ whose risk $R(\widehat{g})$ is small. Since the risk is relative to the problem difficulty, it is customary to compare it with the best possible risk achievable via some reference class of functions; in this work, we mainly focus on the class of all linear functions $\mathcal{F}_{\mathrm{lin}} = \{\langle w, \cdot\rangle : w \in \mathbf{R}^d\}$. We therefore consider the *excess risk* of the estimator $\widehat{g}$, defined by

$$\mathcal{E}(\widehat{g}) = R(\widehat{g}) - \inf_{g \in \mathcal{F}_{\mathrm{lin}}} R(g). \tag{1}$$

One can assume without loss of generality that the infimum above is attained by some linear function $\langle w^*, \cdot\rangle$, where $w^* \in \mathbf{R}^d$. Note that, since $\widehat{g}$ depends on the random sample S_n, the excess risk (1) is also random. In this paper, we study non-asymptotic bounds on $\mathcal{E}(\widehat{g})$, both in expectation and with high-probability, for suitable choices of estimators $\widehat{g}$.

*CREST, ENSAE, Institut Polytechnique de Paris, France, jaouad.mourtada@ensae.fr

†Department of Statistics, University of Oxford, United Kingdom, tomas.vaskevicius@stats.ox.ac.uk

‡Department of Mathematics, ETH Zürich, Switzerland, nikita.zhivotovskii@math.ethz.ch

Arguably the most natural and commonly studied procedure is the linear least squares estimator, which selects a linear function $\widehat{g} \in \mathcal{F}_{\text{lin}}$ minimizing the (possibly regularized) *empirical risk* $\widehat{R}(g) = \frac{1}{n}\sum_{i=1}^{n}(g(X_i) - Y_i)^2$. Estimators based on empirical risk minimization (ERM) are known to achieve optimal d/n excess risk in expectation under well-behaved covariates X and assuming that the noise random variable $\xi = Y - \langle w^*, X\rangle$ is not correlated with X (see, for example, [9, 11, 60]). The work of Oliveira [63] highlights that the usual sub-Gaussian assumption on the distribution of X can be significantly relaxed in the context of linear regression. For example, an L_4–L_2 norm equivalence of the form

$$\big(\mathbf{E}\,\langle X, w\rangle^4\big)^{1/4} \leqslant \kappa\big(\mathbf{E}\,\langle X, w\rangle^2\big)^{1/2}, \quad \text{for all} \quad w \in \mathbf{R}^d, \tag{2}$$

for some constant $\kappa > 0$ is sufficient to achieve the d/n excess risk rate under additional assumptions on the noise. Indeed, [63] shows that under (2), we have a high-probability control over the lower tail of the sample covariance matrix, used in the analysis of linear least squares. This moment equivalence assumption and its variations have become standard tools in the recent literature on robust linear regression (see, for example, [43, 34, 14, 38, 47, 17, 65]). However, as several authors have recently pointed out, the kurtosis constant κ satisfying the inequality (2) may depend on the dimension d, leading to suboptimal bounds [14, 63, 43]. In particular, Saumard [68] shows that the slightly weaker small-ball condition fails to hold (with a dimension-free constant) for dictionaries consisting of many classical function bases, such as histograms and wavelets, leading to bounds with a suboptimal dependence on the dimension d. In fact, in some cases this behavior is inherent to empirical risk minimization, which has recently been shown by the second and the third authors of this paper [76] to incur suboptimal excess risk even in a favorable setup where both X and Y are almost surely bounded. This naturally brings the question of whether distributional assumptions on X such as the condition (2) can be relaxed, and if so, what a corresponding minimal assumption on the distribution of the response variable would be. It is a priori unclear whether non-trivial guarantees are at all possible without imposing any assumption on X.

To better contextualize our aims, we turn to a recent line of work initiated by Catoni [13], concerned with the design and analysis of statistical estimators robust to heavy-tailed distributions. The ERM strategy is known to fail in this setting due to its sensitivity to the large fluctuations and atypical samples arising from heavy-tailed distributions. Thus, different techniques and procedures are required to handle such distributions. We call the excess risk $\mathcal{E}(\widehat{g})$ the accuracy of an estimator $\widehat{g}$; the confidence of $\widehat{g}$ for an error rate of ε is equal to $\mathbf{P}(\mathcal{E}(\widehat{g}) \leqslant \varepsilon)$. Robust statistical learning aims to design procedures with optimal accuracy/confidence trade-off under minimal distributional assumptions. In the context of linear regression, the optimal trade-off is usually achieved via the bounds on $\mathcal{E}(\widehat{g})$ of order $(d+\log(1/\delta))/n$ that hold with probability at least $1-\delta$; in particular, such bounds match the performance of ERM for sub-Gaussian distributions. Using either PAC-Bayesian truncations [5, 14] or the median-of-means tournaments [47], it has been shown that the optimal accuracy/confidence trade-off can be achieved under the L_4–L_2 moment equivalence assumption (2) together with some additional assumptions on the noise variable $\xi = Y - \langle w^*, X\rangle$. We remark that existing procedures for heavy-tailed regression select a function within the class $\mathcal{F}_{\text{lin}}$. However, as we shall shortly explain, any such procedure fails in our distribution-free setting. We can now formulate the question studied in this paper.

> Is it possible to predict as well as the best linear predictor in $\mathcal{F}_{\text{lin}}$ without any assumption on the distribution of the covariates X, while maintaining the optimal accuracy/confidence trade-off? If so, what is the minimal assumption on the response variable Y allowing this?

Independently of the literature on robustness to heavy-tails, two existing results provide non-asymptotic guarantees without assumptions on X, albeit only in expectation. Of course, once all the assumptions on X are dropped, the conditional distribution of Y given X, consisting of a probability kernel $(P_{Y|X=x})_{x\in\mathbf{R}^d}$, needs to be restricted. We now state the only assumption considered in this work; it is satisfied when Y is bounded, but also allows to consider heavy-tailed distributions.

Assumption 1. The conditional distribution of Y given X satisfies, for some $m > 0$,

$$\sup_{x\in\mathbf{R}^d} \mathbf{E}[Y^2|X = x] \leqslant m^2.$$

The first result not involving any explicit restrictions on the distribution of X is a classical bound for the truncated linear least squares estimator $\widehat{g}_m$ due to Györfi, Kohler, Krzyżak, and Walk [29, Theorem 11.3] (we defer the exact definition to Section 3). Under Assumption 1, their result states that

$$\mathbf{E}\, R(\widehat{g}_m) - \inf_{g\in\mathcal{F}_{\mathrm{lin}}} R(g) \leqslant c\,\frac{m^2 d(\log n + 1)}{n} + 7\Big(\inf_{f\in\mathcal{F}_{\mathrm{lin}}} R(f) - R(f_{\mathrm{reg}})\Big). \tag{3}$$

Here the expectation is taken with respect to the random sample S_n, $c > 0$ is an absolute constant and f_{reg} is the regression function given by $f_{\mathrm{reg}}(x) = \mathbf{E}[Y|X = x]$. The bound (3) is a standard benchmark for several communities. Applications of this result are known in mathematical finance [83], optimal control [8] and variance reduction [27, 28]; there are known improvements of this result under different assumptions [19, 20].

The second bound does not depend on the distribution of X and is due to Forster and Warmuth [25]; this estimator originates in the online learning literature and is obtained via a modification of the renowned non-linear Vovk-Azoury-Warmuth forecaster [78, 6]. The Forster-Warmuth estimator, denoted by $\widehat{g}_{\mathrm{FW}}$, satisfies the following expected excess risk bound

$$\mathbf{E}R(\widehat{g}_{\mathrm{FW}}) - \inf_{g\in\mathcal{F}_{\mathrm{lin}}} R(g) \leqslant \frac{2\|Y\|_{L_\infty}^2 d}{n}. \tag{4}$$

Of course, the assumption $\|Y\|_{L_\infty} \leqslant m$ is stronger than Assumption 1. However, an inspection of the proof in [25] shows that Assumption 1 suffices to obtain the above in-expectation performance of this algorithm with $\|Y\|_{L_\infty}^2$ replaced by m^2.

We are now ready to present informal statements of our main findings. In our first result, we prove that the term $7\,(\inf_{f\in\mathcal{F}_{\mathrm{lin}}} R(f) - R(f_{\mathrm{reg}}))$ as well as the excess $\log n$ factor appearing in the bound (3) for the truncated linear least squares estimator can be removed.

Theorem A (Informal)**.** *Suppose that Assumption 1 holds and let $\widehat{g}_m$ denote the truncated least squares estimator of Györfi, Kohler, Krzyżak, and Walk. Then, we have*

$$\mathbf{E}R(\widehat{g}_m) - \inf_{g\in\mathcal{F}_{\mathrm{lin}}} R(g) \leqslant \frac{8m^2 d}{n}.$$

Moreover, Assumption 1 ensures the same guarantee for the Forster-Warmuth estimator $\widehat{g}_{\mathrm{FW}}$.

One may notice that even though the bound of Theorem A scales as d/n, the usual dependence on the variance of the noise variable as is in, for example, [9] is replaced by the dependence on m^2. It can be shown (see Proposition 2) that if only Assumption 1 holds, then the dependence on m^2 is unavoidable in general even if the problem is noise-free so that the variance of the noise is equal to zero. Moreover, if we only impose Assumption 1, then any statistical estimator that

selects predictors from $\mathcal{F}_{\text{lin}}$ (such an estimator is called *proper*) is bound to fail. This fact can be established using the recent result of Shamir [69, Theorem 3], and it remains true even when $d = 1$ and the response variable Y is bounded almost surely. This observation separates our setup from the existing literature where only proper estimators are studied for convex classes such as $\mathcal{F}_{\text{lin}}$ even in the heavy-tailed scenarios (see, for example, [14, 47, 53, 54]).

The bounds of Theorem A guarantee that the *expected* excess risk is at most of order d/n under Assumption 1. However, it is also desirable to obtain *high-probability* upper bounds on the excess risk, with logarithmic dependence on the confidence level δ. It is not unreasonable to expect that the in-expectation guarantees of either $\widehat{g}_m$ or $\widehat{g}_{\text{FW}}$ transfer to analogous high-probability bounds, at least whenever Y is bounded almost surely. Our second result shows that this is, unfortunately, not the case. Indeed, both algorithms fail to achieve high probability upper bounds in a strong sense: they both fail with constant probability. This does not contradict the previous in-expectation bounds, since neither $\widehat{g}_m$ nor $\widehat{g}_{\text{FW}}$ belong to the linear class $\mathcal{F}_{\text{lin}}$, so the random variable $R(\widehat{g}) - \inf_{g \in \mathcal{F}_{\text{lin}}} R(g)$ can take negative values. Consequently, Markov's inequality cannot be applied to obtain deviation bounds of $m^2 d/(\delta n)$ with probability $1 - \delta$.

Theorem B (Informal)**.** *Let $\widehat{g}$ denote either $\widehat{g}_m$ or $\widehat{g}_{\text{FW}}$. There exist universal constants $p \in (0,1), c > 0$ such that the following holds. For any $d \geqslant 1$ and $m > 0$, there exists a distribution P satisfying $\|Y\|_{L_\infty(P)} \leqslant m$ such that, with probability at least p,*

$$R(\widehat{g}) - \inf_{g \in \mathcal{F}_{\text{lin}}} R(g) \geqslant c\, m^2.$$

Theorem B raises the question of whether achieving high-probability guarantees in our distribution-free setting is at all possible. Indeed, all known high-probability guarantees on linear aggregation problems impose some restrictions on X. We show that there is, in fact, a procedure that achieves an optimal excess risk guarantee (up to a logarithmic factor) with a sub-exponential tail. The following theorem is the main positive result of this paper.

Theorem C (Informal)**.** *Suppose that Assumption 1 holds. There exists an absolute constant $c > 0$ such that the following holds. For any confidence level $\delta \in (0,1)$, there exists an improper estimator $\widehat{g}$ (depending on δ and m) such that*

$$\mathbf{P}\left(R(\widehat{g}) - \inf_{g \in \mathcal{F}_{\text{lin}}} R(g) \leqslant c\, \frac{m^2 (d \log(n/d) + \log(1/\delta))}{n} \right) \geqslant 1 - \delta.$$

Theorem C demonstrates that robust learning of linear classes is possible with no restriction on the distribution of X, and under weak tail assumptions on the conditional distribution of the response variable Y given covariates X. Moreover, we show in Section 6 that Assumption 1 is necessary to obtain any non-trivial guarantee without assumptions on X. The estimator of Theorem C naturally leverages the ideas of the analysis of truncated linear functions [29, Chapter 11], skeleton estimators [21, Section 28.3], [66], the deviation optimal model selection aggregation procedures [2, 42, 53], min-max estimators [5, 41], and the median-of-means tournaments [47]. An extended discussion is deferred to Section 5.

1.1 Summary of contributions and structure of the paper

- In Section 2, we discuss known results on distribution-free learning of linear classes.
- In Section 3, we show that the classical bound of Györfi, Kohler, Krzyżak, and Walk [29, Theorem 11.3] for the truncated linear least squares estimator can be improved to achieve the optimal $m^2 d/n$ bound in expectation.

- In Section 4, we establish that the truncated least squares and Forster-Warmuth estimators are both deviation-suboptimal. In particular, we construct a distribution with almost surely bounded response variable Y, under which both estimators incur an excess risk of order m^2 with constant probability.
- Section 5 is split into three parts. In Section 5.1, we consider a simplified setting with a known covariance structure. Combining Tsybakov's projection estimator [70] with the robust mean estimator of Lugosi and Mendelson [45], we provide an estimator attaining the optimal rate d/n with the optimal dependence on the confidence parameter. In Section 5.2, we drop the simplifying assumption of known covariance structure and present our main positive result – a distribution-free deviation-optimal estimator robust to heavy-tailed responses. In Section 5.3, we discuss possible extensions of this result. In particular, we show that an adaptation of our linear regression procedure yields an estimator with deviation-optimal rates for heavy-tailed model selection aggregation under Assumption 1.
- Section 6 is devoted to establishing the necessity of Assumption 1. We show, in particular, that if $\mathbf{E}[Y^2|X]$ is unbounded, no estimator can achieve non-trivial excess risk guarantees. In addition, we establish that the dependence on m^2 in our upper bounds is unavoidable.
- Section 7 contains deferred proofs of lemmas appearing in the previous sections.

1.2 Related work

Analysis of least squares estimators. The most standard approach to regression problems is the least squares principle, where one selects the predictor achieving the best fit to data within some predefined class of functions. A large body of work is devoted to analyzing and obtaining guarantees on its performance, in its most classical form, relying on the fact that the empirical risk provides a good approximation of its population counterpart. This is typically established when the underlying distribution is sufficiently well-behaved (for instance, bounded or light-tailed), using tools from empirical process theory. For this point of view to statistical learning, we refer to the standard textbooks [73, 50, 40, 79]. It should be noted that statistical analysis of linear regression has also been treated via a complementary approach of stochastic approximation; see, for instance, the works [80, 30, 24] and references therein.

A recent line of research has established that empirical minimization can perform well under significantly weaker assumptions. Our starting point is the work of Oliveira [63], where in the context of linear regression the usual sub-Gaussian assumption on X is replaced by a significantly weaker L_4–L_2 moment equivalence assumption (2). In particular, such an assumption does not even require the existence of any moments of X higher than the fourth. Variations of this assumption have became the standard tool in the recent literature on linear regression [43, 34, 14, 47, 38, 60, 17, 65]. The seminal work of Mendelson [52] introduced a more general condition, called the *small-ball* assumption. In most of the aforementioned papers, the analysis is performed for empirical risk minimization, which usually does not lead to the optimal accuracy/confidence trade-off. The papers [4, 5] provide the optimal confidence for ERM, albeit under stronger moment equivalence assumptions than that of (2). The L_4–L_2 moment equivalence is also important in the robust covariance estimation problem [14, 57, 64].

It has been recently observed that the absolute constants involved in the moment equivalence and the small-ball assumptions can behave badly in some cases. First, Saumard [68] shows that the small-ball condition is unsuitable for some important classes leading to suboptimal performance of ERM. Further, the work [14] (see also the discussion in [63] and [43]) discusses that the kurtosis constant κ in the moment assumptions similar to (2) can depend on the dimension and affect the bounds negatively. The recent paper [76] shows this suboptimal behavior in the

context of linear regression, even in a favorable setup where both X and Y are bounded. There is a growing interest in further relaxing these assumptions and refining the underlying methods [68, 15, 55, 54, 18, 60, 56]. In particular, the works [15, 56] replace moment equivalence assumptions by the bounds on the L_p moments for $p \geqslant 4$. This is closer to the setting we are aiming for in this paper.

Robustness to heavy-tailed distributions. In a broad sense, robustness encompasses the study and design of statistical estimation procedures exhibiting certain stability properties under the existence of "outlier" points in the observed sample. For a classical perspective on robustness, originating from the work of Tukey [71] and building on the ideas of contaminated models, influence functions and breakdown points, we refer to the standard books [31, 35, 67].

In contrast to the classical perspective, our work falls within the recent body of work initiated by Catoni [13], where the term robustness is to be understood specifically as robustness to heavy-tailed distributions (rather than, for example, adversarial contamination of the sample). The starting point of this direction is the question of mean estimation, where informally, one aims to construct statistical estimators performing as well as the sample mean does for Gaussian samples, all while making as weak distributional assumptions as possible. Several ways of constructing such estimators (called sub-Gaussian estimators) have been proposed in the literature. The most widespread approach is based on the median-of-means estimators, which appear first independently in [62, 36, 1] and were further developed in the works of [58, 22, 46, 48]. Other techniques include the Catoni's estimator and its extensions [13, 15] or the trimmed means [49]. We refer to the survey [45] for further details and references. For a complementary survey focusing on the computational aspects see [23].

The central ideas behind the robust mean estimation found their applications in many related problems such as regression [34, 10, 18, 47, 41, 18, 59, 54], covariance estimation [14, 57, 64] and clustering [10, 39]. In the context of linear regression, the first works showing the optimal accuracy/confidence trade-off under weak assumptions are attributed to Audibert and Catoni [4, 5] and were further extended in [14, 15]; these papers are based on PAC-Bayesian truncations.

Distribution-free linear regression. Distribution-free non-asymptotic excess risk bounds take their roots in the PAC-learning framework [75, 72], where historically the binary loss is studied the most. Because of its boundedness, excess risk bounds in such setups can be obtained without any assumptions on the distribution of (X, Y). In the context of non-parametric regression with the squared loss, only asymptotic consistency results are possible under truly minimal assumptions on the underlying distribution (see the book [29]). In fact, the standard notions of universal consistency [29, Section 1.6 and Chapter 10] involve only the assumption $\mathbf{E}Y^2 < \infty$ and no assumptions on the distribution of X. The distribution-free nature of this notion is one of our motivations. A notable non-asymptotic result in this direction is [29, Theorem 11.3], where an inexact oracle inequality (3) is proved without any explicit assumptions on the distribution of X.

Another direction originates from the online learning literature (see [16] for background on this topic). For instance, when both X and Y are bounded, the renowned Vovk-Azoury-Warmuth forecaster [78, 6] can be used to provide excess risk bounds of order d/n in our setup even when the aforementioned moment equivalence constants behave badly with respect to the dimension. This observation has been recently explored in [76]. For linear regression, the Forster-Warmuth algorithm [25], which is in turn a modification of the Vovk-Azoury-Warmuth forecaster, leads to the only known exact oracle inequality without imposing any assumptions on X.

1.3 Notation

We now set the notation. We let $P = P_{(X,Y)}$ be the joint distribution on $\mathbf{R}^d \times \mathbf{R}$ (with $d \geqslant 1$) of a random pair (X, Y). The joint distribution P itself can be decomposed into two components, namely the marginal distribution P_X of X (a distribution on $\mathbf{R}^d$), as well as the conditional distribution of Y given X, consisting of a (measurable) probability kernel $(P_{Y|X=x})_{x\in\mathbf{R}^d}$, where for $x \in \mathbf{R}^d$, $P_{Y|X=x}$ is a distribution on $\mathbf{R}$.

For a real random variable Z and $p \geqslant 1$, we denote $\|Z\|_{L_p} = \mathbf{E}[|Z|^p]^{1/p}$, while for a measurable function $f : \mathbf{R}^d \to \mathbf{R}$, we set $\|f\|_{L_p} = \|f\|_{L_p(P_X)} = \|f(X)\|_{L_p}$.

The risk of a measurable function $f : \mathbf{R}^d \to \mathbf{R}$ is by definition $R(f) = \mathbf{E}(f(X) - Y)^2 = \|f(X) - Y\|_{L_2}^2$. It is known that the risk is minimized by the regression function f_{reg} given by $f_{\mathrm{reg}}(x) = \mathbf{E}[Y|X = x] = \int_{\mathbf{R}} y P_{Y|X=x}(\mathrm{d}y)$.

Absolute constants are denoted by $c, c_1, \ldots$ and may change from line to line. For a real square matrix A, let $\mathrm{Tr}(A)$ denote its trace, $\|A\|_{\mathrm{op}}$ its operator norm, A^{T} its transpose and $A^{\dagger}$ its Moore–Penrose inverse. In what follows, $\langle \cdot, \cdot \rangle$ denotes the canonical inner product in $\mathbf{R}^d$ and $\|\cdot\|$ stands for the Euclidean norm. For any two functions (or random variables) f, g the symbol $f \lesssim g$ (or $g \gtrsim f$) means that there is an absolute constant c such that $f \leqslant cg$ on the entire domain. For a pair of symmetric matrices A, B, the symbol $A \preccurlyeq B$ means that $B - A$ is positive semi-definite.

We consider the class $\mathcal{F}_{\mathrm{lin}} = \{\langle w, \cdot \rangle : w \in \mathbf{R}^d\}$ of linear functions. Throughout, our assumptions will imply that $R(0) = \mathbf{E}Y^2$ is finite (regardless of P_X); hence, so is the minimal risk in $\mathcal{F}_{\mathrm{lin}}$, namely $\inf_{f\in\mathcal{F}_{\mathrm{lin}}} R(f)$ is finite. In this case, for $f \in \mathcal{F}_{\mathrm{lin}}$ given by $f(x) = \langle w, x \rangle$ its risk $R(f)$ is finite if and only if $\|\langle w, X\rangle\|_{L_2} < +\infty$, and the set of such $w \in \mathbf{R}^d$ is a subspace of $\mathbf{R}^d$, which coincides with $\mathbf{R}^d$ itself if and only if $\mathbf{E}\|X\|^2 < +\infty$. When the latter condition holds, one can define the covariance of X as $\mathrm{Cov}(X) = \mathbf{E}(X - \mathbf{E}X)(X - \mathbf{E}X)^{\mathsf{T}}$ and the Gram matrix of X as $\Sigma = \mathbf{E}XX^{\mathsf{T}}$; the minimizers f of the risk in $\mathcal{F}_{\mathrm{lin}}$ are then the functions $\langle w, \cdot \rangle$, where w are solutions of the equation $\Sigma w = \mathbf{E}[YX]$. The last quantity is well-defined since $\mathbf{E}|Y|\|X\| \leqslant \|Y\|_{L_2}\mathbf{E}[\|X\|^2]^{1/2}$.

Given the observed sample $S_n = (X_i, Y_i)_{i=1}^n$, the aim is to construct a predictor (usually called an estimator) $\widehat{g}$ whose risk $R(\widehat{g})$ is small. A learning procedure is a measurable function mapping a sample in $(\mathbf{R}^d \times \mathbf{R})^n$ to a measurable function $\mathbf{R}^d \to \mathbf{R}$. In what follows, we avoid measurability issues and use a standard convention that all events appearing in the probabilistic statements are measurable. Given a sample $S_n = (X_i, Y_i)_{i=1}^n$, we usually write $\widehat{g}$ for the function $\widehat{g}(S_n)$. Finally, we remark that since the sample S_n is random, the function $\widehat{g} = \widehat{g}(S_n)$ is also random and so is $R(\widehat{g})$.

2 Distribution-free linear regression: known results

In this section, we set the context for the rest of this work, by reviewing relevant existing results on distribution-free linear prediction, and framing them in our setting (through minor modifications). We remark that the bounds we are about to discuss hold in expectation, whereas we will also be concerned with high-probability guarantees. As will be seen in Section 4, the distinction between the two is not innocuous, as existing procedures achieving distribution-free expected excess risk bounds do not possess matching guarantees in deviation.

Limitations of proper estimators. Recall that in the context of our work, a learning procedure is called *proper* if it always returns an element of the class $\mathcal{F}_{\mathrm{lin}}$ (that is, a linear function); otherwise, it is called *improper* or *non-linear*. The importance of considering improper estimators stems from a fundamental limitation of proper procedures in our distribution-free setting.

Specifically, it follows from the work of Shamir [69, Theorem 3] that for any proper estimator $\widehat{g}_{\text{proper}}$, there exists a distribution of (X, Y) with the response variable Y almost surely bounded by m, for which

$$\mathbf{E}R(\widehat{g}_{\text{proper}}) - \inf_{g \in \mathcal{F}_{\text{lin}}} R(g) \gtrsim m^2. \tag{5}$$

Thus, even when the response is bounded, no proper learning procedure can improve (up to universal constants) over the risk trivially achieved by the zero function, without some restrictions on the distribution of covariates. As discussed in the introduction, this negative result already rules out many procedures introduced and analyzed in the statistical learning and robust estimation literature, including empirical risk minimization and refinements thereof.

Learning with known covariance structure. We now discuss a simplified setting, in which guarantees can be obtained quite directly. Specifically, assume that the *covariance structure* of the distribution P_X, namely, the map $w \mapsto \mathbf{E}\langle w, X\rangle^2$ (which can take infinite values), is known. As noted in Section 1.3, we can restrict our attention to the linear subspace where the above map takes finite values. Thus, we may assume without loss of generality that the covariance matrix $\Sigma = \mathbf{E}XX^\mathsf{T}$ exists. In addition, up to restricting to the orthogonal of the nullspace $\{w \in \mathbf{R}^d : \Sigma w = 0\}$, we may assume in what follows that the covariance matrix Σ is invertible. Hence, the unique minimizer of the risk $R(f)$ in $\mathcal{F}_{\text{lin}}$ is given by $f^* = \langle w^*, \cdot\rangle$, where $w^* = \Sigma^{-1}\mathbf{E}[YX]$. In addition, the excess risk of any linear function $f = \langle w, \cdot\rangle$ is given by the following identity:

$$R(f) - \inf_{g \in \mathcal{F}_{\text{lin}}} R(g) = \|\Sigma^{1/2}(w - w^*)\|^2. \tag{6}$$

The key simplification provided by the knowledge of Σ is that random-design linear regression reduces to multivariate mean estimation. To see this, consider the change of variables $\theta = \Sigma^{1/2}w$ and notice that the excess risk (6) is then equal to $\|\theta - \theta^*\|^2$, where $\theta^* = \mathbf{E}U$ for $U = Y\Sigma^{-1/2}X$. Using Σ, an i.i.d. sample $(X_i, Y_i)_{i=1}^n$ can be turned into an i.i.d. sample $(U_i)_{i=1}^n$, with $U_i = Y_i\Sigma^{-1/2}X_i$ distributed as U. One can thus estimate $\mathbf{E}U$ by the sample mean $\frac{1}{n}\sum_{i=1}^n U_i$. This leads to the *projection estimator* for our original problem, defined as

$$\widehat{g}_{\text{proj}}(x) = \langle \widehat{w}, x\rangle \quad \text{where} \quad \widehat{w} = \Sigma^{-1} \cdot \frac{1}{n}\sum_{i=1}^n Y_i X_i. \tag{7}$$

Under Assumption 1, we have

$$\mathbf{E}UU^\mathsf{T} = \mathbf{E}\big[\mathbf{E}[Y^2|X]\Sigma^{-1/2}XX^\mathsf{T}\Sigma^{-1/2}\big] \preccurlyeq m^2 I_d,$$

and in particular $\operatorname{Tr}(\operatorname{Cov}(U)) \leqslant m^2 d$ and $\|\operatorname{Cov}(U)\|_{\text{op}} \leqslant m^2$. Applying the first inequality to the empirical mean estimator of $\theta^* = \mathbf{E}U$ leads to the following guarantee for the projection estimator, which corresponds up to minor changes in assumptions[1] to the result of Tsybakov [70, Theorem 4]:

$$\mathbf{E}R(\widehat{g}_{\text{proj}}) - \inf_{g \in \mathcal{F}_{\text{lin}}} R(g) \leqslant \frac{m^2 d}{n}.$$

It is worth noting that there is no contradiction between the lower bound (5) and the above upper bound. Indeed, the projection estimator, while proper, relies on the a priori knowledge of Σ, which is unavailable in the typical statistical learning setting. This implies in particular that the knowledge of Σ is sufficient to avoid the previous failure of proper procedures. In this work, the simplified setting with known covariance serves as a benchmark that we aim to match in the general case, where nothing is known a priori about the distribution of X.

[1] Specifically, [70] assumes that the noise $Y - f_{\text{reg}}(X)$ is independent of X, but the same proof applies when replacing this assumption by the conditional moment bound of Assumption 1.

Upper bounds in expectation via non-linear predictors. As mentioned in the introduction, and with the exception of the aforementioned known covariance setting, there are two known results stating non-trivial in-expectation guarantees without restrictions on the distribution of X. These guarantees are achieved, respectively, by the truncated linear least squares estimator $\widehat{g}_m$ and the Forster-Warmuth estimator $\widehat{g}_{\mathrm{FW}}$, which we now define formally.

First, consider the linear least squares estimator $\widehat{g}_{\mathrm{erm}} = \arg\min_{g \in \mathcal{F}_{\mathrm{lin}}} \widehat{R}(g) = \langle \widehat{w}_{\mathrm{erm}}, \cdot \rangle$, where

$$\widehat{w}_{\mathrm{erm}} = \Big(\sum_{i=1}^n X_i X_i^{\mathsf{T}} \Big)^{\dagger} \Big(\sum_{i=1}^n Y_i X_i \Big) = \widehat{\Sigma}_n^{\dagger} \cdot \frac{1}{n} \sum_{i=1}^n Y_i X_i \,,$$

with $\widehat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^{\mathsf{T}}$. Given a threshold $m > 0$, the truncated least squares estimator $\widehat{g}_m$ returns the prediction of the linear function $\langle \widehat{w}_{\mathrm{erm}}, \cdot \rangle$, truncated to $[-m, m]$. That is,

$$\widehat{g}_m(x) = \max(-m, \min(m, \langle \widehat{w}_{\mathrm{erm}}, x \rangle)). \tag{8}$$

We now turn to the Forster-Warmuth estimator. Given the sample $S_n = (X_i, Y_i)_{i=1}^n$, define the *leverage score* of a point $x \in \mathbf{R}^d$ by $h_n(x) = \langle (n\widehat{\Sigma}_n + xx^{\mathsf{T}})^{\dagger} x, x \rangle$. The Forster-Warmuth estimator is then defined by reweighing predictions of $\langle \widehat{w}_{\mathrm{erm}}, \cdot \rangle$ by a function of the statistical leverage of the input point x:

$$\widehat{g}_{\mathrm{FW}}(x) = \big(1 - h_n(x)\big)^2 \cdot \langle \widehat{w}_{\mathrm{erm}}, x \rangle. \tag{9}$$

Recall the guarantees on the risk of these procedures, stated in the introduction. Specifically, under Assumption 1, the truncated least squares estimator satisfies the oracle inequality (3), while the Forster-Warmuth estimator achieves the excess risk bound (4). Note that both procedures are improper, as they introduce non-linearities in the prediction function, either through truncation or through the leverage correction.

As discussed in the recent works [60, 76], the risk of the least squares procedure is large when leverage scores are uneven and correlate with the noise. While this configuration is ruled out under distributional assumptions such as moment equivalences, it can actually occur even under boundedness constraints, leading to poor performance [76]. Both non-linearities partially mitigate the shortcomings of the least squares estimator, by adjusting its predictions at high-leverage points, which are the most unstable and lead to large errors. These corrections allow these procedures to achieve in-expectation bounds, even for unfavorable distributions on which ordinary least squares fail.

3 An improved bound for truncated least squares

As discussed at the end of Section 2, the non-linearities introduced by the truncated least squares and Forster-Warmuth estimators aim to mitigate the instability of ERM predictions at high-leverage points. The more sophisticated Forster-Warmuth procedure (which relies on an explicit leverage correction), however, leads to a better excess risk guarantee. Indeed, the risk guarantee of $\widehat{g}_m$ takes the form of an inexact oracle inequality, suffering from the approximation error term $\inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f) - R(f_{\mathrm{reg}})$. This type of guarantee only ensures that the procedure approaches the performance of the best linear function in the nearly well-specified case, where the true regression function is almost linear. While reasonable in low-dimensional nonparametric estimation [29] (with appropriate linear spaces), such an assumption is generally restrictive in high-dimensional problems and is not satisfied in our setting. Unfortunately, the proof technique employed in [29] can only yield inexact oracle inequalities, and hence, no straightforward modification to their argument can match guarantees of $\widehat{g}_{\mathrm{FW}}$ given by (4).

A natural question remains of whether the gap between the existing in-expectation performance guarantees given by (3) and (4) is intrinsic to the estimators $\widehat{g}_m$ and $\widehat{g}_{\mathrm{FW}}$, or whether it is a byproduct of suboptimal analysis of the performance of the simpler procedure $\widehat{g}_m$. In the theorem below, we show that truncated least squares estimator indeed matches the statistical performance of the Forster-Warmuth algorithm. Our proof is based on a leave-one-out argument akin to the one used to prove the upper bound (4) in [25, Section 3]. We remark that leave-one-out arguments have a long history; see the references [75, Chapter 6] and [33].

Theorem 1. *Suppose that Assumption 1 holds and let $\widehat{g}_m$ denote the truncated least squares estimator (8). Then, we have*

$$\mathbf{E}R(\widehat{g}_m) - \inf_{f\in\mathcal{F}_{\mathrm{lin}}} R(f) \leqslant \frac{8m^2 d}{n+1}.$$

Proof. To simplify the presentation, we introduce additional notation. Let $S_{n+1} = (X_i, Y_i)_{i=1}^{n+1}$ denote an i.i.d. sample of size $n+1$. For any $j \in \{1,\dots,n+1\}$, let $S_{n+1}^{(j)} = (X_i, Y_i)_{i=1, i\neq j}^{n+1}$ be the dataset obtained by removing the j-th sample. On the sample S_{n+1} (respectively $S_{n+1}^{(j)}$), we define the minimal norm empirical risk minimizer $\widetilde{g}$ (respectively $\widetilde{g}^{(j)}$) and its truncated variant $\widetilde{g}_m$ (respectively $\widetilde{g}_m^{(j)}$).

Since S_{n+1} is an i.i.d. sample, for every $j \in \{1,\dots,n+1\}$, $S_{n+1}^{(j)}$ has the same distribution as $S_n = S_{n+1}^{(n+1)}$ (so that $\widetilde{g}_m^{(j)}$ has the same distribution as $\widehat{g}_m = \widetilde{g}_m^{(n+1)}$), and is independent of $Z_j = (X_j, Y_j)$. This implies that the expected excess risk of $\widehat{g}_m$ can be bounded as follows:

$$\begin{aligned}
\mathbf{E}\,\mathcal{E}(\widehat{g}_m) &= \mathbf{E}_{S_{n+1}}\big(\widetilde{g}_m^{(n+1)}(X_{n+1}) - Y_{n+1}\big)^2 - \inf_{g\in\mathcal{F}_{\mathrm{lin}}} \mathbf{E}_{Z_{n+1}}\left(g(X_{n+1}) - Y_{n+1}\right)^2 \\
&= \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\Big(\widetilde{g}_m^{(j)}(X_j) - Y_j\Big)^2\right] - \inf_{g\in\mathcal{F}_{\mathrm{lin}}} \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(g(X_j) - Y_j\right)^2\right] \\
&\leqslant \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\Big(\widetilde{g}_m^{(j)}(X_j) - Y_j\Big)^2 - \left(\widetilde{g}(X_j) - Y_j\right)^2\right], \qquad (10)
\end{aligned}$$

where the last line follows from the definition of $\widetilde{g}$. Now, define the leverage h_j of the point X_j among $X_1,\dots,X_{n+1}$ by

$$h_j = \left\langle \left(\sum_{i=1}^{n+1} X_i X_i^{\mathsf{T}}\right)^{\dagger} X_j, X_j \right\rangle \in [0,1].$$

An explicit computation—postponed to the end of the proof—shows that for every j,

$$\widetilde{g}(X_j) = (1-h_j)\,\widetilde{g}^{(j)}(X_j) + h_j Y_j. \qquad (11)$$

Plugging (11) into the bound (10), we obtain

$$\mathbf{E}\,\mathcal{E}(\widehat{g}_m) \leqslant \mathbf{E}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\Big(\widetilde{g}_m^{(j)}(X_j) - Y_j\Big)^2 - (1-h_j)^2\Big(\widetilde{g}^{(j)}(X_j) - Y_j\Big)^2\right]. \qquad (12)$$

By Assumption 1 and Jensen's inequality we have $\sup_{x\in\mathbf{R}^d}|f_{\mathrm{reg}}(x)| \leqslant m$. It follows that $(\widetilde{g}_m^{(j)}(X_j) - f_{\mathrm{reg}}(X_j))^2 \leqslant (\widetilde{g}^{(j)}(X_j) - f_{\mathrm{reg}}(X_j))^2$, so that

$$\begin{aligned}
&\mathbf{E}\Big[(1-h_j)^2\big(\widetilde{g}^{(j)}(X_j) - Y_j\big)^2 \,\Big|\, S_{n+1}^{(j)}, X_j\Big] \\
&= (1-h_j)^2\Big((\widetilde{g}^{(j)}(X_j) - f_{\mathrm{reg}}(X_j))^2 + \mathbf{E}\big[(f_{\mathrm{reg}}(X_j) - Y_j)^2 | S_{n+1}^{(j)}, X_j\big]\Big) \\
&\geqslant \mathbf{E}\Big[(1-h_j)^2\big(\widetilde{g}_m^{(j)}(X_j) - Y_j\big)^2 \,\Big|\, S_{n+1}^{(j)}, X_j\Big].
\end{aligned}$$

Plugging the above in the upper bound (12), we proceed as follows

$$\begin{aligned}\mathbf{E}\,\mathcal{E}(\widehat{g}_m) &\leqslant \mathbf{E}\bigg[\frac{1}{n+1}\sum_{j=1}^{n+1}\big(\widetilde{g}_m^{(j)}(X_j)-Y_j\big)^2-(1-h_j)^2\big(\widetilde{g}_m^{(j)}(X_j)-Y_j\big)^2\bigg]\\ &\leqslant \mathbf{E}\bigg[\frac{1}{n+1}\sum_{j=1}^{n+1}2h_j\big(\widetilde{g}_m^{(j)}(X_j)-Y_j\big)^2\bigg]\\ &\leqslant 8m^2\mathbf{E}\bigg[\frac{1}{n+1}\sum_{j=1}^{n+1}h_j\bigg]\leqslant 8\frac{m^2d}{n+1},\end{aligned}$$

where the penultimate step follows from Jensen's inequality combined with Assumption 1 and the last step follows from the bound $\sum_{j=1}^{n+1}h_j=\operatorname{Tr}\big[\big(\sum_{i=1}^{n+1}X_iX_i^\mathsf{T}\big)^\dagger\big(\sum_{i=1}^{n+1}X_iX_i^\mathsf{T}\big)\big]\leqslant d$.

We now conclude by showing the identity (11). First, define

$$\widetilde{\Sigma}=\sum_{i=1}^{n+1}X_iX_i^\mathsf{T},\quad \widetilde{\Sigma}^{(j)}=\widetilde{\Sigma}-X_jX_j^\mathsf{T},\quad b=\sum_{i=1}^{n+1}Y_iX_i,\ \text{ and }\ b^{(j)}=b-Y_jX_j,$$

so that

$$\widetilde{g}(X_j)=\langle\widetilde{\Sigma}^\dagger b,X_j\rangle,\quad \widetilde{g}^{(j)}(X_j)=\langle\big(\widetilde{\Sigma}^{(j)}\big)^\dagger b^{(j)},X_j\rangle,\quad\text{and}\quad h_j=\langle\widetilde{\Sigma}^\dagger X_j,X_j\rangle.$$

Note that (11) is an identity, and up to restricting to the linear span of $(X_1,\dots,X_{n+1})$ we may assume that $\widetilde{\Sigma}$ is invertible. In addition, if X_j does not belong to the linear span of $(X_i)_{i=1,i\neq j}^{n+1}$, namely, if $\widetilde{\Sigma}^{(j)}$ is singular, then it can be shown that $h_j=1$ and $\widetilde{g}(X_j)=Y_j$ (since $\widetilde{g}$ minimizes the empirical risk on S_{n+1}, and $g(X_j)$ can be set freely without affecting the other predictions), so that (11) holds. Therefore, we may assume that $\widetilde{\Sigma}^{(j)}$ is invertible. Using the definition and the Sherman-Morrison formula, as $h_j\in[0,1)$, we obtain

$$\begin{aligned}\widetilde{g}^{(j)}(X_j)&=\left\langle\left(\widetilde{\Sigma}^{-1}+\frac{\widetilde{\Sigma}^{-1}X_jX_j^\mathsf{T}\widetilde{\Sigma}^{-1}}{1-h_j}\right)\left(b-Y_jX_j\right),X_j\right\rangle\\ &=\widetilde{g}(X_j)+\frac{h_j}{1-h_j}\widetilde{g}(X_j)-h_jY_j-\frac{h_j^2}{1-h_j}Y_j\\ &=\frac{1}{1-h_j}\widetilde{g}(X_j)-\frac{h_j}{1-h_j}Y_j;\end{aligned}$$

rearranging the last equality yields (11), concluding the proof. □

4 Failure of previous estimators with constant probability

As discussed in Section 2, Assumption 1 suffices to ensure that the Forster-Warmuth estimator [25] achieves an expected excess risk bound of order m^2d/n irrespective of the distribution of X. Our results established in Section 3 demonstrate the same conclusion for the truncated least squares estimator of [29, Theorem 11.3]. In addition to the guarantees in expectation, high-probability or tail bounds are desirable, as they provide a control on the probability of failure of the estimator. The following theorem shows that in fact, none of the two procedures satisfy meaningful high-probability guarantees, in a rather strong sense.

Theorem 2. *Fix the dimension $d = 1$. There exist absolute constants $c > 0$ and $n_0 \geqslant 2$ such that the following holds. For any $n \geqslant n_0$, there is a distribution $P = P(n)$ of (X, Y) with $\|Y\|_{L_\infty} \leqslant m$, such that if $\widehat{g}$ is either the truncated least squares estimator (8) or the Forster-Warmuth estimator (9), computed on an i.i.d. sample S_n, then*

$$\mathbf{P}\Big(R(\widehat{g}) - \inf_{g \in \mathcal{F}_{\mathrm{lin}}} R(g) \geqslant c\, m^2\Big) \geqslant c\,.$$

Note that under Assumption 1, the trivial, identically zero function has risk at most $\mathbf{E}Y^2 \leqslant m^2$. Theorem 2 states that, with constant probability, the truncated least squares and the Forster-Warmuth estimators incur a constant excess risk of the same order. At the first sight, this property may seem incompatible with expected excess risk bounds of order d/n. However, one should keep in mind that the estimators in question are improper (returning predictors outside of the class $\mathcal{F}_{\mathrm{lin}}$), so that the excess risk may well take negative values; the expected excess risk remains small due to the fact that positive and negative values essentially compensate in expectation, regardless of the distribution.

A related phenomenon was observed in the context of model selection-type aggregation by Audibert [2], who showed that the (improper) progressive mixture rule [82, 12], known to achieve fast rates in expectation, exhibits slow rates in deviation. In our context the failure in deviation is even more severe, as the excess risk is of constant order, rather than exhibiting slow rates.

Proof. For any $n \geqslant n_0$, let $P = P(n)$ be the distribution of (X, Y) satisfying

$$(X, Y) = \begin{cases} (1, m) & \text{with probability } 1 - \frac{1}{n}\,; \\ (\sqrt{n}, 0) & \text{with probability } \frac{1}{n}\,. \end{cases}$$

By homogeneity, we may assume that $m = 1$. For any $w \in \mathbf{R}$, set $g_w(x) = w \cdot x$. We have

$$R(g_w) = \Big(1 - \frac{1}{n}\Big)(w-1)^2 + \frac{1}{n}(w\sqrt{n})^2 = \Big(1 - \frac{1}{n}\Big)(w-1)^2 + w^2\,.$$

It follows that the risk of the best linear predictor is equal to

$$\inf_{w \in \mathbf{R}} R(g_w) = \frac{1 - 1/n}{2 - 1/n} \leqslant \frac{1}{2}\,. \tag{13}$$

In addition, let $K = K_n$ denote the number of indices $i = 1, \dots, n$ such that $X_i = \sqrt{n}$. The empirical risk writes

$$\widehat{R}_n(g_w) = \Big(1 - \frac{K}{n}\Big)(w-1)^2 + Kw^2, \quad \text{and so} \quad \widehat{w}_{\mathrm{erm}} = \operatorname*{arg\,min}_{w \in \mathbf{R}} \widehat{R}_n(g_w) = \frac{1 - K/n}{K + 1 - K/n}\,.$$

In particular, $0 \leqslant \widehat{w}_{\mathrm{erm}} \leqslant 1/(K+1)$. Now, note that if $\widehat{g}$ denotes either the truncated least squares (8) or the Forster-Warmuth estimator (9), then $\widehat{g}(1) \leqslant \widehat{w}_{\mathrm{erm}} \cdot 1 \leqslant 1/(K+1) \leqslant 1$, and thus, denoting the sample $(X_i, Y_i)_{i=1}^n$ by S_n, we have

$$R(\widehat{g}) \geqslant \mathbf{E}\big[(\widehat{g}(X) - Y)^2 \mathbf{1}(X = 1) \,|\, S_n\big] \geqslant \Big(1 - \frac{1}{n}\Big) \cdot \Big(\frac{K}{K+1}\Big)^2\,. \tag{14}$$

Thus, under the event $E_n = \{K_n \geqslant 4\}$, it follows from (13) and (14) that for $n \geqslant 16$,

$$R(\widehat{g}) - \inf_{g \in \mathcal{F}_{\mathrm{lin}}} R(g) \geqslant \Big(1 - \frac{1}{n}\Big) \cdot \Big(\frac{K}{K+1}\Big)^2 - \frac{1}{2} = \Big(1 - \frac{1}{16}\Big) \cdot \frac{16}{25} - \frac{1}{2} = \frac{1}{10}\,.$$

Finally, since K_n follows the binomial distribution $\mathsf{Bin}(n, 1/n)$, the probability $\mathbf{P}(E_n)$ is positive for $n \geqslant 16 \geqslant 4$. Further, since K_n converges in distribution to the Poisson distribution $\mathsf{Poi}(1)$ as $n \to \infty$, $\mathbf{P}(E_n) \to \mathbf{P}(\widetilde{K} \geqslant 4) > 0$ with $\widetilde{K} \sim \mathsf{Poi}(1)$, so that setting $p_0 = \inf_{n \geqslant 16} \mathbf{P}(E_n)$, we have $p_0 > 0$. This concludes the proof with $c = \min(p_0, 1/10)$ and $n_0 = 16$. □

5 An optimal robust estimator in the high-probability regime

In this section we present our main positive result. We show that there is an estimator achieving an optimal accuracy and sub-exponential tails for the linear class $\mathcal{F}_{\mathrm{lin}}$ under Assumption 1. We first consider a simplified setup where the covariance structure of X is known.

5.1 Warm-up: known covariance structure

Following the discussion on the learning model with known covariance structure in Section 2, we assume in this section that $\Sigma = \mathbf{E}XX^{\mathsf{T}}$ exists, is invertible and also known. Recall the definition of Tsybakov's projection estimator $\widehat{g}_{\mathrm{proj}}$ (7). Since this estimator always returns a linear predictor, its excess risk is non-negative and we may apply Markov's inequality to show that for any $\delta \in (0,1)$, it holds that

$$\mathbf{P}\left(R(\widehat{g}_{\mathrm{proj}}) - \inf_{g\in\mathcal{F}_{\mathrm{lin}}} R(g) \leqslant \frac{m^2 d}{n}\cdot\frac{1}{\delta}\right) \geqslant 1-\delta.$$

An argument similar to the one used in [13, Proposition 6.2] can be used to show that this bound is essentially the best we can hope for the projection estimator, even when $|Y| \leqslant m$ almost surely.

Fortunately, there is a way to modify this estimator and obtain a guarantee with sub-exponential tails. The result of Lugosi and Mendelson [48, Theorem 1] shows for any $\delta \in (0,1)$, there exists an estimator $\widehat{\mu}_\delta : (\mathbf{R}^d)^n \to \mathbf{R}$ such that, for any sequence $U_1,\dots,U_n$ of i.i.d. random vectors in $\mathbf{R}^d$ with mean μ and covariance matrix $\widetilde{\Sigma} = \mathrm{Cov}(U)$, $\widehat{\mu}_\delta = \widehat{\mu}_\delta(U_1,\dots,U_n)$ satisfies

$$\mathbf{P}\left(\|\widehat{\mu}_\delta - \mu\|^2 \leqslant c\,\frac{\mathrm{Tr}(\widetilde{\Sigma}) + \|\widetilde{\Sigma}\|_{\mathrm{op}}\log(1/\delta)}{n}\right) \geqslant 1-\delta, \tag{15}$$

where $c>0$ is an absolute constant. Now, introduce the *robust* projection estimator

$$\widetilde{w} = \Sigma^{-1/2}\cdot\widehat{\mu}_\delta\left(Y_1\Sigma^{-1/2}X_1,\dots,Y_n\Sigma^{-1/2}X_n\right), \tag{16}$$

and consider the following result.

Proposition 1. *There is an absolute constant $c>0$ such that the following is true. Suppose that Assumption 1 holds. Then, the robust projection estimator $\widehat{g} = \langle\widetilde{w},\cdot\rangle$ (which is a proper estimator) defined in (16) satisfies*

$$\mathbf{P}\left(R(\widehat{g}) - \inf_{g\in\mathcal{F}_{\mathrm{lin}}} R(g) \leqslant c\,\frac{m^2\big(d+\log(1/\delta)\big)}{n}\right) \geqslant 1-\delta.$$

Proof. We have shown in Section 2, that under Assumption 1, $\mathrm{Cov}(Y\Sigma^{-1/2}X) \preccurlyeq m^2 I_d$. Combining the deviation bound (15) and the definition (16) with the identity (6), we finish the proof. □

Proposition 1 serves as a benchmark result for the performance that we aim to establish in the more realistic setting where the covariance matrix Σ is unknown. This is achieved in the next section.

5.2 Deviation-optimal robust estimator

The theorem below is the main positive result of our paper. It demonstrates that Assumption 1 is a sufficient condition for the existence of linear regression estimators satisfying an excess risk deviation inequality with logarithmic dependence on the confidence parameter. In Section 6, we show that Assumption 1 is also necessary.

Theorem 3. *There is an absolute constant $c > 0$ such that the following holds. Assume that $n \geqslant d$. Suppose that Assumption 1 holds and fix any $\delta \in (0,1)$. Then, there exists an estimator $\widehat{g}$ depending on δ and m such that the following holds:*

$$\mathbf{P}\left(R(\widehat{g}) - \inf_{g \in \mathcal{F}_{\mathrm{lin}}} R(g) \leqslant c \frac{m^2 (d \log(n/d) + \log(1/\delta))}{n} \right) \geqslant 1 - \delta.$$

Moreover, the above bound also holds if the class $\mathcal{F}_{\mathrm{lin}}$ is replaced by an arbitrary VC-subgraph class $\mathcal{F}$ of dimension d.

Before presenting our estimator, we briefly comment on the above theorem. First, in contrast to existing work on robust linear regression, our estimator $\widehat{g}$ is improper, even though the underlying linear class is convex. Second, unlike our previous results presented in this paper, the bound of Theorem 3 is not specific to the linear class. In particular, our proof extends without changes to the family of VC-subgraph classes (see [26, Definition 3.6.8]). Some recent results in the robust statistics literature apply to more general classes of functions, including non-parametric classes (see, for example, [47, 54, 18]). However, as discussed in Section 1.2, such results are only known to be valid under additional assumptions on P_X. Extending our results to more general classes presents some challenges; we discuss them in more detail in Section 5.3. Finally, we note that our estimator depends on the value of m. This assumption simplifies the analysis and is standard in similar contexts (see, for example, [29, Theorem 11.3] and [56]).

We now introduce some additional notation needed to define our estimator. For any $\varepsilon > 0$ and any class of real-valued functions $\mathcal{G}$ let $\mathcal{G}_\varepsilon$ denote the smallest ε-net of $\mathcal{G}$ with respect to the empirical L_1 distance $\frac{1}{n}\sum_{i=1}^n |f(X_i) - g(X_i)|$. We only consider ε-nets that are subsets of $\mathcal{G}$. For the standard definition of an ε-net we refer to [77, Section 4.2]. Assume that we have a sample $S = (X_i, Y_i)_{i=1}^{3n}$ of size $3n$ and denote $S_1 = (X_i, Y_i)_{i=1}^{n}$, $S_2 = (X_i, Y_i)_{i=n+1}^{2n}$ and $S_3 = (X_i, Y_i)_{i=2n+1}^{3n}$. Fix any $1 \leqslant k \leqslant n$, and assume without loss of generality that n/k is integer. Split the set $\{1, \ldots, n\}$ into k blocks $I_1, \ldots, I_k$ of equal size such that $I_j = \{1 + (j-1)(n/k), \ldots, j(n/k)\}$. Fix any function $\ell : \mathbf{R}^d \times \mathbf{R} \to \mathbf{R}$, any sample S' of size n, and denote the i-th element of S' by $Z_i = (X_i, Y_i)$. The median-of-means estimator (see also [45, Section 2.1], [62]) is defined as follows:

$$\mathrm{MOM}^k_{S'}(\ell) = \mathrm{Median}\left(\frac{k}{n}\sum_{i \in I_1} \ell(Z_i), \ldots, \frac{k}{n}\sum_{i \in I_k} \ell(Z_i) \right).$$

Finally, for any predictor $f : \mathbf{R}^d \to \mathbf{R}$, denote the associated loss function by $\ell_f(Z_i) = (f(X_i) - Y_i)^2$. We are now ready to present our estimator.

The estimator of Theorem 3

1. Split the sample S of size $3n$ into three equal parts S_1, S_2 and S_3 as defined above. Use the value m to construct the truncated class

$$\overline{\mathcal{F}} = \left\{ f_m : f \in \mathcal{F}_{\mathrm{lin}} \right\},$$

where recall that f_m denotes the truncation of a function f (see (8)).

2. Fix $\varepsilon = \frac{md}{n}$. Using the first sample S_1, construct an ε-net of $\overline{\mathcal{F}}$ with respect to the empirical L_1 distance and denote it by $\overline{\mathcal{F}}_\varepsilon$.

3. Let $c_1, c_2 > 0$ be some specifically chosen absolute constants. Fix the number of blocks $k = \lceil c_1 d(\log(n/d) + \log(1/\delta)) \rceil$ and set $\alpha = c_2 \sqrt{\frac{m^2(d\log(n/d)+\log(1/\delta))}{n}}$. If $k > n$, then set $\widehat{g} = 0$. Otherwise, using the second sample S_2 define a random subset of $\overline{\mathcal{F}}_\varepsilon$ as follows:

$$\widehat{\mathcal{F}} = \Big\{ f \in \overline{\mathcal{F}}_\varepsilon : \forall g \in \overline{\mathcal{F}}_\varepsilon,\ \mathrm{MOM}^k_{S_2}(\ell_f - \ell_g) \leqslant \alpha \sqrt{\frac{1}{n}\sum_{X_i \in S_2} (f(X_i) - g(X_i))^2 + \alpha^2} \Big\}.$$

4. Define the set $\widehat{\mathcal{F}}_+$ consisting of all the mid-points of $\widehat{\mathcal{F}}$, that is, $\widehat{\mathcal{F}}_+ = (\widehat{\mathcal{F}} + \widehat{\mathcal{F}})/2$. Using the third sample S_3, define our estimator $\widehat{g}$ as

$$\widehat{g} = \operatorname*{arg\,min}_{g \in \widehat{\mathcal{F}}_+} \max_{f \in \widehat{\mathcal{F}}_+} \mathrm{MOM}^k_{S_3}(\ell_g - \ell_f).$$

5. Return $\widehat{g}$.

Our estimator involves a combination of several seemingly disconnected ideas in the literature. The truncation step is inspired by the analysis in [29, Chapter 11], with the difference that we use the truncation as a preliminary step, rather than as a post-processing of the ERM prediction (see Theorem 1). The second step replaces the original class by an empirical L_1 ε-net of the truncated class. In many situations, such a construction leads to suboptimal results. However, since we work with a particular parametric class, this step does not affect the resulting performance. The use of the ε-net $\overline{\mathcal{F}}_\varepsilon$ is needed for technical reasons; we explain the technical aspects in detail in Section 5.3. Our third step is inspired by the median-of-means tournaments introduced in [47]. The main difference with the latter work is that our truncated class is now non-convex, and to obtain the correct rates of convergence, we need to adapt the arguments used in the model selection aggregation literature. This motivates our fourth step that can be seen as an adaptation of the *star algorithm* [2] and the two-step aggregation procedure developed in [42, 53] to our specific heavy-tailed setting combined with the idea of min-max formulation of robust estimators [5, 41]. We remark that the idea of combining model selection aggregation techniques with the median-of-means tournaments has also recently appeared in [54], but under different assumptions. As we mentioned, the key distinction therein is that the suggested learning procedure collapses to a proper estimator for convex classes of functions, such as $\mathcal{F}_{\mathrm{lin}}$ considered in our work; as discussed in Section 2, for such procedures some restrictions on the distribution of covariates are required to obtain performance bounds.

The rest of this section is devoted to proving Theorem 3. First, the truncation at the level m can only make the risk smaller whenever Assumption 1 is satisfied. Indeed, this follows from the identity

$$R(g) = \mathbf{E}(g(X) - f_{\mathrm{reg}}(X))^2 + \mathbf{E}(f_{\mathrm{reg}}(X) - Y)^2,$$

and the fact that f_{reg} is absolutely bounded by m. Therefore, we may focus on bounding

$$R(\widehat{g}) - \inf_{g \in \overline{\mathcal{F}}} R(g).$$

We will now state and comment on some technical lemmas that will be used in our proof. The proofs of the below lemmas are deferred to Section 7.

Next, we provide a uniform deviation bound on the L_1 distances between the elements of $\overline{\mathcal{F}}$.

Lemma 1. *Assume that $n \geqslant d$. There is a constant $c > 0$ such that simultaneously for all $f, g \in \overline{\mathcal{F}}$, with probability at least $1 - \delta$, it holds that*

$$\mathbf{E}|f(X) - g(X)| \leqslant \frac{2}{n}\sum_{i=1}^{n}|f(X_i) - g(X_i)| + c\left(\frac{md\log(n/d) + m\log(3/\delta)}{n}\right).$$

To simplify the statements of the lemmas to follow, for any finite class $\mathcal{G}$ and for any confidence parameter $\delta \in (0,1)$ define:

$$\alpha(\mathcal{G}, \delta) = 32\sqrt{\frac{m^2(\log(2|\mathcal{G}|) + \log(4/\delta))}{n}}, \tag{17}$$

where the sample size n and the value m (of Assumption 1) will always be clear from the context. The next technical lemma provides basic concentration properties of the median-of-means estimators, the proof of which follows from a combination of uniform Bernstein's inequality and a median-of-means deviation inequality for mean estimation [45, Theorem 2].

Lemma 2. *Suppose that Assumption 1 holds and let $S_n = (X_i, Y_i)_{i=1}^n$ denote an i.i.d. sample. Let $\mathcal{G}$ be any finite class of functions whose absolute value is bounded by m. Fix any $\delta \in (0,1)$, let $k = \lceil 8\log\frac{2|\mathcal{G}|^2}{\delta}\rceil$ and let α denote any upper bound on $\alpha(\mathcal{G}, \delta)$ defined in (17). Then, with probability at least $1 - \delta$, the following inequalities hold simultaneously for any $f, g \in \mathcal{G}$:*

$$\left|R(f) - R(g) - \mathrm{MOM}_{S_n}^k(\ell_f - \ell_g)\right| \leqslant \alpha\sqrt{\mathbf{E}(f(X) - g(X))^2},$$

$$\left|R(f) - R(g) - \mathrm{MOM}_{S_n}^k(\ell_f - \ell_g)\right| \leqslant \sqrt{2}\alpha\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2 + \alpha^2},$$

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2 \leqslant 2\mathbf{E}(f(X) - g(X))^2 + \alpha^2.$$

For any class $\mathcal{G}$, define its L_2 diameter by:

$$\mathcal{D}(\mathcal{G}) = \sup_{f,g\in\mathcal{G}}\sqrt{\mathbf{E}(f(X) - g(X))^2}.$$

As a corollary of the above lemma, we are able to derive some basic properties of the random set $\widehat{\mathcal{F}}$. In particular, we show that with high probability the set $\widehat{\mathcal{F}}$ contains the population risk minimizer over the ε-net $\overline{\mathcal{F}}_\varepsilon$. At the same time, we establish a uniform Bernstein-type bound on the excess risk of the elements of $\widehat{\mathcal{F}}$, with the role of the variance term played by $\mathcal{D}(\widehat{\mathcal{F}})$.

Lemma 3. *Suppose that Assumption 1 holds and let $S_n = (X_i, Y_i)_{i=1}^n$ denote an i.i.d. sample. Let $\mathcal{G}$ be any finite class of functions whose absolute value is bounded by m. Fix any $\delta \in (0,1)$, $k = \lceil 8\log\frac{2|\mathcal{G}|^2}{\delta}\rceil$ and let α denote any upper bound on $\alpha(\mathcal{G}, \delta)$ defined in (17). Define the random subset of $\mathcal{G}$:*

$$\widehat{\mathcal{G}} = \left\{f \in \mathcal{G} : \text{for every } g \in \mathcal{G},\ \mathrm{MOM}_{S_n}^k(\ell_f - \ell_g) \leqslant \sqrt{2}\alpha\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2 + \alpha^2}\right\},$$

Then, the following two conditions hold simultaneously, with probability at least $1 - \delta$:

1. *The function* $g^* = \arg\min_{g\in\mathcal{G}} R(g)$ *belongs to the class* $\widehat{\mathcal{G}}$.

2. *For any* $f, g \in \widehat{\mathcal{G}}$, *we have* $R(f) - R(g^*) \leqslant 4\alpha\mathcal{D}(\widehat{\mathcal{G}}) + 5\alpha^2$.

Finally, we prove an excess risk bound for the min-max estimator in terms of the L_2 diameter of the set over which the estimator is computed. The intuitive implications of the following lemma are the following. First, if $\mathcal{D}(\widehat{\mathcal{F}})$ is of order $1/\sqrt{n}$, the lemma below immediately yields the fast rate of convergence for our estimator $\widehat{g}$. If, on the other hand, the diameter of $\mathcal{D}$ is much larger than $1/\sqrt{n}$, then we can exploit the curvature of the quadratic loss and the gain in the approximation error (due to considering the larger class $\widehat{\mathcal{F}}_+$ instead of $\widehat{\mathcal{F}}$) to prove the desired rate of convergence.

Lemma 4. *Suppose that Assumption 1 holds and let* $S_n = (X_i, Y_i)_{i=1}^n$ *denote an i.i.d. sample. Let* $\mathcal{G}$ *be any finite class of functions whose absolute value is bounded by* m. *Fix any* $\delta \in (0,1)$, *let* $k = \lceil 8 \log \frac{2|\mathcal{G}|^2}{\delta} \rceil$ *and let* α *denote any upper bound on* $\alpha(\mathcal{G}, \delta)$ *defined in* (17). *Let* $\widehat{g}$ *be any estimator satisfying*

$$\widehat{g} \in \arg\min_{g\in\mathcal{G}} \max_{f\in\mathcal{G}} \mathrm{MOM}_{S_n}^k(\ell_g - \ell_f).$$

Let $g^* \in \arg\min_{g\in\mathcal{G}} R(g)$. *Then, with probability at least* $1-\delta$, *it holds that*

$$R(\widehat{g}) \leqslant R(g^*) + 2\alpha\mathcal{D}(\mathcal{G}).$$

We are now ready to prove Theorem 3.

Proof of Theorem 3. Our proof is split into two parts. First, we approximate the truncated linear class $\overline{\mathcal{F}}$ with a finite class, namely, an empirical L_1 ε-net constructed using the first third of the dataset denoted by S_1. Then, conditionally on S_1, we show that our estimator $\widehat{g}$ achieves the optimal rate of model selection aggregation over the finite class $\overline{\mathcal{F}}_\varepsilon$, in spite of the lack of assumptions on the covariates and the presence of heavy-tailed labels. Finally, we note that if the number of median-of-means blocks k is equal to 0 (*i.e.*, $n \lesssim d(\log(n/d) + \log(1/\delta))$), then we may the 0 function which satisfies the desired bound for such sample sizes. Thus, in what follows we assume that $n \gtrsim d(\log(n/d) + \log(1/\delta))$.

The approximation step. Recall that $\overline{\mathcal{F}}_\varepsilon$ is an empirical L_1 ε-net of the truncated linear class $\overline{\mathcal{F}}$ constructed using the sample S_1. Let $f^* = \arg\min_{f\in\overline{\mathcal{F}}} R(f)$ and let f_ε^* be any element of $\overline{\mathcal{F}}_\varepsilon$ minimizing the empirical L_1 distance to f^*, that is, we have

$$\frac{1}{n}\sum_{X_i\in S_1} |f^*(X_i) - f_\varepsilon^*(X_i)| \leqslant \varepsilon. \tag{18}$$

Let E_1 denote the event of Lemma 1 applied with respect to the sample S_1 (that contains n points) with the choice of the confidence parameter set to $\delta/3$ (thus, $\mathbf{P}(E_1) \geqslant 1-\delta/3$). It follows

that on the event E_1 we have

$$
\begin{aligned}
&R(f_\varepsilon^*) - R(f^*) \\
&= 2\mathbf{E}Y(f^*(X) - f_\varepsilon^*(X)) + \mathbf{E}(f_\varepsilon^*(X)^2 - f^*(X)^2) \\
&\leqslant 2\mathbf{E}(\mathbf{E}[Y|X](f^*(X) - f_\varepsilon^*(X))) + 2m\mathbf{E}|f_\varepsilon^*(X) - f^*(X)| \qquad (\text{since } |f_\varepsilon^*(X) + f^*(X)| \leqslant 2m) \\
&\leqslant 2\mathbf{E}(\sqrt{\mathbf{E}[Y^2|X]}|f^*(X) - f_\varepsilon^*(X)|) + 2m\mathbf{E}|f_\varepsilon^*(X) - f^*(X)| \qquad (\text{by Jensen's inequality}) \\
&\leqslant 4m\mathbf{E}|f_\varepsilon^*(X) - f^*(X)| \qquad (\text{by Assumption 1}) \\
&\leqslant 8m\varepsilon + 4mc_1\left(\frac{md\log(n/d) + m\log(9/\delta)}{n}\right) \qquad (\text{by (18) and Lemma 1}) \\
&\leqslant 12c_1\left(\frac{m^2 d\log(n/d) + m^2\log(9/\delta)}{n}\right), \qquad (\text{by the definition of } \varepsilon)
\end{aligned}
$$

where c_1 is an absolute constant. Observe that on the event E_1, any estimator $\widehat{g}$ satisfies

$$
\begin{aligned}
R(\widehat{g}) - R(f^*) &\leqslant R(\widehat{g}) - \min_{f\in\overline{\mathcal{F}}_\varepsilon} R(f) + R(f_\varepsilon^*) - R(f^*) \\
&\leqslant R(\widehat{g}) - \min_{f\in\overline{\mathcal{F}}_\varepsilon} R(f) + 12c_1\left(\frac{m^2 d\log(n/d) + m^2\log(9/\delta)}{n}\right).
\end{aligned}
$$

From this point onward, we work on the event E_1. It thus remains to prove that with probability $1 - 2\delta/3$, the estimator $\widehat{g}$ computed using the remaining $2n$ points split into samples S_2 and S_3 satisfies

$$
R(\widehat{g}) - \min_{f\in\overline{\mathcal{F}}_\varepsilon} R(f) \lesssim \frac{m^2 d\log(n/d) + m^2\log(1/\delta)}{n}. \tag{19}
$$

Since $\overline{\mathcal{F}}_\varepsilon$ is a finite class of functions, we now turn to the aggregation part of this proof.

The aggregation step. By the L_2 covering number bound stated in [29, Theorem 9.4, Theorem 9.5], which also holds for the empirical L_1 distances, we have (see the proof of Lemma 1)

$$
\log|\overline{\mathcal{F}}_\varepsilon| \lesssim d\log\frac{me}{\varepsilon} \lesssim d\log(n/d).
$$

Note that $|\widehat{\mathcal{F}}_+|$ and $|\widehat{\mathcal{F}}|$ are simultaneously upper bounded by $|\overline{\mathcal{F}}_\varepsilon|^2$. For an arbitrary finite class $\mathcal{G}$, recall the definition of $\alpha(\mathcal{G},\delta)$ stated in (17). It follows that there exists some absolute constant $c_2 > 0$ such that $\overline{\alpha}$ defined below satisfies

$$
\max\Big(\alpha(\widehat{\mathcal{F}},\delta/3), \alpha(\widehat{\mathcal{F}}_+,\delta/3)\Big) \leqslant \overline{\alpha} = c_2\sqrt{\frac{m^2 d\log(n/d) + m^2\log(1/\delta)}{n}}. \tag{20}
$$

Thus, $\overline{\alpha}$ defined above will be used in the applications of Lemmas 2, 3 and 4 to follow.

Let E_2 be the event of Lemma 3 applied for the set $\widehat{\mathcal{F}}$ with confidence parameter $\delta/3$. In particular, on the event E_2 we have

$$
\arg\min_{f\in\overline{\mathcal{F}}_\varepsilon} R(f) \in \widehat{\mathcal{F}}, \quad \text{and for any } f\in\widehat{\mathcal{F}} \text{ it holds that } R(f) \leqslant \min_{f\in\overline{\mathcal{F}}_\varepsilon} R(f) + 4\overline{\alpha}\mathcal{D}(\widehat{\mathcal{F}}) + 5\overline{\alpha}^2. \tag{21}
$$

Conditionally on the sample S_2, let the set $\widehat{\mathcal{F}}$ defined in the third step of our algorithm be fixed. Denote $g^* = \arg\min_{g\in\widehat{\mathcal{F}}_+} R(f)$, where recall that $\widehat{\mathcal{F}}_+ = (\widehat{\mathcal{F}} + \widehat{\mathcal{F}})/2$. Observe that the L_2 diameters of $\widehat{\mathcal{F}}$ and $\widehat{\mathcal{F}}_+$ are equal, that is $\mathcal{D}(\widehat{\mathcal{F}}_+) = \mathcal{D}(\widehat{\mathcal{F}})$. Let E_3 be the event of Lemma 4

applied to the third part of our sample S_3 and the finite class $\widehat{\mathcal{F}}_+$ with the confidence parameter set to $\delta/3$. Thus, on E_3 our estimator $\widehat{g}$ satisfies:

$$R(\widehat{g}) \leqslant R(g^*) + 2\overline{\alpha}\mathcal{D}(\widehat{\mathcal{F}}). \tag{22}$$

Now choose any $g, h \in \widehat{\mathcal{F}}$ such that $\sqrt{\mathbf{E}(g(X) - h(X))^2} \geqslant \mathcal{D}(\widehat{\mathcal{F}})/2$ (such a choice always exists by definition of the diameter). Since $(g+h)/2 \in \widehat{\mathcal{F}}_+$, the parallelogram identity yields

$$\begin{aligned} R(g^*) &\leqslant R((g+h)/2) \\ &= \frac{1}{2}R(g) + \frac{1}{2}R(h) - \frac{1}{4}\mathbf{E}(g(X) - h(X))^2 \\ &\leqslant \frac{1}{2}R(g) + \frac{1}{2}R(h) - \frac{1}{16}\mathcal{D}(\widehat{\mathcal{F}})^2. \end{aligned} \tag{23}$$

On the event E_2, applying (21) for the functions g and h we obtain

$$\frac{1}{2}R(g) + \frac{1}{2}R(h) \leqslant \min_{f \in \overline{\mathcal{F}}_\varepsilon} R(f) + 4\overline{\alpha}\mathcal{D}(\widehat{\mathcal{F}}) + 5\overline{\alpha}^2.$$

Combining the above with equations (22) and (23) we have

$$R(\widehat{g}) - \min_{f \in \overline{\mathcal{F}}_\varepsilon} R(f) \leqslant 6\overline{\alpha}\mathcal{D}(\widehat{\mathcal{F}}) + 5\overline{\alpha}^2 - \frac{1}{16}\mathcal{D}(\widehat{\mathcal{F}})^2 \leqslant 149\overline{\alpha}^2,$$

where the last step follows by maximizing the quadratic equation with respect to $\mathcal{D}(\widehat{\mathcal{F}})$. Plugging in the definition of $\overline{\alpha}$ (see (20)) we obtain the desired inequality (19). The proof is complete by taking the union bound over the events E_1, E_2 and E_3 defined above. □

5.3 Some extensions of Theorem 3

We begin by noting that Theorem 3 holds not only for linear classes, but more generally, for VC-subgraph classes (without any changes to our argument presented in the previous section). Indeed, the structure of the underlying function class only enters our proof though the control on the empirical covering numbers of its truncated elements; sharp bounds for such covering numbers are available in [29, Theorem 9.4]. As a special case, our analysis covers finite classes and hence, provides new results for the problem of model selection aggregation, where a learner is tasked with constructing a predictor as good as the best one in a given finite class (also called dictionary) of functions [61, 70]. It is arguably the most straightforward problem manifesting statistical separation between proper and improper learning algorithms (see, for instance, [12, 37]). Procedures based on exponential weighting were shown to attain optimal rates in expectation [81, 82, 12, 3], yet they were later shown to be deviation suboptimal [2], in close similarity to our results presented in Section 4.

We can now formulate the following result, which from the statistical point of view, generalizes the best known results for the problem of model selection aggregation [2, 42].

Theorem 4. *There is an absolute constant $c > 0$ such that the following holds. Grant Assumption 1, fix any $\delta \in (0,1)$ and let $\mathcal{F}$ be a finite class of possibly unbounded functions. Then, there exists an estimator $\widehat{g}$ depending on δ and m such that the following holds:*

$$\mathbf{P}\left(R(\widehat{g}) - \min_{g \in \mathcal{F}} R(g) \leqslant c\,\frac{m^2(\log|\mathcal{F}| + \log(1/\delta))}{n}\right) \geqslant 1 - \delta.$$

Proof. The aggregation algorithm is the same as the estimator of Theorem 3 with only two differences. First, we skip the step with ε-net discretization of the truncated class $\overline{\mathcal{F}}$. The second difference is that the number of blocks in median-of-means estimators is of order $\log(|\mathcal{F}|/\delta)$ and similarly, the parameter α is redefined to be of order $\sqrt{\frac{m^2(\log|\mathcal{F}|+\log(1/\delta))}{n}}$. The proof follows via the "aggregation step" part of the proof of Theorem 3. □

Concerning aggregation with a heavy-tailed response variable, the above result can be compared with the bounds of Audibert [3] and Juditsky, Rigollet and Tsybakov [37]. Assuming that the functions in $\mathcal{F}$ are absolutely bounded by 1, and that $\mathbf{E}|Y|^s \leqslant m_s$ for some $s \geqslant 2, m_s > 0$, they prove an in-expectation bound on $\mathbf{E}R(\widetilde{f}) - \min_{f\in\mathcal{F}} R(f)$ for some estimator $\widetilde{f}$ with the rate of convergence slower than $1/n$. In contrast, in Theorem 4 we do not assume the boundedness of $\mathcal{F}$, but require that the conditional second moment of Y is bounded. As a result, we provide a deviation bound with the $1/n$ rate of convergence and logarithmic dependence on the confidence parameter δ. We emphasize again that due to the necessity of improperness for optimal model selection aggregation, in-expectation results are not easily transferable to deviation bounds; the in-expectation guarantees of [37, 3] are in fact obtained for variants of the progressive mixture or mirror averaging rule, which is shown by Audibert [2] to exhibit suboptimal deviations. Finally, an argument of Section 6 shows the necessity of Assumption 1 in our distribution-free setting for model selection aggregation.

Further extensions of Theorem 3, particularly, going beyond VC-subgraph classes present technical challenges. First, obtaining distribution-free empirical covering number guarantees for truncations of general classes (as done for $\overline{\mathcal{F}}$ in our case) might be a non-trivial task. Second, it is well-known (see the discussion in [66]) that even when only bounded functions are considered, replacing the original function class by its empirical ε-net (as done via the function class $\overline{\mathcal{F}}_\varepsilon$ in our algorithm) usually renders the recovery of the correct excess risk rates impossible. This in turn leads to the final and the most technical problem: if $\overline{\mathcal{F}}$ is not replaced by $\overline{\mathcal{F}}_\varepsilon$, there are no known ways to obtain an analog of the concentration Lemma 2, while only imposing Assumption 1.

To expand on the last point, an analog of Lemma 2 for general classes can be approached via the analysis of suprema of localized quadratic and multiplier processes (see [54] for related arguments); specifically, the supremum of the localized process $\mathbf{E}\sup_{h\in\mathcal{H}_r}(\sum_{i=1}^n \varepsilon_i Y_i h(X_i))$ is difficult to control for general classes under our assumptions (here $\mathcal{H}_r$ denotes localized subsets of the class $\overline{\mathcal{F}}-\overline{\mathcal{F}}$, see the proof of Lemma 1 for more details). However, even if the response variable Y is independent of X, the standard in this context application of the multiplier inequality [74, Lemma 2.9.1] introduces the dependence on the moment $\|Y\|_{2,1} = \int_0^\infty \sqrt{\mathbf{P}(|Y|>t)}\mathrm{d}t$ in the resulting bounds, instead of the desired moment $\mathbf{E}Y^2$, as we obtain in Lemma 2 for finite classes. It is known that the dependence on the $\|\cdot\|_{2,1}$ norm is unavoidable in some cases [44]. More importantly, we refer to the recent work [32] discussing that the multiplier inequality can lead to suboptimal rates (see [32, Section 2.3.1] for more details).

6 Statistical lower bounds and the necessity of Assumption 1

The statistical guarantees obtained in the previous sections hold under no assumptions on the distribution of X and under Assumption 1 on the conditional distribution of Y given X. In this section, we show that Assumption 1 is necessary to obtain non-trivial guarantees on the excess risk without restrictions on P_X and that our risk bounds are unimprovable, in a precise sense.

Proposition 2. *Fix any $n \geqslant 1$, $\delta \in (e^{-n}, 1)$ and any measurable function $f : \mathbf{R} \to \mathbf{R}$ satisfying $f(0) = 0$ and $\sup_{x\in\mathbf{R}} f(x)^2 \geqslant 1$. Then, there exists a distribution P_X of X such that for any*

estimator $\widehat{g}$ (possibly improper and P_X-dependent), setting $Y = f_{\mathrm{reg}}(X)$ (where $f_{\mathrm{reg}} \in \{f, -f\}$) the following three conditions hold:

- *there exists $w^* \in \mathbf{R}$ such that $R(g_{w^*}) = 0$;*
- $\mathbf{E}[Y^2] \leqslant 1$;
- *denoting $\|f_{\mathrm{reg}}\|_\infty = \sup_{x \in \mathbf{R}} |f_{\mathrm{reg}}(x)| = \|f\|_\infty \in [1, +\infty]$ we have*

$$\mathbf{P}\left(R(\widehat{g}) \geqslant \min\left(\frac{\|f_{\mathrm{reg}}\|_\infty^2 \cdot \log(1/\delta)}{4n}, 1 \right) \right) \geqslant \delta.$$

Before providing the proof, let us comment on the implications of this lower bound. First, note that if the conditional second moment bound $\mathbf{E}[Y^2|X] \leqslant 1$ of Assumption 1 is relaxed to the weaker unconditional bound $\mathbf{E}Y^2 \leqslant 1$, then (taking $\delta = 0.9$, and any f such that $\|f\|_\infty \geqslant \sqrt{n}$) the worst-case excess risk of any estimator $\widehat{g}$ is lower-bounded by an absolute constant c with probability 0.9, matching up to constants the risk of at most 1 trivially achieved by the identically zero function. Second, without Assumption 1 our upper bounds cannot be improved even in the "realizable" case where the linear class $\mathcal{F}_{\mathrm{lin}}$ contains a perfect predictor (that is, when $R(g) = 0$ for some $g \in \mathcal{F}_{\mathrm{lin}}$), and in particular $\mathrm{Var}(Y|X) = 0$ almost surely. As a result, the quantity $\sup_{x \in \mathbf{R}^d} \mathbf{E}[Y^2|X = x]$ in our assumption cannot be replaced by $\sup_{x \in \mathbf{R}^d} \mathrm{Var}(Y|X = x)$. Finally, when $Y = f_{\mathrm{reg}}(X)$, then the worst-case dependence on f_{reg} can be no better than $\|f_{\mathrm{reg}}\|_\infty^2$, as shown in the last part of the above proposition. The dependence on m^2 in our upper bounds is thus unavoidable, recalling that $m^2 \leqslant \|f_{\mathrm{reg}}\|_\infty^2$ whenever $Y = f_{\mathrm{reg}}(X)$.

We point out that the same argument as in Proposition 2 shows that a dependence on $\sup_{x \in \mathbf{R}^d} \mathbf{E}[Y^2|X = x]$ is unavoidable for any conditional distribution $(P_{Y|X=x})_{x \in \mathbf{R}^d}$ (possibly known up to its sign), beyond the case $Y = f_{\mathrm{reg}}(X)$. We considered the latter special case for simplicity, and because it allows to simultaneously impose that $R(g_{w^*}) = 0$ for some $w^* \in \mathbf{R}^d$. We also remark that Proposition 2 is stated in dimension $d = 1$ for simplicity. The same lower bound construction can be used for general dimension d (assuming, for example, that f is continuous, and imposing $|f_{\mathrm{reg}}| \leqslant |f|$), allowing one to replace the $\log(1/\delta)$ term by $d + \log(1/\delta)$.

Proof. Let $p \in (0, 1)$ be such that $(1 - p)^n = \delta$; using that $1 - e^{-u} \geqslant (1 - e^{-1})u \geqslant u/2$ for $u = \log(1/\delta)/n \in [0, 1]$, we have

$$p = 1 - \delta^{1/n} \geqslant \frac{\log(1/\delta)}{2n}. \tag{24}$$

Let $x_0 \in \mathbf{R} \setminus \{0\}$ be such that $|f(x_0)|$ is larger than $\min(\|f\|_\infty/\sqrt{2}, 1/\sqrt{p})$ and let $p_0 = \min(p, 1/f(x_0)^2)$. Fix the distribution of the covariates P_X as follows:

$$X = \begin{cases} 0 & \text{with probability } 1 - p_0, \\ x_0 & \text{with probability } p_0. \end{cases}$$

Up to replacing f by $-f$, assume that $f(x_0) > 0$. For $\varepsilon \in \{-1, 1\}$, let P_ε denote the joint distribution of the random pair $(X, \varepsilon f(X))$ (where the marginal distribution of X is given by P_X defined above), and let R_ε denote the risk functional associated to the distribution P_ε. Note that P_ε satisfies the first condition of the proposition with $w^* = \varepsilon f(x_0)/x_0$. Also, the second condition holds since $\mathbf{E}Y^2 = p_0 f(x_0)^2 \leqslant 1$.

We now turn to proving the third condition of this proposition. Let $\widehat{g}$ be an arbitrary procedure, possibly improper and depending on P_X. Let $S_0 = \big((0,0), \ldots, (0,0)\big)$ denote a sample of n points equal to $(0, 0)$. Since the quadratic loss function is convex, we may assume

without loss of generality that $\widehat{g}$ is a deterministic procedure and let $g : \mathbf{R} \to \mathbf{R}$ denote the output of $\widehat{g}$ on the sample S_0, that is, $g = \widehat{g}(S_0)$. By symmetry of the problem, assume that $g(x_0) \leqslant 0$ and fix the distribution P of (X, Y) to P_1 (if $g(x_0) \geqslant 0$, we may fix $P = P_{-1}$ instead). Consider the event $E = \{X_1 = \cdots = X_n = 0\}$ and note that $\mathbf{P}(E) = (1 - p_0)^2 \geqslant (1 - p)^n = \delta$. Since $f(0) = 0$, on the event E the observed sample is S_0, so that by (24) we have

$$\begin{aligned} R(\widehat{g}) &\geqslant \mathbf{E}[(g(X) - Y)^2 \mathbf{1}(X = x_0)] = p_0 \cdot (g(x_0) - f(x_0))^2 \geqslant p_0 \, f(x_0)^2 \\ &= \min\left(p \, f(x_0)^2, 1\right) \geqslant \min\left(\frac{p\|f\|_\infty^2}{2}, 1\right) \geqslant \min\left(\frac{\|f_{\mathrm{reg}}\|_\infty^2 \cdot \log(1/\delta)}{4n}, 1\right), \end{aligned}$$

which completes our proof. □

7 Deferred Proofs

This section contains the proof of lemmas appearing in Section 5. Note that rescaling the response Y by $1/m$ affects the excess risk by a multiplicative factor $1/m^2$. Thus, without loss of generality, in all the proofs of this section we may assume that Assumption 1 holds with $m = 1$.

7.1 Proof of Lemma 1

The proof of this lemma is based on a combination of the classical localization via empirical Rademacher complexities argument of [7] and the covering number bounds for truncated VC-subgraph classes due to [29].

First, define the star-hull of $|\overline{\mathcal{F}} - \overline{\mathcal{F}}| = \{|f - g| : f, g \in \overline{\mathcal{F}}\}$ by $\mathcal{H}$, and for $r \geqslant 0$, define its localized subsets by $\mathcal{H}_r$:

$$\mathcal{H} = \Big\{\beta|f - g| : \beta \in [0, 1], \;\; f, g \in \overline{\mathcal{F}}\Big\}, \quad \mathcal{H}_r = \Big\{h \in \mathcal{H} : \;\; \frac{1}{n}\sum_{i=1}^n |h(X_i)|^2 \leqslant 4r\Big\}.$$

Let $\widehat{\psi}_n(r) : [0, \infty) \to \mathbf{R}$ denote any sub-root function with unique positive fixed-point $\widehat{r}^*$ (that is, a positive solution to the equation $\widehat{\psi}_n(\widehat{r}^*) = \widehat{r}^*$ (see [7, Definition 3.1, Lemma 3.2]). Suppose that $\widehat{\psi}_n$ satisfies the following inequality for any $r \geqslant \widehat{r}^*$:

$$\frac{1}{n}\mathbf{E}_{\varepsilon_1, \ldots, \varepsilon_n} \sup_{h \in \mathcal{H}_r} \left(\sum_{i=1}^n \varepsilon_i h(X_i)\right) + \frac{\log(3/\delta)}{n} \lesssim \widehat{\psi}_n(r), \tag{25}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ is a sequence of i.i.d. Rademacher random variables. Notice that for any $r \geqslant 0$ and any $h \in H_r$ we have $\sup_x |h(x)| \leqslant 2$ and $\mathbf{E}h(X)^2 \leqslant 4\mathbf{E}h(X)$. Hence, by the first part of [7, Theorem 4.1], with probability at least $1 - \delta$, the following holds simultaneously for all $f, g \in \overline{\mathcal{F}}$:

$$\mathbf{E}|f(X) - g(X)| \leqslant \frac{2}{n}\sum_{i=1}^n |f(X_i) - g(X_i)| + c\left(\widehat{r}^* + \frac{\log(3/\delta)}{n}\right), \tag{26}$$

where $c > 0$ is some universal constant.

In the rest of the proof we show that a suitable value of $\widehat{r}^*$ can be obtained by upper bounding the empirical Rademacher complexity terms via Dudley's entropy integral. To do so, we first need to obtain an upper bound on the covering numbers of the class $\mathcal{H}$ with respect to the empirical L_2 distance, defined between any $h, h' \in \mathcal{H}$ by $\sqrt{\frac{1}{n}\sum_{i=1}^n (h(X_i) - h'(X_i))^2}$. In what follows, for any class $\mathcal{G}$ and any $\gamma > 0$, an empirical L_2 γ-net of $\mathcal{G}$ will be denoted by

$N(\mathcal{G},\gamma) \subseteq \mathcal{G}$. Thus, the covering number of $\mathcal{G}$ with respect to the empirical L_2 distance at scale γ is at most $|N(\mathcal{G},\gamma)|$.

Since $\mathcal{H}$ is a star-hull of the class $|\overline{\mathcal{F}} - \overline{\mathcal{F}}|$, it follows from [51, Lemma 4.5] that for any $\gamma > 0$ we have

$$|N(\mathcal{H},\gamma)| \leqslant \left|N(\overline{\mathcal{F}} - \overline{\mathcal{F}}, \gamma/2)\right| \cdot \frac{4}{\gamma}. \tag{27}$$

Further, noting that the Minkowski sum of $\gamma/4$ covers of $\overline{\mathcal{F}}$ forms a $\gamma/2$ cover of $\overline{\mathcal{F}} - \overline{\mathcal{F}}$ it follows that

$$|N(\overline{\mathcal{F}} - \overline{\mathcal{F}}, \gamma/2)| \leqslant |N(\overline{\mathcal{F}}, \gamma/4)|^2. \tag{28}$$

Let $\overline{\mathcal{F}}_+ = \{x \mapsto \max(0, f(X)) : f \in \overline{\mathcal{F}}\}$ and $\overline{\mathcal{F}}_- = \{x \mapsto \min(0, f(X)) : f \in \overline{\mathcal{F}}\}$. By the same argument, it holds that

$$|N(\overline{\mathcal{F}}, \gamma/4)| \leqslant |N(\overline{\mathcal{F}}_+, \gamma/8)| \cdot |N(\overline{\mathcal{F}}_-, \gamma/8)|. \tag{29}$$

Finally, plugging in the upper bounds on the covering numbers of $\overline{\mathcal{F}}_+$ and $\overline{\mathcal{F}}_-$ due to [29, Theorem 9.4, Theorem 9.5][2], the chain of inequalities (27), (28) and (29) yields

$$\log|N(\mathcal{H},\gamma)| \lesssim d\log(e/\gamma).$$

Plugging in the above inequality into Dudley's entropy integral [26, Theorem 3.5.1] upper bound on Rademacher complexities, we obtain

$$\begin{aligned}
\frac{1}{n}\mathbf{E}_{\varepsilon_1,\ldots,\varepsilon_n} \sup_{h\in\mathcal{H}_r}\left(\sum_{i=1}^n \varepsilon_i h(X_i)\right) &\lesssim \frac{1}{\sqrt{n}}\int_0^{2\sqrt{r}} \sqrt{d\log(e/\gamma)}d\gamma \\
&\lesssim \sqrt{\frac{d}{n}}\sqrt{r\log(e/r)}\left(\mathbb{1}_{\{r\geqslant d/n\}} + \mathbb{1}_{\{r<d/n\}}\right) \\
&\lesssim \sqrt{\frac{dr\log(n/d)}{n}} + \frac{d\sqrt{\log(n/d)}}{n}.
\end{aligned}$$

In particular, the inequality (25) is satisfied by the choice:

$$\widehat{\psi}_n(r) = c\left(\sqrt{\frac{dr\log(n/d)}{n}} + \frac{d\sqrt{\log(n/d)} + \log(3/\delta)}{n}\right).$$

Solving the fixed-point equation $\widehat{\psi}_n(\widehat{r}^*) = \widehat{r}^*$ yields $\widehat{r}^* \lesssim \frac{d\log(n/d)+\log(1/\delta)}{n}$. The claim follows by the localization theorem stated in (26).

7.2 Proof of Lemma 2

Fix any $f, g \in \mathcal{G}$ and recall that $\mathbf{E}(\ell_f - \ell_g) = R(f) - R(g)$. By the standard bound [45, Theorem 2], for any $\delta' \in (0,1)$, the choice $k(\delta') = \lceil 8\log(1/\delta')\rceil$ guarantees that with probability at least $1-\delta'$ we have

$$\left|R(f) - R(g) - \mathrm{MOM}_{S_n}^{k(\delta')}(\ell_f - \ell_g)\right| \leqslant \sqrt{\frac{32\mathrm{Var}(\ell_f - \ell_g)\log(1/\delta')}{n}}. \tag{30}$$

To upper bound the variance term, first notice that

$$\ell_f(X,Y) - \ell_g(X,Y) = 2Y(g(X) - f(X)) + f(X)^2 - g(X)^2.$$

[2] See also the proof of [29, Theorem 11.3] where the same bound on covering numbers is used.

Combining the above identity with the inequality $(a+b)^2 \leqslant 2a^2+2b^2$ for any a, b, Assumption 1 (with $m=1$) and the boundedness of f, g, we obtain

$$\begin{aligned}\mathrm{Var}(\ell_f - \ell_g) &\leqslant 8\mathbf{E}Y^2(g(X)-f(X))^2 + 2\mathbf{E}(f(X)^2-g(X)^2)^2 \\ &\leqslant 8\mathbf{E}(g(X)-f(X))^2 + 2\mathbf{E}(f(X)-g(X))^2(f(X)+g(X))^2 \\ &\leqslant 16\mathbf{E}(g(X)-f(X))^2.\end{aligned}$$

Since the class $\mathcal{G}$ is finite, taking $\delta' = \delta/(2|\mathcal{G}|^2)$ the upper bound (30) extend uniformly to all pairs $f, g \in \mathcal{G}$, with probability at least $1-\delta/2$. In particular, for any $f, g \in \mathcal{G}$ it holds that

$$\begin{aligned}\left|R(f) - R(g) - \mathrm{MOM}_{S_n}^{k(\delta')}(\ell_f - \ell_g)\right| &\leqslant \sqrt{\frac{512\mathbf{E}(g(X)-f(X))^2 \cdot (2\log(|\mathcal{G}|)+\log(2/\delta))}{n}} \\ &\leqslant \alpha\sqrt{\mathbf{E}(g(X)-f(X))^2}. \end{aligned} \tag{31}$$

This completes the proof of the first inequality.

We will now simultaneously prove the second and the third inequalities appearing in the statement of this lemma. Note that $m=1$ ensures that for any $f, g \in \mathcal{G}$ we have $(f(X)-g(X))^2 \leqslant 4$ and $\mathbf{E}(f(X)-g(X))^4 \leqslant 4\mathbf{E}(f(X)-g(X))^2$. Hence, for any $\delta'' \in (0,1)$ and any $f, g \in \mathcal{G}$, Bernstein's inequality ensures that with probability at least $1-2\delta''$ it holds simultaneously that

$$\mathbf{E}(f(X)-g(X))^2 \leqslant \frac{2}{n}\sum_{i=1}^n (f(X)-g(X))^2 + \frac{12\log(1/\delta'')}{n}, \tag{32}$$

$$\frac{1}{n}\sum_{i=1}^n (f(X_i)-g(X_i))^2 \leqslant 2\mathbf{E}(f(X)-g(X))^2 + \frac{12\log(1/\delta'')}{n}. \tag{33}$$

Setting $\delta'' = \delta/(4|\mathcal{G}|^2)$ the above inequalities extend uniformly to all pairs $f, g \in \mathcal{G}$ with probability at least $1-\delta/2$. Noting that $\frac{12\log(1/\delta'')}{n} \leqslant \alpha^2$, the inequality (33) completes the proof of the third inequality of this lemma. Finally, the second inequality appearing in the statement of this lemma is implied (on the event of the first and third inequalities) by plugging in (32) into (31) together with the inequality $\sqrt{a+b} \leqslant \sqrt{a}+\sqrt{b}$ valid for any $a, b \geqslant 0$. The proof of this lemma is thus complete.

7.3 Proof of Lemma 3

Let E denote the event of Lemma 2 (thus, $\mathbf{P}(E) \geqslant 1-\delta$). By the definition of g^*, for any $g \in \mathcal{G}$ we have $R(g^*) - R(g) \leqslant 0$. Hence, on the event E it holds simultaneously for all $g \in \mathcal{G}$ that

$$\begin{aligned}\mathrm{MOM}_{S_n}^k(\ell_{g^*} - \ell_g) &\leqslant R(g^*) - R(g) + |R(g^*) - R(g) - \mathrm{MOM}_{S_n}^k(\ell_{g^*} - \ell_g)| \\ &\leqslant \sqrt{2}\alpha\sqrt{\frac{1}{n}\sum_{i=1}^n (g^*(X_i)-g(X_i))^2 + \alpha^2}.\end{aligned}$$

In particular, on the event E the function $g^* \in \widehat{\mathcal{G}}$, which completes the first part of the proof.

We now turn to proving the second part of this lemma. Since $g^* \in \widehat{\mathcal{G}}$, by the definition of $\widehat{\mathcal{G}}$, for any $g \in \widehat{\mathcal{G}}$ we have

$$\mathrm{MOM}_{S_n}^k(\ell_g - \ell_{g^*}) \leqslant \sqrt{2}\alpha\sqrt{\frac{1}{n}\sum_{i=1}^n (g(X_i)-g^*(X_i))^2 + \alpha^2}.$$

Hence, on the event E, by the third inequality of Lemma 2, for any $g \in \widehat{\mathcal{G}}$ it holds that

$$\begin{aligned} R(g) - R(g^*) &\leqslant \left| R(g) - R(g^*) - \mathrm{MOM}^k_{S_n}(\ell_g - \ell_{g^*}) \right| + \mathrm{MOM}^k_{S_n}(\ell_g - \ell_{g^*}) \\ &\leqslant 2\sqrt{2}\alpha \sqrt{\frac{1}{n}\sum_{i=1}^n (g(X_i) - g^*(X_i))^2} + 2\alpha^2 \\ &\leqslant 4\alpha \sqrt{\mathbf{E}(g(X) - g^*(X))^2} + 5\alpha^2. \end{aligned}$$

By the definition of the L_2 diameter of the class $\widehat{\mathcal{G}}$ and by the fact that $g^*, g \in \widehat{\mathcal{G}}$, it follows that $\sqrt{\mathbf{E}(g(X) - g^*(X))^2} \leqslant \mathcal{D}(\widehat{\mathcal{G}})$ and hence our proof is complete.

7.4 Proof of Lemma 4

First observe that

$$\begin{aligned} R(\widehat{g}) &= R(g^*) + \left(R(\widehat{g}) - R(g^*) - \mathrm{MOM}^k_{S_n}(\ell_{\widehat{g}} - \ell_{g^*}) \right) + \mathrm{MOM}^k_{S_n}(\ell_{\widehat{g}} - \ell_{g^*}) \\ &\leqslant R(g^*) + \sup_{g \in \mathcal{G}} \left| R(g) - R(g^*) - \mathrm{MOM}^k_{S_n}(\ell_g - \ell_{g^*}) \right| + \mathrm{MOM}^k_{S_n}(\ell_{\widehat{g}} - \ell_{g^*}) \\ &\leqslant R(g^*) + \alpha \mathcal{D}(\mathcal{G}) + \mathrm{MOM}^k_{S_n}(\ell_{\widehat{g}} - \ell_{g^*}), \end{aligned} \tag{34}$$

where the last line follows via an application of Lemma 2. Further, notice that by the definition of $\widehat{g}$ we have

$$\mathrm{MOM}^k_{S_n}(\ell_{\widehat{g}} - \ell_{g^*}) \leqslant \max_{g \in \widehat{\mathcal{G}}} \mathrm{MOM}^k_{S_n}(\ell_{\widehat{g}} - \ell_g) \leqslant \max_{g \in \widehat{\mathcal{G}}} \mathrm{MOM}^k_{S_n}(\ell_{g^*} - \ell_g).$$

At the same time, on the event of Lemma 2, for all $g \in \mathcal{G}$ we have

$$\mathrm{MOM}^k_{S_n}(\ell_{g^*} - \ell_g) \leqslant R(g^*) - R(g) + \alpha \mathcal{D}(\mathcal{G}) \leqslant \alpha \mathcal{D}(\mathcal{G}).$$

Combining the above inequality with (34) concludes our proof.

Acknowledgements. We would like to thank Manfred Warmuth for several useful discussions. T.V. is supported by the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). N.Z. is funded in part by ETH Foundations of Data Science (ETH-FDS).

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

[2] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems 20*, pages 41–48, 2008.

[3] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

[4] J.-Y. Audibert and O. Catoni. Linear regression through PAC-Bayesian truncation. *arXiv preprint arXiv:1010.0072*, 2010.

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

[6] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

[7] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[8] D. Belomestny and J. Schoenmakers. *Advanced Simulation-Based Methods for Optimal Stopping and Control: With Applications in Finance*. Springer, 2018.

[9] L. Breiman and D. Freedman. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136, 1983.

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

[11] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[12] O. Catoni. *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg, 2004.

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

[14] O. Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.

[15] O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

[16] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[17] Y. Cherapanamjeri, E. Aras, N. Tripuraneni, M. I. Jordan, N. Flammarion, and P. L. Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020.

[18] G. Chinot, G. Lecué, and M. Lerasle. Robust statistical learning with Lipschitz and convex loss functions. *Probability Theory and Related Fields*, pages 1–44, 2019.

[19] A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics*, 13(5):819–834, 2013.

[20] F. Comte and V. Genon-Catalot. Regression function estimation on non compact support in an heteroscesdastic model. *Metrika*, 83(1):93–128, 2020.

[21] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.

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

[23] I. Diakonikolas and D. M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

[24] A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

[25] J. Forster and M. K. Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.

[26] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40. Cambridge University Press, 2016.

[27] E. Gobet. *Monte-Carlo Methods and Stochastic processes: from Linear to Non-linear*. CRC Press, 2016.

[28] E. Gobet and P. Turkedjiev. Adaptive importance sampling in least-squares Monte Carlo algorithms for backward stochastic differential equations. *Stochastic Processes and their Applications*, 127(4):1171–1203, 2017.

[29] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer Science & Business Media, 2002.

[30] L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.

[31] F. R. Hampel, P. J. Rousseeuw, E. M. Ronchetti, and W. A. Stahel. *Robust statistics: the approach based on influence functions*. Wiley, 1980.

[32] Q. Han and J. A. Wellner. Convergence rates of least squares regression estimators with heavy-tailed errors. *The Annals of Statistics*, 47(4):2286–2319, 2019.

[33] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

[34] D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

[35] P. J. Huber. Robust statistics. *Wiley Series in Probability and Mathematical Statistics*, 1981.

[36] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[37] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

[38] A. Klivans, P. K. Kothari, and R. Meka. Efficient algorithms for outlier-robust regression. In *(Extended abstract) Proceedings of the 31st Conference On Learning Theory*, pages 1420–1430, 2018.

[39] Y. Klochkov, A. Kroshnin, and N. Zhivotovskiy. Robust k-means clustering for distributions with two moments. *The Annals of Statistics (forthcomming)*, 2020.

[40] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

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

[42] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probability Theory and Related Fields*, 145(3-4):591–613, 2009.

[43] G. Lecué and S. Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

[44] M. Ledoux and M. Talagrand. Conditions d'integrabilité pour les multiplicateurs dans le TLC Banachique. *The Annals of Probability*, pages 916–921, 1986.

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

[46] G. Lugosi and S. Mendelson. Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields*, 175:957–973, 2019.

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

[48] G. Lugosi and S. Mendelson. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

[49] G. Lugosi and S. Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

[50] P. Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer-Verlag Berlin Heidelberg, 2007.

[51] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

[52] S. Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):21, 2015.

[53] S. Mendelson. On aggregation for heavy-tailed classes. *Probability Theory and Related Fields*, 168(3):641–674, 2017.

[54] S. Mendelson. An unrestricted learning procedure. *Journal of the ACM*, 66(6):1–42, 2019.

[55] S. Mendelson. Extending the scope of the small-ball method. *Studia Mathematica*, pages 1–21, 2020.

[56] S. Mendelson. Learning bounded subsets of L_p. *arXiv preprint arXiv:2002.01182*, 2020.

[57] S. Mendelson and N. Zhivotovskiy. Robust covariance estimation under $L_4 - L_2$ norm equivalence. *The Annals of Statistics*, 48(3):1648–1664, 2020.

[58] S. Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[59] S. Minsker and T. Mathieu. Excess risk bounds in robust empirical risk minimization. *arXiv preprint arXiv:1910.07485*, 2019.

[60] J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv preprint arXiv:1912.10754*, 2019.

[61] A. Nemirovski. Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

[62] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, 1983.

[63] R. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3-4):1175–1194, 2016.

[64] D. M. Ostrovskii and A. Rudi. Affine invariant covariance estimation for heavy-tailed distributions. In *Conference on Learning Theory*, pages 2531–2550, 2019.

[65] A. Pensia, V. Jog, and P.-L. Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

[66] A. Rakhlin, K. Sridharan, and A. B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

[67] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*, volume 589. John Wiley & Sons, 2005.

[68] A. Saumard. On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, 24(3):2176–2203, 2018.

[69] O. Shamir. The sample complexity of learning linear predictors with the squared loss. *Journal of Machine Learning Research*, 16(108):3475–3486, 2015.

[70] A. B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer Berlin Heidelberg, 2003.

[71] J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

[72] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[73] S. Van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge University Press, 2000.

[74] A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

[75] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka. Moscow., 1974.

[76] T. Vaškevičius and N. Zhivotovskiy. Suboptimality of constrained least squares and improvements via non-linear predictors. *arXiv preprint arXiv:2009.09304*, 2020.

[77] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2016.

[78] V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

[79] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

[80] H. Walk and L. Zsidó. Convergence of the Robbins-Monro method for linear problems in a Banach space. *Journal of Mathematical Analysis and Applications*, 139(1):152–177, 1989.

[81] Y. Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74(1):135–161, 2000.

[82] Y. Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87, 2000.

[83] D. Z. Zanger. Quantitative error estimates for a least-squares Monte Carlo algorithm for American option pricing. *Finance and Stochastics*, 17(3):503–534, 2013.